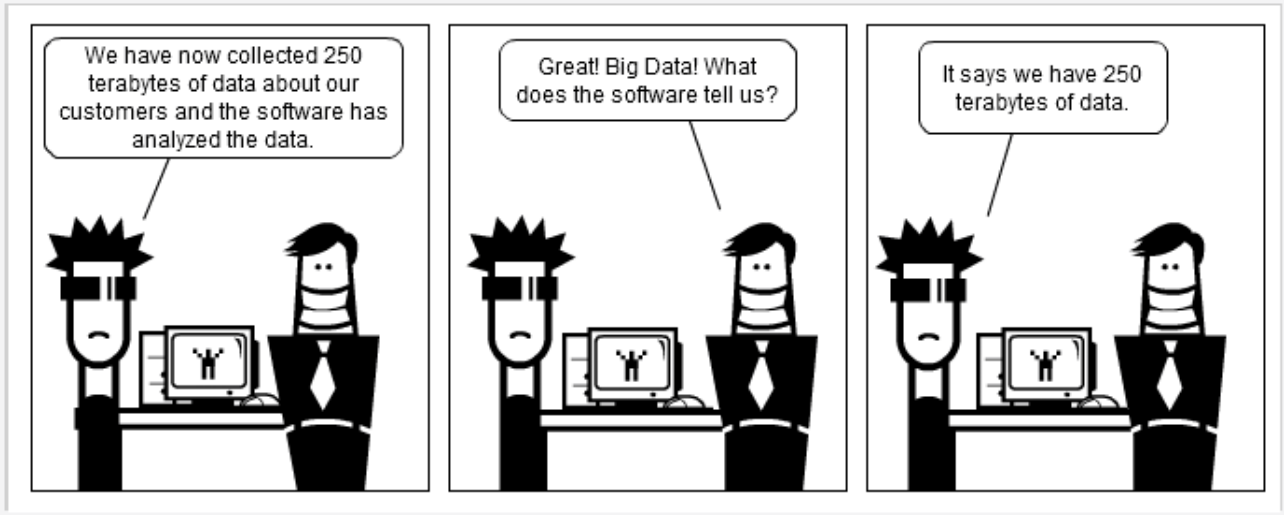
Big data

Interim report in the context of the joint inquiry on "Big data" launched by the AGCOM deliberation No. 217/17 / CONS

Department of Economics and Statistics

AUTORITÀ PER LE GARANZIE NELLE COMUNICAZIONI
AGCOM



"The big data Challenge"

Sean R. Nicholson
www.socmedsean.com

"Big data is like teenage sex: everyone talks about it, nobody really knows how to do it, everyone thinks everyone else is doing it, so everyone claims they are doing it…"

Dan Ariely,
Center for Advanced Hindsight
Duke University

"You talk to a kid these days and they have no idea what a kilobyte is. The speed things progress, we are going to need many words beyond zettabyte."

Adrian McDonald
President, Dell EMC EMEA

Department of Economics and Statistics

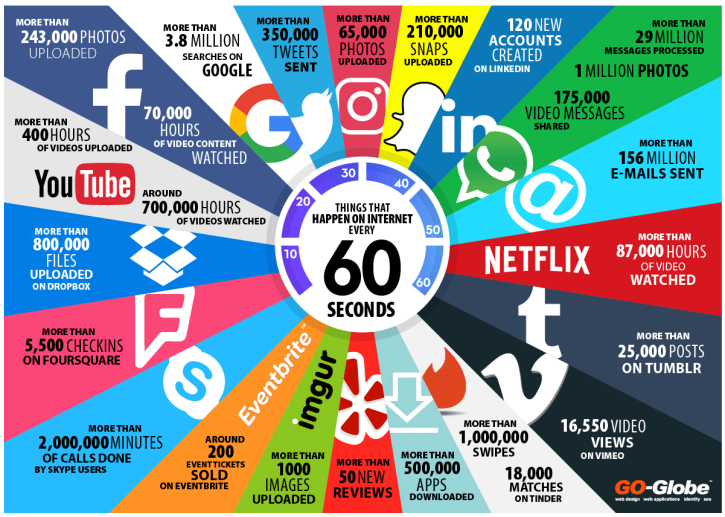
AUTORITÀ PER LE GARANZIE NELLE COMUNICAZIONI
AGCOM

June, 2018

EXECUTIVE SUMMARY

The characteristics of big data

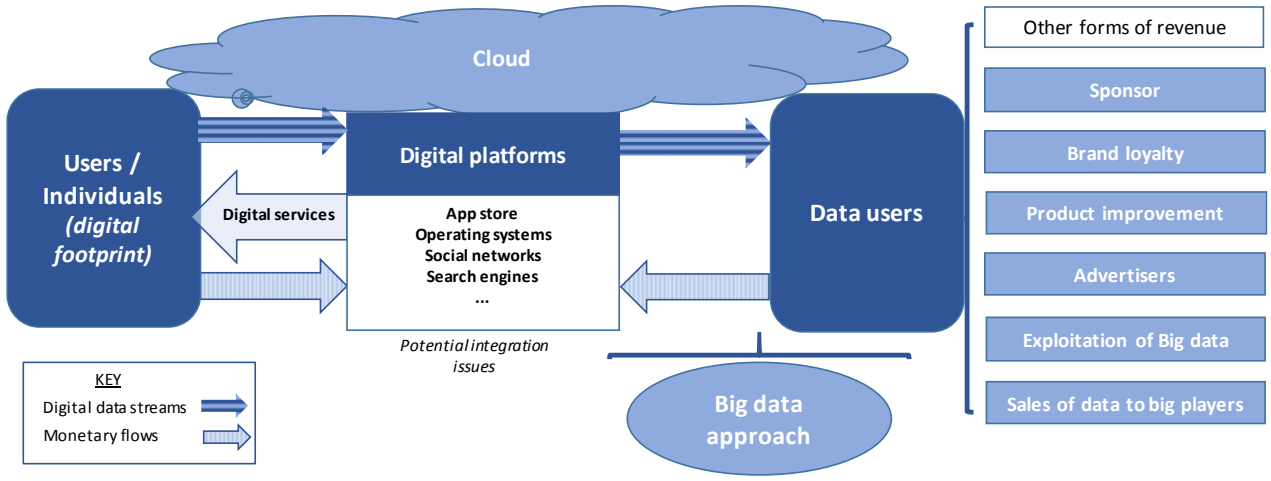
- We are in a period when the use of data is now essential in the decision-making processes of companies, institutions and, increasingly, even of individual citizens. **The current technologies allow, in fact, the ever greater diffusion of the "datafication" processes, a neologism through identifying that set of techniques allowing the conversion in digital format - that is in data - of anything** (films, books, vocal messages, body movements, etc.).
- Words turn into data, geographical position turn into data, social interactions turn into data, even things, if connected through electronic communications networks (Internet of Things - IoT), become data. Sources can be found in any device, sensor, operating system, search engine, social network.
- **The increasing use of the internet by individuals, particularly via mobile devices, is an unlimited source of data**; tracks are left on the web at any time (the so-called ***online footprint***), in moving from one place to another, in sharing photos or comments, in making payments, in doing sports, etc.
- ***Big data* represent the key productive factor in a data-driven economy**; there are several areas, both private and public, where the use of analysis techniques of *big data* has allowed to create new services, improve existing ones, innovate production and distribution processes, make the offer of all products and services (even non-digital) responding to the needs of consumers and citizens.
- *Big data* refer to a jump of **interpretive paradigm of economic and social reality** through data analytics methods (*data mining*) carried out on **huge amounts of data (volume), characterized by very different formats (varieties), stored and processed at an increasingly rapid speed (velocity) often in real time**.



Big data: Volume, Variety and Velocity of data (the data flow on the internet in 60 seconds)

The big data ecosystem

- In the *big data* ecosystem, it is possible to identify, among others, the following main actors:
  - ✓ **subjects generating data** (data "providers" or "data subjects" according to the GDPR definitions);
  - ✓ the **suppliers of technological equipment**, typically in the form of data management platforms;
  - ✓ the **users**, i.e. who use the *big data* to create added value;
  - ✓ **data brokers** that is, organizations collecting data from a set of sources, both public and private, offering them, upon payment, to other organizations;
  - ✓ the **companies and research organizations**, whose activity becomes fundamental for developing new technologies, new algorithms through which to explore data and extract value;
  - ✓ the **public bodies**, both as market regulators bodies, and with reference to the activities of public administration aimed at improving the products and services offered to the citizens and able to increase the social welfare.
- The *big data* ecosystem shows a **degree of interconnection between the different parties involved making it difficult to identify single well-defined markets**; the resulting complexity determines a scenario in which the various segments of the system are often closely interrelated. All this leads to a context in which **(few) large multinational companies, characterized by a high degree of integration in all phases of the ecosystem, work together with a myriad of small specialized companies**.

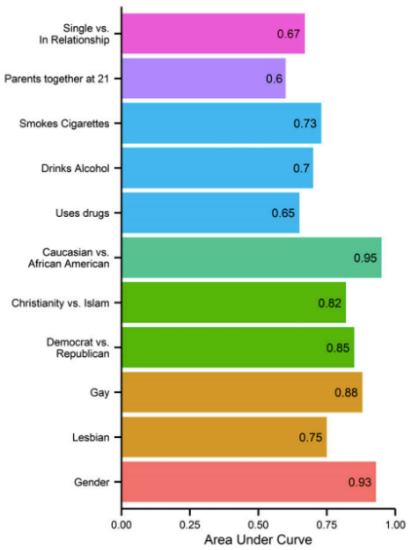


Synthetic representation of the double-side market applied to digital data

- **Market failures** are related to the existence of **barriers to entry and to develop, which can be found at all stages of the value chain**.
- One of the main segments that will rapidly evolve is the one related to **data centress**. As the size of the data collected grows, the need to invest in data acquisition, storage and analysis technologies increases. In this context, the world market is converging towards concentrated assets where the positions of online platforms such as Amazon and Google stand out.
- These effects can occur simultaneously (in parallel) in different components of the ecosystem, reinforcing each other and facilitating the rise **of highly concentrated market areas** (for example, markets for operating systems, search engines, social networks).

The individual as a data source

- Whenever individuals are connected to the network, they leave several "tracks", which are given to the operators either in an informed way or, more often, unconsciously. **The digital footprint of each individual** is made up of numerous information, some of which can be directly associated with him (name, surname, age, etc.), others associated with the activities carried out by individuals (payments, research, etc.), others which, although not having direct ties with the individual can be associated easily with specific persons through their processing.
- The phenomenon of *big data* made the traditional **distinction between "personal data" and "non personal data" completely obsolete** since it is extremely difficult to establish *ex ante* among all the information gathered about an individual, whether they constitute personal data, or not. These take on different nature depending on the amount of data stored, the context, as well as the analysis technologies. For example, **some psychometric techniques can easily gather sensitive individual information (such as political orientation, drug addiction, etc.) from a set, now also reduced, of non-personal data**.



Predictions of psychometric models (results from 68 *like*)

- The choices of an individual relating to the transfer of his data in order to obtain a service are directed according to the balance between benefits, often immediate (e.g. access to a service) and costs (often uncertain and unknown). In this context, **the information asymmetry between users and operators is pervasive and structural**: the consumer does not have all the information he needs to make an informed choice, but many of the behaviors, to be efficient, would require a degree of technical knowledge that goes far beyond the skills widespread among the population.
- A **higher degree of transparency is often useless where consumers fail, due to a structural gap in technological knowledge, to understand this information**. Furthermore, choices such as those relating to the transfer of personal data are carried out very frequently impulsively and without an evaluation of the real consequences of the implicit exchange.

Data exchange: incomplete contractual relationships and implicit markets

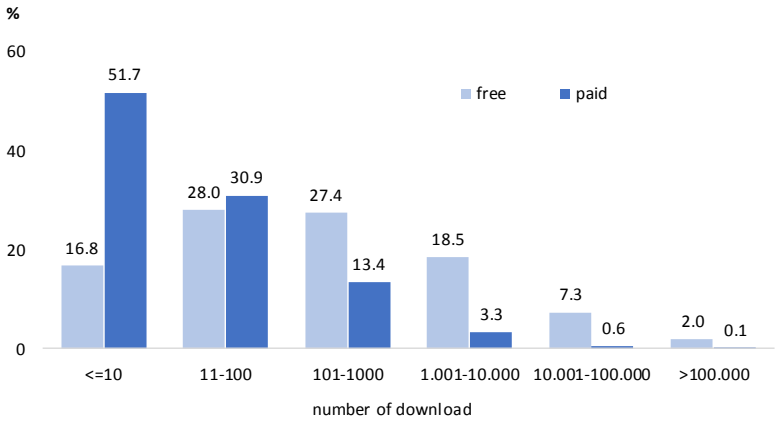
- The **data exchange** often provokes **structural market failures**. On one side, because the investments made by companies to collect data on individuals, which do not internalize social costs, are likely to lead to an **over-investment in gathering information**. On the other, in the presence of transaction costs and uncertainty regarding the assignment of property rights to data, probably market forces are not able to guarantee the achievement of an efficient situation. **The possibility that the interests of those who hold wider technical knowledge and information about the data will prevail, materializes**.

- The *download* and the subsequent use of applications is one of the main mechanisms used by the consumers to transfer data. The **APP *stores*** are an important example of **methods through which digital data are exchanged.**

- In this report, **the Italian Communications Authority (AGCOM) has analysed a dataset with over a million applications**. It emerged how **free apps require a significantly higher number of individual data compared to paid** apps. There is, in essence, an **implicit exchange of data between users and operators, which is part of the commercial relationship concerning the APP.**

- The absence of a real market mechanism can only make these relationships incomplete and inefficient. **The consumer does not have a clear perception of which data are transferred, of their real value (price) and how they are treated, both for primary and, even more so, secondary uses.** It is about **a *one-off* transaction concerning other goods (the APPS), against the dynamic use of users data.** It is, therefore, the same structural configuration of the market and of the related transactions to be distorted and, consequently, to lead to incomplete markets, which inevitably produce inefficient and unbalanced results.

- **The trend of downloads, moreover, detects a "long tail" phenomenon.** This determines that only a few APPS, 2%, is installed by a considerable number of users. **Only 6 APPS are installed more than 1 billion times: *Facebook*, *Google Gmail*, *YouTube*, *Google Maps*, *Google Search* and *Google Play Services*. With a very large number of applications and operators, the market is concentrated in a few large platforms.**

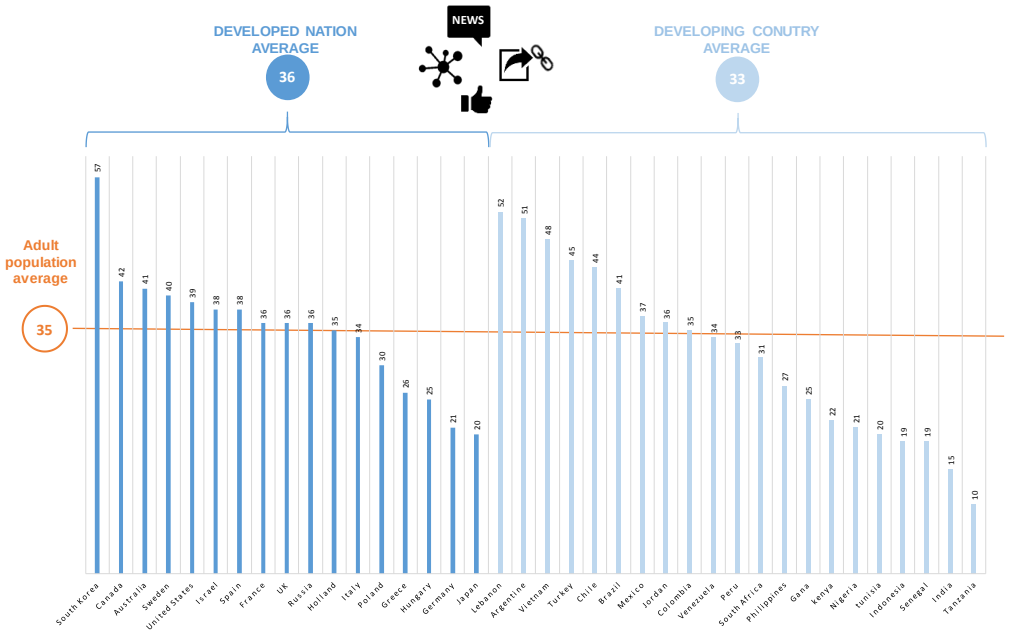


Distribution of APPS according to the number of *download*

Big data in the communications sector, and the online news media system

- **The use of *big data* from search engines and *social networks* is an aspect of particular importance due to the increasingly important role played by these platforms in the information system, at international and national level.** On one hand, their capacity of gathering personal information and extracting value from data by means of accurate profiling, *inter alia*, makes these actors the world *leaders* in the online advertising sector - a resource that is still the main source of funding for online information -; on the other, they now represent the main distribution channel for online news, provided that they operate as gatekeepers for access and distribution of online contents.
- **The spread of *big data* is structurally changing the global information ecosystem**. In particular, ***social networks*** - due to the time spent by users within the same, the multiple actions that individuals perform and the reactions they express through their profiles / pages /accounts as well as the social relationships they establish - **are certainly among the operators able to acquire the greatest variety and the largest volume of data on individuals, including those related to ideological and political preferences and information content read, viewed, appreciated, commented and shared.**

- **The *social networks* have definitely become an integral part of the daily informative diet of citizens in Italy and in the world.**



Use of *social networks* to keep up in the news in Italy and all over the world (2017; %)

- Despite the growing importance attributed by citizens to *social networks* as information tools, **pathological forms have recently emerged such as those relating to the polarization of citizens** (i.e. the tendency to acquire mainly information consistent with their ideological preferences) **and to phenomena of disinformation** (such as *fake news*).
- Through *social networks*, automatic customization systems (operating on the basis of algorithms and *big data* acquired), on one hand, and the actions of sharing information contents made by users, on the other, help the proliferation of fake news and the viral widespread of polarizing contents.

