Ca' Foscari
University
of Venice

International Management

Masters' degree

Big Data in Tourism Industry:

How Online Reviews can affect Hotel Performance

Supervisor

Prof. Stefano Micelli

Assistant supervisor

Prof. Fabrizio Gerli

Graduand

Ilaria Marinosci

Matriculation Number 842489

Academic Year

2017 / 2018

INDEX

INTRODUCTION

If we think of the new Industrial revolution (Industry 4.0), the first sector that we come up with is manufacturing, but of course it is not the only one. This new revolution has been so disruptive to affect every single aspect of our modern society, by changing the industry in a total different way compared to the previous Industrial revolutions.
This revolution brought some new technologies with it by bringing to life a revolutionized system of automation and digitalization that affected the interconnection among the different units of production.

One of the best characterizing element of this new revolution were Big Data and Analytics. They are called "big" data because the amount of data available for the new economic paradigm was so big and variated than before that it even required new techniques and technologies in order to process this data. This bigger amount is characterized by both structured and unstructured data, and since structured data were easier to process and analyze the unstructured ones were kind of new for all the analysts at the beginning. This new form of data came from a big array of sources and were no longer simple text documents but it could also include images, emails, transactions, video and audio. This led to unavoidable advances to all the information and analytics systems companies used to process previous data in order to make better decisions and more efficient business strategy formulations.

This information and communication technology revolution has been so innovative that affected the most diverse sectors in the world economy. One of the sector most affected is undoubtedly the Tourism sector, and in particular the hospitality industry. Thanks to the bigger amount of data available to touristic companies, such as Airlines companies and Accommodation facilities (Hotel chains), it will be easier for them to make more focused strategic decisions. In fact one of the first users have been the big airlines companies like Air France or British Airways and big hotel chains like Hilton and Marriott.
In particular, the analysis of big data in Tourism led to important benefits such as better decision support for the management of travel and touristic companies; new and customizable products and services for customers that could reflect at best their preferences and buying behavioral patterns; it helps to build stronger relationships between the service

providers and their customers in order to develop a brand loyalty; a new, faster and cheaper data processing that allows to save time and make better decisions according to more valuable and rich data that come from multiple sources. All of these benefits can be crucial for the companies that will be able to best exploit the data available, but it is necessary for them to make investments into the new technologies required to process such data and use them wisely in order to translate them into realistic strategic formulations. Furthermore, big data in Tourism industry allows to companies to collect data coming directly from customers and this can be a realistic portrait of their preferences and of what they would their travel experiences look like. For this reason the right analysis of this data can be so valuable for companies in this industry.

The costumer nowadays is more aware of the product and service he or she wants during the trip, and together with the price they want a better quality. In addition, they have access to a wider and cheaper source of information where to search for the product or service they wants, they have the chance to assess all the information available to them thanks to Internet, they can compare many different solutions for their trip and choose the most cost- and quality- effective. These new means available to them have shifted the purchasing power in their hands and made the travel industry more competitive.

This paperwork has the aim to focus the attention on the hospitality industry where it can be very valuable to exploit the potential of big data in order to improve the performance of facilities, and consequently to translate it into an increase in revenues. Big data can be crucial either for big hotel chains or smaller hotels and accommodation facilities such as B&B, that thanks to cheaper and available data can make better strategic decision to be more competitive and productive. In order to do that, hotels should monitor and enhance their online visibility and reputation, aimed at attracting new potential customers and retaining the current ones.

One of the best mechanism to do that is to monitor the online reviews that satisfied and unsatisfied customers release on the web, in particular on online customer reviews platforms like TripAdvisor. Thanks to the monitoring of such opinions about them, hotels can learn about travelers preferences and about what aspects of their structures should be improved or fixed. It has been proved that online reviews can affect other travelers' purchasing behavior because the reviews writers are considered to be more reliable and less biased than the service providers and the pure advertising. Online customer reviews can now become a good example of big data content hotels can employ to increase their online reputation.

In this paper, there will be presented the results emerged from some interviews made to a dozen of Venetian hotels in order to verify if data extrapolated from online reviews could be a real help for hotels managers to increase their performance.

To begin with, in the first chapter the topic of Big data and Big data analytics, as one of the main outcome of Industry 4.0, will be further explained with a particular focus on its employment in Tourism industry. In order to understand the big potential big data has for tourism, their main features will be presented and explained. Thanks to that, it will be easier to realize all the benefits it brings to this sector, and to a bigger extent to the overall industry.
In consideration of that, in the second chapter some example of companies, in particular travel companies (hotel chains and airline companies), which started to employ big data in order to increase their performance and strategic decision making processes, will be presented. Furthermore, it will be possible to acquaint how worldwide companies conceive this information revolution and the ways in which they are practically employing big data in their business activity.
In the third chapter, some techniques and methodologies employed to process big data in the touristic domain will be explained. This will give the chance to understand how this bigger availability of data can be easily extrapolated from all the contents present across the web. The primary methodology explained in this chapter, Sentiment analysis, is the main one employed by many software companies that deals with data science and artificial intelligence to analyze online contents and to extrapolate the most useful information from them in order to assist firms in the hospitality industry. Sentiment analysis is mostly used to analyze online user generated contents such as online reviews in order to extract the sentiment of the overall statement and all the words/element repeated more frequently. This may be useful to hotel managers for a better understanding of their service strengths and weaknesses.
In the fourth, and last chapter, the focus will be on user generated contents (i.e. online reviews) and their role in the hospitality industry. It will be explained how they can significantly impact hotels management and consumers' behavior and the reasons why they can be useful tools in the hands of hotel managers to increase the online visibility and reputation of their facility. In the last part of the chapter, it will be possible to understand how these contents can affect hotel performance from the point of view of the managers. It will be presented the results of a dozen of interviews made with the purpose of understanding how one of the basic form of big data can be useful for managers in the hospitality industry and how much they really know about it and its biggest potential. From the results it will be

possible to understand the main difference in big data usage between hotel chains and small independent hotels in the city of Venice.

CHAPTER 1

1. Big Data in Tourism Industry

In order to better understand what is Big Data and why we talk about a revolution of Big Data, it is appropriate to first introduce the general context of the Industry 4.0. Only in this way, it will be easier to deeply understand what Big Data is and why it is so important and influential across so many industries nowadays, especially in the tourism industry.

1.1. INDUSTRY 4.0: The revolution of Big Data

The name Industry 4.0 means that it takes part in the fourth Industrial revolution characterized by a new system of automation and digitalization that affect the interconnection of all the units of production existing in an economic system (Roland Berger, 2014). The wave of automation that characterize this late revolution started in the 2000s, but was taken even further. For this reason new terms were created such as *Smart Factories*, *Smart Industry* and the *Industrial Internet of things*.

According to the Boston Consulting Group there are nine technological pillar in the Industry 4.0. They are:

- **Big Data and Analytics**: the gathering and the analysis of a big amount of data sets originated from different sources that will become a standard to support real-time decision making

- **Autonomous Robot**: the new evolution of the old robots already in use in the industry, now they are becoming more autonomous, flexible and cooperative. In the future they could interact with each other and work side by side with humans and learn from them. In addition they can be used in many functions, from the production phase to the logistics, and the office management, since they are easily controlled remotely. These new robots are less costly and will have a greater range of capabilities than the previous ones by leading to more efficiency.

- **Simulation**: the 3D simulations of products, materials and processes will be also extended to the plant operations level. These simulations will allow to leverage real-time data to mirror the physical world into a virtual model. This allow then to test and improve the processes and the machine settings in the virtual model before the physical changeover. Thanks to this the setup time lower and the quality increases.

- **Horizontal and vertical system integration**: one of the main aim of the industry 4.0 is to provide a complete integration among companies, departments, functions and capabilities and to enable truly automated value chains.

- **The industrial Internet of things**: this new connection will allow to many devices to communicate and interact with one another. It also allow to decentralize analytics and decision making enabling real-time responses.

- **Cybersecurity**: with the increased connectivity and the use of standard communications protocols there is now the need to protect the critical industrial systems from cyber threats. As a result, secure and reliable communications together with sophisticated identity and access management will be essential.

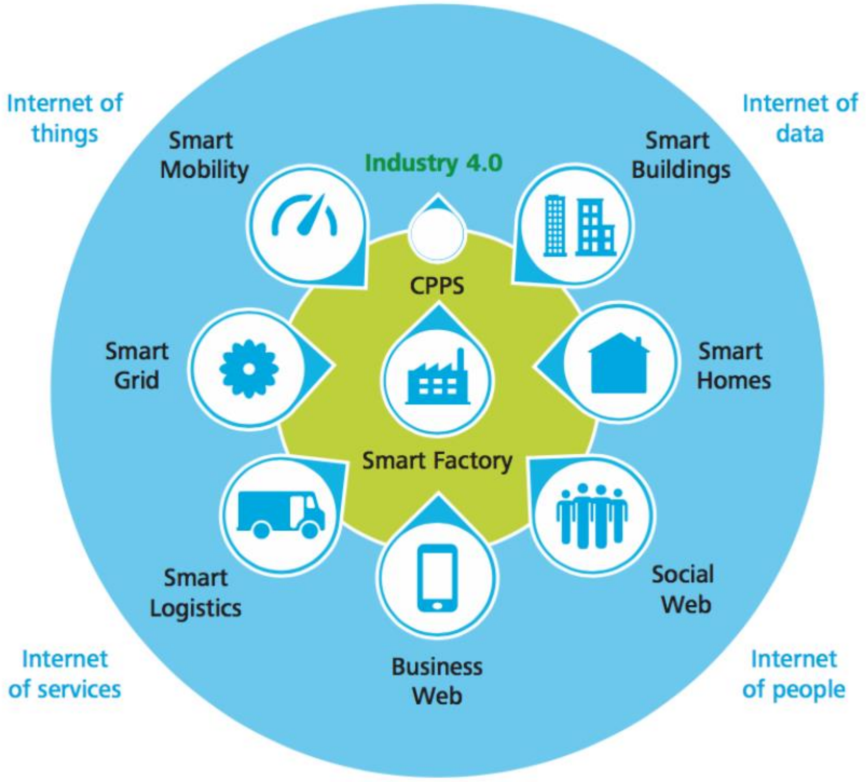
- **The Cloud**: with Industry 4.0 more production-related firms will require an increasing data sharing across the company boundaries. This cloud-based will automatically improve, enabling more data-driven services.

- **Additive manufacturing**: with more of these additive manufacturing methods there will be the chance to produce batches of customized products that will offer production advantages such as complex but lightweight designs. As a result, this new method will reduce transport costs and distances and the stock on hand.

- **Augmented reality**: this tool can support a variety of processes and services but it is still at its beginning. In their future they will provide workers with real-time information in order to improve the decision making process and the work procedures.

Some of these technologies already existed before the Industry 4.0 but, they were mainly focused on improving the production phase in manufacturing. Now they are intended to affect the whole value chain and the supply chain, including the relationship among producers, suppliers and customers.

The new single feature of the Industry 4.0 will be the interaction among the factory and all the infrastructures existing around it, which will become "smart infrastructures", e.g. smart buildings, homes, logistics, mobility and grid, and connectivity to business and social web.



1.1.1. BIG DATA: What is it?

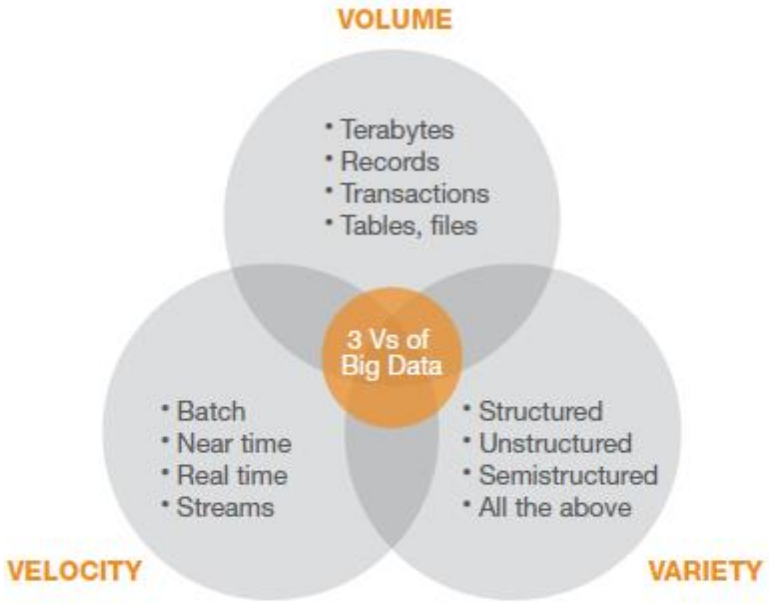
Big data is one of the most crucial feature of this 4th revolution. In particular is the new way to indicate the large volume of data, both structured and unstructured, that flows into a business on a daily basis. This big data can be analyzed for insights that can lead to better decisions and business strategy formulations.

The action of gathering and analyzing a large amount of data has always been fundamental for companies in the competitive environment, but the term "big data" is relatively new and was coined in the early 2000s when the analyst Doug Laney articulated this late definition as the three Vs:

- **VOLUME**: when thinking about volume, most people define big data in terabytes, sometimes even petabytes. But in addition to that, organization quantify data by considering the diverse sources it comes from, including business transactions, social media and the information coming from sensors or machine-to-machine data. Some of them also find it more useful to quantify data in terms of time. For example, due to the seven-year statute of limitations in the U.S., many firms prefer to keep seven years of data available for risk, compliance, and legal analysis. The quantification of big data is also affected by the scope of the data. In fact, in many organizations, the data collected for general data warehousing is different from data collected specifically for analytics. Furthermore, different forms of analytics may have different data sets. For example, some practices require the creation of specific analytic data sets per analytic projects, while the whole enterprise has its own large and complete scope of big data.

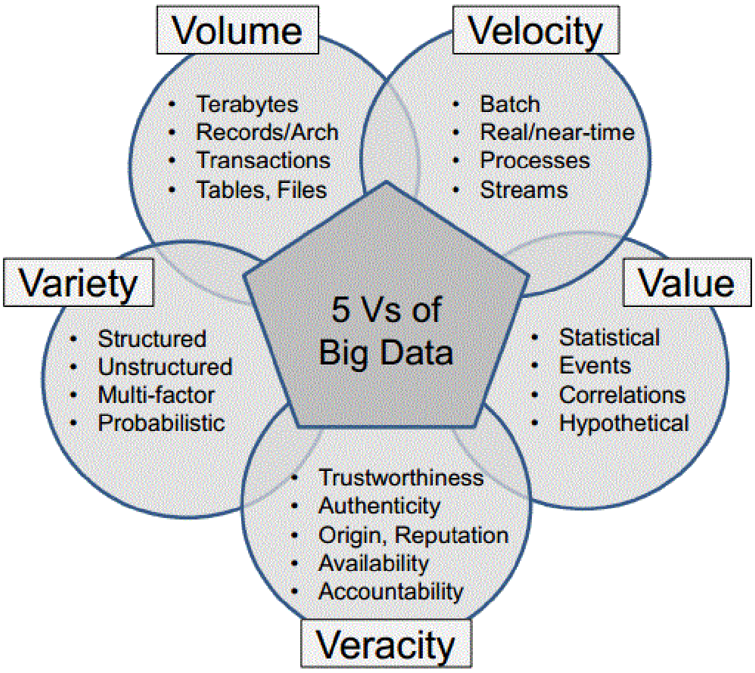
- **VARIETY**: one of the most complex characteristic is that data comes in all types of formats, from structured, numeric data in traditional databases to unstructured text documents, email, financial transactions to some that is hard to categorize such as audio and video. The few organization that started to analyze this data years before now they are doing so at a more complex and sophisticated level. In fact, the really new thing about big data is the way in which more advanced analytics are leveraging it. In addition, variety and volume tend to fuel each other.

- **VELOCITY**: velocity or speed can be thought either as the frequency of data generation or the frequency of data delivery. Now, data streams in at an unprecedented relentlessly speed and must be managed by the analytic tools in a timely manner to be ready to take real-time actions.



Recently two more main feature have been added to this definition, these are: Value (Chen, Mao, Zhang, and Leung-2014), and Veracity (Beyer and Laney-2012):

- **VALUE:** insights that can be revealed through big data and that can be valuable for those organizations employing big data. For example in tourism, big data can be valuable in providing personalized marketing strategies and targeted product and service design.
- **VERACITY:** trustworthiness and authenticity of the data used and analyzed given the context, the variety of communication "touch points", and the speed at which things flow. Big data veracity, in particular, refers to the biases, "noise", and abnormality in data, that means to detect if big data being stored and mined are meaningful to the problem is being analyzed. Compared with volume and velocity, veracity in data analysis is the biggest challenge. In developing a strategy thanks to Big data analysis, it is important to keep the data "clean" and to develop processes to keep "dirty data" from accumulating in the systems.



We cannot forget to consider that data flows can be inconsistent with periodic peaks. Daily, seasonal and event-triggered peak data can be difficult and challenging to manage. Even more with the unstructured data.

It's important to remember that the primary value from big data comes not from the data in its raw form, but from the processing and analysis of it and its insights, products and services emerging from the analysis. A couple of examples are **Siemens** and **Long** (2011), who define big data as "*datasets whose size is beyond the ability of typical database software tools to capture, store, manage and analyze*". This definition underlines that big data are thought in terms of how it gets analyzed, not how much they are and how many specific terabytes of space it fills.

The sweeping changes in big data technologies and management approaches need to be accompanied by similarly dramatic changes in how data supports decisions and product/services innovation (*Thomas H. Davenport in Big Data In Big Companies*).

Furthermore, since data comes from multiple and diverse sources, it is difficult to link, match, cleanse and transform data across systems. For this reason, it is necessary to connect and create relationships, hierarchies and multiple data linkages, otherwise data can quickly run out of control.

Before discovering how big data can work, it is important to understand where it comes from. The sources of big data generally fall into one of these three categories:

- **Streaming data**: this category includes data that reach IT systems from a web of connected devices. You can analyze data as it arrives and make decisions on what data to keep, what to discard and what requires further analysis.

- **Social media data**: the data coming from social interactions is an increasing attractive set of information, especially for marketing, sales and support functions. This data is often in unstructured or semi-structured forms, for this reason it poses a unique challenge for what regards its analysis and consumption.

- **Publicly available sources**: massive amounts of data are available through open data sources like the US government's data.gov, the CIA World Factbook or the European Union Open Data Portal.

1.1.2. Why is Big Data important?

The importance of big data doesn't lay on the amount of data organizations have, but on how they use and analyze it. The analysis can enable 1) cost reduction, 2) time reduction, 3) new product development and optimized offerings, and 4) smart and fast decision making. In order to do that big data need to be combined with high-powered analytics. This combination allows to accomplish some important business-related tasks such as:

- Identifying root causes of failures, issues and defects in near-real-time
- Generating coupons at the point of sale based on customers' buying habits
- Recalculating entire risk portfolios in few minutes
- Detecting fraudulent behavior before it

In addition to that, thanks to data analytics it is possible to accomplish activities such as website design, product placement optimization, customer transaction analysis, market structure analysis and product recommendation.

1.1.3. Who are the main users of Big Data?

Big data are employed in every industry and the main ones that can benefit from it are:

- **Banking**: with large amounts of information streaming in from many sources, banks are now dealing with finding new and innovative ways to manage big data. For them is really important both to understand customers' needs and increase their satisfaction, and to minimize risks and frauds while maintaining regulatory compliance. Big data brings with it big insight, but it also requires financial institution to stay updated and one step ahead with advanced analytics.

- **Education**: educators can make a significant impact on school systems, students and curriculums thanks to data-driven insights. By analyzing big data, they can detect at-risk students, make sure they are making adequate progress, and can implement a better system for evaluating and supporting teachers and principals.

- **Government**: when governmental agencies are able to harness and apply analytics to their big data, they gain significant power when it comes to manage utilities, to run agencies, to deal with traffic congestion or to prevent crime. In this case, governments must handle big data paying attention to the issues of transparency and privacy.

- **Health Care:** the collection of big data can enable patient records, treatments plans and prescription information. In Health Care industry everything must be done quickly but accurately, and in some cases, with enough transparency in order to satisfy strict regulations. When big data is managed effectively, health care providers can uncover hidden insights that improve patient care.

- **Manufacturing:** as we have said above, thanks to big data, manufacturers can boost quality and output while minimizing waste and failures in today's highly competitive market.

- **Retail Companies**: according to the fact that industry 4.0 affect the whole value chain, big data can enable retail companies to pay attention and improve the customer relationship. In fact, retailers need to know the best way to market the products, the most effective way to handle transactions and the best strategic way to bring back lapses business.

1.2. Techniques and Technologies to “manage” Big Data

Since big data are varied and fast growing many technologies and analytics techniques are needed in order to extract relevant information.

1.2.1. Techniques

There are lots of analytics techniques that could be employed when attempting to interpret big data. Which one is used depends on the type of data being analyzed, the technology available and the research question they are trying to solve.

The most frequent tools used are:

- **Association rule learning**: it’s a way of finding relationships among variables. It is often used in data mining and it can support recommender system like those employed by Amazon and Netflix. (Chen, Chiang and Storey 2012)
- **Data mining**: it’s considered combining methods from statistics and machine learning with database management in order to exactly identify patterns in large datasets. Picciano (2012) indicate it as one of the most important terms related to data-driven decision making and describes it as “searching or digging into a data file for information to understand better a particular phenomenon”.
- **Cluster analysis**: it’s a type of data mining process that divides a large group into smaller groups of similar “objects” whose “similarities are not known in advance” and try to discover what the similarities are among the smaller groups, and if they are new groups try to find what caused these qualities.
- **Crowdsourcing**: it’s a way to collect data from a large group of people through an open source. This tool is more useful for collecting data than for analyzing it.
- **Machine learning**: traditional computers only know what we “tell” them, but in machine learning, a subspecialty of computer science, we have to try to create algorithms that allow computers to evolve based on empirical data. A big focus of machine learning

research is to automatically learn to recognize complex patterns and make intelligent decisions based on data of those patterns.

- **Text analytics**: a large portion of data is generated in text form, such as emails, internet searches, web page contents, corporate documents, etc. and they can be a really good sources of information. Text analysis can be used to extract information from a large amount of textual data.

- **Regression analysis:** regression analysis is a mathematical tool for revealing correlations between one variable and several other variables. Based on a group of experiments or observed data, regression analysis identifies dependence relationships among variables hidden by randomness. Regression analysis may change complex and undetermined correlations among variables into simple and regular correlations.

- **Factor Analysis**: factor analysis is basically targeted at describing the relation among many indicators or elements with only a few factors, i.e., grouping several closely related variables and then every group of variables becomes a factor (called a factor because it is unobservable, i.e., not a specific variable), and the few factors are then used to reveal the most valuable information of the original data.

- **Correlation Analysis**: correlation analysis is an analytical method for determining the law of correlations among observed phenomena and accordingly conducting forecast and control. There are a plentiful of quantitative relations among observed phenomena such as correlation, correlative dependence, and mutual restriction. Such relations may be classified into two types: (1) function, reflecting the strict dependence relationship among phenomena, which is also called a definitive dependence relationship, among which, every numerical value of a variable corresponds to one or several determined values; (2) correlation, under which some undetermined and inexact dependence relations exist, and a numerical value of a variable may correspond to several numerical values of the other variable, and such numerical values present regular fluctuation surrounding their mean values.