A new policy paradigm

- The technology shift related to the advent of *big data*, and *of data-driven economy*, needs a **change of paradigm also at the policy-approach level**.
- First of all, *big data* make it necessary to overcome the traditional distinction between different types of data (personal, sensitive, etc.). The **new approach must refer simply to the data "*per se*"**.
- In addition to the undisputed economic and social benefits deriving from the advent *of the data-driven economy*, **some risk factors exist**. The existence of **causes of market failure** (such as incomplete contracting, implicit markets, information asymmetries, market power positions) has been accounted. Furthermore, new possible **discriminatory practices** emerge, among which those linked to the price are the most widespread. The **price discrimination**, which with the modern **online profiling techniques** becomes "perfect", involves a sure effect of social redistribution prerogative of online operators and, in a system on several sides, to the detriment of specific categories of users (which from time to time may be consumers, workers, publishers, etc.). These practices, even when theoretically efficient, **present very significant social risks**. For example, discrimination, often on an algorithmic basis, risks extending, even involuntarily, to differences in the population based on ethnicity, race, sexual orientation, and health condition.
- The market **failures** have repercussions on the whole social context, including the **information system**, the **pluralism of sources**, and the same **methods of social aggregation** and **of public opinion creation**.
- As a consequence of the existence of structural and lasting market failures, it is necessary, especially where social and political rights are under discussion, to adopt an ***ex ante* approach to the data regulation** (and to possible regulation of related algorithms).
- Moreover, this new paradigm must take into account that the **information asymmetries between users and operators are pervasive and structural**. In this context, **it is difficult to restore conditions of efficiency through mechanisms of transparency and informed consent**. In fact, these regulatory tools appear, in many cases, insufficient to guarantee a cognitive rebalancing between operators and consumers, in a situation in which subjects such as experts in the sector, specialized institutions and research centres often do not have at their disposal sufficient knowledge to understand the entity and the very nature of the phenomena. In line with what is already happening in high-tech contexts (such as those of electronic communications), it appears necessary to accompany the new regulation towards **technical forms of direct regulation of operators using *big data***.
- Preliminarily, the new paradigm needs to **open the black box** regulating the processes taking place within the *big data* ecosystem, such as, among others, **moments and methods of data acquisition (*data gathering & storage*), functioning of the algorithms (*algorithm accountability*), methods of conservation and analysis (*data analytics*), derived information, and deriving (primary and secondary) uses**. With respect to these, and others, aspects we still know too little.
- The **new approach must therefore be based on facts, information and knowledge**. In this sense, AGCOM has already started research cooperation with the most prestigious national and international universities (in the case of this Report, the Department of Computer Engineering of "La Sapienza" University of Rome) and carried out analysis with experts in the field (in the case of online disinformation, with Prof. Walter Quattrociocchi).

- Furthermore, AGCOM has already launched the new strategy, in its areas of expertise, through the **Establishment of the Technical Roundtable to guarantee pluralism and fair information on online platforms**. On the basis of the current national and community regulatory context, this initiative tries to describe some of the principles of the new approach: open the black box with analysis and surveys based also on information requested to online platforms; analyse *newsfeed* and recommendation algorithms; identify and bring out collective and shared solutions to the focused market problems; define *ex ante* rules within the operators.

TABLE OF CONTENTS

INDEX OF TABLES AND FIGURES

Foreword

On several occasions, AGCOM has highlighted the importance of big data also in terms of media pluralism since the phenomenon is closely linked to the role of online platforms and, therefore, to the impact of big data on the functioning of the mechanisms they adopted in spreading information. The use of news on the internet is increasingly occurring through these new digital intermediaries (social networks, search engines,...) which, like other platforms, use data as a strategic asset, according to the logic of multilateral markets, to offer online services and contents, with the consequent need to reconcile the trade-off between the commercial value of information and respect for fundamental individual and collective rights such as privacy, protection of competition and guarantees of media pluralism.

The presence of such a complex phenomenon in its nature, which is characterized by a strong interdependence and transversal contents, and whose effects may relate to information, competition, protection of consumers and their privacy, necessarily presuppose a deepening of the issues through an interdisciplinary and joint investigation with the Competition and Market Authority and the Italian Supervisor for the protection of personal data, launched with deliberation no. 217/17 / CONS.

This Report, therefore, is part of the work of the aforementioned Joint Sector-Inquiry on Big Data and represents a crucial step, at least an intermediate one, through which we highlight the main problems and opportunities arising from the use of big data, with particular reference to the markets (those of communications) and to subjects (media and political pluralism, consumer protection) of strictly within the institutional competence of Agcom. This temporary report, therefore, is functional to the future work of the ongoing joint investigation, representing indeed a "guide" simultaneously offering tools, mainly economic concepts, and ideas for the identification and dissertation of further aspects in the subsequent phases of the investigation, also in view of the problems observed.

Introduction

Throughout the world, the innovative use of data in decision-making processes is creating a deep change involving every aspect of economy and society. This change of pace, determined by new technologies and techniques for data collection, storage and analysis, is producing several benefits occurring both at the individual level (consumers and companies) and at the aggregate level (local and national), improving the quality of life, opening up new economic and social opportunities.

According to a report by *IDC* and *Open Evidence*, the value of the data market in Europe will reach, in 2020, 106 billion euros, against an estimate of 60 billion, for 2016, with a direct impact on the whole continental economy that will reach 4% of GDP.[1]

The phenomenon of "datafication", i.e. the transformation of any information (films, books, vocal messages, body movements, etc.) into data, the progressive increase in the use of online communication tools by citizens and companies, as well as the consequent growth of the digitalisation of production processes, not only give rise to a large amount of economic and social data, available and processed at ever-increasing speed, but also with a growing variety of formats, or, in short, to what has been defined as the phenomenon of the *big data*.

The *big data* represent the key productive factor in the *data-driven* economy; there are many areas, both private and public, in which the use of *big data* analysis techniques has allowed creating new services, improving existing ones, innovating production and distribution processes. Making the offer of all products and services (including non-digital ones) increasingly responding to the needs of consumers and citizens.

This tendency appears being incontrovertible and strengthened by the fact that, for the vast majority of individuals, a significant part of private life, as well as that of work life, "moved" to the network thus becoming one of the main sources of data.

Despite the *big data* present rip-roaring potentialities, many of which are still unexplored, it is necessary to underline the presence of some associated risks. First, the ecosystem of the *big data* is characterized by the presence of several forms of incomplete contracting, by implicit markets (i.e. in which the bargaining of the asset takes place in a spurious manner), as well as by notional areas (i.e. characterized by perfect vertical integration and by merely potential market demand). All these risks being the source of deep market failures affecting social, static and dynamic efficiency of the whole system. Secondly, there are collective risks, linked, *inter alia*, to the lack of incorporation of positive and negative externalities in the market. This result appears being very relevant in case of diseases recently detected in the field of networked information, and related to phenomena such as the so-called online misinformation.

[1] IDC and Open Vision, European Data Market SMART 2013/0063 Final Report, February 2017. This report outlines three medium-long term scenarios: a *baseline scenario (or reference scenario)*, whose main hypothesis is the maintenance of current growth trends and the evolution of the current conditions of the European regulatory framework; the second scenario is the high growth scenario in which data market enters in a trajectory of rapid growth (*High Growth scenario*), thanks to more favourable regulatory and macroeconomic conditions; and finally a challenge scenario (*Challenge scenario*) in which data market grows slower than in the reference scenario, due to the less favourable framework conditions and a less favourable macroeconomic environment.

Big data represent a recent field of study and research; as shown by the **Figure (i)**, reporting the analysis of billions of searches carried out in the world by users of the search engine *Google*, the term "*big data*" presents a strongly increasing *trend*.[2]

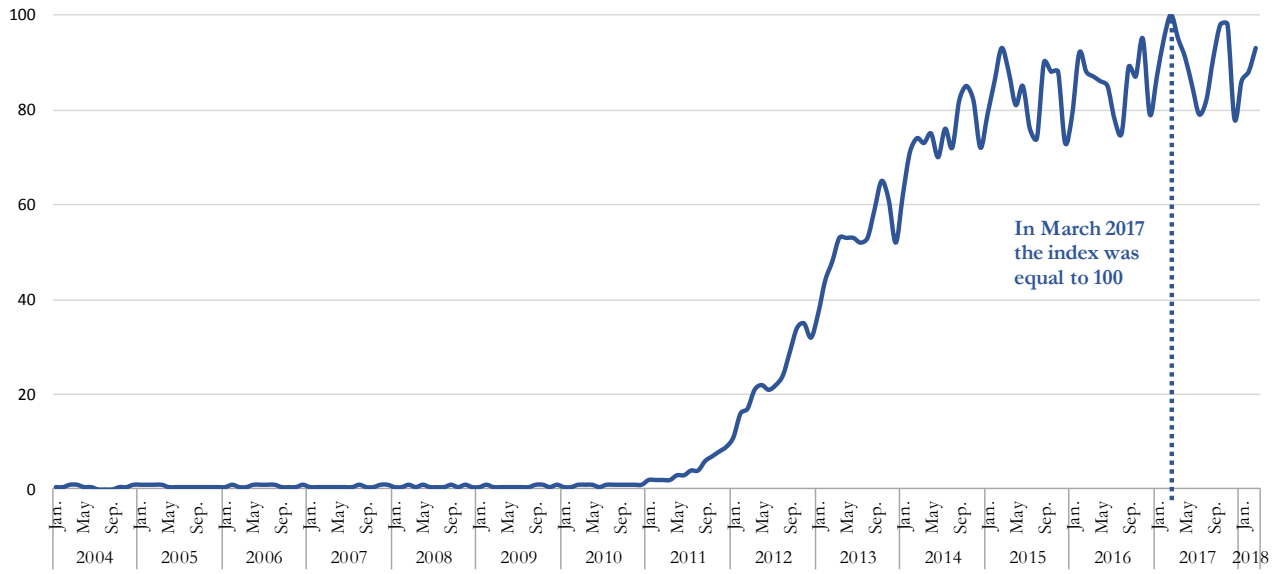


Figure (i) - *Google trends*: trend of the term "big data" in the search engine of Google
Source: Agcom on Google.trends *data*

The spread of the use of the term has, however, the consequence of simplifying the phenomenon. *Big data* on the contrary, are a very complex phenomenon and their impact on the economic and social system must be assessed based on rigorous and accurate analysis.

Hence the need to provide all *stakeholders* with an analytical contribution structured in three parts. Therefore, the first part of the report (Chapter 1) highlights the main characteristics of *big data*, attributable to the growth of volume, variety and speed of data creation, acquisition, storage and analysis, plus an evaluation of the structural characteristics of the various market sectors characterizing the ecosystem of *big data*. The complexity of this field makes it difficult to delimit precisely the perimeters of single markets whose borders are often overlapping; moreover, many companies are vertically or diagonally integrated, and therefore present in several segments at the same time. The presence of network, scale, variety and time economies, as well as the effects deriving from network externalities, engender concentrated market structures.

The second part of the report (Chapter 2) puts at the centre of the big data ecosystem the individual, paying particular attention to cross the traditional, and now obsolete, distinction between personal and non-personal data. The analytical, and therefore regulated approach must focus its purpose on data as such; this by virtue of the fact that it is now impossible to identify *ex ante* a categorization of data: these take on a different nature depending on the amount of data accumulated, the context, as well as the analysis technologies. For example, from a set, now even reduced, of non-personal data, some techniques can easily gather sensitive individual information (such as political orientation, drug addiction, etc.). The individual as a source of digital data is also at the centre of the analysis proposed at the end of the Chapter analysing, through a quantitative approach using *big data*, the relationship between the increasingly massive use of mobile device applications and the release of digital data.

[2] Every second on the search engine of *Google* it is possible to count about 67.194 queries. http://www.internetlivestats.com/one-second/#google-band

In the context of a commercial relationship that does not appear to be structurally well-contextualised and codified, i.e. struggling to have a well-defined contractual structure, the market fails due to the presence of huge information asymmetries between consumers and online service operators. Incompleteness of contracts disciplining property rights over data, absence of explicit markets regulating price formation, as well as informational asymmetries, compromise the possibility that the system converges towards a static and dynamic, socially efficient balance.

Last part (Chapter 3) analyses the effects of these issues related to *big data* on the information system, and therefore on modern processes shaping public opinion. In an informative context where online platforms, especially "social" ones, have an increasingly decisive role, *big data* and algorithms, that formed the basis of the mechanisms of the platforms, become fundamental elements of advanced democracies. Big data *have indeed* a fundamental effect on information pluralism, both on the demand and on the supply sides.

The ultimate purpose of the Report is, therefore, to redesign the reading of *data-driven* economy and society, explaining opportunities and risks of the current context, so as to favour both a sustained growth of the economic context and a social, efficient and deeply democratic progress.

1

THE ECOSYSTEM OF BIG DATA

1.1. The characteristics of big data

The term *big data* refers to a new approach of organizations (companies, public bodies, research institutions and governments), which extract value from data, by combining different databases and using appropriate statistical tools and other *data mining* techniques. It is therefore, **a process of radical reconsideration and evolution of traditional approaches to data analysis requiring, also as a consequence of technological advances, a new interpretative paradigm.**

There is not a single definition of the term *big data*;[3] for the strategic consulting firm *Gartner*, to which many attribute a first attempt to define the concept in 2001, *big data* are *"High-volume, high-velocity and / or high-variety information assets that require new forms of processing to enable enhanced decision making, insight discovery and process optimization"*.[4] With the advent of *big data*, therefore, a new approach to data management is needed to take into account of a scale (in terms of **volume**, **velocity**) and a complexity (**variety**) which is difficult, if not impossible, to face with traditional data analysis techniques.

By referring to the three relevant data characteristics - volume, velocity and variety - it is possible to highlight this radical change in the approach to data analysis (**Figure 1.1**). In the time of *Small-data*, i.e. when data represented a scarce resource, typically it was necessary to ask a research question and consequently collect data ("*data-is-scarce-model*"), or to acquire data on small parts of a reference universe (a sample); the deriving problems were treated separately depending on whether it was necessary to build databases, repeat data collection, get answers in a short time or, also, to make data in different formats "interact".

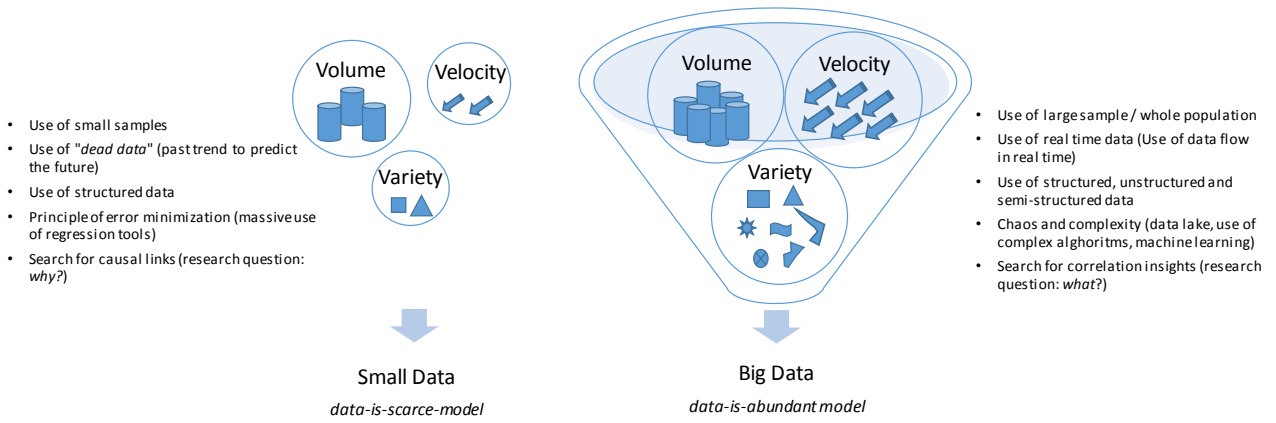


Figure 1.1 – Big data paradigm shift

Source: AGCOM

The computational development has allowed, with an ever increasing intensity, being able to face the greater complexity of database management deriving from the overlapping of at least two of the aforementioned characteristics. In the *big data* era, data are often collected regardless of specific problems in need of an answer: first we collect, we store, we analyse data and, subsequently, also based on what the data themselves communicate, we define the research and commercial issues (*data-is-abundant model*).

With big data, therefore, it is the need to process data that simultaneously present the three above mentioned characteristics with an unprecedented intensity, which makes the traditional management architectures obsolete and unsuitable for data analysis. In fact, the simultaneous

[3] T. HARFORD, (2014): *Big data: are we making a big mistake?* Financial Times, 28.03.2014.

[4] As early as 2001, the company *META Group*, later become *Gartner*, highlighted in a report the critical aspects related to data management by focusing on three dimensions: volume, speed and variety. D. LANEY, (2001), "*3D Data Management: controlling data Volume, Velocity and Variety* ", META Group Report, File 949. Subsequently, in 2012, the definition was coined in a new report. MA BEYER and D. LANEY, (2012), *"The importance of Big data: a Definition"*, Gartner, Analysis Report ID: G00235055.

management of these three characteristics caused the proliferation of data analysis techniques (*analytics*) different from traditional ones and through which it is possible to generate value from *big data* both for the resolution of specific problems and for the identification of new business opportunities and for society as a whole.

Big data, is a disruptive phenomenon, whose range, in terms of economic and social changes, is not yet well defined; like all the phenomena of a radical nature and of global scope, **big data too are emerging, bringing with them significant view to growth, economic and social, associated, at the same time, with doubts and perplexities, related to the phenomenon of destruction of markets, businesses, jobs and workplaces, inevitably accompanying (and often preceding) the creation.**

The flow of data draws its reach from the rapid spread of **new forms of "data sources"**; for example, the spread of the Internet of Things (*IoT*) and, more generally, of the increasingly invasive spread of sensor of any species, are sources of data that are rapidly spreading and integrating with one of the primary sources represented by the massive use by the world population of the latest generation of mobile phones (see section 2 and the Introduction).

Moreover, the dizzying development of technological equipment for the collection, storage, classification and processing of data, allows companies producing more and more information, engendering at the same time the need to make use of specialized personnel, able to extract value from data, and capacity related to the collection and storage of this huge amount of information. Professionalism allowing extracting value from *big data*, therefore, represent a very important competitive stimulus for companies, which need the availability of data *storage* and scalability services. This affects the cost structure of the ecosystem of *big data*, and therefore on its degree of competitiveness as well as on market sectors affected by this technological and commercial revolution. New challenges, therefore, emerge as a consequence of the specific characteristics of *big data.*

Moving on to the analysis of the characteristics of *big data*, the so-called *3V's*, it is important to explain that it is difficult to isolate the single characteristics, precisely because, as anticipated, there is a strong interrelation between them.