- **Statistical Analysis**: Statistical analysis is based on the statistical theory, a branch of applied mathematics. In statistical theory, randomness and uncertainty are modeled with Probability Theory. Statistical analysis can provide description and inference for large-scale datasets. Descriptive statistical analysis can summarize and describe datasets and inferential statistical analysis draws conclusions from data subject to random variations. Analytical technologies based on complex multi-variate statistical analysis include regression analysis, factor analysis, clustering, and recognition analysis, etc.

- **Time series analysis:** there are also techniques that work with time series, other than with cross-sectional data. It comprises methods for analyzing time series data in order to extract meaningful statistics and other characteristics of the data. In this case, time series data has a natural temporal ordering.

1.2.2. Technologies

Together with the analytical techniques, there are several software and technologies available to help big data analytics. Some of the most common are:

- **EDWs**: this stands for Enterprise Data Warehouses that are databases used in data analysis. The biggest question for many companies is whether their current or planned EDW can handle big data and advanced data analytics without degrading the performance of other workloads. Some institutions manage their analytic data in their EDW itself while other use a separate platform, which can help relieving some of the stress resulting from managing data on the EDW.

- **Visualization products**: one of the most important difficulties with big data analytics is finding ways to virtually represent the results. Many new visualization products aim to solve this problem, devising methods for representing data numbering up into the millions. Visualization can also allow to compare models and datasets so that to enable quantitative and qualitative decision-making.

- **MapReduce & Hadoop**: MapReduce is a programming model used to handle lots of data simultaneously, while Hadoop is one of the most popular open-source implementations of that model. According to Szalay and Blakeley, Hadoop works very well for big data, but for smaller projects they underline that it's not as effective as "a good index which might provide better performance by orders of magnitude."

- **NoSQL database**: this database is designed specifically to deal with very large amounts of information that don't use a relational model. It scales very well and is often useful for tracking and analyzing real-time lists of data that grow quickly.

- **Apache Flume:** Apache Flume is a distributed, reliable, and available system for efficiently collecting, aggregating and moving large amounts of log data from many different sources to a centralized data store. Flume deploys as one or more agents, each contained within its own instance of the Java Virtual Machine (JVM). Agents consist of three pluggable components: sources, sinks, and channels. Flume agents ingest incoming streaming data from one or more sources. Data ingested by a Flume agent is passed to a sink, which is most commonly a distributed file system like Hadoop. Multiple Flume agents can be connected together for more complex workflows by configuring the source of one agent to be the sink of another. Flume sources listen and consume events. Flume agents may have more than one source, but at the minimum they require one. Sources require a name and a type; the type then dictates additional configuration parameters. Channels are the mechanism by which Flume agents transfer events from their sources to their sinks. Events written to the channel by a source are not removed from the channel until a sink removes that event in a transaction. This allows Flume sinks to retry writes in the event of a failure in the external repository (such as HDFS or an outgoing network connection). For example, if the network between a Flume agent and a Hadoop cluster goes down, the channel will keep all events queued until the sink can correctly write to the cluster and close its transactions with the channel. Sink is an interface implementation that can remove events from a channel and transmit them to the next agent in the flow, or to the event's final destination and also sinks can remove events from the channel in transactions and write them to output. Transactions close when the event is successfully written, ensuring that all events are committed to their final destination.

- **Apache Sqoop:** Apache Sqoop is a CLI tool designed to transfer data between Hadoop and relational databases. Sqoop can import data from an RDBMS (relational database management system) such as MySQL or Oracle Database into HDFS (Hadoop distributed file system) and then export the data back after data has been transformed using MapReduce. Sqoop connects to an RDBMS through its JDBC connector and relies on the RDBMS to describe the database schema for data to be imported. Both import and export utilize MapReduce, which provides parallel operation as well as fault tolerance. During import, Sqoop reads the table, row by row, into HDFS. Because import is performed in parallel, the output in HDFS is multiple files.

- **Apache Pig:** Apache's Pig is a major project, which is lying on top of Hadoop, and provides higher-level language to use Hadoop's MapReduce library. Pig provides the scripting language to describe operations like the reading, filtering and transforming, joining, and writing data which are exactly the same operations that MapReduce was originally designed for. Instead of expressing these operations in thousands of lines of Java code which uses MapReduce directly, Apache Pig lets the users express them in a language that is not unlike a bash or Perl script.

  Pig was initially developed at Yahoo Research around 2006 but moved into the Apache Software Foundation in 2007. Unlike SQL, Pig does not require that the data must have a schema, so it is well suited to process the unstructured data. But, Pig can still leverage the value of a schema if you want to supply one. PigLatin is relationally complete like SQL, which means it is at least as powerful as a relational algebra. Turing completeness requires conditional constructs, an infinite memory model, and looping constructs.

- **Apache Hive**: Hive is a technology developed by Facebook that turns Hadoop into a data warehouse complete with a dialect of SQL for querying. Being a SQL dialect, HIVEQL is a declarative language. In PigLatin, you specify the data flow, but in Hive we describe the result we want and hive figures out how to build a data flow to achieve that result. Unlike Pig, in Hive a schema is required, but you are not limited to only one schema. Like PigLatin and SQL, HiveQL itself is a relationally complete language but it is not a Turing complete language.

- **Apache ZooKeeper:** Apache Zoo Keeper is an effort to develop and maintain an open-source server, which enables highly reliable distributed coordination. It provides a distributed configuration service, a synchronization service and a naming registry for distributed systems. Distributed applications use ZooKeeper to store and mediate updates to import configuration information. ZooKeeper is especially fast with workloads where reads to the data are more common than writes. The ideal read/write ratio is about 10:1. ZooKeeper is replicated over a set of hosts (called an ensemble) and the servers are aware of each other and there is no single point of failure.

- **MongoDB:** MongoDB is an open source, document-oriented NoSQL database that has lately attained some space in the data industry. It is considered as one of the most popular NoSQL databases, competing today and favors master-slave replication. The role of master is to perform reads and writes whereas the slave confines to copy the data received from master, to perform the read operation, and backup the data. The slaves do not participate in write operations but may select an alternate master in case of the current master failure. MongoDB uses binary format of JSON-like documents underneath and believes in dynamic schemas, unlike the traditional relational databases. The query system of MongoDB can return particular fields and query set compass search by fields, range queries, regular expression search, etc. and may include the user-defined complex JavaScript functions. As hinted already, MongoDB practice flexible schema and the document structure in a grouping, called Collection, may vary and common fields of various documents in a collection can have disparate types of the data. The MongoDB is equipped with the suitable drivers for most of the programming languages, which are used to develop the customized systems that use MongoDB as their backend player. There is an increasingly demand of using MongoDB as pure in-memory database; in such cases, the application dataset will always be small. Though, it is probably easy for maintenance and can make a database developer happier; this can be a bottle neck for complex applications that require tremendous database management capabilities.

- **Apache Cassandra:** Apache Cassandra is the yet another open source NoSQL database solution that has gained industrial reputation which is able to handle big data requirements. It is a highly scalable and high-performance distributed database management system that can handle real-time big data applications that drive key systems for modern and successful businesses. It has a built-for-scale architecture that can handle petabytes of information and thousands of concurrent users/operations per second as easily as it can manage much smaller amount of data and user traffic. It has a peer to peer design that offers no single point of failure for any database process or function, in addition to the location independence capabilities that equate to a true network-independent method of storing and accessing data, data can be read and written anywhere. Apache Cassandra is also equipped with a flexible/dynamic schema design that accommodates all formats of big data applications, including structured, semi-structured, and unstructured data. Data is represented in Cassandra via column families that are dynamic in nature and accommodate all modifications online.

- **Apache Splunk:** Splunk is a general-purpose search, analysis and reporting engine for time-series text data, typically machine data. Splunk software is deployed to address one or more core IT functions: application management, security, compliance, IT operations management and providing analytics for the business. The Splunk engine is optimized for quickly indexing and persisting unstructured data loaded into the system. Specifically, Splunk uses a minimal schema for persisted data – events consist only of the raw event text, implied timestamp, source (typically the filename for file based inputs), source type (an indication of the general type of data) and host (where the data originated). Once data enters the Splunk system, it quickly proceeds through processing, is persisted in its raw form and is indexed by the above fields along with all the keywords in the raw event text. Indexing is an essential element of the canonical "super-grep" use case for Splunk, but it also makes most retrieval tasks faster. Any more sophisticated processing on these raw events is deferred until search time. This serves four important goals: indexing speed is increased as minimal processing is performed, bringing new data into the system is a relatively low effort exercise as no schema planning is needed, the original data is persisted for easy inspection and the system is resilient to change as data parsing problems do not require reloading or re-indexing the data.

  These are just some of the more common techniques and technologies used in big data analytics, but not every organization will use all these techniques and technologies for every project. Everyone will have to evaluate its needs and choose the tools that are the most appropriate for it. These needs obviously change, not only from business to business, but also from sector to sector.

1.3. The potential of Big Data

As we have seen, big data are the best representation of this new Industrial revolution. Big data come in several forms and now it can be used in many fields and industries, not only in manufacturing.

One of the main feature of the industry 4.0 is the internet of things, which implies that the advances in technology will allow more connections among different devices .

These advances in information technology has led to an increasing number of insights included in datasets, and Big data is the word used to indicate this huge amounts of data. This big amounts is composed by structured and unstructured data which are transferred with a great velocity, for this reason they cannot be managed only by traditional methods, as they used to process the old type of datasets. As the size of this data is really big, its sources of generation are several and very differentiated from each other, starting from mobile phone data going to credit cards used in economic transactions, from television to the storages for the computer applications, from the intelligent urban infrastructures to the sensors or cameras installed in buildings, to private and public transports, etc.

The first organizations which started to process and employ this data were the online companies such as Google, eBay, LinkedIn and Facebook. These are completely build around big data, which represents for them the main instrument to run their business rather than a simple aim to reach.

One important feature of this data is the great amount of sources, which are clearly miscellaneous, thus this data is often unstructured, such as images, emails, GPS and social networks. This feature implies that the methods used to process it are no longer the traditional ones but new methods and technologies must be developed to do it and in order to produce a significant value from this big amount of data. Consequently, this big amount enables to extract more and more information than before.

Big data is increasingly important because it can reveal the behaviors and the preferences of the users that creates it. Alexander Jaimes, a Yahoo Research researcher stated that " data is us", this means that data represents what we are and it is the best way to get to know what we like, our needs and our preferences. Thanks to this, companies can develop and produce customized products in order to retain their customers. Naturally, these "tailor made" products increase the overall competitive advantage.

"Big data" is called in this way to indicate such a big set of data in terms of volume, speed and variety that it requires specific techniques and technologies in order to extrapolate a real value from it. But, otherwise, putting the accent on the term "big" could be misleading , because we might think exclusively in terms of wideness of the quantity of information deriving from it. On the other side, the emphasis on the feature of volume is unavoidable because of the huge size of data comparing to the past, that must be processed with more advanced software and hardware.

But, what we have to keep in mind is that the volume is not the only important feature of this new datasets. For many companies it is crucial the structure of the data they use and analyze, more important than the mere size . The structure, in fact, allows to them to analyze data in real-time and this can be fundamental for many companies in fast-changing environments.

Another important parameter is thus the "variety" of the insights. An important focus regarding Big data is the research of information on several fronts , without the need that it must necessarily materialize in structured dataset represented exclusively by statistical figures, but also by images, video, text documents, chat, etc.

Another important feature is the speed or velocity, which is meant not only as the rate at which the data grows, but mostly the speed needed to analyze and process it before it becomes obsolete.

Besides the three main features of big data, *volume*, *variety* and *velocity* (3V) , two new parameters has recently been introduced, the *variability* and the *complexity*. Variability means that there is the possibility of inconsistency in its generation, i.e. social media trends.

Complexity indicates that even the simplest operation to collect and analyze a specific dataset implies considerable efforts and great resources in terms of software and hardware.

At last another feature of big data can be considered the "value" it produces for businesses, for example Amazon estimates that 30 % of its revenue comes from sales figures derived thanks the use of big data.

Big data can be represented mainly by two categories of data: the first one is composed by the traditional techniques to collect statistical figures; the second one is composed by the majority of web-based data. Examples of big data can be:

- Data structured in relational tables (SQL);
- Semi-structured data: in general it is about the business-to-business data generated by XML;
- Data deriving from planning events and machinery: messages, sensors, Mfid;
- Unstructured data: human language, images, video;
- Unstructured data deriving from social media: tweet, blog, social networks;
- Data deriving from the web surfing: web Logs, tag javascript, etc. that allows to extrapolate the preferences of web users;

- GIS data: geospatial data that can be generated by several applications which allows to get information about business, social issues and security issues;
- Scientific data: astronomical, physical and genetic data.

The term “big data” can be also referred to indicate the algorithms able to process so many variables in a short time and with computational resources quite limited. Until recently, a scientist who wanted to analyze a small or medium size of data he would spend a long time to do so and he would employ very costly computers; today with all the advances in technology and thanks to innovative algorithms and programs that amount of data can be processed and analyzed in a very short time, and sometimes even in real-time.

Thus, the real revolution of big data is the birth of all of these innovative instruments and methods able to connect all the information in order to present a wider view of it, by suggesting interpretational methods unimaginable until now.

We could assume that this revolution is affecting only, or mostly, the IT sector, but this is only the starting point and on the other hand the revolution of big data may affect the most diverse sectors and industries, from the automotive, medicine, commerce, astronomy, biology, chemistry, and finance to the insurance sector, and we will further see to the tourism industry as well. No sector which is based on marketing and analysis of data can be considered left out from this revolution.

Unfortunately, this revolution is not taking place without some difficulties for the potential users. The first obstacle to overcome , beyond the technical difficulties, is the reluctance of the companies to share such information necessary to this approach. In fact, the most data coming from the web are often unreliable, and its dynamism makes any attempts of study or analysis more complex. Moreover, errors and defaults may result magnified when many diverse types of data are analyzed together. Without a clear research goal and an adequate collecting plan, the danger of obtaining less significant or deceptive results is quite high.

These difficulties and costs of using big data risk to produce a digital divergence among those who can afford to process this data and create value from it and those who cannot take advantage of the information deriving from this data from the economical point of view. When you have the availability of such a potentially unlimited amount of data, the chance to find a crucial correlation between two sets can reach the 90%.

For this reason, it is fundamental to take care of the processing phase of the data derived from any survey or research that uses information coming from an online environment enormously populated by several and very different users. If companies are able to overcome this initial mistrust, big data can enable them to better collect, classify, analyze and process data from a specific sector, by providing them with valuable information which are further to the raw data.

1.4. Big Data for Tourism

Among the sectors where the revolution of big data has its biggest effects, one of most significant is the Tourism industry. In fact, the first companies to employ big data have been the Airports and the Airline companies. For example, British Airways in order to be more competitive decided to invest in knowing at the best its customers thanks to the collection of online and offline information about its fidelity plans. In this way, they can understand the most frequent needs and issues of travelers and how to develop more efficient and effective offers and solutions for customers. Other companies, such as Swiss Air, Air France-KLM and Lufthansa employ big data with the aim to improve their revenue management strategies, even some hotel chains have started to implement operations based on the use of Big Data. For example, Hilton introduced the strategy of the balanced scorecard to better understand which are the variables that drives the performance. Thanks to this initiative, it has been able to detect the correlation between the satisfaction of the customers and their behavior. Some other hotels, on the other hand, use platforms capable to encourage an effective management of the energy, in order to reduce the costs of 10-15 %.

Also the main OTA (online travel agencies) don't disregard the importance of Big Data: for example, Expedia is increasingly investing in this field because it is going to be the key point for the travel industry.

The potential for the Tourism industry is quite big: to collect, extract and well interpret the dataset which represents the behavior, the choices and even the "feelings" of the tourists will be operations focused on the analysis of further data which are not influenced by the habits and lifestyles, preferences and the real touristic flows. These will be very valuable information for those who will be able to exploit it.

Big Data for Tourism can also provide important information, not only about collective behavior of tourists, but even about the relationship among places, objects and people.

The Online Social Networks, for example, are not only a powerful instrument to market and commercialize touristic offers, but they also are an extraordinary source of information about the tastes of tourists, about their activities, or how they rate what they are offered. We can easily think of the value of the analysis of all those online platforms that host comments and posts of the users about their travel experiences, or of the surveys of their implicit "traces" they leave on the web during their travels. The aptitudes and the behavior of tourists are even more "social" and "digital". Among all the users who dispose of an internet access, 91 % (Ciccarelli, Scarsella) booked online at least one product or service during the last 12 months, and he or she uses the research engines as the main source through which to search and plan an holiday or travel; 42% of them use a mobile device (smartphones, tablets, etc.) to plan, book or look for info (33% only in 2012); 68% make researches online before choosing the place and the itinerary of its holiday. Not only: an internet access is confirmed to be fundamental for the tourist not exclusively in the choosing phase (61%), but especially during the planning phase (80%); once they arrived at their destination 58% of tourists uses online sources to rate activities and services around them, while 40% of them directly creates new contents and share them with other users. Furthermore, thanks to the available online information, it is possible to provide in a more precise way an evaluation on the effective consistency of the touristic flows, by analyzing their activities on the social media. It is not so difficult to understand that it is a great opportunity if we think that nowadays some institutions still measure the touristic flows in a traditional way, i.e. through the figures of visitors hosted in the "classic" accommodation, while they neglect the visitors staying in alternative accommodation such as private houses, couchsurfing, farmhouses or religious accommodation.

Being able to dispose of real-time figures allows to promptly intervene, find alternative solutions or to correct flawed situations, and even to predict possible future configurations. Tourists, as every human being, are aware or unaware "producers" of Big Data and of digital hints: a structured analysis of this data could represent a great utility predictive mean.

It's undeniable that this analysis cannot be exempt from criticism, in fact, the analysis of data coming from social media must be rigorous in order to avoid any kind of fallacy. Nonetheless,

it is clear that the great value added in terms of knowledge that is derived from the correct processing of such an amount of data: this means that thanks to this analysis complex events can also be understood by combining all the information derived from all the sources available. It is clear that the advantage for the travel industry is very huge: every hotel booking, every flight ticket, every car renting, every transaction or train ticket bought, almost every activity that comprise the use of a smartphone, GPS, credit card, etc., leave behind a great number of data which is really important. Even more, we witness that the planning phase of a travel is often discussed among travelers in dedicated areas on blogs, where people have the possibility to tell their personal experiences, to underline positive and negative aspects of a place or a service and they can share all of this with everyone around the world.
Big Data are indeed considered a great way to predict or affect behaviors, opinions and perceptions; after all, to better understand the travel experience of a tourist is critical to understand what a touristic offer must include, better or remove to be more desirable. The greatest advantage will be the opportunity to take real-time decisions, a very decisive resources in a sector such as the travel industry in which time can be a really precious factor.

Shortly, all the advantages offered by such an analysis are both a strategic approach , because Big Data allow to know the reputation of a specific accommodation, of a destination or of a specific service; and an operational approach, because all the information collected and analyzed can lead to the maximization of the tourists' satisfaction, by customizing the experience and the offer as well. All this information is crucial in order to improve the side of *reputation* for destinations and infrastructures.

1.4.1. OTA (Online Travel Agencies), Channel Manager, Metasearch

This new IT wave, that is characterized by Big Data, has totally revolutionized the travel industry in many ways, but the most visible one is the booking phase. Not only travelers use the web to search for the information about their destination, some of them organize their travels totally online. If before, people booked their accommodation by directly calling the structure or by simply sending an email or fax, now, thanks to all these new technological instruments allow to relay on intermediaries, such as *OTA, Channel Manager* and *Metasearch*, in order to book.

- ## OTA (online travel agencies)

OTA are intermediaries that act as physical travel agencies, and allow to hotels to increase their visibility among travelers and to customers to rate the best offers in a specific destination. They are a key point for the accommodation in order to sell their offers and services, for this reason it is critical to be present on the most influencing OTA. To achieve this, accommodation managers must put a great effort to timely update all the details of their offers, especially the price and availability. To support these activities, there are some instruments that help accommodation in succeeding in the OTA where they are subscribed, such as the channel manager. Thanks to this instrument, accommodation managers can update on only one channel, and they are automatically updated everywhere on the web, so that managers can save time in increasing the visibility.

The birth of OTA lead to the transition from a traditional vertical approach of supply chain management, characterized by few interaction between accommodation and customers, to a horizontal approach where OTA and social networks allow to much more interactions between seller and customer.

OTA allow even to customers to save time because they help them to compare prices and services offered by accommodation following their preferences and needs. Furthermore, these agencies offer to their users efficient booking engines, manage the credit cards banking commissions and invest in various types of online advertisement.
Obviously, accommodation managers have to pay commissions in order to collaborate with the OTA (the range of the commission is between 13 % and 20% out of the price. This fact is one of the reasons why some accommodation are reluctant to employ these online agencies, beyond the fact that some of their customers may prefer the traditional way to book their trip.

- ## Channel Manager

With the technology advances, OTA has increased in number. Now, accommodations on OTA has to update an increased number of online information about prices, offers and availability, called “extranet”. The increasing number of “extranets” can lead to an increasing possibility of error, for this reason there have been created specific software, *Channel Manager*, in order to help to reduce errors, to increase sales and to solve the updating issues for accommodation.

Channel managers can be defined as operating platforms that thanks to only one interface they can automatically manage availability, rooms, and prices on all the online reservation portals through a centralized database, connected to the OTA servers.
With this technology, accommodation are able to manage its offers and details through just one channel manager, and in this way they can also avoid the risk of overbooking in few steps, because after any room reservation the availability is automatically and timely updated.

- ## Metasearch

With the increasing online travel information, it is ever more difficult to compare data, prices and feedback. In this context, *Metasearch*, are here to help users to evaluate and rank reviews and tariffs on more websites at the same time.

The most famous and used metasearch are:

- ***Trivago:*** this is a metasearch focused on hotels and accommodation. The website is able to compare about 900.000 offers from 250 different sites, such as Expedia, Booking and Priceline. While, Trivago Hotel Manager is the platform used by hotel owner in order to improve their profiles, to read the clients’ reviews and even to be advised by expertise.

- ***TripAdvisor:*** this famous website provides people with all the reviews and feedbacks and allows them to compare prices and services. This metasearch also provides the hotel owners with an instrument, TripConnect, a subscription service that allow them to join the price comparison with the OTA.

- ***Kayak:*** this metasearch allows to compare hundreds of travel websites aimed at finding flights, hotels and holiday packages. One of the most interesting instrument it offers is Kayak Explore, that helps to choose the entire travel through the filling of search form with all the desired information and then it shows the results with the filters applied by the user. Also Kayak provides a service for hotel owners to directly propose their offers.

- ***Google hotel Ads:*** it shows hotel ads in its search engines in the foregrounds in order to help travelers to find information more easily and quickly and to allow to generate more transactions and reservations on the hotels' websites. It is possible to directly enter availability and prices in the search engines through the channel manager in partnership with Google.

1.4.2. Social Media Analytics for Tourism

The technological changes related to Internet, such as smartphones and tablets, have revolutionized the tourism industry from the brick-and-mortar and person-to-person service industry to a heavily digitally supported travel service networks.

Travelers now have much more control and more options in the planning, organizing and personalizing phases of their travels. They not only interact with a range of platforms and online intermediaries to increase their knowledge in relation to travelling and decision making in tourism, but they are also connected with other travelers who share their experiences. Travelers have access to online platforms to provide feedback and make recommendations for others.

Examples of new successful platforms that deal with tourism and travel are: TripAdvisor, Expedia, VirtualTourist and LonelyPlanet.

Information provided through these independent platforms has been found to be superior and more trustworthy compared with companies' websites and professional reviews (Akehurst, 2009; Gretzel et al., 2007; Rabanser & Ricci, 2005). In addition to professional systems, online social media, such as Twitter, Instagram, Facebook, FourSquare, and Google Plus, play a significant role in sharing travelers' experiences and points of view.

It’s apparent that online social media, travel professional websites and platforms, and blogs present inexpensive means to gather real, authentic, and unsolicited data on travelers’ opinions and preferences. Up to now, online social media are one of the best instrument to create a big amount of “Big Data”, that can be used in the Tourism Industry.

Social media data are increasingly used as the source of research in a variety of domains, and thanks to them our society is greatly connected, social media has 2.206 Billion active users with 30% global penetration (http://wearesocial.com/uk/special-reports/global-statshot-august-2015).

Social media activities are therefore an important class of daily activities performed by people worldwide to fulfill their social needs. These activities have generated lots of “social data”, which can provide meaningful and even possible real-time insights to a variety of studies. Social media are also leveraged as means of marketing strategies for many multinational companies, in order to improve their image and credibility worldwide. One of the best sector which can take advantage of them is the Urban science.

Unfortunately, the potential of social media data has been much less investigated so far in the study of tourism, despite the fact that tourism plays a key role in economic and social development of many cities. (Dhiratara et al.). Basically social media are different from the other data sources used for tourism studies, such as visitor surveys, transportation statistics, or online reviews. The latter require a considerable effort for data acquirement because they are infrequently updated, or require a large amount of volunteer inputs from online users, thus are highly sparse. Compared to them, social media data are more easily accessible in big size, in particular they are more precise because they are topically, spatially and temporally tagged, thus providing a unique opportunity for tourism researches.

On the other hand, the high availability of social media data raises some challenges. In fact, in order to retrieve relevant data, you should consider effective parameters and techniques to filters them, for example hashtags, keywords and geographic coordinates. For this purpose, the selection of parameters is fundamental and needs to be carefully designed to avoid bias in the interpretation of the results. For example, to capture the popularity of Eiffel Tower in Paris w.r.t. the number of visitors, we should avoid only using keywords or hashtags (e.g. #Eiffel Tower, #TourEiffel, or #Eiffel) to filter the data, as we cannot assume whether people posting tweets with these hashtags are indeed currently visiting that location. In this case, geographic coordinates become necessarily to be used as an additional parameter for filtering social media data. (Dhiratara et al.)

As we have said before, social media are useful in lots of domains and studies, in particular they can also be considered important in modeling users' attributes, including personal traits, their personality, their interests, etc.. in addition, social media are apparently effective for identifying social trends, such as hot topics, fashion trends, etc.. For this reason, social media are able to show the different ranks in popularity of monuments and of the main attractions of a touristic destination.

1.5. The Benefits of Big Data in Tourism

The main benefits of the use of Big data in tourism industry include:

- **Better decision support:** many travel companies are using big data no only to speed up decision-making process and data processing , but also to make better internal and customer-driven decisions. In some cases, these firms also benefit from an increased speed of data processing thanks to the new technologies employed. Most of the times, the relevant data is internal in the organization. The systems contain a variety of customer data, for example, that can be used to improve marketing strategies and service process (i.e. British Airways). On the other side, external data also offer the possibility of improving other types of decisions, with the benefits of efficiency and safety. For example, forecasting consumer demand could be better through the analysis of macroeconomics and weather data. In fact, many airlines companies consider the possibility of predictive maintenance of planes, engines and other equipment based on sensor data, but they are not yet pursuing such applications.

- **New products and services for customers:** one of the most fascinating possible benefits from the use of Big data is the creation of new customized products and services. Within the travel industry, the most likely creators of big data-based products and services are OTA (online-travel-agencies), travel search firms, and leading technology providers. For example, the travel meta-search website KAYAK has developed a predicted price offering. Since travel distribution is one of the most data-intensive area of the industry, it is possible that many products and services will address that area. It is also quite likely that external vendors will provide data-derived products and services that address operational process in travel industry. For example, vendors of energy management systems to the hotel industry could gather and manage “smart building” data, and optimize energy consumption. While, travel management service vendors could provide new data –based products and service to individuals and corporations

- **Big data helping players in the travel industry from better customer relationship:** since customer relationships have always been fragmented across a variety of systems and databases, data aggregation should create better customer relationships, and increased revenue from better-targeted products and services. Through predictive analytics, the most favored destinations, accommodation and dining preferences, service needs, and travel experiences can be identified for each passenger. Online analytical services like price prediction and desirability ranking can increase the likelihood of purchase. The most advanced approaches to customer targeting involve various forms of online travel advertising. Most of the travel providers employ some type of intermediary such as Facebook, to place their online ads and offers. In fact, a new development in online travel advertising is social-based advertising. For example, Facebook has not only a very large user base, but also the possibility to target ads across other social networks.

- **Cheaper, faster data processing:** new generations of information technology have always been adopted in part because they offer better price/performance ratios. Given the great amount of data travel companied have to look at, and the relative thin profit margins in the industry, the appeal of cheaper, faster data technologies is apparent. There are several cluster of community servers and open-source software that can process data at costs fifteen to twenty times lower than previous generations of data warehousing technology (Davenport). However, this kind of benefit is not easy to adopt for travel companies, at least outside of online businesses. For example, airlines and hotel chains depend heavily on Big Data for operations, and the new architectures may not seem as reliable as the previous technology generations. Furthermore, integrating the new technology with the existing “legacy” technology is really challenging for travel companies. Some of them, such as Air France-KLM, have begun to experiment with it and plan to use it.

- **Seeking multiple benefits:** while many organizations seek a singular benefit, there are, of course, firms that want all of them. Unfortunately, it is quite early in the Big Data era to undertake multi-pronged big data programs, but the most "aggressive" hotel chains which can afford it are doing so. (i.e. Marriot).

1.5.1. Consumer Behavior

At this moment, we are in time of unprecedented flux in consumer behavior, customer expectations, and company business models created by technologies that is disrupting and at the same time establishing new businesses. Tourism Big Data is showing significant changes in the relationship between businesses and their customers, in fact, they can use big data to provide superior buying experiences with a view to enhance customer choice and expectations. The "catalyst" of using big data to recognize the customer behavior is the pervasive use of mobile devices, apps, and other social media, which is playing and ever increasing role in the collection of raw information.

Big data keeps many insights into customers' behavior. The potential created by big data is particular acute in retail since businesses can exploit new communication channels, more service delivery options, and unprecedented sources (Marko,2015). Collecting, correlating, and analyzing tourism big data created from the interactions across all the different channels is the key to transform the customer experience from a nightmare to the key of success. The aim of using big data is to create an authentic emotional connection between customer and partners in the tourism industry in order to lead to a significant improvement in customer service and support. The exploration of tourism big data has huge implications and provides opportunities for the continuous meshing of customer experiences across mobile devices, websites, and personal interactions in multiple communication channels, such as mobile phones, pc and social networks. The real goal of using tourism big data is for the travel industry to be proactive, not just to provide an integrated service. They need to forecast customers' behavior and needs and provide them with solutions and various options.

1.5.2. Feedback Mechanisms

Nowadays, in the tourism industry feedback is essential in identifying customer preferences and deliver positive experiences. Soliciting feedback is extremely important to achieve company growth and to build a strategy around better meeting customer needs. The feedback based on tourism big data coming from tourists, employees, partners, suppliers and communities has also improved the capabilities of big data analytics. Data-driven businesses and consumer apps are the most common ways to collect feedback anywhere at any times.

The increase in gathering feedback using modern and advanced techniques has led traditional feedback marketing being progressively replaced by new commercial messages that are quick, unique, focused and personal. One of the applications of the feedback mechanisms is price adjustment, in which a change in travel demand detected from the big data analysis and forecasting can provide useful information for a timely and effective price adjustment.

Machine learning is one of the major techniques used in tourism industry to construct the feedback mechanism between customers and tour operators. (Bajari, Nekipelov, Ryan and Yang, 2015) For example, through cooperation between tour operators, financial institutions, and telecommunication operators, machine learning can identify whether a person has just changed her/his residential address or travel internationally by checking for unusual charges. Machine learning with big data on customer experience can enable travel companies and tour operators to proactively send text messages or calls to customers with new offers after they purchased some of their services. In particular, machine learning could modify the feedback system by identifying the users tasks and measuring their rates of success. By using this information, tourism companies can then provide solutions to correct inefficiency, customer frustration, and cross-channel breakdowns.

Predictive analytics are often as an effective solution for all problems for companies and can be incredibly useful. The predictive analytics with tourism big data used in modern feedback mechanisms represent a major improvement over the traditional human feedback. Predictive analytics can give marketing professionals more insights into customer preferences, which can be used to better understand customers and improve targeted sales. However, the success of this analytics depends on both the quality of big data and on the efficiency of feedback mechanisms.

Customer feedback mechanisms must be well designed and comprehensive to deliver useful data in a timely manner in order to act upon immediately. Timely and reliable tourism big data can provide a rich portrait of customers and potential customers, and consequently lead to marketing efforts more precisely targeted toward the most fruitful channels. (H. Song and H. Liu).

1.6. Pros and Cons in the use of Big Data

Considering all we have described so far, it is apparent that a correct use of Big Data requires, besides an intensive training of the personnel, the planning of an investment aimed at fill the generational, technological and mental gap which affect all the users. Knowing and being able to interpret this size of data available on the web lead to considerable advantages from the point of view of competitiveness and efficiency of the companies' management in the tourism sector.

However, even though all these datasets are available it doesn't mean that they automatically have always reliable statistical figures in the economic field; for this reason it is very important that this data must be treated in a very rigorous manner and that its value must be verified, both from the technical and ethic point of view. In fact, it is only thanks to complex analysis that it is possible to detect hidden models, unknown correlations and to obtain further information that is not immediately recognizable simple looking at the array.

There also are some misleading ideas around the concept of Big Data: any analysis which can lead to accurate results; any piece of data can be analyzed by rendering the traditional techniques of processing obsolete; it is not necessary to detect the cause and effect of the variables, because the their correlations are enough to tell everything we need; we don't need to use sophisticated models because the big size of data is already revealing lots of information. But, as the best statistician know it is better to avoid sampling errors than to have a big-sized sample to analyze. Fortunately, it is possible to measure both the figures in order to select which is the most relevant in a sampling operation.

We often forget that, even though we could extract all the helpful information referred to a specific object from social network like Facebook, it is not consequential that the more the size of data is big, the more reliable it is. This happens because all the Facebook users cannot be really considered a realistic sample displaying the real population and for this reason is not a reliable representative.

The real biggest issue, maybe, is that the actual methods don't seem to be able to allow an effective and efficient use of Big Data. We are still far from quitting the traditional collection and procession techniques of statistical data that will lead to fulfill requirements of relevancy, reliability and replicability, which are fundamental in order to get quality information, important for the decision-making process.

What is written above doesn't actually mean that Big Data analysis must not be taken in the right account; on one hand the advances in technologies will lead to an improvement in the techniques and technologies useful for a correct analysis; on the other hand it is a big error to neglect the big importance of this huge amount of data, that from the point of view of some aspect cannot be treated with traditional methodologies. It is increasingly important to get to know both the potentiality and the limits of Big Data, because only in this way it is possible that qualitative leap of the simple collection of information.

1.7. Big Data Challenges for Tourism Industry

Big Data is arguably the biggest opportunity for travel businesses to adopt the changing structure of data and maximizing its use (Davenport). It offers the potential for an important shift for all travel companies, leading them to enhance both the business and the experience pf travel. As any other technological change, this revolution will also drive to a significant disruption, which consequently brings many challenges for the industry.

What's make Big data such a powerful idea? First, we can say that big data provides insights that could help to deliver a more intelligent and efficient travel experience to tourists than has ever been possible before. In the past, structured data has always been divided between different "silos", systems or companies, but now the availability of both structured data and unstructured data promises a more integrated view of this industry. Thanks to this, travel companies have the chance to enhance their offers, prompt innovation and build better relationships with their customers. For sure, Big data can help to make travel more responsive and focused around travelers' needs and preferences.

In order to enhance this industry, Big data needs creative ideas and courage. Managing and analyzing big data is no longer an issue for IT department alone, indeed it need the collaboration of all travel industry departments to address all the biggest challenges of this new technological wave: from the technology complexity; data accuracy and rights of use; business and technological alignment; to the need for data specialists and training. (Hervé Couturier)

Many of the specific challenges in Big data that the companies have to face results from its long-term usage of information systems for key processes. One consequence of that is the fragmentation of key data across multiple functions and units of the organization. For example, airline data on passengers experience is spread across flight operations, baggage, loyalty programs, complaint databases, and external sources like social media. In order to make effective decisions on how to promote their offers to customers and to recover from service failures, airline companies need to combine all of this information into one data warehouse and one dataset. This would require a considerable investment over technology.

Unfortunately, creating an integrated source of customer information is not only costly, but somehow difficult no matter how large is the available budget. A further challenge is also the privacy issue. Individual customers typically have several different identities across different systems. However, third party data providers can assist with this data integration problem.

Another result of the long-term usage of information systems in large travel companies is that Big data technology architectures have to coexist with existing hardware, software and databases. These “legacy” tools and the data they contain are still necessary, and will still be useful in analyzing and improving travel operations and customer relationships. Big data technologies may be the only technologies available for some startup and purely online travel companies, but large and established companies will have a kind of “hybrid” environment. This, of course, will lead for challenges of IT architectural cohesion and efficient functioning of all new and old systems.

Some other big data challenges are not specific for the travel industry, but will nonetheless pose obstacles to firms which plan to take advantage of the Big data revolution. Another challenge/ issue is the skills shortage for people who are skilled in big data manipulation. They are often called “data scientists”, they have not only data management and programming skills, but also the ability to analyze and understand business processes problems. Since there have been no formal training programs for such specialists in the past, many who currently perform the role have scientific or IT backgrounds and Ph.Ds.

Although mainstream travel companies have employees with analytical skills in areas such as revenue management, they may not be familiar with the technologies used for Big Data.
Lastly, another challenge for businesses is the difficulty to maintain a sustained competitive advantage from the use of Big data. For example, some U.S. based airlines developed an early advantage from the exploitation of Big Data in areas such as revenue management and customer loyalty. But nowadays, such programs are widely distributed throughout the airline industry and very common in hotel chains and railways as well. For this reason, maintaining a competitive advantage requires endless innovation, unique data types and a creative and updated experimentation with the new technologies used for the analysis of the data.

CHAPTER 2

2. Examples of big data use

Since the revolution of big data has touched many sectors, it brought several advantages for many companies worldwide. In the following part, some example of how companies, and in particular some among the biggest international travel companies, benefit from the employment of big data in their business activities and processes.

2.1. Companies using big data and big data analytics

Before presenting some example of big companies that started to use big data to increase their performance, we would like to emphasize the potential of big data for so many companies by describing the findings of a survey conducted by one of the biggest consulting and technology services company, Accenture.

The survey was carried out in 2014 with more than 1000 respondents from companies operating across seven industries and headquarters in 19 countries that had completed at least one big data implementation. More than 4,300 targets have been screened, but 36% did not complete nor were still implementing their first big data project. As a result, a total of 1,007 organizations completed the survey.

According to this survey "Big Success with Big Data", organizations that are using Big Data today reports overwhelming satisfaction with their results, and see Big Data as a catalyst for their transformation as digital enterprises.

The main key points of this survey are:

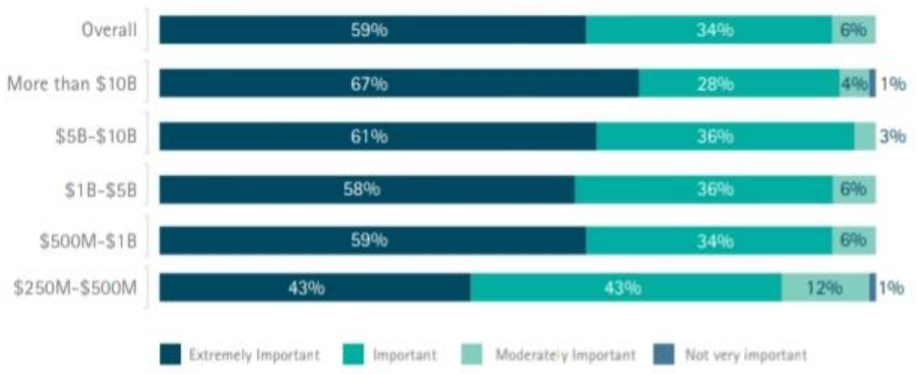
- Big data is taking off: users that have completed at least one project are very satisfied with their initial adventure with big data. The vast majority of them that are satisfied with the business outcomes they are having, also report that their big data initiative is meeting their needs.

- Bigger companies are getting more from big data: in fact, the bigger the company, the better the results, because they have more availability in terms of resources. Larger organizations are leading the way by starting with focused initiatives, rather than trying to do everything at once.
- Big data demands broad learning: users often begin their projects with big data thinking it will be easy, only to discover that there is a lot to learn about data as an asset and especially about analytics.
- Help needed: with big data analysts in short supply, successful users source skills wherever they can find them, leaning heavily on external expertise.
- Big data is definitely disruptive, and potentially transformational. The consensus is apparent: big data brings disruption that can revolutionize business in every form.

The vast majority of all users in this survey (92%) report they are satisfied with the business outcomes, and 94 % feel their big data initiative meets their strategic needs. Organizations perceive big data to be critical for a wide range of strategic corporate goals, such as new revenue generation and new market development to enhance the costumer experience and improve the overall performance. Especially larger companies are more likely to consider big data as extremely important for their digital strategy (Figure 1)

Figure 1: Importance of big data

How important is big data to your organization?



'Source: http://smartdatacollective.com/bigdatastartups/201286/why-ups-spends-over-1-billion-big-data-annually

Source: Accenture Big Success with Big Data Survey, April 2014

2.1.1. Winning big by thinking big

Large companies appear to be the biggest beneficiaries of big data implementations for some reasons:

- They have a deeper understanding of big data's scope and value.
- They put a serious focus on practical applications and business outcomes.
- They have greater commitment in budget and talents.
- They have a keener appreciation of the importance and disruptive power of big data

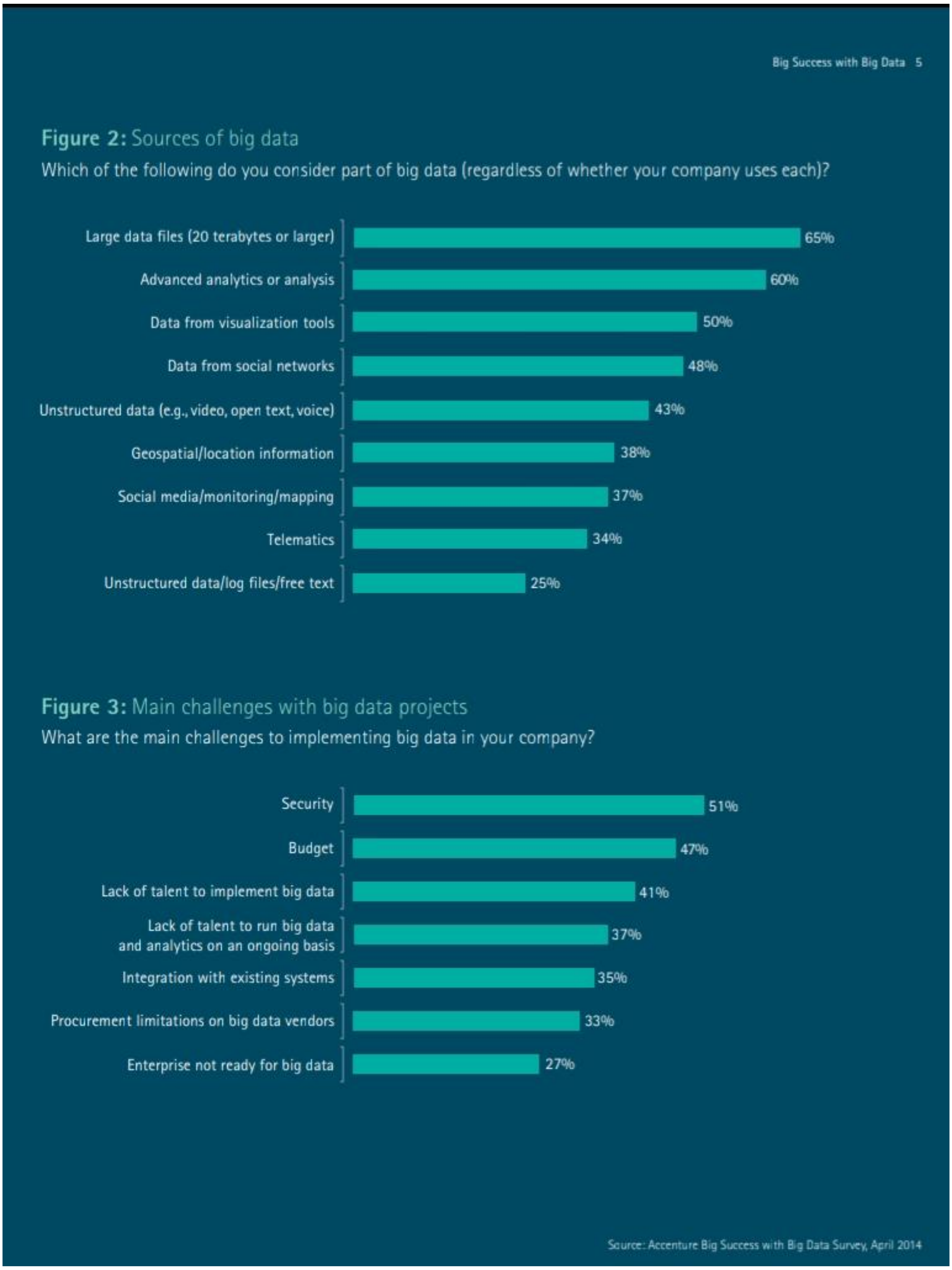
Another plus for big companies is the fact that they start thinking small, i.e. they initially focus on realistic objectives and feasible expectations. That means they focus their resources around a specific area at a time, rather than attempting to do all at once. The mantra for them could be "start local, end global".

2.1.2. Big data demands broad learning

Big data clearly poses some challenges for companies. For example many companies may have different definitions of big data and may not be patient expecting outcomes that need more time or more effective initiatives.

There also are companies that are reluctant to implement big data analytics because they distorting perceptions about the scope and benefits of big data. More than one third of users (36%) think big data always requires an extremely big investment. A roughly equal percentage (37%) thinks organizations can for sure achieve extremely large cost-savings with big data. One in four (26%) believe companies are required to implement big data all at one across the enterprise.

Many initial users imagine big data initiative will be easier until they face multiple challenges, from security and budget to talent (Figure 3). In fact, more than four in ten (41%) reported a lack of appropriately skilled personnel, and almost as many (37%) felt they did not have the talent to employ big data and analytics on an ongoing basis.