1.1.1. Volume

Volume certainly represents the characteristic that can be easily related to *big data*, since this term **explicitly refers to the size of the phenomenon, i.e. the large amount of data available today, composing the so-called "datasphere"**. Numerous studies and statistics try to measure this characteristic.[5] However, the difficulty of measuring such a phenomenon is clear because it is impossible to know exactly the amount, given the existence of quite different estimates. In any case, all statistics agree on an exponential growth trend: a growing trend, which does not seem to want to stop in the coming years. A summary of these results, which *inter alia* is linked to another feature of *big data*, i.e. velocity, is well captured by the speed with which the unit of measurement of information is updated;[6] in a short time, in fact, the *Exabyte* proved to be no longer suitable for photographing the evolution of the phenomenon, and we therefore passed to next measurement, namely the *zettabyte*, which is equivalent

[5] As part of a pioneering project begun in 2000, the economists P. Lyman and H. Varian looked at the issue of quantifying the information produced in the world, with particular reference to the production of original information; see. http://www2.sims.berkeley.edu/research/projects/how-much-info.

[6] The units of measurement are established by the *International Bureau of Weights and Measures*; in particular, the Conference for Weights and Measures which took place in 1991 introduced the *Zettabyte* and the Yottabyte, to date the last known threshold to quantify large volumes. In 2010, the *Product Manager* of *Google*, Jonathan Effrat, during the announcement of *Google Instant*, stated that on the web the measure of digital content in the world was now close to *zettabyte.*

to 1000 *Exabyte* (see **Table 1.1**). In practical terms, 1 *zettabyte* corresponds to a storage capacity of more than 36,000 years (in terms of duration) of HD video or a stack of 250 billion DVDs.[7]

Table 1.1: Units of measurement of information

Unit and simbol	Size	In practical terms ....
Bit (b)	1 or 0	The elementary unit of information measurement that can have a value of either 0 or 1, as it corresponds to a choice between two equally possible alternatives. Diminutive of "binary digit".
Byte (B)	8 bits	The byte is a unit of digital information that most commonly consists of eight bits; the byte was the number of bits used to create a single character in the computer code; it is the basic unit of calculation.
Kilobyte (KB)	1,000 or 2^{10} bytes	A text page is equivalent to 2KB. A photographic image in low resolution is equivalent to 100KB
Megabyte (MB)	1,000KB or 2^{20} bytes	An MP3 file of a "typical" piece of music is 4MB. 100MB is equivalent to a stack of books equal to one meter.
Gigabytes (GB)	1,000MB or 2^{30} bytes	A film lasting about two hours can be compressed in 1-2 GB. A text of 1GB contains approximately 1 billion characters, or approximately 4,500 books with 200 pages or 240,000 characters.
Terabyte (TB)	1,000GB or 2^{40} bytes	1TB is equivalent to 262,144 MP3 files (with an average duration of 4 MB). 1TB is equivalent to about 4,580,000 books of 200 pages. All books listed in the *American Library of Congress* amount to 15TB. All the *tweet* sent before 2013 are equivalent to a text file of 18.5TB; to print such a text (at a speed of 15 A4 pages per minutes) it would take 1,200 years.
Petabyte (PB)	1,000TB or 2^{50} bytes	1PB corrisponds to about 4,691,000,000 books of 200 pages. The NSA (*National Security Agency*) analyses about 1.6% of global internet traffic, around 30BP per day. If you want to listen to 30PB of music without interruptions, it would take more thna 60,000 years.
Exabyte (EB)	1,000PB or 2^{60} bytes	1EB of date corresponds to a data storage capacity corresponding to 33,554,432 iPhone 5 with a memory of 32GB. In 2018, the montly volume of data traffic via mobile telephony is estimeted at 1EB; if this amount of data was stored in 32GB *iPhone5* smartphones, it would be necessary to form a stack of *iPhone 5* with a height of 239 times the *Empire State Building.*
Zettabyte (ZB)	1,000EB or 2^{70} bytes	1ZB corrisponds to 281,474,977,500,000 MP3 files with an average size of 4MB, or 250,000,000,000 of 4.38 GB DVD's.
Yottabyte (YB)	1,000ZB o 2^{80} bytes	The content of genetic code belonging to a single person can be stored in less than 1.5GB; 1YB can therefore contain the genoma of 800 trillion individuals, or 10,000 times the population of the planet.

Source: AGCOM elaboration from Economist, www.computerhope.com, Cisco e Emmanuel Letouzé

[7] *Goodbye petabytes, hello zettabytes*, https://www.theguardian.com/technology/2010/may/03/humanity-digital-output-zettabyte, The Guardian 2010;
From Bits to Brontobytes, The Oxford Math Centre, http://www.oxfordmathcenter.com/drupal7/node/410;
The Zettabyte Era: Trends and Analysis, CISCO Public white paper, June 2017.

At the aggregate level, according to ICD (*International Data Corporation*), a mass of data amounting to 163 *zettabytes* is expected in 2025 (**Figure 1.2**), with a volume growth of about ten times compared to that recorded in 2016. In particular, a growing amount of data will derive from the consumption of online video and the presence of sensors connected to the IoT (*Internet of Things*).

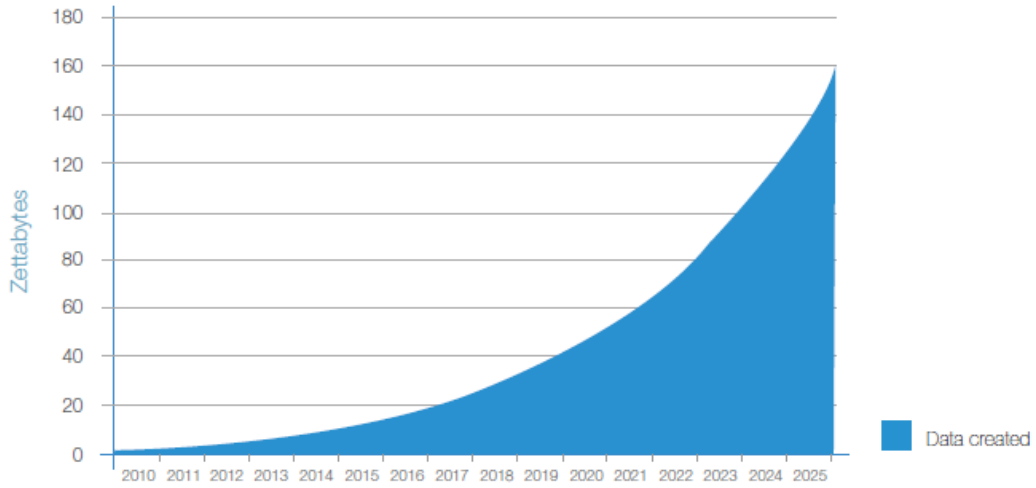


Figure 1.2 – The growth of the datasphere (in zettabyte)
Source: IDC Data Age 2025 - April 2017

It should be emphasized that many data that are collected are redundant; the same techniques used in *big data* provide for the duplication of data in order to preserve its functionality.[8] In a *big data* approach, redundancy is not synonymous with uselessness. Consider, for example, the set of information that each user generates when he carries out his activities on the internet (considered as a real individual footprint, so-called *online footprint,* see section 2). For a significant part of this mass of raw data, the term "*data exhaust*" has been coined:[9] that is information (cookies, temporary files, log files, typed words, etc.) which at the same time have a huge volume, must be acquired at great speed, and are composed of the most varied formats. From the joint analysis, and often in real time, of these data it is possible to extract a huge value, since the habits and characteristics (even sensitive) of the users are integrated (see Chapter 3 and Chapter 5).

The use of *exhaust data* allowed, for example, *Google* increasingly improving its search engine. More generally, *Google* undoubtedly represents an example of a web services company which has created huge value from a "gigantic and growing mass of information", many of which, at first glance, seem unimportant.

Furthermore, exploiting digital data generated by users not only makes it possible to respond to specific research demands (first use or primary uses), when these exist (see above), but also to exploit, over time, their "optional value" (subsequent uses or secondary uses), almost always not even known at the time of data collection. **The actual value of data, therefore, is significantly higher than the one deriving from their first use**.[10] The reuse of data, in fact, is the basis of numerous projects that *Google* and other

[8] For example, the software *Hadoop*, one of the main tools allowing the management of several *petabytes* of data, as well as their processing with high reliability and scalability of the data themselves, is based on a reproduction of a triple copy (3 *factor replication*) of the data source *files*, in order to reduce, *inter alia*, the risk of data loss (see section 1.2).
[9] The term is inspired by the ways in which these data are generated and collected; similarly to the exhaust gases of a car, which escape from the exhaust pipe (Exhaust) located at the rear of the vehicle, *exhaust data* hide behind the activities carried out by a user on the Internet. https://whatis.techtarget.com/definition/data-exhaust
[10] For example, *Google* in 2008 presented for the first time a system (available online at https://www.google.org/flutrends/about/) to predict the trend of seasonal flu in most of the world; the algorithm, based on

companies in the network are projecting (and are often released to the public in so called *beta* versions, i.e. experimental).

The use of "data-driven economy" in public policy today indicates the increasingly important role that the primary and secondary use of data has in the decision-making processes of the different economic and social actors. The growing use of the network by both citizens and businesses has undoubtedly fostered this growth process. *Social networks*, for example, have certainly given a strong boost to the increase in the amount of information in circulation.[11] Think, for example, about *Facebook*, the *social network* more widespread in the world with over two billion unique users; most of these users use the platform by uploading images, text documents, music and video, as well as expressing views and sharing contents they like or dislike. Any passage / activity that the user makes on the *social network* is traced and transformed into data.

Numerous commercial consequences arise as a result of the exponential increase in the volume of data; first of all, as it is easy to imagine, the mass of data, in order to generate value, must be archived (*gathering*) and efficiently stored (*storage*). **As data volume increases, therefore, costs for data filing and storage increase**, but also those related to the extraction of value (data *performance*) since it is necessary to use sophisticated algorithms and software and highly specialized professionals to manage complexity.

1.1.2. Variety

In the *big data* world, volume, as described above, intertwines with the other two characteristics represented by variety and speed of data. **Variety refers to heterogeneity in sources of data, in formats, with which information (traditional / structured and, above all, unstructured) is acquired and in the representation and analysis (even semantic) of the stored data.** The traditional *small data* approach provides for the use of structured data; most of the data is therefore organized into structures composed of rows and columns that can be easily ordered and processed according to techniques referring to relational data base (RDBMS, spreadsheets, data warehouse, Customer Relationship Management System, etc.), from which it is certainly easier to extract value thanks to the use of well-established techniques.

Within *big data*, instead, **the heterogeneity of the data has grown exponentially and the presence of unstructured data has been more and more spread**; which are data that are not organized according to a precise structure and, as a consequence, require very sophisticated techniques to transform the data itself into information (images, photos, texts, e-mails, RSS feed, videos, sensors, Social media, etc.).

a research carried out by users on the search engine on seasonal illnesses, is able to create a map updated in real time on the spread of influenza virus. This is a classic case of data reuse represented by the keywords typed by users. For completeness of information, it is necessary to underline the existence of a series of studies questioning the reliability of the *Google* algorithm which on several occasions overestimated the peaks of influenza. D. LAZER, R. KENNEDY, G. KING and A. VESPIGNANI, (2014), The Parable of Google Flu: Traps in Big Data Analysis, Science, Vol. 343, Issue 6176, pp. 1203-1205.

[11] Not surprisingly as a source of data, *social networks* are considered to be firefighters fire hydrant (*fire hose data source*) or inversely, it can be argued that *"Getting information off the internet is like taking a drink from a fire hydrant"* Mitchell David Kapor, an American businessman known for supporting and promoting the diffusion of the first spreadsheet for PC, VisiCalc, and later founder of Lotus.

Although structured data contain a higher information density, the current trend is that about 80% of the data available today has an unstructured nature (**Figure 1.3**).

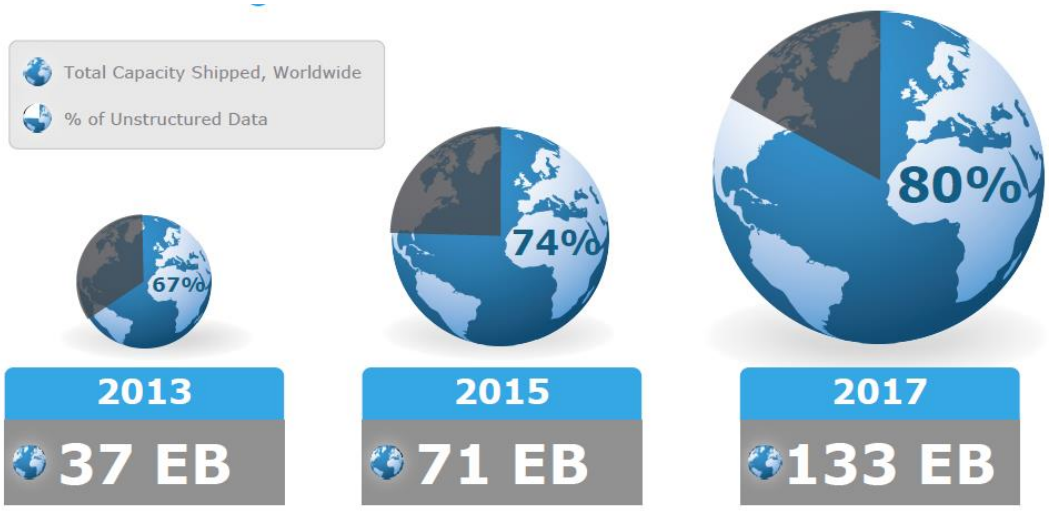


Figure 1.3 – The growth of unstructured data (in Exabyte)
Source: *IDC* Structured Versus Unstructured Data: The Balance of Power Continues to Shift, March 2014

A considerable amount of data is then defined as semi-structured, since although there are possibilities to separate some elements, much of the information contained in the data remains unstructured. The typical case is the *e-mail*; in fact, any e-mail service presents a series of structured data (which can be collected and organized into tables) connected to easily identifiable information (date and time of sending, sender and recipient, etc.). The body of the *e-mail*, instead, is generally composed of an unstructured text, which can contain, particularly in the attachments, data with very different formats: images, video, audio, etc.

Since unstructured data are the main part of information available today, it is clear that the problem of managing heterogeneity, and its complexity, becomes crucial. It is precisely the diversity of formats that today, much more than in the era of the *small data*, mainly characterizes the data. In fact, not all companies have to face the problems caused by the growth of volume and speed, while all, even the smallest ones, must be able to manage the variety of data, both due to the presence of different data sources, and for the great opportunities deriving from the possibility of combining data of different formats. It is no coincidence that access to a great variety of data - new and old data, small samples and large samples, structured and unstructured data, *social media data*, data on consumer choices - undoubtedly represents a distinctive feature of modern online platforms.

This is confirmed by the results of a research carried out by NVP (*New Vantage Partners*)[12] showing that 40% of the interviewed companies- belonging to the *Fortune 1000* ranking drawn up by the business magazine *Fortune*- feels the need to integrate data with different formats and sources, compared to 14.5% identifying in volume and 3.5% indicating speed as the driving factors in the investment choices for the management of *big data*.

[12] NVP, (2016), *An update on the adoption of Big data in the Fortune 1000*. NewVantage Partners defines itself as a strategic consultancy company especially for what concerns the setting of a *business data strategy* http://newvantage.com/about/.

1.1.3. Velocity

Velocity is related to, firstly, the timing with which databases are fed, in particular **at the high frequency with which data circulate from a point of origin to a collection** point; this is also thanks to the increasing availability of technologies allowing data collection in real time.

Figure 1.4 provides a summary of how much data are produced, then acquired, on the internet in just 60 seconds, giving an idea of the joint operation of the three main characteristics of the *big data* (volume, variety and speed).

However, speed is not just about the flow of data, but also about the need to process data quickly and to make decisions at an ever faster pace, often in real time (so called real time action). This aspect requires skills, technological infrastructures, and sophisticated software solutions.

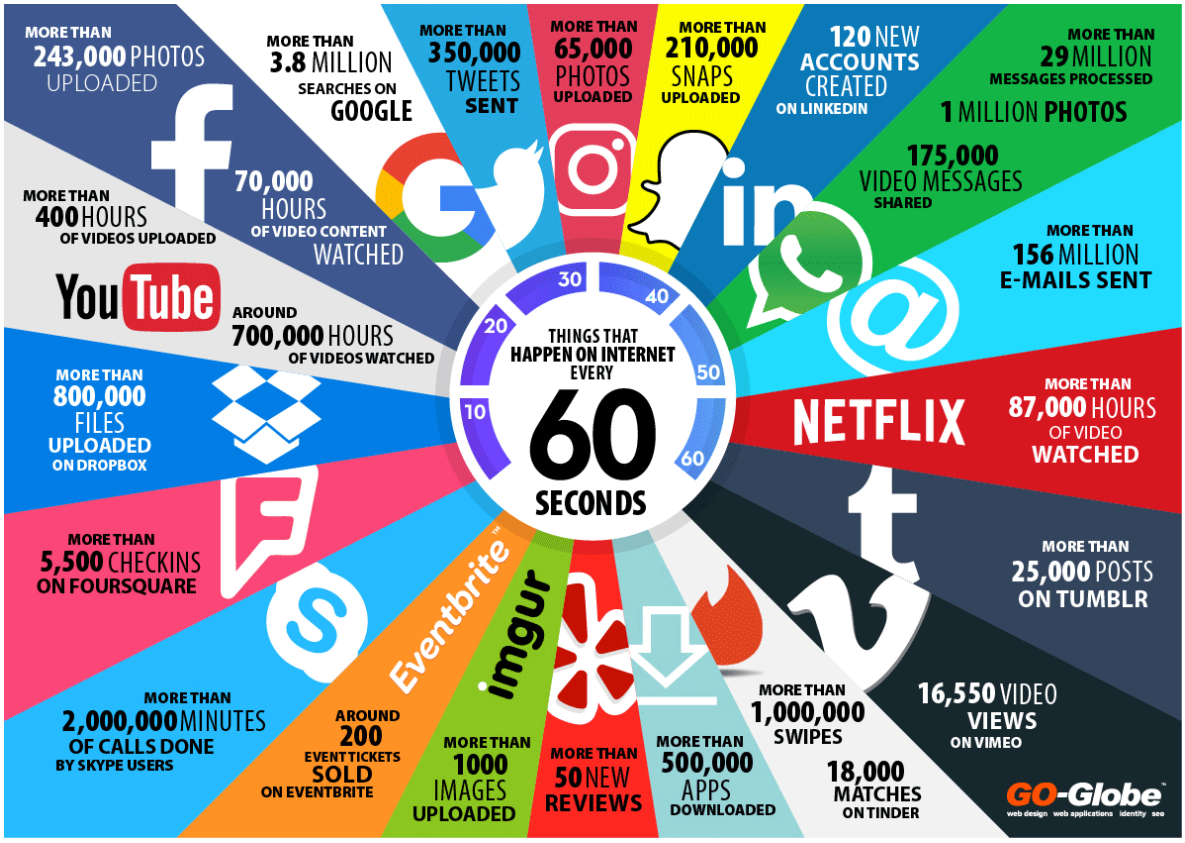


Figure 1.4 – Internet data flow in 60 seconds
Source: *Go-Globe.com* – 2017
https://www.go-globe.com/blog/things-that-happen-every-60-seconds/

Despite the fact that for many types of data, the saying "data are like wine, improve with age", meaning that after its first use data do not lose value but can be reused for many other purposes are still valid, it is equally true that many business opportunities are linked to the ability to quickly and promptly exploit the available data.