Big Success with Big Data 5

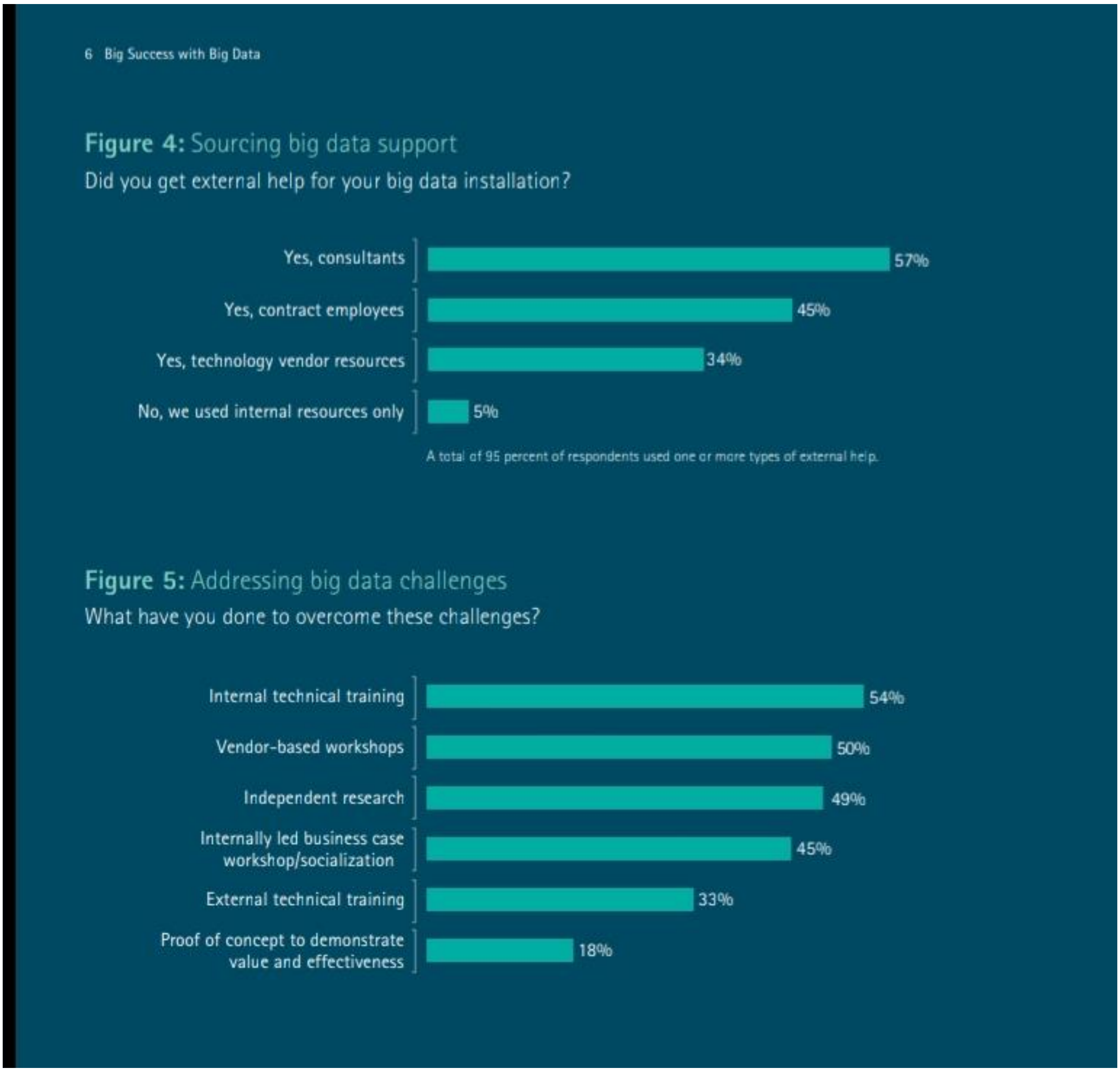
Figure 2: Sources of big data

Which of the following do you consider part of big data (regardless of whether your company uses each)?

Figure 3: Main challenges with big data projects

What are the main challenges to implementing big data in your company?

Source: Accenture Big Success with Big Data Survey, April 2014



6 Big Success with Big Data

Figure 4: Sourcing big data support

Did you get external help for your big data installation?

A total of 95 percent of respondents used one or more types of external help.

Figure 5: Addressing big data challenges

What have you done to overcome these challenges?

2.1.3. Help needed

With so many companies simultaneously competing in big data skilled resources, sourcing talent is undeniably hard. In fact, more than half of respondent (57%) leveraged the help of consultants, 45% used contract employees and 34% used technology vendor resources (Figure 4).

Organizations that relied on consultants, contractors or other external resources found their big data installations to be easier than those using internal resources.

In order to face the talent shortage and other challenges, many companies seek new kinds of strategies (Figure 5).

Nearly all (91%) companies expect to increase their data science expertise within the following year.

Understanding business use cases and data usage patterns provide a crucial help to appropriate solutions, technologies and approaches that will be used to deliver results. Multiple solutions exist for each big data challenge, so that it is fundamental for companies to be open to new possibilities and scenarios, and become a learning enterprise by testing extensively, learning what is better and then refining to go forward.

2.1.4. Big data's disruptive potential from the survey

Analyzing the expectations among survey respondents the main idea that conveys is the "potentially life-or-death" competitive threat. A vast majority of users (89%) believe big data will revolutionize business operations in the same way the Internet did (Figure 6). Nearly as many (85%) feel big data will drastically change the way they do business.

Almost eight in ten users (79%) agree that "companies that do not embrace big data in their strategy will lose their competitive position and may even face extinction". Even more (83%) of them have pursued big data projects in order to reach a competitive edge.

Perceptions about the disruptive power of big data lead early adopters to rapidly move in that direction as they see a competitive advantage in it, and they will move to disrupt their own data practices rather than letting their competitors beat them.

All of this helps to understand why users have such a strong expectation to increase their data science expertise as soon as possible withstanding huge investments. 91% of users report plans to build out or increase their current data science expertise soon, and the larger the company the sooner they plan to invest, 69% within the coming year for companies greater than $10 billion (Figure 7).

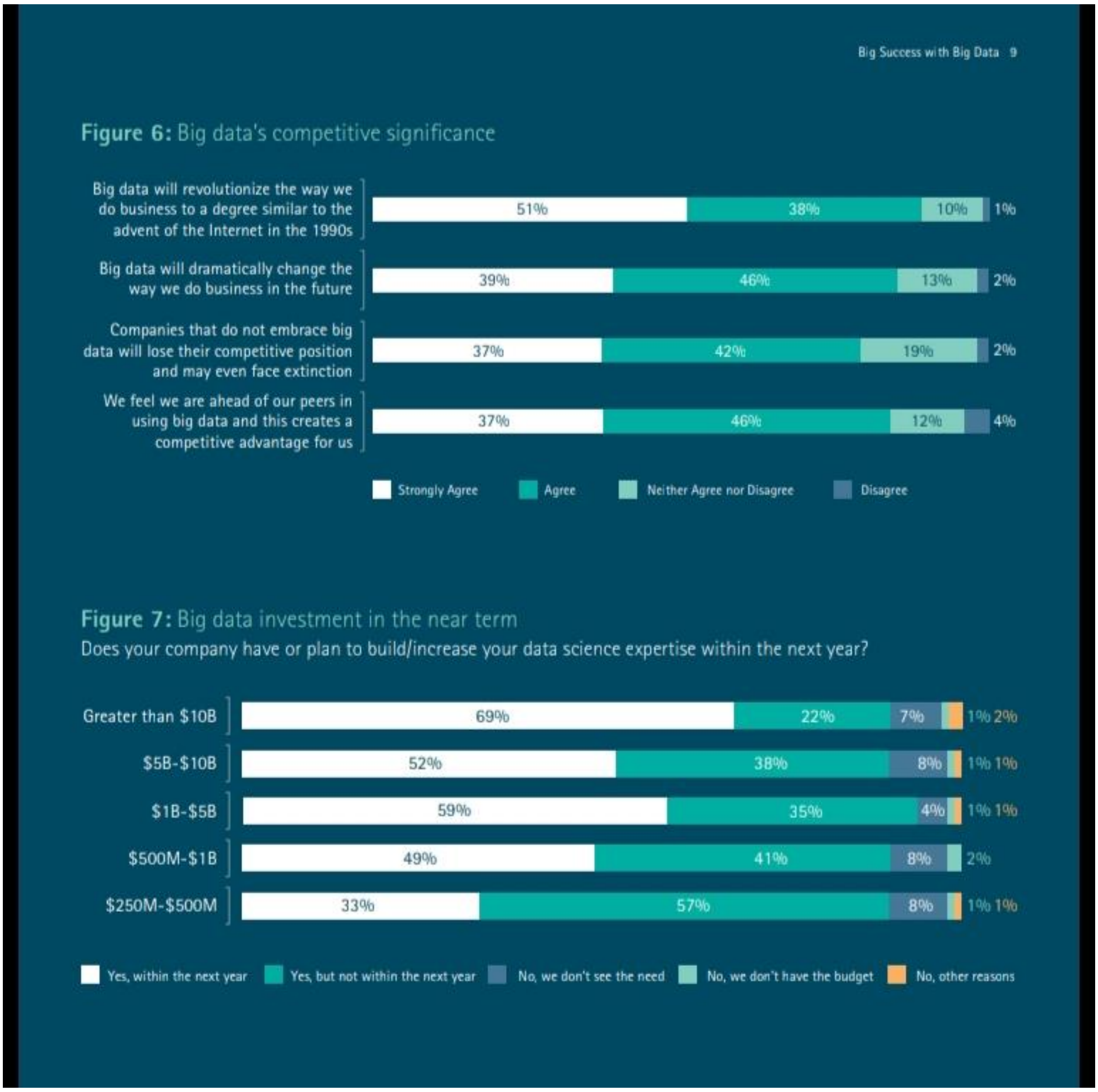


Figure 6: Big data's competitive significance

Figure 7: Big data investment in the near term

Does your company have or plan to build/increase your data science expertise within the next year?

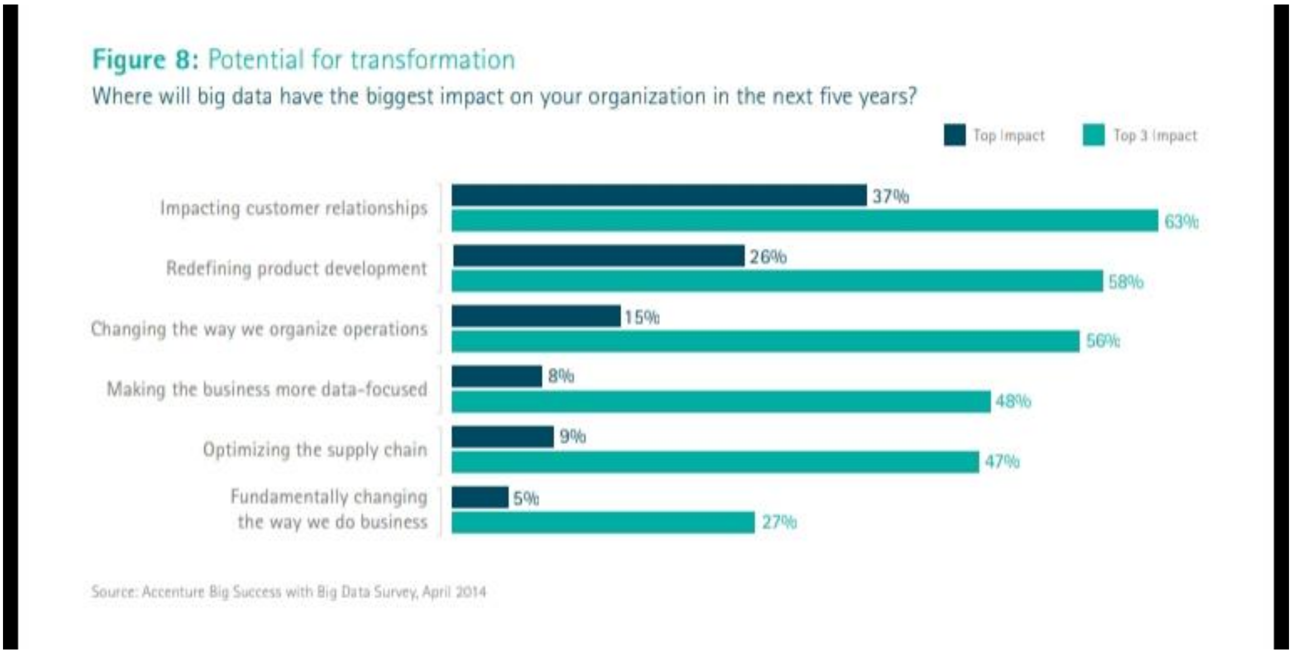


Figure 8: Potential for transformation

Where will big data have the biggest impact on your organization in the next five years?

Source: Accenture Big Success with Big Data Survey, April 2014

2.2. Travel companies use Big Data

After seeing in the survey conducted by Accenture that the big data revolution is impacting on several industries and companies, which think to improve their performance thanks to this new and crucial Big data analytics, we want to present some example of big travel companies who are acting as early users in this field.

We are now going to present some business case of travel companies that started to use big data analytics in their operations.

2.2.1. British Airways

British Airways is the flag airline and the largest airline company in the United Kingdom based on fleet size, or the second largest after Easy Jet if we consider the number of passenger carried.

Facing competition from low-cost travel companies on the low end, and country companies backed by sovereign wealth on the high end, British Airways has focused on achieving a competitive advantage through customer insight. Thanks to its technology, it had accumulated

substantial customer information from a program called Executive Club Loyalty program and its official website, and then it had incorporated the data into a customer data warehouse for analysis. The company decided to put customer data to work in its Know Me program. The specific goal of this program is to understand customer needs and features better than any other airline, and to put the customer knowledge accumulated across tens of millions of touch points to work for the customer's benefit.
BA data analytics team is enhancing its big data program with the support of a big data analytics firm, Opera Solutions. The airline company is extracting and using data in order to apply it to customer decision making, stating that the three key points of the Know Me program are:

-Personal recognition: this aspect of the program involves the process of recognizing customers for being loyal to BA, and expressing appreciation with targeted benefits and recognition activities;

-Service excellence and recovery: BA will track the service it provides to its loyal customers and attempt to keep it always at a high level. Given the problems in air travel, BA also wants to understand what problems and particular situations its customer experience, and do its best to recover a positive overall result;

-Offers that inspire and motivate: BA's customers are busy people who don't have time for irrelevant offers, so this pillar of the program analyses customer data to construct relevant and targeted "next best offers". BA hopes that its customers will consider it more like a service to better their travel experience than a mere marketing program.

The information to support these objectives is integrated across a variety of systems, and applied to real-time customer interactions at check-in locations and lounges. Even on BA planes, crew has iPads that display customer situations and authorized offers. Some aspects of this new program have already been rolled out, while others are still under development. Initial results are very positive and customers are pleased by BA's understanding of their air travel needs.

2.2.2. Marriott

Marriott is an American multinational diversified hospitality company that manages and franchises a broad portfolio of hotels and related lodging facilities. With its 6,500 properties in 127 countries, it is the largest hotel chain in the world. (Wikipedia)

Marriott is also one of the first hotel chains that adopted analytics in the form of revenue management, beginning about 25 years ago. Revenue management is the process by which hotels establish the optimal price for their rooms. If an hotel is able to predict the optimal price at which to fill all its rooms, it will make more money and become more efficient. And if an hotel management company like Marriott can persuade property owners that they will gain more revenues using the Marriott brand than with that of its competitors, they will tend to adopt it even more.

In order to improve its revenue management processes, Marriott combined two separate systems, made revenue management accessible over the Internet, improved revenue management algorithms, made the system work faster so that revenues could be optimized more frequently, and extended revenue management practices into restaurant, catering, and meeting space areas. These capabilities are used by a global team of corporate, regional, and local “revenue leaders” who have the tools to measure the effectiveness of their decision making and overrule the system’s recommendations when there are local factors that can’t be predicted.

The hotel chain also uses analytical approaches to the offers it makes to its frequent customers, and to understand their possibility of staying with Marriott or defecting to competitors. The company was also an early adopter of web analytics, and uses A/B and multiple testing to improve its website. At last, Marriott has experimented for several years with a variety of customized options for visitors to the website. This is its first foray into Big data usage.

Recently, Marriott has been analyzing big data from its website activity to create a strong marketing attribution model. Its ultimate goal is to understand which sales and marketing activities really drive the sale to a customer.

2.2.3. Munich Airport

Among travel companies using big data to increase their efficiency and performance, we may find Airports as well who attempt to exploit the potential of big data , as we will see in the case of Munich Airport.

Munich Airport is the second busiest airport in Germany and a hub for Lufthansa. The airport's management has a goal of facilitating "seamless travel" for its passengers. Since passengers are customer for airline companies but are not generally known to airports, and since almost half of airport revenues come from retail, food and beverage sales, and parking, the airport would like to know more about its passengers and eventually hopes to individualize services for them. Actually, this would constitute a loyalty program for airport customers.

In addition to being an airline hub, the airport is also a center of rail and automobile service. Michael Zaddach, the Chief Information Officer of Munich Airport, would like to integrate information on the multi-modal travel plans of passengers. The airport hopes to be able to offer navigation from passengers' home to his gate, or from the gate to other modes of transport. These initiatives are being planned but unfortunately are not yet implemented.

For sure, every trip that starts at an airport ends at another one. Nowadays, every airport is developing its own smartphone apps. Munich would like to develop solutions that can be used at or with other airports and transport modes. In order to do that, Munich airport is working with other airports to develop standards for passenger and journey information. The airport is also working with the airline company Lufthansa and Amadeus (Amadeus IT Group is a major Spanish IT provider for the global travel and tourism industry) to explore approaches to sharing statistical data to implement the customer information system planned.

2.2.4. France Airways-KLM

Air France-KLM is a world leader in its three main business line: passenger transportation, cargo transportation and aeronautics maintenance. With its 90 million customer per year and nearly 2.5 million visitors on the website each month, customer data processing is a key issue for the Air France-KLM group.

Even for Air France-KLM , it has proven difficult for the airline to set itself and its prices apart from low-cost companies. Making its products stand out against those of tis Asian and Gulf competitors has also been a challenge. The real challenge no longer lies with customization of the customer experience, but with a hyper customization. The ultimate goal has become to meet passengers specific travel needs, as it happens with BA as well.

Within a few years, the amount of data available to airline companies has exploded. Sites and applications also generate several interactions. While the group began collecting customer data many years ago through call centers, social networks and its staff at airports, lounges and on planes too, the data collected until now has not been yet centralized. Therefore, the first big challenge was to combine all customer data on a common platform for all Air France-KLM businesses so they can be redistributed in real time to all customer service points.

The second big challenge for the French company was data management. They wanted to ensure data quality and to respect the privacy of their customers, at the same time offering them a clear benefit. For this purpose Air France-KLM relied on an external resource, Talend, that is an American enterprise data integration software vendor

Thanks to this collaboration, the airline company may collect and process personal data concerning passengers (PII-personally Identifiable Information) who use the services available on its website, its mobile site and its mobile applications, while respecting the privacy of its customers. " Air France-KLM can locate customer data, determine its origin and destination, and share the information within the company ten times faster than before" (Damien Trinité-CRM Big data Project manager). In the field, call centers agents were the first to take advantage of this data management solution.

For what regards the improvements on the website side, an engine has been designed to recommend destinations to its visitors. Based on the pages consulted on the website, Air France-KLM created algorithms able to help offer customers promotional rates for their next preferred destinations.
But the biggest revolution to the customer experience could come from "bots". In fact, KLM is one of the first travel company to have launched its conversational robot on Facebook

Messenger, which gives access to all travel information (boarding passes, flight status, etc.) the idea is “to be where customers are” (Gauthier le Masne-Chief Customer data Officer).

CHAPTER 3

3. Techniques to process big data in tourism

One of the main technique used by software companies to "process" Big data is the Sentiment analysis, in order to detect general customer satisfaction for each facility. These companies possess a *semantic engine* able to "read" each type of insight found online, to detect the subjects included, and to understand the opinions linked to each detected subject.

An opinion or judgment affects the Sentiment rate on the basis of various elements: **1)** quality and the subject which is referred (room, personnel, Wi-Fi, services, etc.); **2)** connotation of adverbs, verbs and adjectives used in reviews; **3)** rating given by the reviewer; **4)** weight based on the positivity or negativity of the connotation of verbs, adverbs and adjectives used.

Insights are dissembled in multiple clusters of topics, in which it is possible to analyze the elements that positively or negatively affect the user's experience. An insight may be considered to be positive when the "Sentiment" of all the expressed opinions reaches at least 55% of positivity; while it is negative when the "Sentiment" is below 55%. The final semantic analysis is synthetized in the percentage of the positive "Sentiment" and in the general value of customer satisfaction.

Another technique used to extract data from websites or webpages in order to collect the most useful information from a big array of data is the so-called Screen Scraping. This can be a very useful tool to manage big data found online in a very easy and accessible way and it further allows to categorize huge amounts of data.

3.1. Sentiment Analysis

Big data analytics aims to generate new insights that can be very valuable and often in real time, complement traditional statistics, surveys, and archival data sources that often remain largely static.

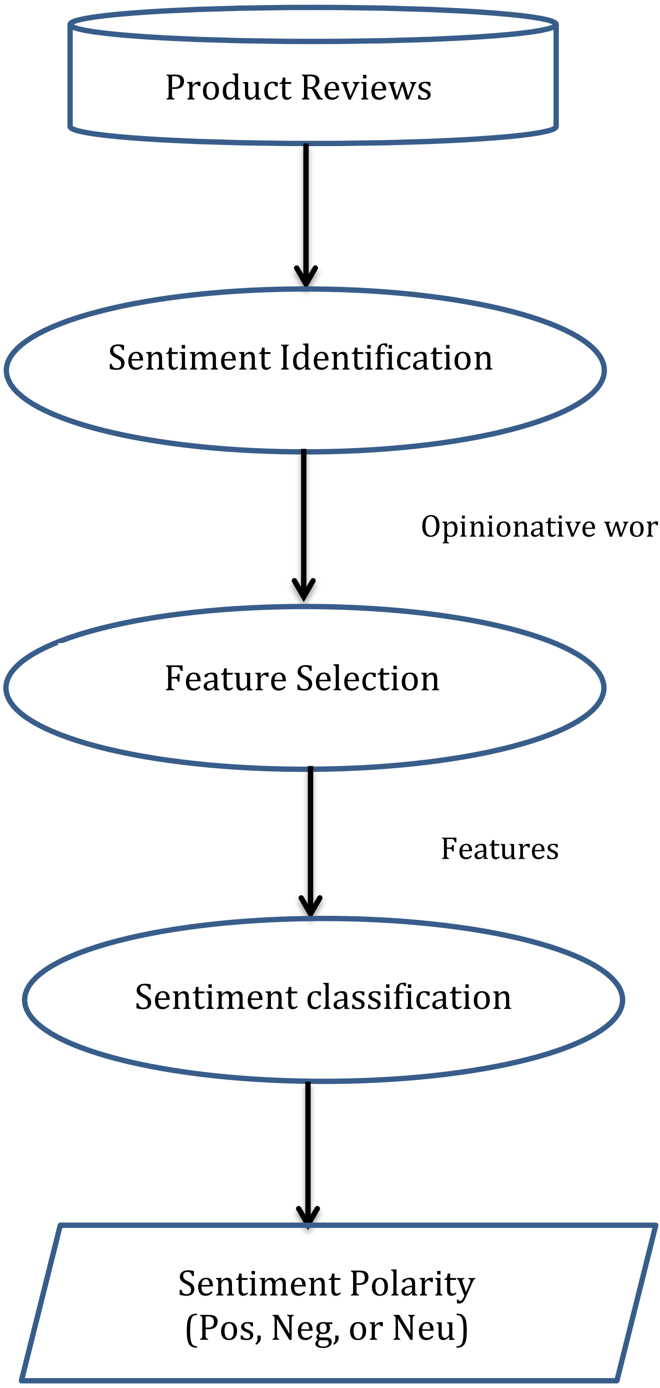
Mining social media and consumer-generated contents on the internet has attracted much attention for their value as public and community data (*George et al.,2014*). For example, according to some research in the literature, online consumer reviews can be used to predict product quality (*Finch, 1999*), stock market volatility (*Shumaker and Chen, 2009*), and box office sales for the movie industry (*Duan et al.,2008*). It also has been found that online news postings have sufficient linguistic content to be predictive of a firm's earnings and stock returns (*Tetlock et al., 2008*). Few years ago, *Ghose* and *Ipeirotis* (2011) used text content and reviewer characteristics to estimate the helpfulness and economic impact of online hotel product reviews.