Consider, for example, the importance of making quick decisions in a highly competitive sector such as retail, where there is the pressing need to know in real time which and how many types of products are needed to efficiently supply the shelves of the store.[13]

The speed of data has led to a reconsideration of traditional business models, typically based on data block processing (*batch processing*) mostly related to the past performance of the activity (so-called "*dead data*"). This approach does not suit for the characteristics of the internet. In a recent letter to the shareholders, the founder of *Amazon*, Jeff Bezos, has emphasized the role of speed, not only in data and in business management: "*speed matters in business*" and also *"Most decisions should probably be made with somewhere around 70% of the information you wish you had. If you wait for 90%, in most cases, you're probably being slow.*[14] Nobody wants to make wrong choices; however, temporise waiting for "perfect information", leads to decision-making delays and, therefore, to potential losses of *business* opportunities; make corrections on the run, using data information dynamically, on the other hand, can be more efficient in the current context.

It is important to stress that using data to make decisions in real time (real-time processing) requires particular software architectures allowing you to live with the timing constraint. In this sense, as above mentioned, some typical activities of filing, storage and cleaning of data, typically carried out within companies, are increasingly outsourced to companies specialized in supplying these services (see section 1.5.3).

Speed, among other things, is also very important in decision-making processes that do not involve business activities; in politics, for example, the exploitation of data is also used to try to obtain an increase in consensus, especially during the election period (see Chapter 3). But even in the health sector, making decisions in real time can be a source not only to eliminate waste in the use of resources, but above all to improve the health of individuals, for example, anticipating the onset and spread of diseases.

1.1.4. Others features

To the three main dimensions of the *big data*, over time, many others have been added; every dimension, represented by an additional *V*, identifies a specific feature of *big data* to which specific risks and opportunities are associated. We moved in a short time from *3V*, to *4V,* to *7V*, until arriving, in 2017, to the *42V's of big data*.[15] This inflationary process, among other things, does not seem to stop, although any further identified *V* seems to respond increasingly to niche issues.

[13] In this regard, an example of how data velocity is an extremely relevant feature in modern forms of business comes from the American retail giant *Walmart*. During the Halloween period, the company collected meaningful data concerning the sale of a novelty cookie produced for the occasion, except in two stores. From a quick analysis of the warehouses, it simply turned out that these cookies had not been correctly placed on the shelves. The fact that a warehouse is managed through the *Big data* paradigm allows a real-time insights regarding the whole situation of individual points of sale and, when in a certain place sales below certain parameters emerge, there is an *alert* allowing a quick and targeted intervention which, mainly, allows the commercial conditions to be restored to efficiency parameters. MARR B., (2017), *Really Big Data At Walmart: Real-Time Insights From Their 40+ Petabyte Data Cloud*, www.forbes.com

[14] BALAKRISHNAN A., (2017), *Bezos shareholder letter: Don't let the world push you into becoming a 'Day 2' company*, www.cnbc.com, 2017.

[15] T. SHAFER, (2017), "The 42 V's of Big Data and Data Science", Elder Research - Data Science and Predictive Analytics, https://www.elderresearch.com/company/blog/42-v-of-big-data. In this article, *inter alia*, an interesting chronology of the addition of V to the characteristics of the *big data* is presented.

Nevertheless, four other characteristics of *big data* deserve to be mentioned; one refers to the ability to extract economic **value** from *big data*.[16] **It is not just a matter of generating value for companies, but more generally of exploiting the growing amount of data to increase the overall social welfare.**

The second one, **veracity** (or truthfulness) instead, focuses on the importance of the **qualitative aspects linked to data and, consequently, to the trust that can be placed in them**.[17] As the volume, the variety and the velocity of the information flow increases, and by virtue of the ever greater spread of *machine learning* processes, it becomes crucial for organizations to work with "truthful" data so that analysis lead to correct results.

The third refers to the **valence** of data; the term is borrowed from chemistry to indicate the number of electrons an atom gains, loses, or shares when it creates bonds with other atoms, that is, the ability of an atom to create bonds. Transposing this concept in the field of *big data* simply means that the **more data are connected to other data, the greater is their valence**. Two *Facebook* users for example, are directly connected to each other if they are friends, just as a worker is connected to his workplace; the data can also be connected indirectly, as in the case of two scientists belonging to the same scientific community even if they do not know each other. The valence of data grows over time, making the connections between data more and more dense and complex, in this case also establishing new challenges.

Finally, a last *V* deserving to be mentioned concerns the **visualizations** of data; **being able to extract synthetic information from a vast amount of data undoubtedly represents one of the most difficult challenges**. In fact, with the right analysis and visualization, information gains value, otherwise they will always remain at the level of raw data. However, the term "visualization", does not refer to the classical graphics used hitherto (histograms, pie charts, etc.); in fact, these are complex graphs (infographics) that must have the ability to synthesize different information without, however, reducing its informational scope. Therefore, data visualization is not technologically complex. In fact, many platforms offer today services to carry out infographics; it is, still, a crucial and complex activity being able to tell an event, a story through diagrams is an activity at the same time difficult and essential.

In conclusion, the different dimensions of *big data* confirm that we are facing an extremely complex phenomenon characterized by a very rapid evolutionary dynamic; each of the characteristics identified, as shown in **Figure 1.5**, implies precise challenges to be faced, associated with risks and opportunities, for businesses, citizens, and the society as a whole.

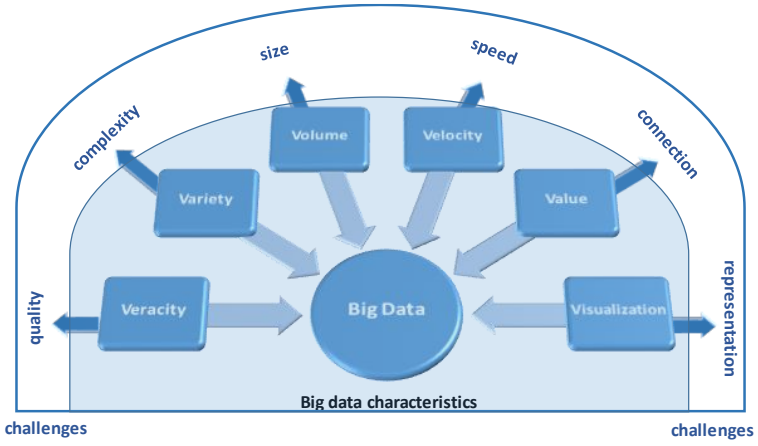


Figure 1.5 – The characteristics of big data
Source: AGCOM

[16] S. GOGIA, (2012): *The Big Deal about Big data for customer engagement*, Forrester Research.
J. GANTZ, D. REINSEL, (2013): *The digital universe in 2020: Big data, Bigger Digital Shadows, and biggest growth in the Far East*, IDC's Digital Universe Study.
[17] M. WHITE, (2012): *Digital workplaces: Vision and reality*, Business Information Review, 29(4) 205–214.

1.1.5. A new approach to the analysis of social phenomena

The advent of big data requires a new approach to data processing; the new philosophy, in fact, provides that it is sufficient to create powerful algorithms able to explore data in order to discover correlations and regularities (so-called *correlation insights*).[18] Regardless of any analysis seeking a causal link, as typically occurred in the so-called "*small data*" period (**Figure 1.1**), the algorithms will identify predictions and actions to be taken; therefore, it is no longer necessary to identify a behavioural model with the related hypotheses to be tested and the consequent statistical approach.[19] *Big data* in fact, increase the specific weight of correlation analyses, causing a turn upside down of the research structure: first of all a link between the variables must be sought and then, if necessary, we will try to establish a plausible interpretation of the phenomenon. In other words, data themselves "speak". The overcoming of an approach based on the shortage of information, and on the sample analysis, produced the rise of correlation analysis, with an epistemological overthrow.

However, the main problem of a correlation-based approach lies in the fact that, as the data availability increase, the probability of finding variables with a completely random link is higher;[20] the possibility that these bonds are completely random, therefore, is very high in large data sets, or increases as the volume and variety of data grows. Scientific literature for a long time highlights that relying exclusively on correlations exposes the analysis to a series of tricks leading to misinterpretation.[21]

Very often, correlation analysis can create bonds between phenomena having no meaning: if two phenomena have a high correlation between them, it does not necessarily mean that there is a causal relationship between them, since the correlation deriving from a third phenomenon, in common with the two analysed, or even be due to chance (so-called spurious correlation), as in the case presented in **Figure 1.6**.

According to what is shown in the figure, wanting to take the concept to the extreme, we could deduce from the analysis of the correlations that, by implementing initiatives aimed at encouraging the consumption of mozzarella the number of individuals with a specialization in civil engineering grows, that is, to increase the consumption of mozzarella, it would be appropriate to expand admissions to the doctorate in civil engineering.

[18] C. ANDERSON, (2008), *The End of Theory: The Data Deluge Makes the Scientific Method Obsolete*, Wired, "*There is now a better way. Petabytes allow us to say: "Correlation is enough." We can stop looking for models. We can analyze the data without hypotheses about what it might show. We can throw the numbers into the biggest computing clusters the world has ever seen and let statistical algorithms find patterns where science cannot*". It is useful to remember, in this regard, that the correlation between variables does not explain why two variables move in a direction rather than another, rather it merely verifies the existence of a trend. Precisely for this reason, the correlation is a very used tool because it is characterized by a strong predictive power since it allows to intervene on a variable to predict (or modify) the value of the correlated one.

[19] CLAUDE C.S., LONGO G., (2017), *The Deluge of Spurious Correlations in Big data*, Foundation Science, Volume 22, n. 3.

[20] GRAHAM R., SPENCER J.H., (1990), *Ramsey theory*, Scientific American, 262.

[21] FERBER R., (1956), *Are correlations any guide to predictive value?* Journal of the Royal Statistical Society Series C (Applied Statistics), 5(2).

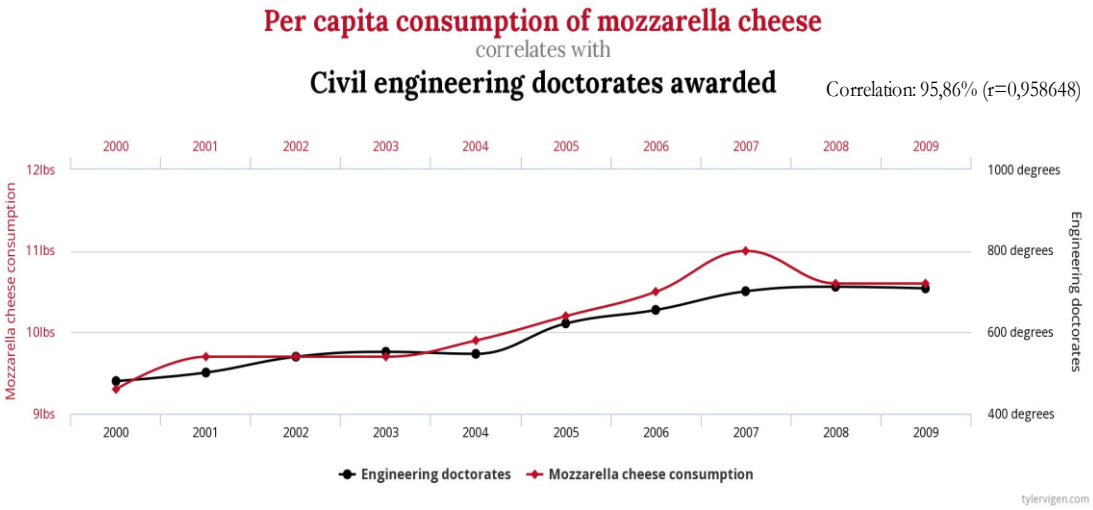


Figure 1.6 – A case of spurious corr4elationUn caso di correlazione spuria
Source: http://tylervigen.com/spurious-correlations

However, when it comes to making decisions concerning companies, the focus on spurious correlation decreases. For a company operating under the assumption of economic rationality, in fact, the possibility of increasing profit is fundamental, even if this increase occurs by identification of a spurious correlation.

In conclusion, the advent of *big data* has deeply changed also the analysis of society. The social sciences in the past have often been subject to constraints related to the scarcity of data in the hands of researchers. The arrival of *big data* has therefore made it possible to quantitatively analyse subjects that in the past were qualitatively analysed, often carried out with a certain degree of arbitrariness. **Data revolution involved both the commercial (business world and institutions) and the scientific side, creating new professional figures and a new approach to the analysis of phenomena. However, some of the methods defined big data need to be rethought because they risk arriving at paradoxical conclusions**, like the one concerning the link between mozzarella consumption and civil engineers.[22]

[22] POPPELAARS J., (2015), OR (Operations Research) at work, https://john-poppelaars.blogspot.it/2015/04/do-numbers-really-speak-for-themselves.html

1.2. *The value chain*

The previous section highlighted the high degree of complexity characterizing the world of *big data*. In this sense, identifying a **value chain** (or *data science process*) allows clarifying the steps following data so that it can extract value from them. As observed so far, in fact, **from the collection to the use, data go through different interdependent phases gradually increasing its value; these phases can be assimilated to a life cycle of the data**.[23]

We should assume that a **single datum in itself has little or no value.** Many researchers have approached the role that data play today for the world economy and society to what oil in the last century represented; [24] to some extent this analogy is appropriate, especially considering that, like oil in the development of modern industry, data today allow a plethora of possible new uses therefore representing a decisive productive factor in an information-based economy. Compared to oil, however, there are some significant differences:[25] among these, one lies in the fact that, while a barrel of oil has its own value, this is not so true for data. Then, **while for a product like oil the tension between supply and demand, due to the presence of a scarce resource, will give rise to a balanced price, this mechanism does not work with data** (see section 2.7, for an example concerning the mobile applications sector). A further **difference is, as mentioned, the possibility of re-use, a characteristic that data have compared to a raw material as oil**. In order to make data acquire value, it is necessary the analysis phase;[26] in turn, analysis is not possible if all the problems generated by the characteristics of *big data* are not faced and overcome (see section 1.1). **Figure 1.7** provides a possible summary representation of the processes (or macro-activities) necessary to extract value from data.[27]

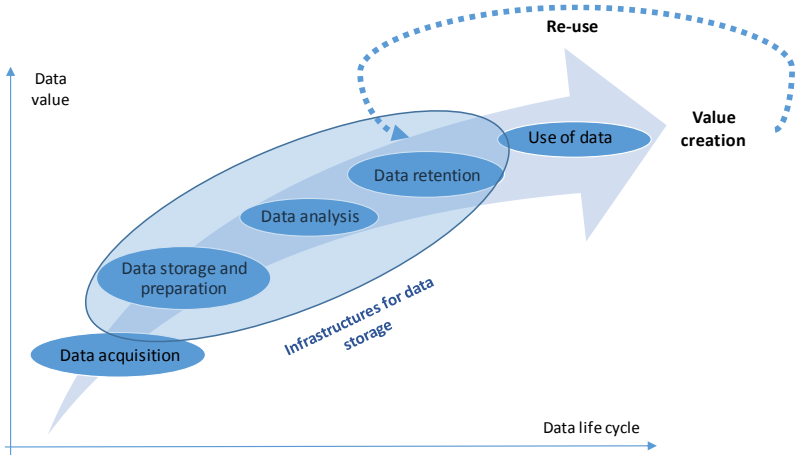


Figure 1.7 – The value chain
Source: AGCOM

First of all, **data must be obtained**; this phase identifies **all activities through which data are collected and aggregated with other data and, therefore, transported from a source to a data distribution system**. It includes also activities carried out by the controller to check data already

[23] FTC REPORT, (2016), *Big data: A tool for inclusion or esclusion? Understanding the issues.*
[24] The Asian Banker, (2016), *From fintech to techfin: data is the new oil*; Fortune, (2016), *Why Data is the New Oil*; Economist, (2017), *The world's most valuable resource is no longer oil, but data.*
[25] For more details on the reasons making the data asset different from the oil asset, see among others: J. GOLDFEIN, I. NGUYENHTTPS, (2018), *Data is not the new oil*, //techcrunch.com/2018/03/27/data-is-not-the-new-oil/; M. MANDEL, (2017), *The economic impact of Data: Why Data Is Not Like Oil*, Progressive Policy Institute.
[26] According to a study carried out in 2013, only 0.5% of available data is analysed. A. REGALADO, (2013), *The Data Made Me Do It*, MIT Technology Review.
[27] For a detailed study of the issue, see, among other things, the Report *European Data Market* written by the IDC and Open Evidence on behalf of the European Commission, February 2017.

available, freely accessible, or accessible upon payment, as well as the necessary tools for their collection, etc. The acquisition of digital data can take place by using different means, such as the APIs of those providing data from social network, sharing applications, importing data using ETL tools and using web scraping tools. In some circumstances, this phase of the value chain requires significant financial efforts for the necessary infrastructure investments, as minimum requirements are required to ensure low inactivity in the data acquisition phase and at the time of their query.

It is worth pointing out that today we collect data from a set of very heterogeneous sources (see section 1.1.2). We think, for example, about tracking cookies, about digital footprint or about the techniques of history sniffing. Currently a significant proportion of data is produced by the activities that individuals perform through mobile devices (see section 2.2) and the production of data in real time under the impetus of technologies IoT (Sensors). The mobile station connection engendered a huge development of mobile web, especially in countries where fixed network infrastructures are lacking, rapidly increasing a specific ecosystem where companies collect data through applications (see section 2.5).

In any case, technological evolution has made available instruments allowing "Cross Device tracking" of preferences and habits of individuals; in this way it is possible to monitor the single consumer by exploiting different devices, desktops, laptops, tablets, wearables, or smartphones. Precisely this widespread system of data collection can lead to a defined phenomenon of over-collection:[28] that is those practices gathering quantities (volume) and types (varieties) of data going beyond the declared purpose.

Companies interested in data collection aim at their quantity and variety; however, as widely repeated, it is not enough to acquire many different data, we need to be able to analyse them. Once collected, then, the data undergoes a second processing phase concerning its own **preparation** and **storage for subsequent uses**. In this phase, **data starts to turn into information** and volume, velocity and variety of data are particularly relevant in choosing the necessary technological infrastructures. In order to make raw data becoming information, in fact, it is necessary to have tools and skills allowing to face problems related to the variety of data and, therefore, create the conditions for different types of data to interact with each other, often in real time. It is a matter of implementing infrastructures (mainly in terms of their processing capacity and software), guaranteeing both an easy data scalability and their proper conservation. **The architecture to be designed, therefore, is very complex, since the need to address these critical issues related to speed, variety and volume of data require technologies and capabilities characterized by a strong ductility**.

In this sense, for a greater data integrity and usability, the **data lake paradigm** has become increasingly popular. The term was introduced for the first time by James Dixon (CTO of Pentaho, a company specialized in business intelligence) who in his blog defined *data lake* as follows: "*If you think of a DataMart as a store of bottled water – cleansed and packaged and structured for easy consumption – the data lake is a large body of water in a more natural state. The contents of the data lake stream in from a source to fill the lake, and various users of the lake can come to examine, dive in, or take samples*".[29]

In a nutshell, data lake were born as a useful paradigm to exploit the potential generated by the characteristics of *big data* with the aim of overcoming the strong rigidities present in traditional approaches to data management (data silos, data warehouse, data marts) typically focused on the creation of

[28] From the report *Big data and Privacy: A Technological Perspective* by the "President's Council of Advisors on Science and Technology": «*Over-collection occurs when an engineering design intentionally, and sometimes clandestinely, collects information unrelated to its stated purpose. While your smartphone could easily photograph and transmit to a third party your facial expression as you type every keystroke of a text message, or could capture all keystrokes, thereby recording text that you had deleted, these would be inefficient and unreasonable software design choices for the default text-messaging app. In that context they would be instances of over-collection*».

[29] DIXON J., (2010), *Pentaho, Hadoop, and Data Lakes*, https://jamesdixon.wordpress.com/2010/10/14/pentaho-hadoop-and-data-lakes/

structured databases, therefore certainly easier to use (in terms of querying, surfing, etc.), but independently managed and characterized by a pre-processing of data that often caused a subjective selection of features deemed most interesting.[30] Data lake, instead, help and accelerate data sharing because they are built on structured, semi-structured and unstructured raw data in their original format, allowing their analysis and, ultimately, value extraction. The contents of data lake can thus be exploited by different users who can examine it and query it searching for relevant information (or insight).

The following phases of the value chain refer to **all those activities allowing data to go from the stage of simple information to the knowledge** of the analysed phenomenon. It is, first of all, the phase that in **Figure 1.7** is indicated as **data analysis**; the inherent activities include exploration, transformation and modelling, in order to highlight the relevant ones and, at the same time, to synthesize the information with the intent, equally important, to bring to light hidden information.

Next phase concerns **data storage**; this process must respect criteria allowing, as repeatedly said, easy scalability, also considering that a growing volume of data increased the problem of their memorisation (see section 1.5.3). **The way in which each organization plans the storage of its own file, consequently, has a significant impact on the speed and efficiency of data access and, consequently, of decision-making processes**.

For big companies or large amounts of data, today's reality implies **distributed databases**. In the world of *big data*, data can be distributed on mass memories of different computers (or hubs) representing the network of an organization, whose hubs can also be physically very distant (think for example about the different seats of a multinational). In fact, for wider organizations it is impossible to keep all the information in one place, both because the access traffic to the single hub in which data are collected is likely to be excessive, with consequent problems of congestion and therefore delays in the answers to the questions (*queries*), both for security issues, since concentrating data in a single point makes the system more vulnerable.

However, **the effectiveness of a distributed file system depends on its correct integrated management, guaranteeing all users of one organization's information system, whatever their geographical position is, the availability of always updated data;** therefore, a logical integration of the archives becomes crucial even if data from a physical point of view can be distant from each other. The lack of integration, in fact, can cause duplication problems (for example the same variable can be differently named depending on the hub of the distributed file system) and risks of anomalies in the update, occurring when the same data is updated in an archive of the system but not in another.

All these phases of the value chain are preparatory to the phase in which data are used to make decisions, the final stage in which **data from simple knowledge is transformed into a vision of facts** (or *wisdom*).

The need to deal with big volumes and variety of data, more and more often in real time, generates a complexity so far unknown. In fact, this requires the various organizations involved

[30] In this sense, a classic example is that of companies *data silos* , i.e. systems for managing isolated data (for example a data *silo* for each company management, such as payroll data, financial data, customer data, seller data, and so on). These databases are typically built for very specific purposes, such as the recovery of a single customer order, payroll processing at the end of each month, and so on, while they are not designed to communicate with each other, and therefore do not allow individual users to explore data in an innovative way. Traditionally, data were organized into tables or folders and files, in many cases even following a hierarchical schema; instead a *data lake* uses a flat architecture where each element composing it has a unique identifier and is marked with a series of *tag* associated with the metadata of the data itself. A first clear advantage is the elimination of the initial costs of inserting and transforming data themselves, i.e. all those operations necessary to turn data into the passage from the source to the *silo*. Through the paradigm of the data *lake*, on the other hand, data collected on a given consumer can be linked to a series of other data coming from other sources (such as data on traffic, weather, etc.) which do not appear to have a direct connection with the consumer, can be used in their combination in order to extract more information in the analysis process.

to have adequate infrastructures, since the use of traditional systems are not functional anymore, as is the case for database management software systems (DBMS) that usually are characterized by an architecture focused on the use of a single dedicated hardware component.

Last phase of the value chain concerns, therefore, the **use of data to support decision-making processes**; these activities basically consist of the need to find a connection between data and the directions undertaken by the organization. The use of data in the decision-making processes may concern the reduction of production costs, the organization of personnel, the invention of new services and / or products, as well as any contribution to the improvement of the *performance* indicators.

It is important to stress that not only the private sector benefits from the use of big data, but also public sector, both in terms of efficiency improvements in the use of resources, both in the creation of new services and in the improvement of the current ones.

The phases of the value chain have a certain degree of overlapping due mainly to the fact that all must take place as quickly as possible (sometimes, a few millionths of a second, as in the case of auctions for the sale of online advertising).[31]

In conclusion, the value chain tool allows to model the *big data* system and, consequently, to identify the different steps necessary to generate value, and, more generally, knowledge from data.

[31] One of the main technological infrastructures currently used by an ever growing number of companies is *Apache Hadoop*. *Hadoop* looks like a real ecosystem made up of a series of tools thanks to which it is possible to process *big data*, or browse in a *data lake*. The complexity of this technology is such that its detailed description is clearly beyond the scope of this work, but it is interesting to note how the scaffolding of the architecture of *Hadoop* is based on the great capacity to quickly process a large amount of data through the *Hadoop Distributed File System* (*HDFS*), that is a system of distributed data storage, with a very high level of flexibility, through which it is possible to manage structured and non-structured data, coming from different sources. Furthermore, *Hadoop* it is built in such a way as to provide also a *fault tolerance* so that, in the event that a hub presents a hardware failure, the architecture is able to redirect the activity to another hub that owns the copy data giving continuity to the computational process. Furthermore, the data replication mechanism also serves to recover more efficiently data. In conclusion, it is a technological tool including all the main characteristics of the value chain reviewed in this section and which for its prerogatives is today a standard used by the largest companies to browse *big data*.

1.3. Active subjects

The complexity underlying the value chain determines a very varied and articulated market scenario for *big data*. **If the actors participating in the market can be identified, even if not always easily, it is much more difficult to dissolve the intricate interweaving of interactions taking place in the world of big data.**

In the ecosystem of *big data*, it is possible to identify, among others, the following main actors:[32]

a) **subjects generating data** (or data providers);
b) **suppliers of technological equipment**, typically in the form of data management platforms;
c) **users**, i.e. who use big data to create added value;
d) **data brokers** that is, organizations collecting data from a set of sources, both public and private, offering them, upon payment, to other organizations;
e) **companies and research organizations**, whose activity becomes fundamental for developing new technologies, new algorithms by exploring data and extracting value;
f) **public bodies**, both as market regulatory authorities, and with reference to the activities of public administration aimed at improving products and services offered to the citizens and able to increase the public interest.

However, the ecosystem of *big data* presents a **degree of interconnection between the various parties involved making it difficult to identify individual well-defined markets**; consequently, the resulting complexity determines a scenario in which the various segments of the system, of which the **Figure 1.8** offers a possible representation, are often closely interrelated. This determines a market structure in which **(few) large multinational companies, characterized by a high degree of vertical, diagonal and horizontal integration in all (or almost all) phases of the ecosystem, operate alongside a myriad of small specialized businesses** that often, after the period of *start-up*, are acquired by the larger ones.

Competitive assets in specific market areas strictly depend on some common structural characteristics; this circumstance justifies an approach, in first instance, of a wider scope, aimed at identifying a series of structural characteristics of *big data*, then deepening the effects of their implementation in some specific areas.

It is also worth noting that, as will be explained in greater detail in Chapter 4, **the ecosystem of big data is characterized by the presence of several forms of incomplete contracting, by implicit markets (i.e. in which the bargaining of the asset takes place in a spurious manner), as well as by notional areas (i.e. characterized by perfect vertical integration and by merely potential market demand).**

This in itself is already a source **of market failures undermining the social, static and dynamic efficiency of the entire big data ecosystem**.

[32] CURRY E., (2016), *The Big data value chain: definitions, concepts and theoretical approaches*, in New Horizons for a Data-Driven Economy, Springer, Cham.

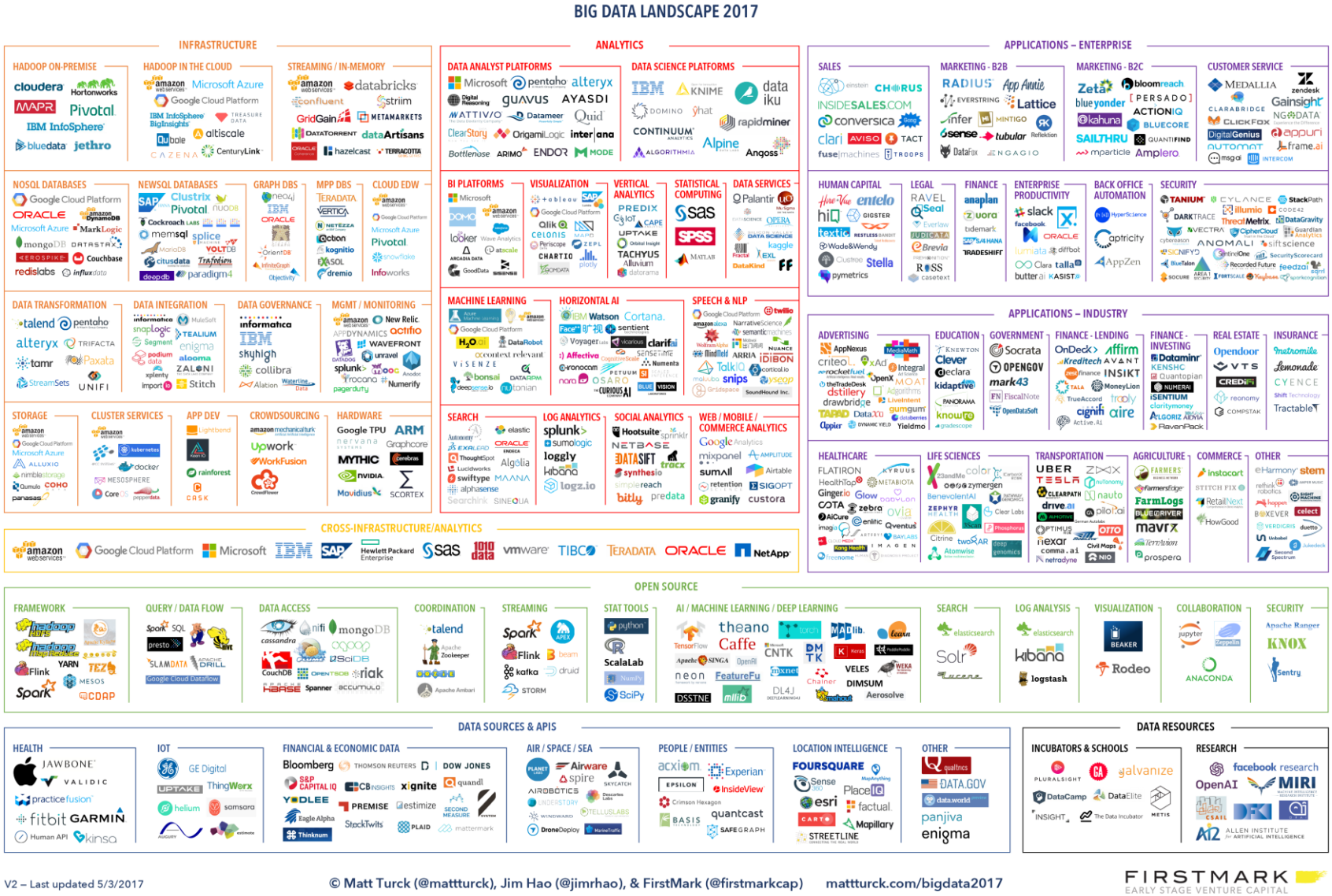


Figure 1.8 – Big data landscape 2017

Fonte: http://mattturck.com/wp-content/uploads/2017/05/Matt-Turck-FirstMark-2017-big-Data-Landscape.png

1.4. *The main features of big data markets*

A first fundamental structural feature applicable to the ecosystem of big data, in particular to the component of digital data related to individuals, is the possibility of framing this context in the economic theory of **two- sided markets** or **multi-sided market**, as it will be explored later by analysing the case of the APP (see Section 2.5).

In general, markets with two (or more) sides are characterized *i)* by the presence of two (or more) distinguished and separated groups of economic agents, whose interactions are mediated by a so-called "platform"; *ii)* by the close interdependence of the choices made on one side compared to those made by the agents operating in the other (or in the others) side.

In the case of digital data related to individuals, as seen from the synthetic representation of the **Figure 1.9**, it is possible to identify on the online platforms offering services to consumers the role of intermediary between the first and the users of data.

In the first place, **online platforms take the technical connotation of "platform" in the sense of the multi-sided market theory**; that is an intermediary between economic agents located in different market areas and that "communicate" through their presence.

Secondly, **on the side of consumers, transactions between consumers and online platforms are characterized by deep and incontrovertible information asymmetries and by a strong and structural incompleteness of transactions, so that the digital exchange of data does not assume a specific value (price) as in all markets, but is ceded in spurious and not contracted form**. Section 2.7.2 discuss in detail about type of relationship and the consequences on the economic efficiency of the whole system, while, in Chapter 3, the perspective will be further enlarged to the political and social aspects.

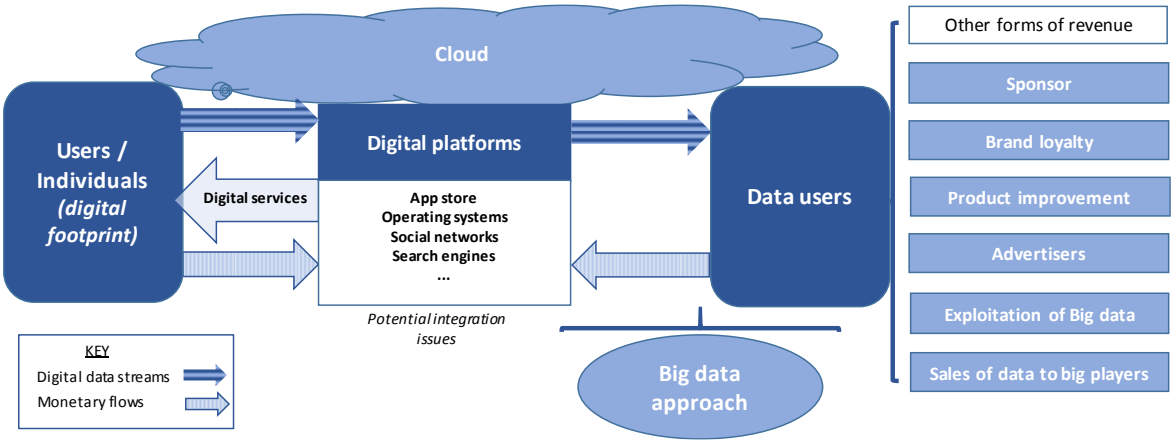


Figure 1.9 – Synthetic representation of the two-sided market applied to digital data
Source: AGCOM

Thirdly, as regards **data users, it is worth observing how this area** is often characterized by the **presence of several phases (and sometimes intermediaries), and, at the same time, by the vertical integration of some large platforms in all the levels**.

In these market contexts, there are both the classic **network externalities (so-called "direct")**, as for the individual consumers the value of the service offered by the platform (e.g. a messaging service, a *social network*...) often increases with the growth of the "network" of consumers using it, and those typical of a multi-sided market, that is cross-network **externalities**, occurring when the decisions made by the actors belonging to one side of the market produce effects on agents that are part of other sides, and whose intensity has a decisive influence on the price structure (for more details please refer to the section 2.7).

These characteristics (also called "demand-side returns to scale"), other things being equal, naturally make the ecosystem converge towards particularly concentrated market outcomes.

Further characteristics concerning the structure of the whole *big data* system, and which consequently affect the competitive assets of the various segments, are a direct consequence of the *3 V* (see section 1.1). It is, in particular, the **presence of economies of scale (on the supply side), as a consequence of the growth of the data volume and of their capacity to produce increasing returns of scale, and of economies of scope (or variety), due to the growing possibility of combining an ever-increasingly variety of data.** This costs structure (with decreasing average costs and low marginal costs, nearly zero), and the associated distribution of enterprises (in literature known as "firm size distribution"), is made even more asymmetric (i.e. skewed) from the presence of large fixed and undeclared costs related to R&D activities (see *infra*).

This cannot help affecting the presence of **entry** (or more often **development**) or **access to big data** barriers. In an era in which decision-making processes are increasingly based on the use of data, the analysis related to access to this relevant *input* is crucial for market dynamics directly (e.g. online advertising) and indirectly (e.g. online information) interested. Companies with a greater chance to collect digital data, and / or those that can efficiently aggregate heterogeneous datasets, and / or possessing the skills and tools of data analytics, enjoy a considerable competitive advantage. The barriers to entry or to development, in fact, identify those circumstances making it difficult to enter or enlarge companies in specific markets, guaranteeing those already operating a greater market power.[33] **Barriers to entry and development can be found at all stages of the value chain and can have a technological, legal and / or strategic nature and can also be presented simultaneously, reinforcing each other.**

In particular, the phenomenon of entry or expansion barriers is straightforward in the first stage of the value chain, namely that of data collection, as a consequence of the strong dependence on it of the subsequent stages, and therefore of the **spillover effects** that the creation of a barrier in the initial stage generates in subsequent ones.[34]

In the light of what has been described so far, it is clear that the issue of data access (methods, operator-consumer relationships, market structure, etc.) becomes of primary importance in order to create socially efficient contexts. Moreover, in these contexts there are also different collective interests, such as privacy, competition, media pluralism.