Due to the volume and unstructured nature of social media and consumer-generated contents, opinion mining or sentiment analysis, the so called text analysis, started to play an important role in big data analytics. In particular, sentiment analysis techniques for extracting opinions from unstructured human-generated documents, that will be explained later in this chapter, can be excellent tools for handling many business intelligence tasks, including reputation management, public relations, tracking public viewpoints, and for market trend predictions too. Since sentiment analysis employs the same techniques that are derived from and based on the natural language processing (NLP), information retrieval (IR), information extraction (IE), and artificial intelligence (AI), we can say that compared to a broader scope of text mining approach, sentiment analysis may be considered a special type of text mining with the focus on the determination and identification of subjective statements.

While research on sentiment analysis goes back to the 1970s, it has received increasing attention from both researchers and practitioners only recently (*Brob, 2013; Pang et al., 2002*). This vivid interest is driven by an escalation in the amount of web- and social media-based information, the evolution of new information technologies, especially the machine learning approaches for text analysis, and by the development of new business models and applications that leverage this type of information.

In the review "Sentiment analysis in Tourism: Capitalizing on Big Data" by Alireza Alaei and Susanne Becken from Griffith University, Australia, it has been argued that sentiment analysis can become a very important tool in tourism research. Moreover, it may be that data-driven research models might be very relevant to tourism research.

Sentiment analysis process on product reviews



3.1.1. What is Sentiment Analysis?

Sentiment Analysis, or Opinion Mining, is a contextual mining of documents which identifies and extracts subjective information in source material, and helps businesses to understand the social sentiment of their brand, product or service while monitoring online conversations.

The basic task of Sentiment analysis is to classify the polarity of a given text at the document , sentence, or feature/aspect level, whether the expressed opinion in a document, a sentence or an entity feature/aspect is **positive**, **negative**, or **neutral**. An advanced version of this type of analysis, "beyond polarity", is the classification looking, for instance, at emotional states such as "angry", "sad", and "happy".

Sentiment analysis, in particular in relation to customer reviews, is built on the premise that information provided through text (e.g. a review) can be either subjective (i.e. opinionated) or objective (i.e. factual). Subjective reviews are based on opinions, personal feelings, beliefs, and judgment about entities or events. Objective reviews are based on facts, evidences, and measurable observations (*Feldman, 2013*). Consumer reviews and social media posts often reflect happiness, frustration, disappointment, delight and other feelings (*O'Leary, 2011*). Tapping into these large volumes of subjective eWOM (electronic word of mouth) is of great value to tourism organizations and businesses who seek to improve customer management and business profitability (*Choi et al., 2007; Kuttainen et al., 2012; Ye et al., 2009*).

From the methodology point of view, sentiment analysis represents a polarity classification. Considering different numbers of classes, sentiment polarity classification can be conceptualized as *binary*, *ternary* or *ordinal* classification. In a binary classification, we initially assume that a given customer review is subjective. In other words, a binary classification assumes that the given text is mainly either positive or negative, and then it determines the polarity of the specific review as 'positive' or 'negative'. The definition of the two poles of sentiment as positive and negative depends on the particular application and domain to which is applied. For example, in the context of tourism, 'positive' and 'negative' may, respectively, refer to “satisfied” and “unsatisfied”, but further research to link sentiment polarity to the theoretical constructs of satisfaction should be investigated.

Reviews may not always be subjective, therefore, the binary classification needs to be extended to a ternary classification that contains a third, ‘objective’ category. In the ternary classification problem, the classifier implicitly performs a classification to differentiate between objective and subjective sentences, providing a class-label as ‘*positive*’, ‘*negative*’, or ‘*neutral’*. Neutral polarity is sometimes interpreted as a polarity between positive and negative. The sentiment analysis can also be treated by the means of a cascaded approach, composed of a binary classifier to differentiate between subjective and objective reviews and a binary polarity classifier to further classify subjective reviews into two groups, namely positive or negative. Objective reviews generally do not contain those words that are clearly defined as positive or negative in a dictionary. They may also contain mixed polarities without a clear perspective of direction. In addition to the simple binary and ternary classification, ordinal classification can be performed by the means of a rating scale (e.g., 1 to 5 stars) of the sentiment strength (*Brob, 2013*).

There are three main classification levels in sentiment analysis: document-level, sentence-level, and aspect-level SA. Document-level sentiment analysis aims to classify an opinion document as expressing a positive or negative opinion or sentiment. It considers the entire document a “basic information unit”(talking about one topic). Sentence-level sentiment analysis aims to classify sentiment expressed in each sentence. As mentioned early, the first step is to identify whether the sentence is subjective or objective; if the sentence is subjective then sentence-level SA will determine whether the sentence contains positive or negative opinions. However, there’s no fundamental difference between document and sentence level classifications because sentences are just short documents. Since classifying text at the document or at the sentence level does not provide the necessary detail needed opinions on all the aspects, to obtain these details we need to consider the third level: the aspect level. Aspect level SA aims to classify the sentiment with respect to the specific aspects of entities considered in opinions. The first step is to identify the entities and their aspects. The opinion holders can give different opinions for aspects of the same entity, and they can express opposite sentiment. For example in the sentence “*the voice quality of this phone is not good but the battery life is long*” the same entity has aspects that are judged with opposite sentiment.

Relatively less research has focused on sentence level analysis, since it is more challenging to accurately extract polarity from a small number of words compared with paragraphs and documents *(Brob, 2013; Choudhury, 2016; Höpken et al., 2016; Schmunk et al., 2014; Ribeiro et al., 2016).*

In sentiment analysis, it is also important to understand what a sentiment relates to. The determination of a target and aspect (i.e. **topic detection**, *Menner et al., 2016*), relates to determining the subject of a sentiment expression. Sentence-level sentiment analysis supports aspect-based review mining. Based on the level of analysis, a sentiment aspect may refer to a concrete or tangible entity or to a more abstract topic. A target or an aspect might be referred to either implicitly or explicitly. Reviews with explicit targets or aspects are easier to analyze than those with implicit ones. A hotel review may be composed of different aspects of a hotel, for example, "the size of the bed was small and there was a noisy refrigerator" is a review, which explicitly describes two aspects of a "hotel room" as "small bed" and "noisy". Whereas in the review "hotel was expensive!", the word "expensive" is an implicit aspect that refers to the "price" of the hotel. In particular, by extracting both implicit and explicit aspects in reviews results in an increase in the accuracy of sentiment analysis results. A comprehensive sentiment analysis also includes data on who provided the information and at what point in time.

For a clear explanation and understanding of the different sentiment analysis methods, the relevant key terms are defined in Table I.

Table 1

Table 1 Key terms and definitions

Key term	Description
Aspect	Every topic or target in sentiment analysis has different features and characteristics. For example in tourism-related text, 'restaurant' as a potential target has various aspects, such as the food and atmosphere, ambiance, cleanness, price, and location.
Bag of Words (BoW)	The BoW is a feature extraction method where the frequency of occurrence of each word in a given text/review, disregarding word order and grammar rules in the text, is used as a feature.
Classification and classifier	In machine learning, classification is the procedure that helps identify to which set of predefined groups a new sample belongs to. The model, which is called classifier, needs to initially 'learn' based on a training set of data that contains instances of text (or individual words) that are representative of a particular group. Once trained, the classifier can then perform the classification task on a new sample.

Confusion matrix It is a table used to describe the performance of a classifier on a set of test data for which the true labelled are known.

Experimental analysis To evaluate the performance of an algorithmic model, a set of tests/experiment is performed using training and testing data. Considering the results obtained from the test data, evaluation metrics are also computed. This process is called experimental analysis.

Feature extraction Feature extraction is the process of building or deriving a set of discriminative, informative and non-redundant values from a set of data, which eventually facilitates the learning process.

Information Gain (IG) IG is a feature selection strategy, which uses more important features or more discriminative features for the classification purposes.

Maximum entropy Maximum entropy is a classifier, which mainly relies on the concepts of data uniformity and entropy. In the maximum entropy classifier, it is assumed that the probability distribution of the prior data that best represents the current state of data/knowledge should have the largest entropy.

N-gram An N-gram is an adjacent order of N items in a given text (review) or speech. In a text (review) the items can be letters or words.

Naive Bayes Naive Bayes classifier is a probabilistic classifier which works based on a strong assumption that features are all independence.

K-Nearest Neighbour (K-NN) K-NN is an instance-based and non-parametric classifier used for classification, where K denotes the K closest training samples. The K-NN algorithm is one of the simplest machine learning algorithms.

Part-of-Speech (POS) POS is a category of words (lexical items) which have similar grammatical properties (syntax, morphology) in English. Noun, verb, adjective, adverb, pronoun, preposition, conjunction, interjection, and sometimes numeral, article or determiner are commonly listed English parts of speech.

Polarity In sentiment analysis, the main problem is to determine to which extent a review is positive or negative. The positivity and negativity of reviews are two main poles of human feeling. Therefore, a review generally belongs to either positive or negative polarity.

Support Vector Machine (SVM) SVM is a supervised machine learning algorithm, which uses a separating hyperplane/line to categorize a given data. The hyperplane/line needs to be trained using labelled data in such a way that optimally segregates the data.

Conditional Random Fields (CRF) CRF is a discriminative undirected probabilistic model which is especially used in NLP to pars a sequential data or predict sequences of class labels for sequences of input samples.

Target In sentiment analysis, the topic (or particular subject of text) against which the analysis is performed is known as target. In tourism context, e.g., restaurants or hotels are targets.

Term Frequency (TF) TF is the number of times an item (letters or words) occurs in a review.

Term Frequency Inverse Document Frequency (TF-IDF) TF–IDF is the product of TF and IDF. The IDF is a measure to show whether a term is . common or rare across all reviews.

Unigram Unigram is a special case of N-gram (defined above) where N=1.

Weakly labelled data Data with the class labels determined heuristically by machine and not manually by human beings (such as star rating).

3.1.2. Sentiment Analysis Methods

Sentiment analysis comprises a multi-step process: a) **data retrieval**, b) **data extraction** and **selection**, c) **data pre-processing**, d) **feature extraction**, e) **topic detection**, and f) **data mining process** (*e.g., Hippner & Rentzmann, 2006; Schmunk et al., 2014*).

Data retrieval requires the identification and definition of the data source, for example, a commercial service provider portal or a social media network. To collect the review data from these sources, a specific web crawling mechanism is necessary to fetch the data and then save them in a database considering the format (*Menner et al., 2016; Schmunk et al., 2014*). After collecting data in a database, the review data needs to be extracted from within a set of heterogeneous data fields. For example, in the case of TripAdvisor data, a review is embedded within a retrieved HTML document, which is composed of different elements, such as footers or headers, tags, and the review text itself. The review text needs to be extracted using appropriate expressions. Each extracted review contains one or several sentences reflecting the reviewer's opinion. 11

Different tasks including splitting a review into sentences, splitting a sentence into words, tokenisation, filtering of stop-words, Part-of-Speech (POS) tagging, stemming and the transformation to lower/upper cases are performed on the reviews in the pre-processing step to prepare them for the next step (i.e. feature extraction) *(Schmunk et al., 2014*). POS tagging is an important pre-processing task that generally forms a part of sentiment analysis by assigning each word a particular label (e.g., noun, verb, and adjective).

Feature extraction is known as the process of deriving a set of discriminative, informative and non-redundant values to numerically represent a review or text. One of the commonly used feature extraction techniques is based on term occurrences, called *term frequency* (TF) or *term frequency-invers document frequency* (TF-IDF). Using the TF feature extraction technique, reviews or sentences are converted into a 'term document matrix' (*Pang et al., 2002; Hippner & Rentzmann, 2006; Menner et al., 2016*).

Topic detection is a multiclass classification problem where a text is classified to an appropriate topic class depending on its content and application. Topic detection research dates back to 1998 where topic identification in the context of broadcast news was studied (Allan et al., 1998). Hu and Liu (2004) later proposed a method to summarize customer reviews based on different product features. Suggested approaches mainly involved word dictionaries, clustering, and similarity measures.

In the data mining process, different types of sentiment analysis methods can be distinguished in the literature; namely (I) **machine learning**, (II) **rule/dictionary-based** and (III) **hybrid approaches** (*Feldman, 2013; Ribeiro et al., 2016*). Machine learning methods are further categorized into supervised and unsupervised approaches. The dictionary-based approach also includes a subcategory called semantic-based approach (*Tsytsarau & Palpanas, 2012*).

3.1.2.1. Supervised machine learning approach

A sentiment analysis method based on supervised machine learning involves creating a model by using annotated data or weakly labelled corpora. In the manually annotation process, for example, “what a wonderful holiday!!!” is annotated as a sentence with “positive” sentiment polarity. Weakly labelled data are those data where the class labels were determined heuristically by the machine. For example, user-generated content on review platforms often contains weakly labelled data when reviewers assign categories (e.g., restaurant) and ratings (e.g., stars) to their reviews (*Brob, 2013*).

Supervised machine learning approaches follow several steps. After applying pre-processing techniques to clean, segment and tokenize the text data, a feature extraction method is applied to characterize the review. Features extracted from the reviews are then fed to a classifier to train the classifier. The trained classifier is finally used to determine the polarity of new text. Support Vector Machine (SVM) and Naïve Bayes are the key machine learning methods used for sentiment analysis in the literature (*Brob, 2013; Kang et al., 2012; Markopoulos et al., 2015; Shi & Li, 2011; Shimada et al., 2011; Ye et al., 2009*), as they were conventionally designed for two-class classification problems. A SVM is a classifier which uses annotated data for training to obtain an optimal separating line to accurately categorize new samples data into different groups. A Naïve Bayes classifier is a probabilistic classifier, which uses Bayes' theorem in the classifier's decision rule, with an assumption that the features are independent.

3.1.2.2. Unsupervised machine learning approach

Cluster analysis, as an unsupervised machine learning approach, has been used for data mining, pattern recognition, and image analysis. Clustering is the task of grouping a set of data in such a way that items in a cluster are more similar to each other compared to those in other clusters. Clustering techniques, such as k-means (*Xiang et al., 2015*), and statistical models based on the probability distribution of reviews in sentiment space (*Rossetti et al., 2015*) were employed in the literature for sentiment analysis of short text data. In addition, Naïve Bayes models were also adapted in an unsupervised method for sentiment analysis (*e.g., Shimada et al., 2011*).

3.1.2.3. Dictionary-based approach

As dictionary-, lexicon- and rule-based approaches were used in the literature interchangeably, this review also uses the terms as synonyms. In this approach, the detection of subjectivity versus objectivity can be integrated into the framework or it can be handled by the sentiment polarity detection process itself. Aspect or topic detection can also be included within the framework based on the specific needs of the application. Dictionary-based systems rely on the use of comprehensive sentiment lexicons and sets of fine-tuned rules. A sentiment dictionary can be created either by humans, by machine or by both humans and machine (semi-automatically). For instance, a dictionary may contain words, such as “good”, “nice”, “fantastic”, “bad”, “worse”, and “ugly”, with their associated values of polarity. While creating dictionaries, the polarities are assigned to the words without considering any contextual information.

Different methods were developed for dictionary-based approaches. SentiWordNet in itself (*Bucur, 2015; Garcia et al. 2012*), and in combination with a simplified Lesk Algorithm, was used in sentiment analysis (*Bjorkelund et al., 2012*). The Lesk algorithm is an algorithm for disambiguating word sense that works based on the hypothesis that words in a given "neighbourhood" have the same topic (*Bjorkelund et al., 2012*). Valence Aware Dictionary for Sentiment Reasoning (VADER) is a method that has provided promising results on Twitter data (*Hutto & Gilbert, 2014*). VADER combines a lexicon and a series of intensifiers, punctuation transformation, and emoticons, along with some heuristics to compute sentiment polarity of text. Five general rules that embody grammatical and syntactical conventions for emphasizing sentiment intensity are used for computing the sentiment polarity. Umigon is another dictionary -based method, which uses a lexicon with heuristics for sentiment detection in Twitter reviews (*Levallois, 2013*). It is a fast and scalable method, which can handle negations, elongated words and hashtags. Umigon provides additional semantic features, such as time or subjectivity (*Levallois, 2013*).

3.1.2.4. Semantic approach

The dictionary-based approach was improved by introducing semantic-based analysis methods (*Tsytsarau & Palpanas, 2012*). The semantic approach is mainly a rule-based linguistic model to obtain a polarity for each text segment. In this approach a dictionary of domain specific terms and their associated polarity values is required.

3.1.2.5. Hybrid approach

In hybrid approaches, dictionary and machine learning-based techniques can work in parallel to compute two sentiment polarities. The results obtained from the dictionary and machine learning based methods are then combined to provide a final sentiment polarity. It is also possible to design a sentiment analysis model by incorporating both dictionary and machine learning based methods at different stages of the model. If the text is objective, then the task of sentiment analysis is over. However, if the text is subjective, it is then further classified as either positive or negative. For text with zero polarity, the neutral label is assigned (*Pappas & Popescu-Belis, 2013*).

According to the tourism-related studies discussed in the above cited article, tourism researchers have typically used two types of online content for their sentiment analysis: reviews of tourism derived from professional websites (e.g. TripAdvisor, Booking.com, etc.), and social media posts (e.g. Twitter). Both types of sources usually contain short text. For example, Twitter allows tweet of up to 140 characters in length, leading to a mostly sentence-level sentiment analysis.

Both supervised and unsupervised machine learning, dictionary based, semantic and hybrid sentiment analysis approaches were used in the tourism literature.

Among the unsupervised machine learning approaches, there is one, Naïve Bayes. This sentiment classification approach was trained using automatically labelled data, and emoticons, such as ☺ and ☹ , were used to represent ‘positive’ and ‘negative’ seeds to label data for training instead of words, such as “excellent” and “poor”. Thus, reviews that contained a smiley face were considered as positive and those with an angry face were classified as negative (*Shimada et al., 2011*).

As mentioned earlier, most sentiment analysis methods provide either a 2-class (positive and negative) or a 3-class (positive, neutral and negative) classification. It is important to evaluate and quantify the performances of different methods.

An easy and unambiguous way to present the prediction results of a classifier is to use a confusion matrix, which is also called contingency table (table 3). Each letter in table 3 denotes the number of reviews instances where class labels are positive ‘Pos’, negative ‘Neg’, and neutral ‘Neu’.

Table 3 Confusion matrix of the results obtained for a general 3-class classification problem

			(Predicted)	
		‘Pos’	*‘Neu’*	*‘Neg’*
	‘Pos’	*a*	*b*	*c*
(Original)	*‘Neu’*	*d*	*e*	*f*
	‘Neg’	*g*	*h*	*i*

The **Accuracy** (*A*) is one of the evaluation metrics commonly used in the literature (Ribeiro et al., 2016). It simply represents the number of correct predictions of sentiment made, divided by the total number of predictions made. The accuracy measures how accurate the method is in its prediction of the correct output. The metric *A*, as shown in Formula (1), assumes that every correct classification of the input reviews independent of the class label has an equal weight.

Formula (1)

$$A = \frac{a+e+i}{a+b+c+d+e+f+g+h+i}$$

Furthermore in the article, the results reported from the literature indicates that most sentiment analysis methods perform better in classifying positive sentences than negative or neutral ones. One reason, according to Dodds. Et al. (2015), could be the natural human language bias towards positivity. In addition, analyzing the negation in reviews is semantically way more complex. It also appears that neutral reviews are difficult to detect in most of the sentiment analysis methods (*Ribeiro et al., 2016*).

Recognizing the embeddedness of social media in people's lives and behaviors will also help the tourism industry to develop better systems for product development and delivery, market research, and risk management.

Advancements in sentiment analysis practices means to focus analyses on specific targets or aspects mentioned in text is meant to be considered. Target-specific polarity detection is a key challenge in this field, as the sentiment polarity of words and phrases may depend on the specific aspect analyzed. For example, considering the adjective "small", in the case of "small room" can be interpreted as negative, but in relation to a "handbag" it might be seen as positive. But as always, further research on the relations between targets, expressions, and implications for sentiment is necessary.

One important problem in aspect-oriented sentiment analysis is to discover implicit aspects. Consider the following two reviews, for example: "our luggage was delivered very quickly", and "it took an hour time to receive our luggage!!!". The first example includes a subjective assessment ("quickly"), the second example merely states a fact (an hour time = late delivery). To provide a negative evaluation of the luggage delivery process in the second example, common sense knowledge is required to interpret that an hour is not acceptable for luggage delivery.

In some cases travel agencies and hotel booking service providers use scalar ratings to rate users' reviews, for example, scores between 1 and 5 stars (e.g. TripAdvisor). Such scores alone cannot help managers or service providers understand what the main issues are and where improvements are necessary. However, from an analytical perspective, the user-provided scores function can improve the classification accuracy, as well as help to verify polarity.

3.2. Screen Scraping

Screen scraping (or Web scraping) is an information technology that allows to collect screen display data from one application or website and then translate it so that another application can display it. This is normally employed to capture data from a legacy application to display it using a more modern user interface. Screen scraping usually refers to a legitimate technology used to translate display data in order to make them more readable. Sometimes, it is even referred to as a terminal emulation. But, it must not be confused with "Content scraping", that is a technique used to detect contents from a website without the approval of the website owner. This latest is obviously not permitted by law.

This technique can be used with big data because it is able to extract large amount of data from websites and then save it to a local file or to a database in table format or spreadsheet. It can be confused or intended as data mining, but there are some differences between these approaches. In fact, screen scraping is basically a process through which a computer program or software extract significant information from websites. This is different from crawling or mining a site because you don't need to index every element on the web page, screen scraper simply extrapolate precise information selected by the user.

Screen scraping is a useful application when you want to make real-time, price and product comparisons, archive web pages, or acquire data sets that you want to evaluate or filter. For example, the manufacturer might want to monitor the market trends and uncover the actual customer attitudes, without always relying on the retailer's monthly reports. By using web scraping the company can collect a huge data set of the product descriptions on the retailer sites, customer reviews and feedback on the websites of the retailers. Analyzing this data can help the manufacturer provide the retailers with better descriptions for their product, as well as list the problems the end users face with their product and apply their feedback to further improving their product and securing their bottom line through bigger sales.

When you perform screen scraping, you are able to scrape data more directly and, you can automate the process if you are using the right solution. Different types of screen scraping services and solutions offer different ways of obtaining information. Some of them look directly at the html code of the webpage to grab the data while others use more advanced, visual abstraction techniques that can often avoid "breakage" errors when the web source experiences a programming or code change.

On the other hand, data mining is basically the process of automatically searching large amounts of information and data for patterns. This means that you already have the information and what you really need to do is analyze the contents to find the useful topics you need. This is very different from screen scraping as screen scraping requires looking for the data, collecting it and then analyzing it.

Data mining also involves a lot of complicated algorithms often based on various statistical methods. This process has nothing to do with how you obtain the data. All it cares about is analyzing what is available for evaluation.

Screen scraping is often mistaken for data mining when, in fact, these are two different things. Today, there are online services that offer screen scraping. Depending on what you need, you can have it custom tailored to meet your specific needs and perform precisely the tasks you want. But screen scraping does not guarantee any kind of analysis of the data. (Gina Cerami, "Are You Screen Scraping or Data Mining?",2012-Connotate)

Screen scraping is in this terms another useful tool to collect important information coming from "big data", and for this reason they also can be used by travel agencies in order to retrieve data they need from the touristic service providers' websites automatically and stay updated with prices and offers with a very small effort and cost. The analysis of such data helps the agencies to understand what kinds of deals and packages their competitors are providing to the market and what costs and locations the market prefers, so that they can modify their business plans accordingly. In this way they can easily keep track of travelers preferences and feedbacks, analyze the cost of operations of their competitors and inspect what they do better and predict the forthcoming changes to the market and prepare for them in advance. (Markrs.co)

The web or screen scraping process consists of parsing the target sites and removing the programming elements to get the text data. Further, the data needs to be cleansed of unnecessary information and the relevant data like costs, timings, location, photos etc. are stored in a tabular format so that it can be synced with the database. This is repeated periodically in order to keep the database up-to-date. In fact, the scraper must be updated frequently to deal with the possible changes that may occur at the target website.

The scraper can also be used to get data from the competitor's website. This data could then be structured into a comparative format with one's own data so that it can be easily analyzed what the competitor is doing in the market and compare its offers or products with their own deals and keep themselves aligned with market prices and desires accordingly. All this data extracted and processed together in more complex ways can yield the market preferences like hotels of which area of the city are booked most, the cost range preferred the most, the places visited the most and to keep track of what the market demands and how much it is willing to pay for it.

As we explained before, sometimes screen or web scraping can be confused with the "illegal" practice of content scraping with the authorization of the owner of the website. But in most cases when this practice is permitted and it is not harmful for others, it can have a very positive impact, especially in the tourism industry. Some tourism service providers like airlines, for example, don't mind being scraped on as being listed on the travel sites would eventually increase their customers. Due to this, web data scraping is completely harmless to one's own business, which makes utilizing it even more significant to the growth and flourish of the agency.

With today's dependency on the internet and technology for the touristic sector, data has become a major advantage for all kinds of services across different sectors. Agencies such Trivago, are already utilizing the web scraping process to the fullest in their business models. For travel agencies, a database of buses, restaurants, attractions etc. along with timings, costs, feedback, monthly visits etc. is a major desire, along with the need to keep track of others in the same sector. Information Retrieval methods help them to achieve this and to analyze the collected data to give the best and most accurate results in the present day. It was only fair that travel agencies utilized this data-driven approach in their model and achieve their goals faster and easier. (datahut, 2018).

CHAPTER 4

4. How online reviews data impact on hotel performance

As we explained at the beginning, big data are mostly unstructured data comprised of images, dialogues, texts, videos, or transactions. In hospitality industry, this amount of unstructured data constitute the majority of information source from which hoteliers can spot trends, make predictions, understand customer needs, and improve decision- making, especially regarding prices. In this sense, online reviews, together with reservation patterns, could be a precious source of this kind of data. This makes big data so much valuable for hotels. Unfortunately, for relative small hotels, this unstructured data are often hard to process and extract the useful information they need. Hotels, in order to do that, need some particular advice from software and data analytics expertise. Techniques such as sentiment analysis, opinion mining and screen scraping are among those used to process data contained in review aggregator online platform like TripAdvisor or booking engine like Booking.com. These tools are able to extract the most valuable information among numbers, photos, chats, post and reviews generated by customers across the web in the simplest way. Then, it's hotels' turn to employ the information extracted by these techniques in the best possible way in order to translate them into performance improvements. The best advantages hotels can take from that is to detect reservation patterns in order to adjust prices more precisely and to know and better understand customer needs and preferences from online reviews users themselves create and spontaneously release on the web through platforms, blog and websites. Big data for hotels can offer them increased efficiency that can be reflected through some performance indicators hotels try to monitor to take their profit and revenues under control.

In this part, I would like to explain how online reviews and big data extrapolated from them can be very valuable for hotel management and in particular for hotel revenue management and their performance.

The first part of this chapter will present the literature found about the importance and the role of online reviews in the improvement of hotel performance and the reasons why hoteliers should take into account these contents in order to increase their presence online and consequently the revenue. While the second part will present how online reviews and user- generated contents can affects hotel performance according to some interviews to a group of Venetian Hotels.

4.1. Big data analytics on online reviews

There are currently more than
100 MILLION REVIEWS
written across the internet
More than
75%
new reviews are written online every minute.
81%
Traveller Reviews are positive
ONLINE REVIEWS
HOTEL
SUBMIT

Big data are revolutionizing the world of every economic sector and with particular interest of the hospitality sector, where it is fundamental to build a strong relationship with the customer and build the so-called guest loyalty. This will be easier for the bigger hotel chains that can count on a strong brand, but it is less true for the smaller hotel and accommodation facilities that should start to take advantage of this information revolution and of the Internet to be more competitive in the hospitality industry.

Tourism organizations and enterprises, especially travel agencies, hotels and destination marketing organizations, have been seriously challenged by the rise of the internet, but at the same time enormous opportunities have opened up to them.

The revolution of internet has brought improvements in communication, distribution channels, and transactions even in the tourism sector. Tourists and visitors have now at their disposal online resources, that enable them to research on destinations, transportation, accommodation and leisure activities, and enable the purchase online of touristic products and services. This could be considered a sort of a consumer revolution which has effectively transferred much power from suppliers to consumers. For this reason, understanding consumers' needs and preferences has been a major source of success for hotels management. Hotel investment will always require evaluation of fundamental metrics such as regional economic growth, revenue per available room, capital expenditure and earnings projections. But there is now a new type of evaluation or analysis that is becoming increasingly important in hospitality industry: using customer intelligence from the social web to further improve the decision process and validate the financial assumption. Surfing the Big data revolution wave, it is interesting to notice that this type of intelligence comes from publicly available sources that millions of customers use every day to share open and honest feedback about their hotel experience.

Social analytics may identify opportunities across the hotel, not just in the marketing department. The most interesting aspect of online reviews and social media monitoring is that they are not biased towards any special interest. The data serves as a "virtual focus group" which comprises tens of thousands of customers giving details on what they want to buy or how they want the service delivered to them. The feedback in aggregate could suggest improvements in many areas such as sales messaging, marketing themes, operational process, etc.. Thanks to this hotel management will be able to distinguish between areas that need capital expenditure and those that need just an operational process change.

Online reviews provide unequalled business intelligence and a competitive benchmarking , highlighting the factors that affect future revenue performance. Owners usually make their money by providing a product that people want to buy, and now they can gather this information directly from their customers.

Compared to other sectors, the hospitality industry has always been one of the first to experience the changes in technologies and in particular the transition from the physical to the Internet channel for managing customer relationship. For example, phenomena like dynamic pricing, infomediation (like OTAs), online reservations, and recommendation systems have taken place in this industry first than in others and then have been expanded in sectors such as retail and services, leading changes and consequences in business models, competitive strategies and customers' behaviors. (Raguseo E., Neirotti P., and Paolucci E., 2017).

In particular two actors have emerged in infomediation between end customer and hotels, and both of them create economic value by ,managing information about travelers and hotels, they are review aggregators of travel-related contents (like TripAdvisor and Trivago) and Online Travel Agencies (like Booking.com and Expedia). A broader array of customers can now access at a lower search cost richer and better information that is customized to their preferences and interests about destinations and accommodation.

According to Raguseo, Neirotti and Paolucci, these two actors of infomediation have a high market concentration due to entry barriers such as network externalities, the high degree of capital intensity and scale economies in data management. In this scenario, OTAs and users' review aggregators play a crucial role in creating economic value in the hospitality industry and in shaping the competition in this sector. In fact, considering that review aggregator and online travel agencies have lowered the costs of reaching more useful information for travelers, they also have brought more competition among hotels and other accommodation facilities. Infomediaries thus create economic value by certifying quality and increasing the online reputation and visibility, and by favoring lower prices and thus a larger sales volume. For this reason, managing the online visibility in the new infomediation platforms in order to attract and retain more travelers is a critical factor for hotels' profitability. They facilitate travelers' search by ranking hotels price or quality and by bringing more competition and they are becoming and even more important competitive factor for hotels and facilities. This equates to a redistribution of economic power towards both digital platforms and travelers. The first ones can capture up to 25%of the price of a room in the form of an intermediation fee charged by OTAs and shared with review websites that convey internet traffic to them. While, for customers the value gained consists in lower prices and richer information to guide their travel decisions. From the point of view of hotels, they face the challenge of capturing the economic value created by these new intermediation over the internet. This challenge consists in balancing the preservation of unit profit margin and the pursue of sales growth through a better online visibility. (Raguseo E., Neirotti P., and Paolucci E., 2017). Being visible on the new intermediaries channel is a great opportunity even for the smaller hotels, that in this way can increase their revenue and their occupancy rate, which would be rather impossible in other ways given their small size and resources. Even though, the increased revenues not always turn out into a greater net profitability for small hotels because of the reduced market power and their reduction in unit profit margin rate due to the commissions paid to the OTAs. For this reason, it is fundamental, in particular for smaller hotels, to increase their online visibility through every means available and even handling new activities, like search engine optimization, and employing new business variables such as search engines, infomediaries, social media, consumer communities, and web marketing agencies.

4.2. User-generated contents

The information exchanged online is termed user-generated content (UGC) or e-WOM, which refers to "any positive or negative statement made by potential, actual or former consumers about a product or company, which is made available to a multitude of people and institutions via the Internet" (Hennig-Thurau et al., 2004). UGC not only includes online reviews, recommendations and opinions exchanged by consumers, but also forms the bases on which consumers revise their purchase decisions and ultimately change their buying behavior (Serra Cantallops and Salvi 2014; Sparks and Browning 2011).

In particular, online reviews and eWOM (electronic word of mouth) are becoming increasingly important in the tourism sector. According to the article by the Cornell University, *Online Customer Reviews of Hotels: "As Participation increases, Better Evaluation is Obtained"* there are some key points regarding online reviews and comments that touristic facilities should take into account to monitor their online reputation and their image to costumers.

First of all, in the article is underlined the importance of the WoM for what concerns the reviews and opinions found on the web. The word of mouth is very important for the reputation of a specific entity because it is able to influence behaviors and preferences of other users who come in contact with these opinions. In particular, with the spread of internet the tourism sector has evolved significantly, and the way the society, i.e. tourists and visitors, interact. Now, together with the traditional word of mouth methods there are the electronic word-of-mouth generated in the web. The opinions that customers express in online review sites represent an important type of eWOM.

Since the great influence online opinions can have on online users decisions , businesses must pay attention to the opinions about them that are released by their own customers. This is also an important aspect why customer satisfaction is fundamental for businesses. The influence of eWOM is especially important in the tourism industry, given the intangibility of what is offered and the high risk perceived by customers (Lewis and Chambers, 2000). Since the intangible component of the service is very prominent, it is apparent that potential guest check online reviews before making decisions about their travel, so that for tourism businesses is very important to avoid negative reviews and increase the positive ones.

Furthermore, it is important to underline, as it is explained in the article mentioned above, that there is a central difference between WOM and eWOM, in fact eWOM is registered in cyberspace. Therefore, electronic word of mouth is an information source that anyone can access easily, quickly and that remains over time. We can also say that eWOM overcomes the barriers of traditional WOM, because it makes the acquisition, transfer and use of this information easier than with the traditional mechanism (Andreassen and Streukens, 2009). According to the studies (Karakaya and Barnes, 2010), it has been shown that eWOM is more effective than communication generated from marketing carried out by companies. In addition to that, it has been shown that customer satisfaction and the quantity of WOM and eWOM generated has a U-shaped relation, i.e. very satisfied and very unsatisfied customers will be the ones who will generate more WOM and eWOM, while customer who are on average satisfied or unsatisfied will tend not to generate as much WOM and eWOM (Bansal and Voyer, 2000; Litvin et al., 2008).

In addition, further studies have shown that people pay greater attention to criticisms or negative reviews than to positive reviews (Lee et al., 2008).
As it is pointed out in the article, recent studies refer to the influence of online reviews in consumer decisions. Park et al.(2007) found that purchasing intentions increase as the quality and quantity of online reviews increase. In the hotel sector, it has been found that online opinions of other consumers are one of the most important variables in accommodation choice (Ye et al., 2011). All of these studies highlight the importance of websites where the customers have the chance to express their opinions and preference (e.g. TripAdvisor), and for this reason it is crucial for tourism business to receive positive reviews and limit the number of negative ones.

As a result of the analysis of 26,439 hotels' online reviews from the website TripAdvisor in more than 200 destinations, in the article it is showed that as the number of reviews increases, the ratings in these reviews are more positive. Thus, when fewer customers evaluate the hotels, negative eWOM/opinion has a bigger value than when more customers evaluate them. In addition to that is has been shown that from the chronological point of view, the first reviews are worse than the last ones. However, as time passes and more reviews are released, the ratio of positive reviews increases, thus leading to a better overall rating for the hotel. This is one reason why hotels should prone their customers to leave even more reviews about them. In fact, in this study, it has resulted that positive reviews are prominent compared to negative ones (more than 70% were positive, i.e. values 4 or 5 on a 1 to 5 scale, while just 15% were negative, i.e. values 1 or 2, for more than 1.28 million reviews examined). By obtaining more reviews from the customers, as this article shows, hotels will get to know more quickly their real average score, since the unjustified excessive weight of negative reviews will be mitigated, and the average of these opinions will better reflect the customers as a whole. This obviously does not mean that negative reviews should not be considered so important; in fact, negative reviews are a very good and helpful source of free and valuable information about areas that need improvement in the overall service granted by the hotel.

4.3. How economic value is created by infomediaries

Economic value is created by both online review aggregators and online travel agencies. In the first stage, value can be said to be created by travelers thanks the ratings, reviews and other contents like pictures, they provide on review aggregator websites such as TripAdvisor. Aggregators contribution to the creation of value lay on the integration of travelers' opinions and ratings, and then drawing on these data by providing ranking of visibility for hotels, restaurants and touristic destinations. It has largely demonstrated the effectiveness of online customer reviews for sales growth, in fact consumer reviews act for companies as a mechanism to build consumer's trust and brand trustworthiness, by helping firms to apply effective differentiation strategies, and by mitigating the effects of information asymmetries between the firm and the customer about the real quality of the product. (A.R Rao, L. Qu, R.W. Ruekert. "Signaling unobservable product quality through a brand", 1999)

Reviews aggregators websites aim at creating greater economic value by presenting their ranking as objective facts that represents people and their opinions and by controlling customers experience as much as possible. In this perspective, TripAdvisor for example uses social media mechanisms, like giving reviewers the possibility to disclose their Facebook identity, to avoid the tendency of people to give offensive statements and to detect fake reviews, and allows travelers to decide upon recommendations provided by their friends or people similar to them in age, interests and social relations. Thanks to its mobile application and the GPS data collected, TripAdvisor is furthermore able to control and support customer experience, as trough localization data of travelers it provides context-aware recommendations of bars, restaurants, locals. In this sense, TripAdvisor can be considered a "quality guarantor" for hotels and destinations and create economic value for both travelers and touristic services providers. For users the value is created in the form of rich information that is presented as objective that support their purchasing decisions and reduce the service and product uncertainty, while hotel can increase their visibility and reputation in order to attract and retain customers.

4.4. The role and impact of online reviews for travelers

Even more consumers read and share travel-related content online which has been created and posted by previous travelers rather than by travel service providers.

According to a TripAdvisor survey back in 2007, most travelers read online reviews before making travelling decisions. According to the data collected from a total of 7,000 users, a majority (82.5%) uses Internet as a source of information and advice for the planning phase of a trip.

Approximately 97.7% of consumer read other travelers' online reviews. About 57.8% of users stated they read other travelers' reviews every time they plan a trip. Out of this number about 36.7% of them frequently take reviews suggestions into account, while 5.5% do so only occasionally. Moreover, about 53% of travelers said they won't book a hotel if it has no online reviews.

63.7% of the respondents read other travelers' reviews at the beginning of their trip planning process in order to get ideas from the web, 64.7% do so in the middle of the process to narrow down their choices, while 40.8% of consumers also used online reviews later in their planning in order to confirm their decisions. Almost 9% use the reviews during the trip, and nearly 30% use reviews after the trip to compare and share their experiences. Others stated that reading reviews is an ongoing process and they read them with no specific trip in mind; or they read reviews through the planning process, way before a trip or only read for accommodations. Specifically, they look for others reviews on: virtual community Websites (92.3%), travel guidebook sites (60.6%), online travel agency sites (58.1%), search engines or portal sites (51.5%) and local destination Websites (44.6%).

Out of the total, a great part (77.9%) of users think other travelers' online reviews are extremely or very important for deciding where to stay. The three main elements that are critical to evaluate a review are: detailed descriptions (71%), the type of Website where the review is posted (64.7%), and the date when it was posted (59.3%).

A majority of respondents reported that they feel the information they find in other travelers' reviews compared to travel service providers are more likely to be: up to date(65.3%), enjoyable (61.2%), and reliable (61.1%). On the other hand, other travelers' reviews seem not to bring a clear advantage for what concerns detailed and relevant information. Nevertheless, review readers still say that reviews are superior along these two dimensions.

Another important aspect of online reviews is that they influence other traveler in the following ways: reviews are very useful for learning about a travel destination, product or service (94.6%), thanks to those readers can evaluate alternatives more easily (91.9%), readers can avoid to choose places or services they would not enjoy in advance (91.8%), and of course as we said before, reviews can provide new ideas about the trip. An interesting point of this survey is that over 80 % of respondents who frequently read online travel reviews indicated that the reviews influence them in all these ways: they increase their confidence in decisions (86.6%), make it easier to imagine what place will be like (85.3%), help reduce risk and uncertainty about a touristic product or service (82.4%), make it easier to reach decisions (81.3%), and help to plan a trip more efficiently (80.2%). Furthermore, over three quarters of online travel reviews readers said that reviews reduce the likelihood of regret (77.6%), make travel planning more enjoyable (77.4%), make them feel more excited about travelling (76.(%) and add more fun to the travel planning process (76.5%). Finally, almost 70% strongly agreed that other travelling reviews save time in the planning process and help to imagine trips more vividly.

Knowing all this information, hotelier can exploit the benefit of online reviews in these ways:

- **Responding to customer online reviews, in particular to the negative ones.** In fact, online reviews are not only a source of insights and information for hotel management, but also a tool to showcase that the values and ethics of hotel are in alignment with their operations and the core business interests. By responding to reviews they can reassure the existing and potential customer about the service or product quality and listen to their personal needs. In particular for what concerns responding to the negative reviews, hoteliers shows their acceptance towards guests feedback and are willing to solve and fix the issues their guests raises in their reviews. Thanks to the maintaining an ethical and transparent review response and mechanism, hotel management can ensure building a better brand loyalty to drive more revenue. This is important especially for small hotels and facilities that cannot count on a strong brand name and they are more willing to build a customer loyalty to compete with the bigger ones.
- **Testimonials of quality.** People tend to be cautious about direct advertising, especially if they are not loyal or familiar with the brand. However, as we said before, potential guests are prone to listen to advice and suggestions from genuine customers who are considered to be less biased and more realistic than the advertising of the services providers. This is why positive reviews from real guests that can be presented as reliable testimonials work like a charm in order to build a public opinion about the facility. When people love the hotel where they stayed during their trip, it is more likely they will talk about it online, leading towards more brand loyalty and by extension, towards more revenue generation.
- **Industry/sector insights.** As we have underlined, online reviews can be a useful tool to shape strategic operations. In fact, by tracking their own online reviews and those of their competitors, it is possible to compare their performance and gain critical insights about the SWOT analysis of hotels doing it. A right implementation of this analysis can be used to position their facility against those of their competitors and kelp to plan the next strategy accordingly.
- **Guest data.** Last but not least important, customer online reviews often provide a vast array of guest data, providing information on why guests choose one specific hotel rather another one, what they liked and what not, what was the reason of their visit, if they will come back or not, etc.. By analyzing and evaluating the guest data provides

valuable insights that are needed to launch "aggressive" and competitive personalized marketing campaigns in order to attract and retain more potential customers.

These four points can summarize the reason why customer online reviews must be considered important and critical to the hotel revenue management and seriously taken into account by hoteliers who wants to improve their positioning in the competitive market in order to materialize all these benefit in a real revenue increase, as the hospitality industry is a very competitive environment.

4.5. Main Hotel performance indicators

As we said before, if online reviews are wisely monitored and employed they can impact hotel performance, strategy and economic metrics. According to an article released by Cornell Hospitality Research (Cornell University of New York), online guest satisfaction has a direct impact on the financial performance of hotels. In fact, the research conducted by the professor Chris Anderson has demonstrated the relation between the Global Review Index (ReviewPro) and some of the figures used to measure the performance of hotels in hospitality industry, i.e. ADR Average Daily Room Rate, Occupancy rate and RevPAR (revenue per available room).
The study shows in particular how 1 point increase in a hotel's 100 point global review index leads to a 0.89% increase in average price of room (ADR), a 0.54% increase in occupancy rate, and a 1.42% increase in RevPar. And the impact is verified to happen across all distribution channels, both online and offline. These findings are based on the analysis of more than 31,000 observations on midscale, upscale and luxury hotels in 11 major metropolitan areas in North America and Europe, such as London, Milan, Rome, Madrid, Berlin, Prague, Chicago, Los Angeles, New York, San Francisco and Miami.
The main indicators of performance, that can be used in hospitality industry by management in order to verify that their operations are generating revenues, are these below:

-ADR (Average Daily Room Rate)

This indicator measure the average sales price of the rooms of a hotel and it is very useful in the budgeting phase. It is computed by dividing the room revenue, minus discounts, taxes and food and beverage revenues, by the total of rooms sold. Unfortunately, this is not a very useful indicator of performance compared to the next ones because it simply indicates the average price a room should be sold at, but as we know the price of a room can fluctuate a lot depending on the season

-Occupancy Rate

It refers to the number of occupied units/rooms at a given time, compared to the total number of available units/rooms in the facility. It is very important in order to know how much of available space in a hotel is actually being utilized at the time considered. It is expressed as a percentage and it is computed by dividing the number of rooms occupied by the total number of rooms available.

-RevPAR

This is the value of the revenue per room available and it is one of the most critical parameter to take into consideration in order to monitor the hotel trend. It is fundamental to evaluate the economic performance of Room Division department and it can be computed into two ways:

-the first one is by dividing the room revenue, after discounts, food and beverage costs, and taxes, by the total number of rooms available;

-the second way is by multiplying the ADR per occupancy rate.

RevPar in necessary in order to know the revenue of every single room, whether it is sold or not. It is particularly important when we consider a small-sized hotel because it is more frequent that the most part of revenues comes from the selling of the rooms. While for bigger-sized hotel it is slightly different because they offer more additional service such as spa or restaurants, or boutique.

All these three metrics are really valuable when it is time to measure the performance of a hotel, but as we said explaining them they are more realistic when we want to consider small to medium size hotels , in which the most part of the revenue comes from the selling price of rooms. But, when we want to consider bigger hotels where the operating costs are bigger and it is more difficult to occupy the total number of rooms available, the value expressed by these figures, especially by the RevPar, is less realistic and it can be less favorable for a bigger hotel than for a small-sized one, even if the total revenues are higher. In fact, these figures do not take into account all the additional services and products a bigger and upscale hotel can offer to its clients, like spa, boutiques, meeting rooms, which most of the case can be more valuable than the mere room revenue. In addition to that, they do not take into account all the revenues coming from food and beverage, which for upscale hotels are always fundamental.