In the remaining part of this Chapter, the aspects related to the functioning of the ecosystem will be examined, referring to the component related to the technological, cost and competitive structure of the sides following the initial phase of data acquisition (and therefore to the user-operator relationship). In the following chapters, this first stage will be investigated, with reference to the economic aspects (see section 2.4), for the efficient and transparent assignment of property rights (see section 2.7), as well as the consequences on the entire political-social context (see Chapter 3).

[33] For an in-depth analysis of the role of access barriers on the market *of Big data* see, among other things, the work of RUBINFELD DL, GAL MS, (2017), *Access barriers to Big data*, 59 Arizona Law Review 339.

[34] As an example, entry barriers can occur in the data collection phase where it is not possible for all operators to replicate data collection in parallel, as happens, for example, for the exclusive information collected by *social network*. As far as the data retention phase is concerned, if it is true that the costs of the technologies allowing their storage have been reduced, it is equally evident that the explosion of the *cloud*, with its related *lock-in* effects, makes it difficult for new companies to enter the specific market segment. Lastly, the effects produced by the existence of access barriers in the phases of analysis and use of *big data* have equal intensity; as repeatedly stated, in fact, these activities require professional figures of the highest profile and very complex (even algorithmic) technologies. With regard to the use of data, a legal barrier, which deserves mention, relates to the correct attribution of property rights.

1.5. *Analysis of specific segments*

The next step of the analysis aims at studying some specific segments of the *big data* ecosystem in order to show how the structural characteristics above described materialize, on the production and technological side, in scenarios and market structures.

1.5.1. The first level: operating system (OS)

A first interesting field, analysed both in the literature and in the regulatory and antitrust cases, concerns **operating systems (OS). Operating systems represent a privileged channel (gatekeeper) for the collection of data, showing itself, in fact, as a real technological entry barrier.**

Operating systems are indeed software controlling the basic functions of a device and allowing the user to use the same device and the installed software applications. They are developed by specialized manufacturers (as in the case of *Microsoft*), or by companies producing also hardware (see *Apple*). Operating systems can control the functionality of any type of web browsing device, from traditional PCs, to new mobile devices (smartphones, tablets...).

In this sense, operating systems are at the top of the hierarchy of levels through which individuals surf the web (**Figure 1.10**), thereby being the most direct and main area for the acquisition of the digital footprint (see Chapter 2).

In general, the operating systems sector is dominated by the presence of a few operators, vertically integrated, and is split into two (or more) distinct commodity markets: operating systems for PCs and for mobile device. These markets have different structures, also considering the fact that they are going through a different phase of the product life cycle.

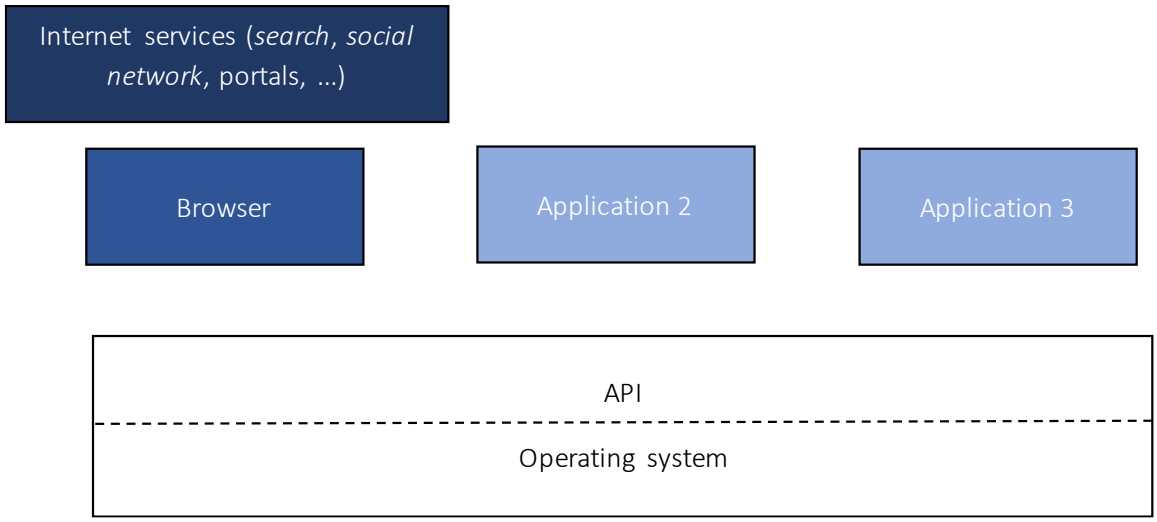


Figure 1.10 – Stages in accessing individual data
Source: AGCOM

The market **of operating systems for mobile devices** is of particular interest for the present analysis for reasons that will be illustrated in more details in this section 2.5 and the following ones. In this context, the dynamics of diffusion of the main mobile operating systems in the world (**Figure 1.11**) clearly show the prevalence of (only) two operating systems, *Android*, produced by *Google*, and the operating system *iOS* produced by *Apple*. The first is a free and open source operating system, the second is the operating system developed by *Apple*, and can only be used through the devices produced and distributed by the Cupertino Company (*iPhone*, *iPod touch* and *iPad*).

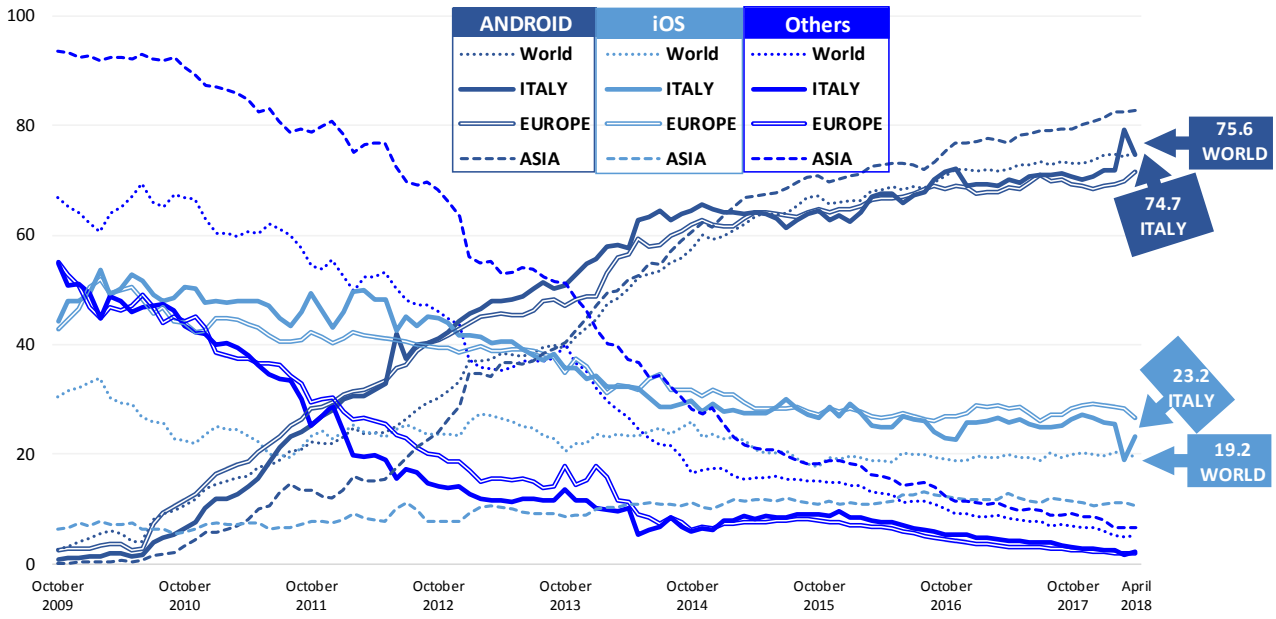


Figure 1.11 – Distribution of mobile operating systems (OS) worldwide (October 2009 - April 2018)

Source: AGCOM elaboration on monthly data from *StatCounter.com*

Google and *Apple*, therefore, are *leading* companies among the operating systems for mobile devices; in April 2018, the cumulative share in the world reached 95% (98% in Italy), increasing compared to the 90% reached in October 2016. Considering the dynamics over time, the market share of the operating system *Android* shows a trends of strong growth in all the geographical areas taken into account, representing 75.6% of the operating systems in the world (74.7% in Italy). The dynamics of the operating system *iOS*, on the other hand, shows a slight downturn which brought the market share to settle around 20%. In general, the market for operating systems of the mobile devices, as well as the desktop workstations, is characterized by a very high level of concentration, which results in a duopolistic structure on all world markets; this concentration is mainly due, as mentioned, to the implementation of the network effects[35].

In this context, all the other operating systems show a declining trend that made them almost disappear; all this starting from the introduction on the market, in 2008, of the first *Apple* smartphone. Until 2011, their overall penetration was still higher than the one of the two most widespread operating systems, while from that year onwards they have had a gradual and rapid reduction, until reaching, in April 2018, just 5% of operating systems in circulation.[36]

Furthermore, on one hand, the level of vertical integration with the devices producers especially the smartphone manufacturers (see the case of *Apple*), and, on the other, the level of openness of the software itself has a decisive influence on the diffusion of operating systems device. The operating system *Android* shows an opening degree higher than the one of *iOS* and is therefore usable for the functioning of several

[35] In this sense, "*Early on, [Microsoft] recognized that consumers would benefit greatly if a wide range of hardware and software products could interoperate with one another. Among other things, (i) the products would be more useful if information could be exchanged among them, and (ii) development costs would fall and a broader array of products would become available if they could be developed for larger customer segments without the need to rewrite software to target narrow platforms. As more products became available and more information could be exchanged, more consumers would be attracted to the platform, which would in turn attract more investment in product development for the platform. Economists call this a "network effect," but at the time we called it the "positive feedback loop*". See direct testimony by Bill Gates, Civil Action No. 98-1233 (CKK) in section 25.

[36] The item "Others" includes numerous operating systems including *BlackBerry OS*, *Series 40*, *Windows Phone*, *Samsung*, *SymbianOS*. Some of these operating systems have disappeared.

mobile devices offered by different manufacturers, while the operating system *iOS* is used only by the products of *Apple*. In any case, the first level in the *big data* ecosystem chain, at least in the part concerning data directly associated with individuals, is characterized by a high and increasing concentration of the markets.

1.5.2. The second level: search engines and social networks

A second level of data acquisition concerns the one related to web browsing that can take place through specific software (browsers) or directly through APP. In this context, search engines and social networks represent platforms concentrating on them an absolute importance in terms of audience (reach and time spent by consumers),[37] of social relevance, and of superiority from the point of view of media pluralism (see Chapter 3). On these specific areas of the market, the AGCOM has already undertaken studies and analysis, given the importance that these tools assume today in the ways in which public opinion is formed.[38]

As regards the **search,** this is one of the first online services offered once the web has become an open system for contents and services to individuals.[39] **Search engines** solve transactional problems, both on the demand side (the user's search for information), and on the supply side (the need for subjects offering services and products to be known by users), thus playing the role of platform. **In particular, the search is characterized by the existence of crossed network externalities (or feedback effects between the two sides of the market), which contribute to determine a particularly concentrated market outcome.** The historical trend of market shares (**Figure 1.12**) makes it possible to observe an evolution typical of the markets where network externalities prevail: a first phase in which several subjects operate, followed by market concentration in which a dominant platform established itself (with shares exceeding 80%).

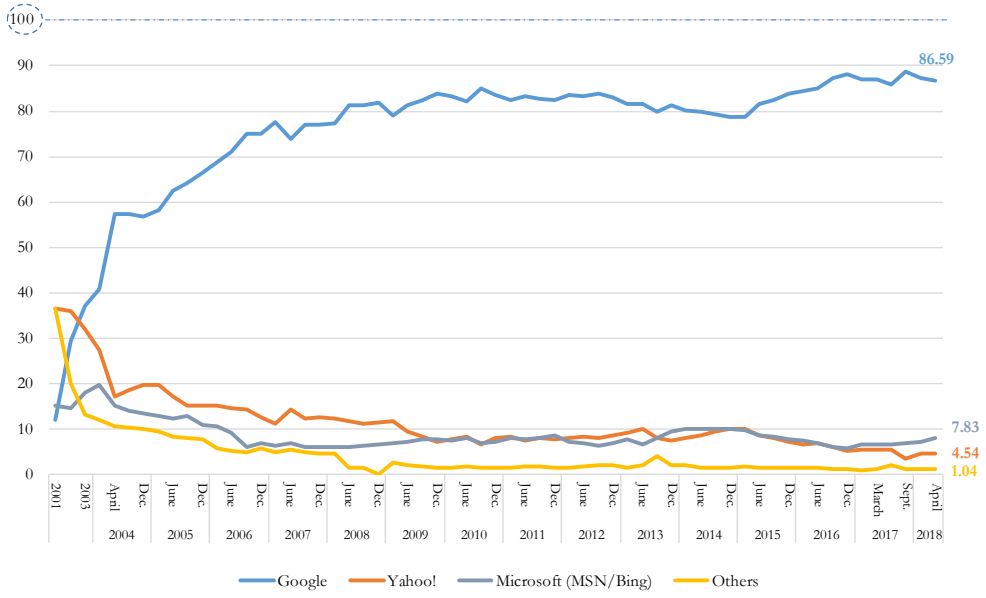


Figure 1.12 – Historical evolution of the market shares of search engines in the world (%)
Source: AGCOM elaboration on data from *SEW/WebSideStory*, *NetApplications*, *NetMarketShare* and *StatCounter*

[37] See Communication markets monitoring system, AGCOM, no 1/2018, (slide 2.4).
[38] *Cognitive survey on Internet services and online advertising*, concluded with resolution no. 19/14/CONS Indagine conoscitiva sui servizi internet e sulla pubblicità online.
[36] [39] For the history of search engines see www.searchenginehistory.com, www.wordstream.com

The market structure of search engines has been therefore developed towards a particularly concentrated structure, in which, at present, an operator like *Google* holds, since almost fifteen years and almost everywhere in the world, over 80% of market shares.[40]

The second case considered concerns **social networks**; as is known, these are online platforms allowing users to build a public or semi-public profile within a predefined system, by creating their own network of contacts (between users connected and subscribed to the same platform), and to view and scroll the lists of users in the other profiles. Nature and definition of relationships within the social network of users can vary from one system to another.

There are many features allowing you to recognise the different social network such as, with no claim to exhaustiveness, the different degree of public visibility of the profile, and of the related contacts, the tools available to interact, the users base (the target) to which the service is addressed. A further distinction can be made in relation to the features offered to its users, including the possibility of leaving comments, messages, suggestions, as well as expressing their reaction, even though using different methods and names. Moreover, among the most widespread technological features we remember the sharing of photos, videos and multimedia contents, the creation of blogs and discussion forums, instant messaging and chats. **All activities involving the generation of digital data, which are collected and processed by social network** (in this sense, see Chapter 3).

With regard to market characteristics, the model of business chosen for social networks is characterized, similarly to other horizontal web services, by the enhancement of contacts in advertising terms, against a completely free (or almost) service for users.

Although there are examples of (partial) enhancement on the side of end users (ex. *LinkedIn*) and / or of the developers of programs and applications (ex. *Facebook*), however, for all social networks the advertising component is still predominant. In this regard, **an evolutionary dynamic very similar to the one of the search** (**Figure 1.13**) in which the acquisition of the critical mass of users necessary for the exploitation of network effects and to trigger positive feedback phenomena represented the indispensable element for the success of the platform, as well as the prerequisite for achieving and exceeding the break-even mainly through (innovative) forms of advertising. Thanks to the operation of the network effects, in a few years, i.e. from the launch of the site (which occurred, as mentioned, in 2006) to date, *Facebook* has rapidly reached the world leadership with a market share that has been steadily higher than 80% for three years (June 2014 - October 2017), then slightly decreasing and settling, currently, over 70%.

[40] For a competitive examination of the *search*, and of *social network*, as well as other areas of the market (operating systems, *browser*, gateways,...) please refer to Chapter 3 of the aforementioned *Cognitive survey on internet services and online advertising*.

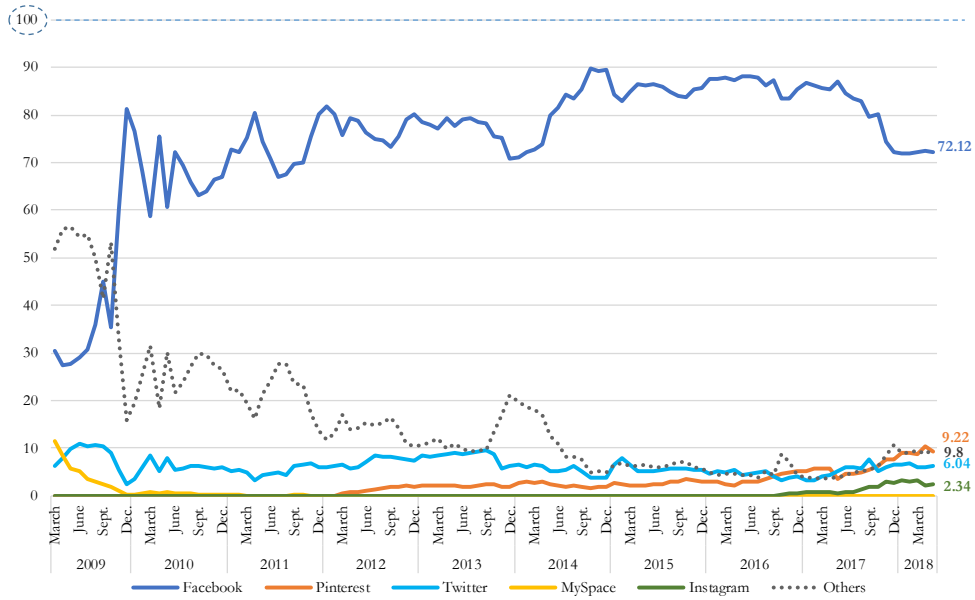


Figure 1.13 – Historical evolution of the market shares of *social network* in Europe (%)
Source: AGCOM elaboration on data from *StatCounter*

This process has been facilitated by the interconnection with other internet sites able to "recall" data traffic and increase user involvement. The interconnection between internet services is, in fact, a key element for the affirmation of evolutionary scale processes, as well as for the reduction of entry barriers, when some operators have already established themselves.

Furthermore, the sudden decline of some social networks - think, for example, about *Friendster*, where the loss of some (groups of) users was enough to cause a mass abandonment of the platform - confirms the absolute importance for this type of web service of direct network effects.

1.5.3. The third level: data centres ("Production capacity")

A last case useful to show how the features of the *big data* market structure may lead to market balances with high levels of concentration is the creation of material infrastructures allowing data acquisition and storage activity*;* this is the market segment linked to **data centres**.

A data centres represents a physical infrastructure (a building or a real estate), which houses data processing equipment (*server* and the infrastructure necessary for their operation) of one or more companies or organizations (co-location).[41] The data centre, therefore, must consist of at least a separate room with independent power supply and relative air conditioning. This definition immediately makes it possible to highlight that this is a market segment where small, medium and large companies operate. In many cases, the same business activity of companies needs data centre, whose dimensions, therefore, depend on the commercial needs of the company.[42]

[41] According to a study carried out by *Microsoft* in 2009, the total costs of a data center are distributed as a percentage as follows; 45% for servers and other components (CPU, memory, storage systems), 25% of the costs are attributable to the creation of infrastructures for the transmission of energy and for the operation and cooling of the servers, 15% is related to the electricity and the remaining 15% relates to the creation of the network through which the network is connected to data center. See also GREENBERG A., HAMILTON J., MALTZ D.A., PARVEEN P., (2009), *The Cost of a Cloud: Research Problems in Data Center Networks*, Microsoft Research, Redmond USA.

[42] For example, a company like Eni in 2013 inaugurated its data centre (in Ferrera Erbognone, in the heart of the Po' Valley) in which there are the fundamental tools (the HPC4 supercomputer capable of carrying out 22.4 million billion of mathematical operations in a second) for Eni's exploratory discoveries around the world and useful, for example, to develop

The data centre represents the main infrastructure of data-driven economy. The offer of any online service is based on server located in a data center. The most impressive data centers are complex structures with numerous mechanical, electrical and communication instruments. For an efficient operation, data centers need electricity to feed the server, and water for their cooling. Also, to carry data to and from data centers, communication infrastructures are necessary (usually fiber-optic links).

Investments in data centers, consequently, are constantly increasing, driven both by the need of private companies to build their own centers for data collection and processing, and by the intense spread of **cloud computing services**, i.e. the outsourced offer of storage and calculation services (servers, filing resources, databases, software, analysis, etc.) via the web.

It is crucial to stress that many digital service providers build data centre mainly to be able to manage data they acquire directly from their users; this is the case of *Google* that is, however, one of the main player in the cloud market, as well as *Facebook*, that at the moment does not take into account (or at least only indirectly), in its offer, the sale of cloud services but that obviously has its own data centre.

According to *lifelinedatacenters.com*, data centres are about 8.6 million in 2017, albeit decreasing due to the increasing spread of the cloud services; this decrease is counterbalanced by the increase in square meters required for each individual plant, i.e. an increase in the average size,[43] associated with strong scale economies.[44]

For example, from 2007 to 2017, *Google* invested about 3.2 billion euros to build the four data centres currently operating in Europe, for a value of around 300 million euros a year;[45] this implies the existence of increasing and high entry costs for those who want to enter these market areas.

The phenomenon of the data centres is not recent, since all companies have always needed physical spaces where to store information. However, it is evident how, under the pressure exerted by *big data*, the need to use more and more data acquisition and management infrastructures (also in cloud computing,) was indeed a new and growing drive towards this type of infrastructure.

3D images of the subsoil. Overall, over 7,000 systems, with more than 60,000 CPU cores. https://www.eni.com/it_IT/innovazione/piattaforme-tecnologiche/aumento-recupero-idrocarburi/hpc.page.
An example concerning a small company, but that still shows how important being the specifications of a data centre, it concerns the company *CriptoMining* (http://criptomining.online/) an innovative start-up producing cryptocurrencies on an industrial scale through an IT process known as mining. The data centre situated in the basement of a building in the centre of Milan, will use 250 24-hours machines (currently 12 machines are operating). The structure is positioned three floors below the ground, to ensure low temperatures allowing the cooling of the machines, monitored 24 hours a day.

[43] https://lifelinedatacenters.com/data-center/emerging-data-center-trends/

[44] See also GREENBERG A., HAMILTON J., MALTZ D.A., PARVEEN P., (2009), *The Cost of a Cloud: Research Problems in Data Center Networks*, Microsoft Research, Redmond USA.

[45] Source: *European data centres* edited by Copenahgen Economics, 2018. These are data centres located in Dublin (Ireland), St. Ghislain (Belgium), Eemshaven (Holland) and Hamina (Finland).

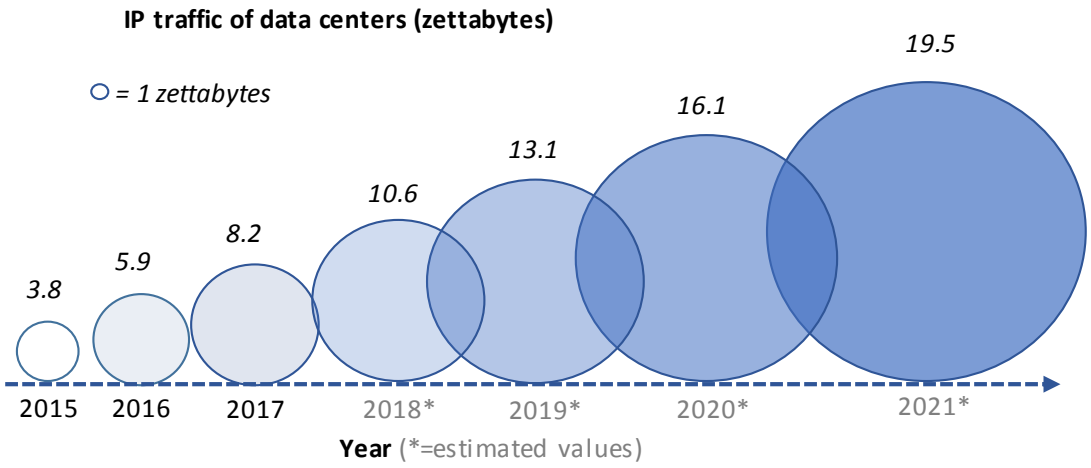


Figure 1.14 – Growth of IP traffic for cloud services
Source: AGCOM elaboration on data from *Cisco*

The huge diffusion of the big data economy, and consequently of the cloud services (**Figure 1.14**)**, produced substantial changes** in a segment previously dominated by proprietary data centres. Furthermore, the segment's development is strongly **influenced by the characteristics of the big data** described above.

The volume, the variety and the velocity of data have pushed the ecosystem towards a cost structure dominated by the presence of scale economies, and therefore by a competitive structure characterized by increasing levels of concentration and by increasing entry barriers.

According to the consulting company *Gartner,* the world market for (public) cloud services this year is expected to grow by 21.4% for a total amount of revenues equal to 186 billion dollars.[46] In this context, the 10 main providers have significantly increased their market share, reaching together 70%.[47]

Similarly, according to a study provided by *Synergy Research Group* (**Figure 1.15**), the overall market (including public and private cloud*)* is led by Amazon, with an increasing share of 34%. By offering the *Amazon Web Services*, Amazon was the first operator, in 2006, to launch a cloud service on a large scale.

[46] The main difference between private and public cloud lies in the fact that, opting for the second solution, data are stored in the data centre of those who offer an outsourced service that, therefore, is responsible for their management and maintenance. If among the large companies the use of public cloud is still not fully widespread, since they prefer maintaining direct control over data, for many small businesses and especially for individual users (which still represent a primary source of digital data) the use of these storage environments has become essential. The public cloud identifies a particular architecture granting each user a personal storage space (cloud storage) which has the characteristic of being accessible in any place and with any *device* through an internet connection, supplemented by data processing services. Through cloud storage, it is possible to synchronize all the files in a single place (the cloud), with the consequent advantage of being able to re-download, modify, delete, update, without having anymore the need to bring the so-called external memories (external hard disk, USB pen drives, etc.). With the advent of cloud computing there has therefore been a radical change in the way in which data are processed both by consumers and operators.

[47] https://www.gartner.com/newsroom/id/3871416.

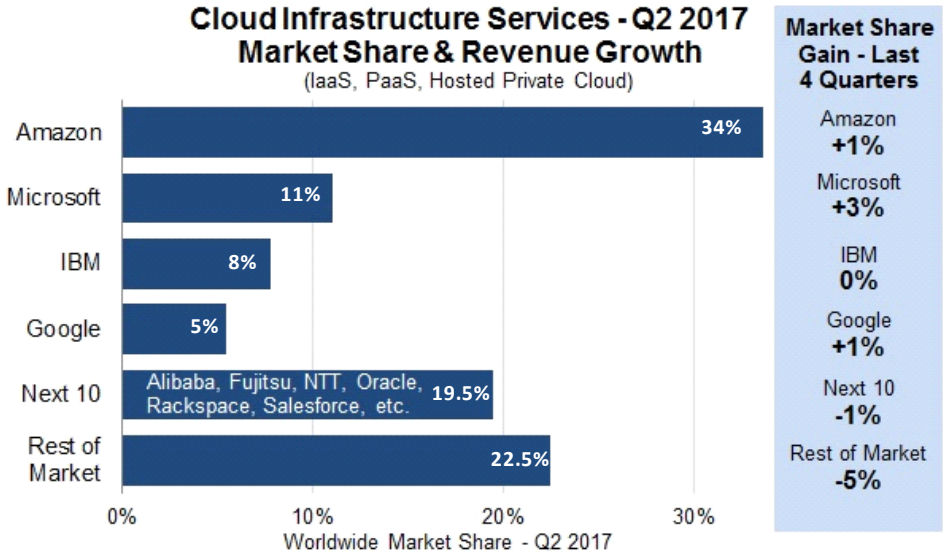


Figure 1.15 – Market shares in the services of cloud (2nd quarter of 2017)
Source: https://www.srgresearch.com/articles/leading-cloud-providers-continue-run-away-market

The infrastructure asset follows then the global dimension of the major providers. The existence of scale (and scope) economies, as well as the need to duplicate data (see above) and to place computer centres closer and closer to consumers have in fact pushed operators to create real global networks (as that of *Google*, **Figure 1.16**).

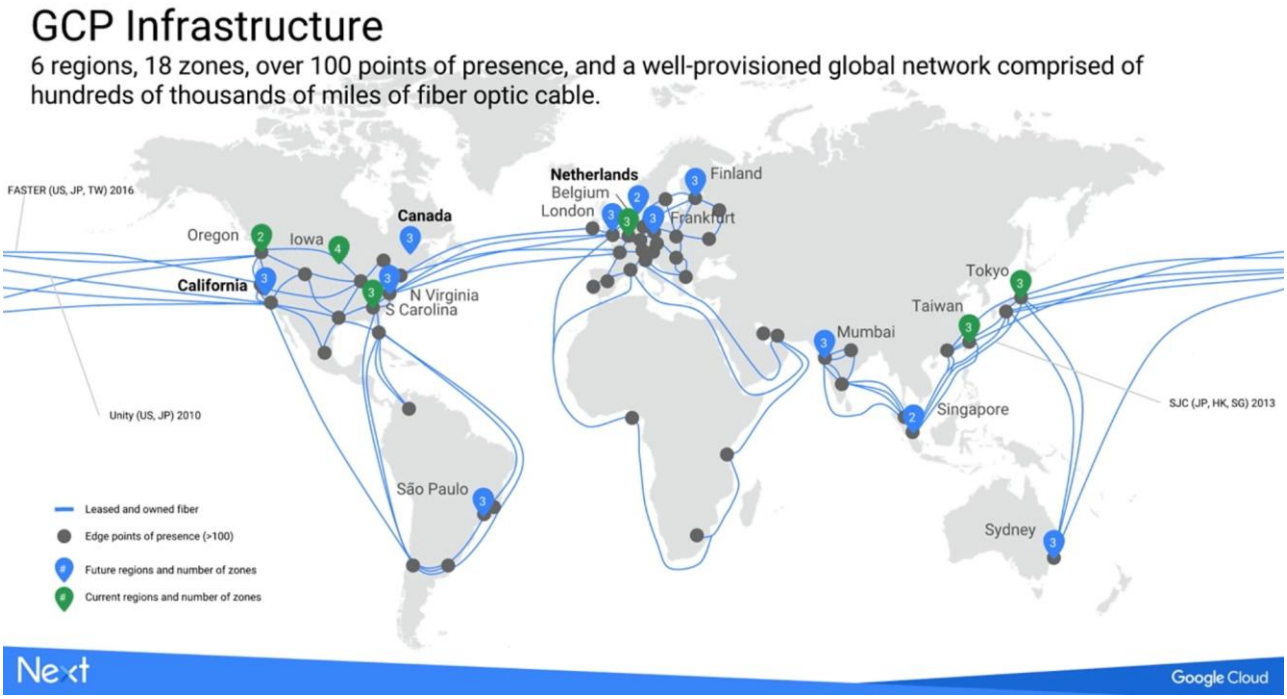


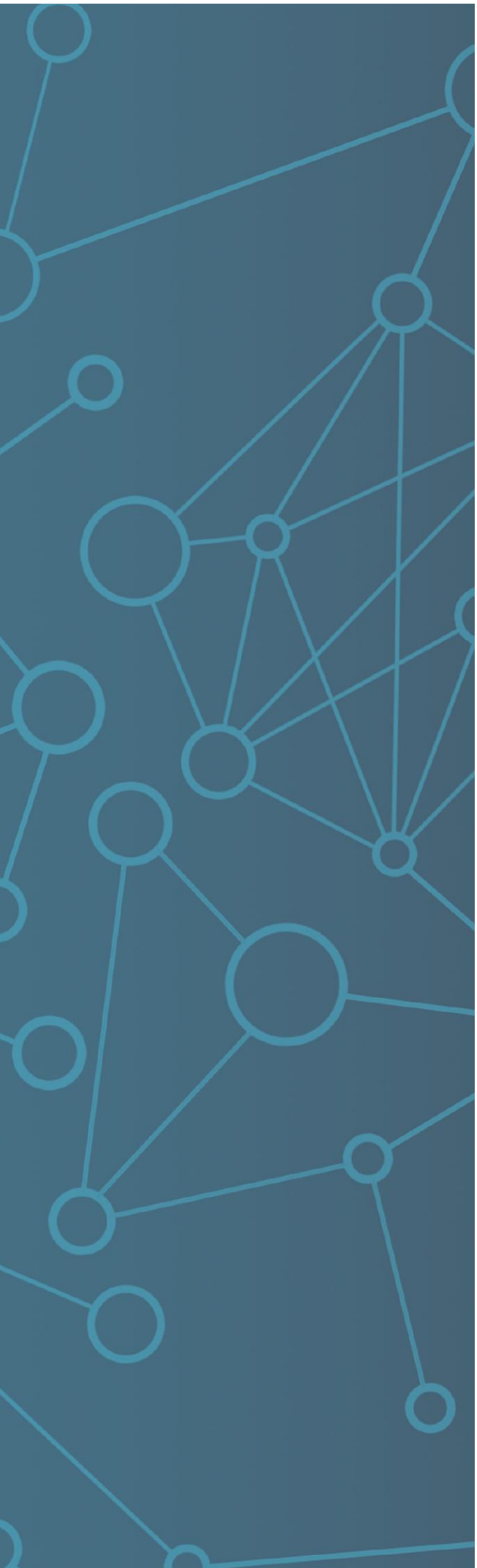
Figure 1.16 – Infrastructure of Google for supplying cloud services - *Google Cloud Platform* (GCP) - (2017)
Source: *Google Cloud Next* 2017, 8 - 10 March 2017 San Francisco - https://www.youtube.com/watch?v=vX92qwNtkFo&t=1362s

Costs are difficult to determine. According to a 2014 study, the average size of *Google* data centres is between 15,000 and 18,500 m2, although some reach 90,000 m2 like *Pryor Creek* in Oklahoma (USA). Moreover, the variability of the dimension represents a specific strategy adopted by the companies,

structuring their own data centre according to the need and the specific geographical area.[48] Not surprisingly, in 2003, *Google* obtained a patent for the implementation of modular data centres (built in classic containers used to transport goods) allowing greater flexibility in carrying out infrastructures. It is also interesting to underline that energy power, in megawatts (MW), required for the operation of data centres of *Google* has been estimated, in a year, equal to 0.01% of the world energy power.[49]

In conclusion, the advent of *big data* has determined acquisition, storage and analysis of a growing number and variety of data. This is having a strong impact on the underlying cost structure and therefore on the related market structures. Despite the transition from a regime of scarcity to one of data redundancy, very concentrated market balances are emerging, which, then, can only reflect on the connected downstream market areas. Because of these and other characteristics (the existence in particular of network externalities), areas such as operating systems, web services (such as search and social network) and online advertising have strong and growing concentration characteristics.

[48] GHIASI A., BACA R., (2014), *Overview of largest Data Centers*, 802.3bs Task Force.
[49] GHIASI A., BACA R., (2014), *Overview of largest Data Centers*, 802.3bs Task Force.



2

THE INDIVIDUAL AS A SOURCE OF DATA

2.1. Digital data and the individual

A particularly relevant field in *big data* economics concerns information on individuals and, consequently, of how these information are protected. Traditionally, there is a distinction between "personal data" and information that is not considered as such. This type of approach is apparently simple to understand from the theoretical point of view, but not in practice, since **the complexity of the big data phenomenon makes it difficult to distinguish, among all the information collected about an individual, between personal and non-personal data.**

Many of the practical problems of the traditional approach derive from the difficulty of establishing a priori what data are used to identify an individual, his habits, even his private sector. This is innate in the very nature of *big data*, which was investigated in the previous chapter.

Technological development and the ways in which individuals consume products and services in the digital economy also make acquisition, storage and analysis of large amounts of both structured and non-structured data, an activity very difficult to brake. In fact, the growth of the internet usage among the world population has produced disruptive effects on the traditional approach to individual information; internet is now a pervasive tool that for many is essential both in the workplace and for aspects related to everyday life (entertainment, fitness, tourism, reading, etc.). The availability of an "*always on*" connection is often considered by citizens an indispensable element of their life.[50]

However, every time an individual is connected to the network (also via sensors), he leaves numerous "traces", which are transferred to online operators either deliberately or, more often, unintentionally. The fingerprint (online footprint) of each individual is made up of numerous information, some of which can be directly associated with him (name, surname, age, etc.), others associated with the activities carried out by individuals (payments, research, etc.), others which, although not having direct ties with the individual, through their processing, can easily be associated with people, nevertheless, the use of technical tools aimed at the anonymization of data.