For this reason it is always better to integrate this metrics with others in order to obtain more realistic values representing the performance of a hotel.

The integrating or edited metrics are:

-ROS

There is a further indicator of performance a hotel can use to evaluate its economic results, and that is ROS (return on sales). It needs to measure revenues in terms of operating revenue and obviously it is highly influenced by the sector and the market where the company operates. It is computed by dividing the operating revenue by sales revenue. ROS is used to calculate the average margin (on sales) obtained by the profit. It can be considered one of the best indicator of performance and the most suitable to the Revenue management practices because the denominator is composed by the factors on which managers can take action and that can be influenced by online visibility and reputation. i.e. sales. In addition, all the operation costs considered include beverage and food and all the extra services offered by hotels.

-MOL/EBTDA

This economic indicator is based only on the characteristic operations and it indicates the earnings before taxes, depreciation and amortization. Thanks to this it is possible for the company to monitor whether it is earning through its core businesses, excluding the financial operations and all the amortization. It can be computed as operating revenue minus operating costs, including employments costs, raw material costs, running costs. EBDTA does not consider taxes and depreciation so it can be misleading from the point of view of revenue increments and it can be useful only to monitor the operating management.

-GopPar

GopPar stands for “Gross Operating Profit per Available Room” and it basically is the gross operating profit for every single room available a day. It is slightly different from RevPar because it includes taxes, food & beverage costs and all the fixed and variable costs, so it offers a more global view of all the expenses and revenues, not only room revenues. In fact, it is computed by dividing the gross operating profit by the total number of available rooms. It can be considered to be a better indicator than RevPar taking into consideration the limitation of the latest that we have explained before.

According to Juston Parker (GOPPAR, Gross Operating Profit Per Available Room - Best Measurement of Success), there might be four ways for hotels to improve their GopPar, such as:

1- To well manage all the income, not exclusively those coming by the sold rooms. In fact, there’s still hotel managers that tend to consider only room revenues and to underestimate or not well manage the other source of income. Managers should measure the revenues and expenses for all the areas of their hotels, including restaurants, meeting rooms and souvenir boutiques insides their hall.

2- To increase the revenues for each customer. This can be done through the cross-selling, that means to offers additional extra service to customers even if it not initially required by them, in order to increase the revenues of the service as much as possible.

3- To limit profit losses as much as possible. Sometimes, hotel managers in order to get more reservations by big groups of people tend to offer them services such as free upgrades, special offers and exclusive conditions. This obviously can lead to a better customer loyalty but this practice must be carried by taking into consideration all the expenses and if they are worthy for the business or not.

4- To seek paired selling, for example to sell rooms and the meeting room together for special events. This fourth point is well linked to the second one, since hotel manager should take advantage of every area of their hotel.

4.6. Sample

The reference sample of this research is composed of twelve Venetian hotels situated in the city center area. These hotels have been selected by looking at the ranks in TripAdvisor , in particular they have been picked among the thirty best hotels according to visitors best choice.

During the research phase, all the reviews and the rates according to the main reviews content platform, TripAdvisor, have been analyzed and subdivided among the classification as terrible, poor, average, good, excellent, in order to better understand how negative reviews were written and how they can impact on the hotel image. Furthermore it has been analyzed the period of time in which the reviews were more frequent, they appears to be from September to November, and from June to August. The category of traveler reviewers has also been analyzed, finding that couples and families travelers are the most prone to write online reviews after their travel experiences. While the travelers for business were less and they mostly wrote positive reviews and gave a rating from good to excellent for most of the hotels selected.

They all are medium, upscale and luxury hotels, and for this reason, as we said before, the mere room price is not the only factor that impact on revenue and earnings. The price range is between 100€ and 500€. It's been decided to analyze this classifications of hotel because they can offer a more precise view of how quality services can improve the total ranking in the reviews and because they offer a more significant number of reviews that can be analyzed.

The best attributes visitors ranked for these hotels are cleanliness, location, service and value/price, and according to these attributes all the hotels have a rank of 4.5 and 5 out 5 points in TripAdvisor, for these reason they appear to be among the best travelers choice.

According to what it has been said at the beginning citing the article "as participation increases, better evaluation is obtained", the hotels selected for this research have 1000 reviews and above, hotels in the list of best travelers choice that has less than 500 reviews have not been considered. All the discarded hotels that have less than 500 reviews are not considered "bad" by tourists as someone could think, but simply they are either relatively new or so exclusive they have very few rooms , but quite expensive, that the number of visitors often likely to write reviews after their vacancy is quite scarce.

With the analysis of the reviews on TripAdvisor, we came up with some percentages. The terrible reviews constitute only 0-1% of the total, the poor ones constitute 1% of the total, the average ones range between 1% and 4% , the good ones range between 6% and 29%, and finally the excellent ones range between 71% and 91%. We can clearly see that the most part o reviews are among the good ones and the excellent ones, and this demonstrate why they are among the best travelers choice of TripAdvisor. Having a very good rank on TripAdvisor is very important for the online visibility and reputation. This is also confirmed by the fact that searching these hotels on Google their official websites appear among the first results after the main OTA and TripAdvisor websites. Obviously it is not always true, especially for small hotels that does not invest too much in their websites (this has been demonstrated by looking at the last hotels found on the same list, the best travelers choice, for which sometimes it has been hard to straight see the official website among Google research results).

4.7. The review-content website analyzed

For the purpose of this research the main platform analyzed has been TripAdvisor. TripAdvisor is the biggest user-generated content platform in the touristic sector, and the biggest online community of tourists, who can share their experiences and opinions.

This has been chosen because it is undoubtedly the main source of online reviews nowadays and it is the most evident example of the strategic importance of online reputation for hospitality industry. (Milano and Tapinassi, 2013)
It was founded by Steve Kaufer , Langley Steinert, Nick Shanny, and Thomas Palka in February 2000. Kaufer stated that the original idea wasn't to create a user-generated social media site to share reviews, it was more intended to be "site where we were focused more on those official words from guidebooks or newspapers or magazines. We also had a button in the very beginning that said, "Visitors add your own review", and boy, did that just take off." (Steve Kaufer)(Wikipedia).
In 2004, the company was purchased by Interactive Corporation, and in 2005 it was incorporated in Expedia.Inc.. In 2011, Expedia.Inc decided to spin off the TripAdvisor brand of travel site.
According to a July 2011 PhoCusWright survey of 3,641 respondents, randomly picked through a pop-up invitation link on TripAdvisor.com and commissioned by the same Trip

Advisor, “98% of participants found that TripAdvisor’s hotel reviews accurately reflect the experience”(Wikipedia). Since that, TripAdvisor becomes the largest travel site with nearly 280 million unique monthly visitors. ("TripAdvisor is now the world's largest social travel website". Traveltradejournal.com. ,2012).

Not only allows it to share advice and opinions about travel experiences but also has several functions to search and personalize with filters the information and the results, in order to better compare solutions and offers, look for the best prices and found the direct link to some of the main OTAs, such as Expedia. Booking.com, Trivago, eDreams and to the hotel official websites too.

In January 2005, TripAdvisor reached 1 million reviews around the world, in April 2008 it reached 15 million, and in March 2011 45 million online reviews. The company counted 260 million of unique visitors per month in 2013, with more than 150 million reviews on more than 3.7 million accommodation facilities, restaurants and touristic attractions. (Google Analytics, 2013). The company operates in 30 countries around the world and it is available in 21 different languages. It also includes other member websites such as cruisecritic.com, holidaywatchdog.com, oyster.com, etc..

According to some TripAdvisor directions, it is possible to review any type of accommodation on the platform but users are asked to be honest about their travel experience and to release truthful reviews complying to the company’s regulation.

Reviews are not published straight forward, but they are approved after 24-48 h later and only if they comply with the regulation. Reviews can stay on the web forever, unless they are removed from the site at the request of the reviewer or, less often when they don’t comply. On the website, it is stated that “only reviews that reflect true personal experience about the accommodation or location are allowed, while generic arguments that are not referred to real personal experience won’t be published. But unfortunately, it has been found that some users release online reviews even if they didn’t stay in that particular accommodation but they just saw that or found/reserved that online. In fact, TripAdvisor requires its users register by simply entering some personal details like name or nickname and an e-mail address, obviously they must be personal address not corporate. In this sense, anyone can register and write reviews and it appears difficult to verify if the travel experience is effectively real and if the person actually stayed in that accommodation.

Users can also rate the accommodation or destination with a ranking between 1 and 5 (1:terrible; 2:poor, 3:average, 4:good, 5:excellent), and he or she can also rate single aspects, such as cleanness, position, room, service, price/value. At the end, the single rating are synthetized in one rate, the " Traveler Rating", always with a ranking from 1 to 5. In addition to that, TripAdvisor provides a ranking of similar accommodation in the same area, or nearby, called "Popularity Index". This essentially is based on the total number of reviews (quantity), the average judgement (quality), and how the review is recent (freshness). The latter has a big importance, in fact the facilities with more positive reviews in the last period are more favored compared to those that have negative reviews among the most recent ones.

On the other hand, for accommodation managers is easy to be included, it simply needs to register by entering their own data, the facility e-mail address, and some information about the accommodation (the most important ones, and the description of the accommodation must be a maximum of 75 words), images and it is also possible to show a promotional video of a maximum of 10 minutes.

The websites also provides a section called "TripAdvisor for business", made of two sub-sections. One is free and can be used by restaurants and accommodation owners, and the other one reserved for hotel managers. The first one allows to the user to promote its activity, handle the profile and manage the relative reviews, while the second one allows to managers to take advantage of some promotions, among which there is the possibility to publish some offers on the official hotel website, among the activities on the same geographical area, and in the newsletter send periodically to subscribers. Obliviously, this last section is more favorable for hotel managers because it allows to obtain a better position on TripAdvisor list of facilities, largely increasing its visibility on the website.

TripAdvisor also provides hotels with some widgets, that can be included on the official website or on the Facebook page, and that show the positioning on the platform, the ranking and reviews, with the possibility to be linked to their profiles. This allows to easily monitor any change in rankings and number of reviews and to easily manage the profile. Moreover, TripAdvisor provides certifications and paper flyer that are very aimed because they are a guarantee instrument for hotels and facilities to show to their customers.

TripAdvisor has also created an application linked to Facebook, Cities I've visited, that allows to users to know the destinations visited by their friends and to derive some suggestions from their reviews and posts.

Not so long ago, the truthfulness of some reviews on TripAdvisor has been questioned and some false reviews were found on the famous platform. In fact, for users is so easy to register and write reviews because it does not require any particular effort or information, in contrast to some sites such as Booking.com, where in order to release reviews the user has to submit the number and some data of the reservation. In this sense, for TripAdvisor is harder to monitor the compliance and accuracy of its reviews and it needs stronger instrument to ensure them. As Kaufer stated, the company employ tree ways to ensure the accuracy: before the publishing of the review it must be analyzed and checked by a group of expertise to detect the mistrust messages, in addition they employ some automatic instruments that help to identify the break-in attempts to the system, and finally the users community and facilities owner can report to the company tricky messages and ask to remove them.

In 2011 in Italy, TripAdvisor concluded an agreement with the main associations of accommodation owners/managers, Confindustria Alberghi and Confindustria AICA in order to safeguard the needs of owners and managers, and it highlight the importance of a mutual dialogue between the rating websites and hotel managers/owners.

4.8. Hotels' interviews

The purpose of this research is to verify how online reviews and data that can be extracted from them can have an impact on the performance of hotels. Considering the sample that is comprised of a group of 12 Venetian hotels, some general considerations must be made about the tourism in Venice.

Venice is primarily a city of art and for this reason touristic flows have some particular features, that make this destination different from others, like seaboard touristic locations. Tourists are not likely to return in the same structure at the same period of time as the years before. In this sense, it is more difficult for specific hotels to build a customer loyalty, as this is easier for hotel chains that can count on a strong brand loyalty and image. For this reason, it is more important for the former kind of hotels to monitor their online visibility and image on platforms such as TripAdvisor or Booking.com.. Unfortunately, for these hotels it is very hard to depend their online visibility by merely counting on their websites, even if it crucial to design a very well made official website in order to convey an idea of quality and customer-focused vision. Moreover, in the era of Internet the competition has dramatically increased, and this kind of hotels have always to be among the top in order to attract more customers, and in particular "quality" customers who can lead to significant revenue increments. Uncontrolled flows of tourists has always been a huge issue for the city of Venice, in particular the mass tourism which does not bring value to the city but only lots of people that require a big effort to be managed.

As other cities of art, tourism in Venice heavily depends on the time of the year. For this reason, prices range of every hotel fluctuate a lot depending on the high and off-season. In most cases the fluctuation may happen day by day. To monitor competitors and the own online visibility is crucial during the off-season, in which competition is even more wild.

Hotels and small hotel chains in Venice are all independent, so that prices must be very competitive during off-season; in fact with equivalent position and service quality, customers will chose the accommodation with a lower price.

After these few considerations, some hotels' interviews and information resulted from them will be exposed in order to explain how online reviews data can be exploited by managers to increase their performance. Some of the interviews were made by phone, others were made in person by talking face to face with hotels managers, who revealed to be willing to expose their personal point of view regarding the touristic scenario in the city of Venice.

Since all the facilities were selected through a research on TripAdvisor, they are all present on this platform, and in fact, as it has emerged from the interviews according to managers, it is unconceivable not to be present on such a platform nowadays, either due to the increasing number of online reviews written or to the fact that to not be present on such platforms means that hotels are not selling. The power such platform has acquired is incredibly high and can affect hotel image and visibility in a way that can significantly impact reservations, and consequently revenues.

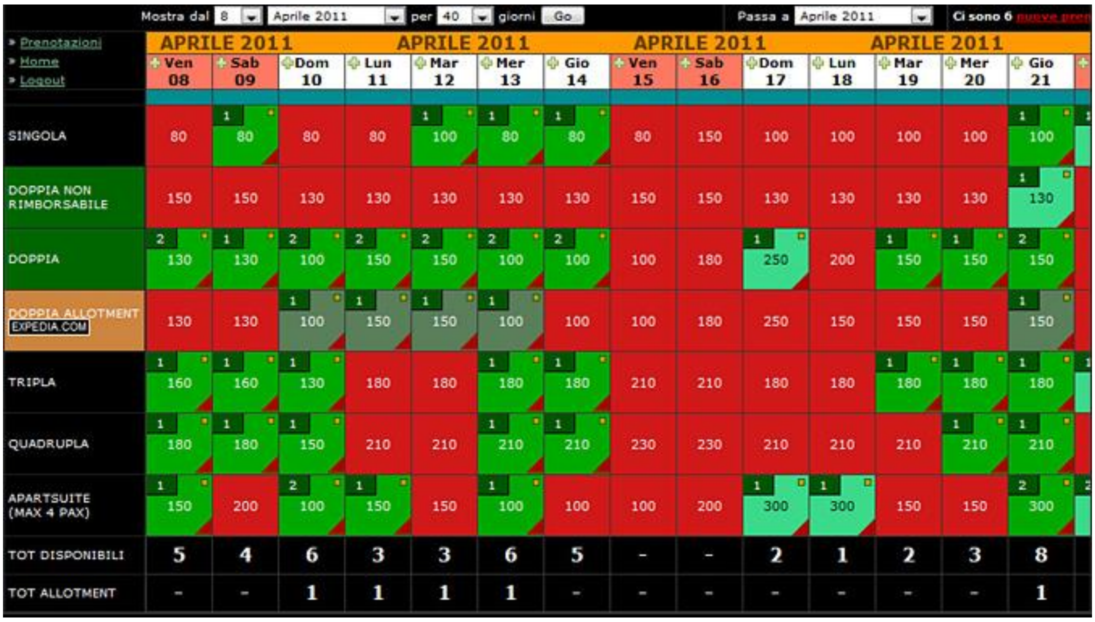
Together with their official website , that in most cases (90%) is personally managed with advice from web marketing companies, all the hotels in the sample use social media in order to be in contact with customers and to promote their accommodation with special offers and social images. The social networks employed to be connected with their customers and promote their visibility are: Facebook, Twitter, Instagram, WhatsApp, google+, YouTube. Some of the interviewees (70%) admitted that showing several images and interactive video on their website and on metasearch platforms can be so much effective to increase the online visibility. The most effective pictures and video for hotel are especially those of the surrounding than those of rooms, as it can give hints about the location where the hotel is situated and attract more visitors, since the position for hotels in Venice is such an important aspect.

From the examination of all these elements, according to all the interviewees, online visibility especially on Google, metasearch websites and reviews aggregator websites is fundamental for their activity. As much important as the visibility online, is the reputation they have especially on the web. Once again, all the respondents agreed on the importance for an accommodation to have a good reputation online, particularly on platforms such as Booking.com, TripAdvisor, and on the main searching engine Google, Google plus and Google Adds. Nowadays, reputation is fundamental for these hotels in order to increase revenues because through it they can deliver an idea of quality and willingness to provide the best level of customer care. All the hotels selected, as we explained before, are 4 stars or above and because of that their main goal is to provide the best quality service to customers. This is the main factor that affect a good reputation and that can build a strong brand loyalty. This is especially true for independent hotels that compared to big hotel chains cannot count on a brand strongly built over time (all the hotels selected are relatively recent, 10 to 20 years of activity). It is here that online reviews step in. With the introduction of reviews aggregator platforms such as TripAdvisor , online reviews are the major aspect that accounts to build a good reputation. As all the respondents agreed on, even few negative reviews can jeopardize the reputation of hotel, even if the review might not be true. More than half of the respondents confirms that they encountered false or biased reviews. In fact, the majority of them (95%) stated that they consider Booking.com to be better for what regards reviews, since to release reviews it is required to submit some details of the reservation. Even though, TripAdvisor is the largest travel site and still many people heavily use it to make research about hotels or touristic destinations, for this reason they consider important to have a good score on it. (they do not consider a lower than 8.6 score to be a good indicator of quality, especially for 4-stars hotels).

Having a good general score and be among the first pages of TripAdvisor can allow them to set prices higher to their minimum standard especially during off-season and sometimes even higher than those of competitors that have a better positon. In fact, according to most of respondents the best factors in order to be competitive is the price and the location. These allow them to have higher scores and more positive reviews, together with the quality that must be at the head of their goals in order to increase their performance.

Furthermore, according to all the respondents online visibility is positively related to occupancy rate, even though the responses have not been further deepened. Occupancy rate is considered by the major part (85%) of the respondent as the main performance indicator they look at, since for them it is very important to occupy the greatest number of rooms, considering the fact that in most cases they have from 10 to 30 rooms available. The occupancy rate for these hotels ranges from 80% to 95%, it is clear that they try to keep it quite stable, with obvious fluctuations during the off season. In addition to that, for most of the respondents (almost all), responding to negative reviews is an uneasy work that must be done. Managers try to respond to the most part of negative reviews themselves because they consider important for their online reputation to demonstrate customer-care focus and if the reviews is a constructive critics they will be willing to fix aspects that don't work.

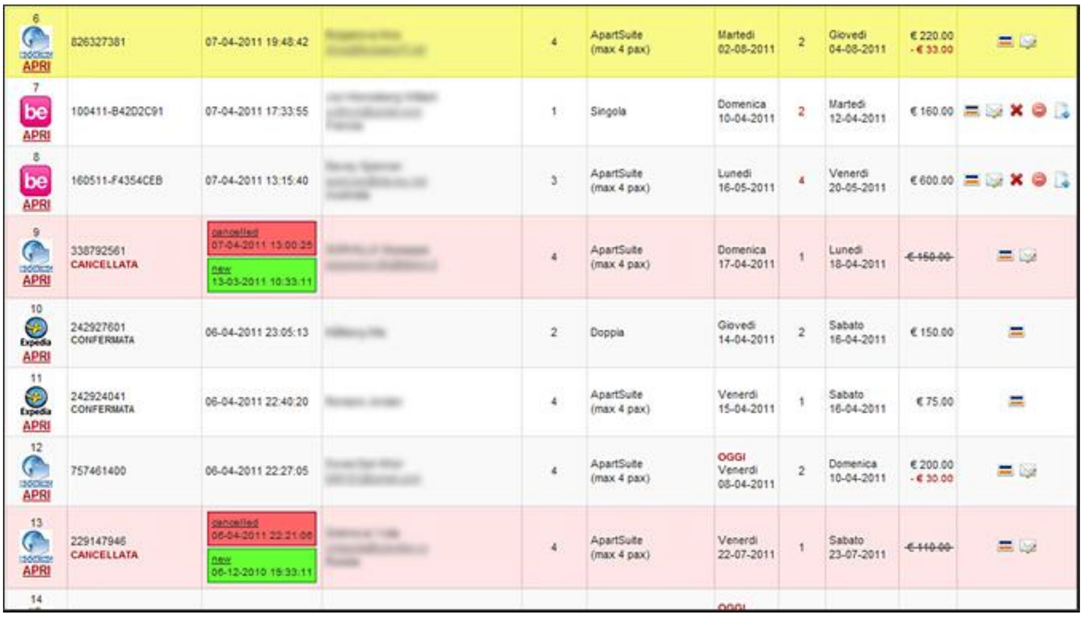
When it came the turn to discuss about the employment of big data analytics, the portrait was less definite. Some of the facilities managers confirmed to have made use of some expertise regarding big data analytics and to know only to some extent how it actually works. The primary way they make use of such technology is to help them in the pricing phase. Through the analysis of data found firstly on the websites above mentioned, they are able to compare all the prices their room are sold at and to monitor them in order to check if they comply with those offered on their official websites. This can be a way to make offers to customers that will decide to book directly through the website. In fact, another crucial aspect that emerged from the interviews and agreed by all the respondents, it's the compelling willingness to reduce the power of such intermediaries, such as the OTA, and to favor the direct reservation. This will not be reflected on evident revenue increasing but rather on the decreasing of commission costs. Hotels have to pay quite high commissions to websites such Booking.com and TripAdvisor, but at the moment it is unthinkable to act differently, even though all the facilities are highly promoting the direct reservation on the website with several offers with price discounts. Obviously, this is possible thanks to big data processors that are able to provide hotels with all the different tariffs present on the web.



Mostra dal 8 Aprile 2011 per 40 giorni Go | Passa a Aprile 2011 | Ci sono 6 nuove pren

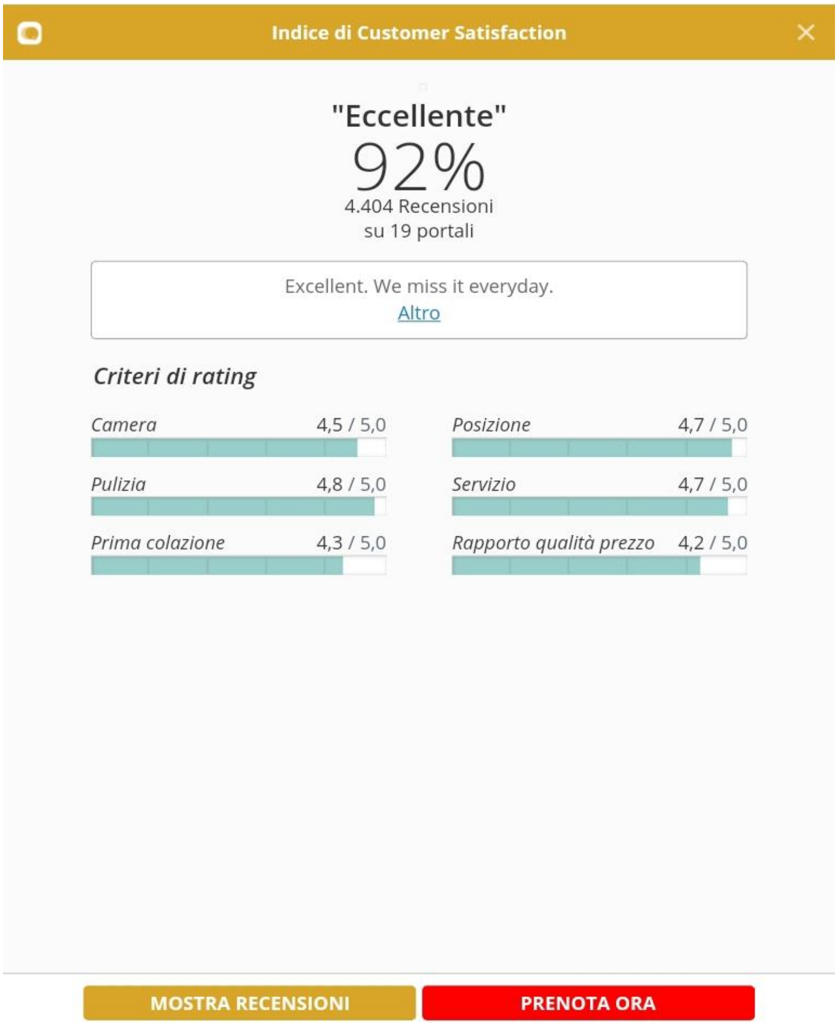
» Prenotazioni
» Home
» Logout

	APRILE 2011													
	Ven 08	Sab 09	Dom 10	Lun 11	Mar 12	Mer 13	Gio 14	Ven 15	Sab 16	Dom 17	Lun 18	Mar 19	Mer 20	Gio 21
SINGOLA	80	1 80	80	80	1 100	1 80	1 80	80	150	100	100	100	100	1 100
DOPPIA NON RIMBORSABILE	150	150	130	130	130	130	130	150	150	130	130	130	130	1 130
DOPPIA	2 130	1 130	2 100	2 150	2 150	2 100	2 100	100	180	1 250	200	1 150	1 150	2 150
DOPPIA ALLOTMENT EXPEDIA.COM	130	130	1 100	1 150	1 150	1 100	100	100	180	250	150	150	150	1 150
TRIPLA	1 160	1 160	1 130	180	180	1 180	1 180	210	210	180	180	1 180	1 180	1 180
QUADRUPLA	1 180	1 180	1 150	210	210	1 210	1 210	230	230	210	210	210	1 210	1 210
APARTSUITE (MAX 4 PAX)	1 150	200	2 100	1 150	150	1 100	100	100	200	1 300	1 300	150	150	2 300
TOT DISPONIBILI	5	4	6	3	3	6	5	-	-	2	1	2	3	8
TOT ALLOTMENT	-	-	1	1	1	1	-	-	-	-	-	-	-	1



6 APRI	826327381	07-04-2011 19:48:42	[redacted]	4	ApartSuite (max 4 pax)	Martedi 02-08-2011	2	Giovedi 04-08-2011	€ 220.00 - € 33.00	
7 be APRI	100411-B42D2C91	07-04-2011 17:33:55	[redacted]	1	Singola	Domenica 10-04-2011	2	Martedi 12-04-2011	€ 160.00	✖
8 be APRI	160511-F4354CEB	07-04-2011 13:15:40	[redacted]	3	ApartSuite (max 4 pax)	Lunedi 16-05-2011	4	Venerdi 20-05-2011	€ 600.00	✖
9 APRI	338792561 CANCELLATA	cancelled 07-04-2011 13:00:25 new 13-03-2011 10:33:11	[redacted]	4	ApartSuite (max 4 pax)	Domenica 17-04-2011	1	Lunedi 18-04-2011	~~€ 150.00~~	
10 Expedia APRI	242927601 CONFERMATA	06-04-2011 23:05:13	[redacted]	2	Doppia	Giovedi 14-04-2011	2	Sabato 16-04-2011	€ 150.00	
11 Expedia APRI	242924041 CONFERMATA	06-04-2011 22:40:20	[redacted]	4	ApartSuite (max 4 pax)	Venerdi 15-04-2011	1	Sabato 16-04-2011	€ 75.00	
12 APRI	757461400	06-04-2011 22:27:05	[redacted]	4	ApartSuite (max 4 pax)	OGGI Venerdi 08-04-2011	2	Domenica 10-04-2011	€ 200.00 - € 30.00	
13 APRI	229147946 CANCELLATA	cancelled 06-04-2011 22:21:06 new 06-12-2010 19:33:11	[redacted]	4	ApartSuite (max 4 pax)	Venerdi 22-07-2011	1	Sabato 23-07-2011	~~€ 110.00~~	
14						OGGI				

The second way, as it emerged from the interviews, "big data" technologies are used is to help hotels to aggregate and analyze all the online reviews present on online platforms and blogs. Big amounts of reviews are created on a daily basis, and as all the respondents stated they try to monitor them as well as possible, but there are often only one person who deals with them and is mostly the own manager to do that. Most of them, a part from two or three of the interviewees, employ the service of information companies that use aggregator software in order to collect the most significant ones and to create an aggregate score that comprised all the average scores present across the web. This allows them to have a general score or a general customer satisfaction index to present on their official websites and to contribute to their personal reputation.



Indice di Customer Satisfaction

"Eccellente"

92%

4.404 Recensioni
su 19 portali

Excellent. We miss it everyday.
Altro

Criteri di rating

Camera	4,5 / 5,0	*Posizione*	4,7 / 5,0
Pulizia	4,8 / 5,0	*Servizio*	4,7 / 5,0
Prima colazione	4,3 / 5,0	*Rapporto qualità prezzo*	4,2 / 5,0

CONCLUSIONS

“Big data” and “big data analytics” in a simplified way refers to the system of new and innovative tools employed to process a bigger amount of data, either structured or unstructured, which is no longer feasible to process with the previous techniques and technologies. This amount of data that can now be processed less costly and in an even faster way, can be so much valuable for many industries because it can bring several advantages to companies management. Having so much information coming from this increasing amount of data can lead to better decision-making processes and help companies to better draw their competitive environment. In fact, big data analysis can help to create more precise and focused strategic plans.
Tourism and hospitality industry are among the main sectors that can take big advantages from it. Especially for companies and activities working in these sector, it is crucial to obtain as much as possible information on their customers and competitors. They can predict tourists’ flows and deliver better targeted and customizable service and products.

The research carried out has highlighted some interesting points. Even though the processing of this big amount of data is now easier and less costly, not many companies in the hospitality industry totally take advantage of “big data” in the most efficient way. It is evident that this source of information is more accessible for big hotel chains or airline companies that have the necessary resources, expertise and technologies to exploit it at the best. It appears not to be the case for small independent hotels that do not often possess the right resources to invest in such technologies or that sometimes prefer to do it by themselves, because they do not confer to these technologies the real weight they have nowadays.

Another particular aspect emerged during the research is that in Italy “big data” and big data analytics are not exploited at their best, and that in most cases hotel managers do not know enough on this topic and do not know the real benefits of this “information revolution”. In addition to that, managers are not able to give the right interpretation to such an amount of data. This might be a weakness for many tourism firms since Tourism is one of the leading sectors of the overall industry in Italy.

According to what Euro Beinat , Data and Computer Science professor at the University of Salzburg, said during a workshop about future perspectives in tourism, “We need to switch to a more intelligent use of this information. This should involve a processing, through artificial intelligence, that allows to anticipate and drive tourists' choices and to organize a more focused and assessed touristic offer". Moreover, "the main challenge is to drive all the phases of the touristic product characterized by managerialism and functional organization. With this array of information at our disposal and the right system and software to employ it, we only need to put courage and willingness to exploit the new opportunities, by changing our way of working and thinking about business. The touristic sector is one among the other industrial sectors that has more of this kind of data available but that exploit it at the lowest". Furthermore during the workshop, Alessandro Nucara, general manager of Federalberghi, stated that "big data form a still long path to go through even and especially for all the small firms that characterize our touristic sector. In a world that is ever more interconnected, they are undoubtedly the tools able to develop the synergy with all the other touristic operators. We need to share more opportunities by enhancing the excellences of our territory and by favoring the complementarity and the economies of scale".

In accordance to a survey conducted by ISNART (Istituto Nazionale Ricerche Turistiche), 53,8% of the Italian touristic firms ask for a greater availability of data in order to increase their business activity. The most appealing data are those about the reasons of destinations or accommodation choice and those about the main indicators of a territory resources. The percentage above mentioned considers that having information on the features and behavioral patterns of Italian and foreign tourists would be very helpful. Of course the percentage of managers that share this consideration vary across regions in the country. North-eastern firms (75,7%) and Central Italy firms (58,9%) consider important the information about the reasons why tourists choose a particular trip and destination, north-western firms consider more important information about the reasons why tourists choose a particular type of accommodation(48,8%), while southern firms and those in the main islands are more interested in personal and economic data about their tourists.
According to what has been explained in the survey above mentioned, Italian chambers of commerce are also starting to employ big data to design a new observing model on economic features of the Tourism industry. This could be very helpful for small touristic firms that are not able to manage this big amount of data, but this will be feasible only if they start to be aware of the real benefit of big data and its analysis.

In consideration of the still long path Italian firms in tourism have to ride for what concerns big data, I decided to focus my paper on one of the basic data found on the web that hotels can exploit, i.e. online user-generated contents data. This can be one of the best example of "big data" because they are mostly unstructured data composed by text, transactions, images, posts, video, not simply statistical figures. I tried to demonstrate though some interviews what I found in the literature and in several articles about online user-generated contents and online reputation. As some articles has highlighted, the best example of user generated content is the review. Reviews are demonstrated to greatly influence customer purchasing behavior, and for this reason they can be used to better online reputation and online visibility. During the interviews it has clearly emerged that they play a big role for the reputation and image a hotel wants to conceive and as a mean to attract more potential customers. Through the processing and the analysis of some data found on these contents it is possible to create a scorecard on the strengths and weaknesses according to what customers think and want . It is possible for hotels and other kinds of accommodation to develop more targeted offers according to their customer preferences, in order to deliver them the best service and quality ever. In fact, in this competitive environment, quality is the must have for every facility in order to see evident revenue increments. A good online reputation and visibility can give hotels the chance to sell their rooms and service at higher prices than their competitors and ensure higher occupancy rates. This is undeniable valuable for all of them which will be able to extract and exploit data at their disposal in the most efficient and effective way.

I decided to conduct these interviews among a group of hotels in Venice because I assumed this city reflects one of the best example of "international" touristic destination, where an analysis of its visitors features thank to big data analytics can be so much valuable for all the facilities located there. This is true especially for many quality hotels and accommodation facilities that have to face a strong competition in order to attract the same quality customers they need.

Unfortunately, during the interviews it has emerged that for most independent hotels this big potential is still underestimated. They have clearly heard of big data but they still do not know how to fully exploit them. In most cases they rely on web marketing companies to help them with their website and online visibility, but they do not merely need numbers or statistics, rather they need a total support.

Furthermore, the research has proved that "big data" issue is still at its infancy in Italy. Many articles found on the web describe how this amount of data and related technologies (artificial intelligence) could be helpful to businesses, and in particular for the tourism industry, that according to some authors (like Euro Beinat, Data Science professor at Salzburg University) is one among the other sectors that have at its disposal the greatest amount of “big data”, but it unfortunately appeared that many managers still have to learn about it. They still have to learn how to interpret this data in the most efficient way and how to use it to be more competitive, in a world where big foreign companies such as TripAdvisor and booking.com earn the major part of economic value created by big data. This is an additional reason why hotel managers should be more informed and prepared on this theme and that’s why they need a complete support by the main authorities and associations in the tourism industry. It is apparent that in Italy the real big data “revolution” should start among hotels managers since in most cases what really lacks in hospitality industry, and at a bigger extent in Italian Tourism, is managerialism.

REFERENCES

1. Accenture Consulting. *"Big Success with Big Data"*, 2014.

2. Alireza Alaei, Susanne Becken. *Sentiment Analysis in Tourism: Capitalizing on Big Data-* December 2017, Griffith University;

3. Allan, J., J. G. Carbonell, G. Doddington, J. Yamron, and Y. Yang. "Topic detection and tracking pilot study final report." Proceedings of the Broadcast News Transcription and Understanding Workshop, 1998.

4. Andrea Ciccarelli, Eleonora Scarsella, Adolfo Braga. *In viaggio con un click, Nuovi strumenti di marketing digitale per il settore turistico*- 2017;

5. Andreassen T.W. and Streukens S. *"Service innovation and electronic word-of-mouth: Is it worth listening to?"*, 2009.

6. Akehurst G." User generated content: the use of blogs for tourism organizations and tourism consumers", 11 December 2008.

7. Aurchana P., R. PIyyappan, and P. Periyasamy. *"Sentiment analysis in tourism."* International Journal of Innovative Science, Engineering & Technology, 2014.

8. Bansal, H. S., & Voyer, P. A. "Word-of-mouth processes within a services purchase decision context", 2000.

9. Bajari P., Nekipelov D. , Ryan S. and Yang M. *"Machine Learning Methods for Demand Estimation",* American Economic Review 2015.

10. Beyer M.A. and Laney D: *"The importance of Big Data: a Definition"*, 2012

11. Bing Liu. *"Sentiment Analysis and Subjectivity",* Department of Computer Science University of Illinois at Chicago, 2010.

12. Brob, J. *"Aspect-oriented sentiment analysis of customer reviews using distant supervision techniques."* PhD Thesis, Department of Mathematics and Computer Science, University of Berlin, 2013.

13. Brown James R., Dev Chekitan. *"Looking beyond RevPAR: Productivity Consequences of Hotel Strategies"*, 1999.

14. Bjorkelund, E., T.H. Burnett, and K. Norvag. "*A study of opinion mining and visualization of hotel reviews."* in Proceedings of the 14th International Conference on Information Integration and Web-based Applications & Services, 2012.

15. Boston Consulting Group*, "Industry 4.0 The future of Productivity and Growth in Manufacturing Industries"*, April 2015.

16. Bucur, C. *"Using opinion mining techniques in tourism."* in Proceedings of the 2nd Global Conference on Business, Economics, Management and Tourism, Procedia Economics and Finance, 2015.

17. Cerami Gina, *"Are You Screen Scraping or Data Mining?"*,2012-Connotate

18. Chareyron Ga¨el, Da-Rugna J´erˆome and Raimbault Thomas, *"Big Data: a new challenge for tourism"*in IEEE International Conference on Big Data, 2014.

19. Chen M., Mao S., Zhang Y. and Leung V. *"Big Data: Related Technologies, Challenges and Future Prospects",* 2014.

20. Chen H, Chiang R.H.L, Storey V. "*Business Intelligence and Analytics: From Big Data to Big Impact"*, December 2012.

21. Choi, S., X.Y. Lehto, and A.M. Morrison. *"Destination image representation on the web: content analysis of Macao travel related websites"*, 2007

22. Choudhury, N. N*. "Sentiment analysis of twitter data for a tourism recommender system in Bangladesh."* Master Thesis, School of Science, Aalto University, 2016.

23. Davenport Thomas H.. *At the Big Data Crossroads: turning towards a smarter travel experience*- 2013 Amadeus IT Group Report.

24. Davenport Thomas H. and Dyché J. *"Big Data in Big Companies"*, May 2013.

25. Deloitte. "*Industry 4.0: challenges and solutions for the digital transformation and use of exponential technologies"*-2014.

26. De Mauro A., Greco M. and Grimaldi M. "*A Formal Definition of Big Data based on its essential Features",* in Library review 2016.

27. Dhiratara Arkka Yang Jie, Bozzon Alessandro, Houben Geert-Jan. *Social Media Data Analytics for Tourism A Preliminary Study*- 2016;

28. Dodds, P.S. et al*. "Human language reveals a universal positivity bias."*, . 2015

29. Feldman, R. *"Techniques and applications for sentiment analysis*." Communication of the ACM, 2013.

30. Garcia, A., S. Gaines, and M. T. Linaza*. "A Lexicon based sentiment analysis retrieval system for tourism domain.*", 2012.

31. Ghose, A., P.G. Ipeirotis, and B. Li. "Designing ranking systems for hotels on travel search engines by mining user-generated and crowd sourced content.", 2012.

32. González Santiago M., Gidumal,Jacques B. and González L.V. Beatriz *." Online customer reviews of hotels: as participation increases, better evaluation is obtained",* 2013.

33. Gretzel, U., Yoo, K.H. & Purifoy, M. (2007), "Online Travel Review Study: Role and Impact of Online Travel Reviews", Laboratory for Intelligent Systems in Tourism, Texas A & M University 2007.

34. Hal R. Varian. *Big Data: New Tricks for Econometrics*- June 2013, Revised: April 14, 2014;

35. Hennig-Thurau T., Gwinner Kevin P., Walsh G., Gremler D. D. "*Electronic word-of-mouth via consumer-opinion platforms: What motivates consumers to articulate themselves on the Internet?",* 2004.

36. Hippner, H., & Rentzmann, R. *"Text mining."*, 2006.

37. Höpken, W., M. Fuchs, T. Menner, and M. Lexhagen. "Sensing the online social sphere – the sentiment analytical approach", 2016.

38. Hutto, C. and E. Gilbert. "Vader: *A parsimonious rule-based model for sentiment analysis of social media text."* In Proceedings of the 8th International AAAI Conference on Weblogs and Social Media, 2014.

39. Josian Mackenzie "Online guest review data is revolutionizing hotel investment", in ReviewPro News, January 27, 2014.

40. Kadri-Liis Kusmin. *IFI8101 - Information Society Approaches and ICT Processes: Industry 4.0;*

41. Kang, H., S. J. Yoo, and D. Han. *"Senti-lexicon and improved Naïve Bayes algorithms for sentiment analysis of restaurant reviews."*, 2012.

42. Karakaya F. *"Impact of online reviews of customer care experience on brand or company selection"*, 2010.

43. Kuttainen, C., M. Lexhagen, M. Fuchs, and W. Höpken. "*Social media monitoring and analysis of a Swedish tourism destination"* in Proceedings of the 2nd Conference on Advances in Hospitality and Tourism Marketing and management, 2012.

44. Laney D. *"3D Data Management: Controlling Data Volume, Velocity and Variety"*, February 2001.

45. Levallois, C. *"Umigon: Sentiment analysis for tweets based on terms lists and heuristics."* In: Second Joint Conference on Lexical and Computational Semantics, 2013.

46. Lewis R. and Chambers R. *"Marketing Leadership in Hospitality: Foundations and Practices"*, 2000.

47. Markopoulos, G., G. Mikros, A. Iliadi, and M. Liontos. *“Sentiment analysis of hotel reviews in Greek: A comparison of unigram features of cultural tourism in a digital era.”*, 2015

48. Menner, T., W. Höpken, M. Fuchs, and M. Lexhagen. *“Topic detection – Identifying relevant topics in tourism reviews.”*,2016.

49. O’Leary, D. *“The use of social media in the supply chain: Survey and extensions.”* Intelligent Systems in Accounting, Finance and Management, 2011.

50. Pang, B., L. Lee, and S. Vaithyanathan. “*Thumbs up?: Sentiment classification using machine learning techniques.”*, 2002.

51. Pappas, N. and A. Popescu-Belis. “*Sentiment analysis of user comments for one-class collaborative filtering over ted talks”*, 2013.

52. Park D., Lee J. and Han I. “*The Effect of On-Line Consumer Reviews on Consumer Purchasing Intention: The Moderating Role of Involvement”,* 2014.

53. Park S., Yang Y., *“Electronic word of mouth and hotel performance: A meta-analysis”*, 2018.

54. Peressotti Veronica, “*Il vero significato di Industry 4.0 Quali impatti avrà sulle aziende”*, 2016.

55. Phillips P., Zigan K., Barnes S. Schegg R. *“Understanding the Impact of Online reviews on Hotel Performance: An Empirical Analysis”*, 2016.

56. Picciano A. “*The Evolution of Big Data and Learning Analytics in American Higher Education”,* June 2012, Journal of Asynchronous Learning Network.

57. Rabanser U., Ricci F. *“Recommender Systems: Do They Have a Viable Business Model in e-Tourism?”* In: Frew A.J. (eds) Information and Communication Technologies in Tourism 2005.

58. Raguseo E., Neirotti P. and Paolucci E. “How small hotels can drive value their way in infomediation. The case of ‘Italian hotels vs. OTAs and TripAdvisor’ “, 2017.

59. Raguseo E., Neirotti P. and Paolucci E. “Are customers’ reviews creating value in the hospitality industry? Exploring the moderating effects of market positioning”, 2016.

60. Rao A.R, Qu L., Ruekert R.W. *“Signaling unobservable product quality through a brand”*, 1999.

61. 2016 Raritan Inc. “British Airways Case Study”.

62. Ribeiro, F.N., M. Araujo, P. Goncalves, M.A. Goncalves, and F., Benevenuto. *“A benchmark comparison of state-of-the-practice sentiment analysis methods.”*, 2016.

63. Roland Berger Strategy Consultants. “THINK ACT Industry 4.0 The new Industrial Revolution”,2014.

64. Rossetti, M., F. Stella, L. Cao, and M. Zanker. *“Analysing user reviews in tourism with topic models.”*, 2015.

65. Russom Philip. “*Big Data Analytics*”, 2011.

66. Schmunk, S., Höpken, W., Fuchs, M. and M. Lexhagen. *“Sentiment analysis: Extracting decision-relevant knowledge from UGC.”*, 2014.

67. Serra Cantallops A. and Salvi F. *“New Consumer Behavior: A Review of Research on eWOM and Hotels”*, 2014.

68. Shi, H.-X., X.-J. Li. *“A sentiment analysis model for hotel reviews based on supervised learning.”,* 2011.

69. Shimada, K., S. Inoue, H. Maeda, and T. Endo. “*Analyzing tourism information on Twitter for a local city.”,* 2011.

70. Siemens G. and Long P. *“Penetrating the Fog: Analytics in Learning and Education”,* 2011.

71. Song Haiyan, Liu Han. “*Predicting Tourist demand Using Big Data”,* 2017.

72. Sparks B.A. and Bowning V. *“The impact of online reviews on hotel booking intentions and perception of trust”,* 2011.

73. Szalay A. and Blakeley J. “*Gray’s Laws: Database-centric Computing in Science*”, 2009.

74. Talend Corporation. *“AirFrance-KLM business case”*, 2016.

75. Tckhakaia Ekaterina, Rodrigues Sergio Assis, Cabras Ignazio *“Knowledge Management in Airline Industry: Case Study from the British Airways”,* 2015.

76. Tsytsarau, M., and T. Palpanas*. “Survey on mining subjective data on the web.”*, 2012.

77. Wang Ye, Rodgers Shelly “*Electronic Word of Mouth and Consumer Generated Content From Concept to Application”*, 2011.

78. Xiang Z., Yufeng M., Qlanzhou D., Welguo F*. “A comparative analysis of major online review platforms: Implications for social media analytics in hospitality and tourism”*, article in Tourism Management , Feb 2017.

79. Xiang, Z., Z. Schwartz, and M. Uysal. *“What types of hotels make their guests (un-) happy? Text analytics of customer experiences in online reviews*.”, 2015.

80. Ye, Q., Z. Zhang and R. Law. *“Sentiment classification of online reviews to travel destinations by supervised machine learning approaches”*, 2009.

81. Zaho X., Wang L. Guo X. and Law R. *“The influence of online reviews to online hotel booking intentions”*, 2015.

82. Zakir Jasmine. *Big Data Analytics*- September 2015, Dell Inc..