It should be remembered, among other things, that the only information concerning the name and surname very often are not enough to identify a specific individual: for example, the name Mario Rossi does not allow to identify a specific individual, since it identifies all the existing Mario Rossi. Very often, on the contrary, the identification of a specific subject occurs through the use of alternative information that make it possible to uniquely identify a person.

Therefore, it becomes more and more important to shift the attention on market operators who collect data online, on how they collect, process and store data, as well as on the purpose of using such information, even if in this case we must take into account, as previously described, that in many situations the usefulness of a data is not known at the time of the collection, or rather, in a second moment, new data deployment opportunities can later emerge however unknown at the time of data acquisition.

It is useful to remember, moreover, that the same data could be considered as "personal data" in the hands of a given subject, while it could lose this characteristic if another subject has got it. The example that can be done is a photo in a public place (for example, a public event): a journalist might be interested only in showing the crowd present, while a public official could use the same information (photo) for identify the people present. This phenomenon, among other things, happens more and more often due to the increasingly pervasive presence of cameras that private individuals and public administrations place in the territory for video surveillance. Much, therefore, depends also on the purpose underlying the

[50] See also AGCOM, "The use of communication services: experiences and perspectives" (published on October 20, 2016), in particular, see Chapter 3, where it is highlighted as *"Consumers confirm the importance of the Internet in everyday life; access to the network is indeed considered an indispensable service for over 90% of individuals.* "(Page 21).

processing of data and, as described above, the possible search for spurious correlations and the increasingly massive use of algorithms and *machine learning* techniques. In other words, **the characteristics of data also depend "on the eyes of the beholder".**

A contribution to the entirety of digital data comes from **all the activities that individuals carry out browsing the internet or in any case if connected to the network** (and in some cases even temporarily off-line): the use of search engines, the reading of news, video viewing, shopping, online games, etc., these are all activities leaving digital traces, that is data, released more or less consciously by individuals. The only fact of connecting one's own device to the network produces a considerable amount of digital data, since an IP address is required containing geographical data; all the pages visited by internet users form a data trails that are recorded. These trails provide for many information about tastes (or preferences) guiding consumption choices, uses and habits of the user, often without user perceiving he is releasing data, against systems of data analytics allowing the collection and processing of these information.

As highlighted above (see section 1.1.2), another activity that creates, over time, a large quantity and variety of digital data comes from the **use of email,** one of the main tools used, especially in the workplace, to communicate; structured and non-structured data, such as text, images and the network of acquaintances, can be easily collected, stored and used. This is true even for those software allowing you to communicate in real time that in addition to the text use images, audio, video, files, etc.

Clearly, a significant amount of data comes from the use of **social networks** (see Chapter 3), so widely spread in recent years among the population. Also in this case, users' awareness appears to be rather reduced: posts, comments, images, contacts of individuals are data that are collected and used, without the user often being aware of ceding them. On top of this, users have often little knowledge about the purposes of their data treatment.

Another fundamental element of this ecosystem concerns **mobility**. Today, in fact, users do not seem able to live without using mobile services, which allow the localisation of devices at any time (*Located-Based Services - LBS*). As a consequence, the daily route (routing) of a device owner is traced with precision.[51] Furthermore, more and more mobile devices replace desktop computer, so-called *Mobile PC* (*mPC*), with related activities (email, messaging, video viewing, social network, etc.) and this, as described above, immediately generated the development of "cross-device" techniques allowing tracking of consumers' preferences and habits (think, for example, about the vision of a video that starts from a desktop computer, then is interrupted, then seen on a mobile device).

Finally, digital data related to individuals are increasingly collected through **sensors and sensor systems** already pervading the daily life of citizens: for example, video surveillance systems and advertising panels with optical sensors have contributed to the development of techniques allowing facial recognition, significantly contributing to the growth of variety, speed and volume of digital data. With the proliferation of sensors and the Internet of Things (IoT), also the use of goods by users is constantly monitored and recorded. In this regard, it is useful to remember that even the modern mobile devices contain a series of sensors originating a variety of relevant information (accelerometer, gyroscope, magnetometer, proximity detector, fingerprint reader and facial, recognition, light sensor, thermometer, GPS, etc.). This evolution cannot but undergo further acceleration with the future implementation of fifth generation mobile networks (so-called 5G).[52]

[51] For example, each owner of a smartphone can, by subscribing to the services of *Google*, verify the accuracy with which information on the travel history is collected: https://maps.google.com/locationhistory.

[52] See also: AGCOM, Cognitive survey concerning the development prospects of wireless and mobile systems towards the fifth generation (5G) and the use of new spectrum portions above 6 GHz, published March 5, 2018.

The sources of digital data closely related to individuals (network connection, use of e-mail, use of mobile telecommunications services, sensors and sensor systems) produce a continuous data flow (velocity), very variegated (variety) and very dense (volume). It is worth remembering that this data flow only partially consists of data consciously provided by the user; an increasingly important part, in fact, is collected without the explicit consent of users, who passively (passive date) become a primary source of information.

The increasingly pervasive collection of data related to the habits and preferences of individuals, together with the ability, through their analysis, to find models unknown to today, have led to the creation of gigantic opportunities for innovation, but also considerable risks. Furthermore, the same notion of big data, or big data analytics, suggests that the value of the information available is not exclusively linked to the increasing ability to collect data or to the quality of the data, but above all to the subsequent possibility of carrying out, or supporting, through those same information, decision-making processes (often in real time): increase in information corresponds to the transformation of them into knowledge (see **Figure 1.7**).

Against the revelation of their own information, the individual tangibly and immediately **benefits** in terms of monetary compensation in the form of discounts or rewards, access to preferential treatment or free and personalized services. On the contrary, a given data "disclosure" towards third parties is associated with **costs** such as those of "invasion" that an individual must support to deal with spam, telemarketing, and advertising (even via email), up to real identity theft, but also costs related to the fact that, by "profiling" the user, it is increasingly probable to receive selective offers (i.e. designed on their spending intentions) and, therefore, being recipients of price discrimination policies.

In a digital environment populated by few gatekeepers (see section 1.5), it happens that few privately-owned companies, through the detection of preferences, desires, interests and habits of millions of people, have made the collection and exploitation of data their own core *business*.

Data are the driving force of an ecosystem at the centre of which some platforms on different sides offer, on one hand, free services to individuals who, in exchange for them, give, more or less consciously, their data. On the other side of the platform, and thanks to the data acquired from individuals, all the agents who wish to use them are offered data analytics services, profiling of potential customers and targeted advertising spaces.

In this context, therefore, data are only "bargaining chip". In fact, in the digital economy era, the widespread practice for which the true value of the transaction between consumers and businesses does not concern the consumption of goods and services, which, as mentioned, are often presented as free, but from the exchange (mostly implicit) of the underlying information (see section 2.7).

In the *big data* era, individuals contribute, most of the time unintentionally, in a decisive manner to the creation of the digital data flow; as two sides of the same coin, on one hand, users benefit from better services; on the other hand, they bear costs deriving from the transfer of information, with a redistributive process typical of digital economies.

2.2. Economic characteristics of data

The interest of economists in the data market was mainly focused on the information dimension that emerged within it. The analysis focused on the trade-off deriving from the indicatively rational choice, based on the knowledge of the relating risks and benefits, of the individual in relation to the transfer of information concerning him / her. If the transfer of data by an individual takes place with a view to costs-benefits analysis, then it is possible to talk about the exchange of data in the same way as the transfer of any commodity, that is to say **"data as a commodity"**.

In this sense, it becomes first of all important to analyse which are the **characteristics of the economic good represented by the data**. As already underlined previously (see section 1.4), the proliferation and use of data for commercial purposes we have seen in recent years have been compared to the discovery of hydrocarbons and their application to modern industry.[53] In fact, the comparison appears to be quite shareable if we think of data as production inputs able to trigger a commercial revolution generating new products and services that increase the wellbeing of consumers and, at the same time, present certain risks (like pollution) able to reduce the scope of the same social benefits.

The distinctive characters of *big data*, as economic assets, can be summarized by using some classic concepts used in the economic field. First of all, data is presented as a **non-scarce** resource. Scarcity is a very important feature of economic assets that is transferred directly to the price level; the more a commodity is scarce, in fact, the greater being the equilibrium price on the market. As described above (see section 1.1), the amount of data not only shows an exponential growth in volume, but also in variety and above all it is spreading at an unprecedented speed; this abundance of information makes the identification of an economic value attributable to the individual data very difficult. While a barrel of oil has an intrinsic value, representing the unit of measurement for identifying the reference price of an asset on the market, a single datum in itself hardly presents an economic value; it is, in fact, through the aggregation of more data and their subsequent analysis that we can extract value from data.

Secondly, data are presented as an asset characterized by **non-rivalry** in consumption, a feature that partly allows comparing data to a public good;[54] in fact, the use of data by an agent does not affect the ability to fully enjoy it by third parties or, in other words, the same data can be re-used without its re-use leading to a reduction in value. Moreover, data is also a partially excludible good, i.e. it is impossible **excluding** third parties from their consumption; despite the interest and the presence of rules to protect the confidentiality and integrity of data relating to individuals, the condition of being non-excludible is not always fully realized (see also in this 2.3). On one hand, in fact, technological development allows extremely easily and at marginal costs close to zero to reproduce the same data several times; this significantly reduces the possibility of exclusion of data exchanges. On the other hand, the more complex being the data collected, the more it is possible to make the data exclusive, by creating trade barriers to their exchange and, therefore, to their widespread use.[55]

Economic analysis uses the concepts of **substitutability** and **complementarity** of the goods (or of the productive factors) regarding the preferences in tastes (or in the combination of the productive factors); in this sense, and in an intuitive way, data can present both characteristics, even if the *big data* world is based predominantly on the concept of complementarity of use. In fact, as repeatedly emphasized, the complementarity between data, of different formats and from different sources, is predominant: the ability to aggregate heterogeneous sources of information is essential to extract value from the data. However, data has also a certain degree of substitutability; for example, for the purpose of carrying out a marketing campaign, data on consumer behaviors relating to consumption choices can be partly replaced by data on the actual consumption of goods and services. Nevertheless, it is from the

[53] *The world's most valuable resource is no longer oil, but data*, Economist, 2017; *Why data is the new oil*, Fortune, 2016; *From fintech to techfin: data is the new oil,* The Asian Banker, 2016.

[54] A good is defined a "pure public good" when everyone can consume concurrently (no rivalry in consumption) the same good and nobody can be excluded (not exclude in the exchange) from the consumption of that good. It is good to remember that in reality the two characteristics combine in such a way as to provide a very varied scenario of public goods, depending on the greater or lesser intensity with which the two characteristics present themselves.

[55] VAN TIL H., VAN GROP N., PRICE, K., (2017), *Big data and Competition Policy*, Ecoryse on behalf of the Dutch Ministry of Economy.

combination of both types of data that we can obtain even more precise information that can make the advertising campaign more effective.[56]

In general, the economic assets typically considered in literature can also be classified according to the speed at which they lose value that is **their perishability**. Also for this characteristic the complexity of *big data* makes it difficult to classify data according to their perishability. In fact, some data lose their value almost immediately after their use; think of the data on the conditions relating to traffic that can have a huge value at a specific time and no value at all in the following moments. However, by virtue of the possibility of data reuse, as well as their optional value, the loss of value of a so strategic asset, like others, takes place very slowly. Indeed, the distinctive feature of *big data* is to provide a collective dynamic information technology value.

This means that the single datum not only provides timely information on an individual, but also allows you to locate social behavioural patterns, which are later exploited in various ways to extract value. In this sense, **the sum of the values of a single datum taken at a given moment is very different (and much lower) to the value of data taken as a whole and in a wider time interval**. This characteristic, which has implications also in terms of structure of the costs of the companies (high fixed and sunk costs and marginal costs close to zero), can only lead to market concentration.

The concise analysis of the economic characteristics of data is particularly interesting as it shows how the application of concepts belonging to the classical economic paradigm is not always useful to explain such a complex phenomenon as *big data*. In particular, the idea of considering data as a traditional good does not allow, for example, the use of traditional national accounting tools to measure the flows of data exchange between countries, as it happens for any other asset.[57] The fundamental identity of international trade, according to which domestic consumptions are equal to domestic production plus imports to which we must subtract exports, complies with goods such as hydrocarbons or classic raw materials, but presents huge application difficulties as refers to data. Just think of the application market for mobile devices, the so-called mobile apps (see section 2.5); in fact, most of the applications are present simultaneously on several geographic markets without this implying a process of exporting the service itself. In fact, if a country wants to increase the exports of a specific good, being equal the total amount produced, it will have to reduce its domestic consumption; when it comes to data, on the other hand, exports do not imply a reduction of the amount of data internally available, also because of the non-rivalry in consumption.

A next step in the economic analysis of *big data* concerns the identification of possible **trade off occurring when individuals are called upon to make decisions about the transfer or not of information**. An individual who decide to give consent to the transfer of his/her data, against a partial knowledge of the environment in which the assessment is carried out and, therefore, of risks and benefits deriving from it, makes a cost-benefit assessment similar to the one made before any purchase decision. More specifically, the individual performs an evaluation aimed at understanding whether it is convenient to cede his/her information in exchange for certain benefits, even if they are not benefits which can be economically evaluated.

It is an exchange taking place based on the presence of significant and structural **information asymmetries** among the agents involved (in this case, among users who cede data and operators who acquire and use them); this turns for individuals into a context in which, not having access to all the information, it is in fact impossible to proceed with a correct measurement of costs (uncertain and

[56] VAN TIL H., VAN GROP N., PRICE, K., (2017), *Big data and Competition Policy*, Ecoryse on behalf of the Dutch Ministry of Economy.

[57] "National Accounting is the quantitative description of the economic activity of a country, in the form of a complete and systematic presentation of the economic and financial flows occurring between significant groups of operators and the final amount of real and financial assets", SIESTO V., (2003), *La contabilità nazionale*, Il Mulino.

potential).[58] Individuals do not have the same quantity and quality of information of those who collect and put them together. Therefore, the cost of disclosure of individual information weighs on the individuals themselves. In addition, individuals (original source of the data) are not aware of how their data will be used, and are therefore excluded from the benefits generated by their use.

The consumers' (individuals) choices, therefore, are performed in an environment characterized by exogenous and endogenous components that are difficult to evaluate, among which the presence of uncertainty and dependence on the context strongly emerges (**context dependence**).[59]

Uncertainty in this case is a function of technological progress and of data collection activity; the individual, in fact, will have to evaluate the fact that technology conceals, firstly, the type and mode of data collection and, secondly, their subsequent use. It is rare that individuals being really aware of what and how much information are acquired by online operators with which they enter, more and more frequently, in contact, but above all it is very difficult to trace the types of use by those who collect data. The difficulty in understanding what the consequences of the possible transfer of data are, gives the analysis of costs and benefits a significant degree of uncertainty.

Another factor determining the intensity of information asymmetries and increasing the uncertainty concerns the dependence of choice from the specific context (**context dependence**); this component makes it possible for the same subject to express diametrically opposed preferences with respect to the choice to cede her / his data according to the context: as the "environmental conditions" change, individuals pass by an attitude of profound distrust with respect to the eventuality of cede their own data, to total indifference to the consequences of such a choice. Given the presence of information asymmetries, moreover, those who are interested in data collection often use every possible strategy to favour attitudes providing for the transfer of data. In practice, context dependence further increases systemic uncertainty.

In conclusion, individual choices relating to the transfer of her / his data in order to obtain a service are directed according to the balance between benefits, often immediate (e.g. access to a service) and costs (often uncertain and unknown). In this context, the information asymmetry between users and market operators is pervasive and structural: consumers do not have all the information they need to make an informed and rational choice. Most behaviours, to be efficient, would require a degree of technical knowledge that goes far beyond the skills widespread among the population. In other words, **a higher degree of transparency is in many cases useless where consumers fail to correctly process such information due to a structural gap in technological knowledge.**

The literature has also shown how human behaviour, especially in conditions of uncertainty, is not rational at all. **Choices, such as those relating to the transfer of personal data, are very frequently carried out impulsively and without an evaluation of the real consequences of the implicit exchange**.[60]

These elements have then effects on the **overall efficiency regarding the functioning of the markets**. The economic literature on the subject does not provide univocal answers to the question of what is the optimal level of disclosing data related to individuals. According to the so-called "Chicago School" approach, an excessive data protection turns into a reduction in market efficiency, as companies do not receive enough signals on how to efficiently allocate their production inputs. In other words, an excess

[58] VARIAN HR, (1999), *Economic aspects of Personal Privacy*, in Privacy and self-regulation in the information age, National Telecommunications and Information Administration.
[59] PURCHASES A., BRANDIMARTE L., LOEWENSTEIN G., (2015), *Privacy and Human Behaviour in the age of information*, Science 347 n. 6221.
[60] KAHNEMAN D., TVERSKY A. (1979), *Prospect Theory: An Analysis of Decision under Risk*, Econometrica 47, n. 2, pp. 263-292.

of restrictions would result in a transfer of costs from individuals to companies that would therefore find themselves operating less efficiently, both from a static (efficiency linked to the allocation of resources) and from a dynamic point of view (efficiency linked to innovations).[61] On the other hand, this type of exchange goes beyond the perspectives of the market, dealing not with simple goods but with aspects concerning fundamental, both individual and collective, human rights. If much of the analysis carried out in this Report refers precisely to the protection of these rights, and in particular to the constitutional one to information (see in particular Chapter 3), the economic aspects are analysed in greater detail here.

From the above arguments, **it is evident that the exchange of data often gives rise to structural market failures**. Firstly, as in the case of the aforementioned hydrocarbon emissions, the investments made by companies to collect data on individuals, not internalizing social costs, risk leading to a situation of overinvestment in the collection of information (in the case of hydrocarbons, pollution overproduction).[62]

As described earlier, the effect produced by the context is very relevant in this debate; for example, it is clear that in the face of the possibility of obtaining discounts or free and / or personalized services immediately, the individual will be required to release individual data such as those relating to their tastes and preferences without considering the costs arising from their disclosure.[63] In a context where there are transaction costs and uncertainty regarding the proper assignment of property rights on data (for example, who has the right in case of resale of data to third parties, or as it happens more often, who has the property right when data related to an individual are aggregated with other data), it is very probable that the market forces are not able to guarantee the achievement of an efficient economic balance. The possibility that the interests of those who hold wider technical knowledge and information about the data will prevail, materializes.

2.3. Discrimination strategies

The growth of the datasphere (see section 1.1.1) brings great opportunities for growth and development, not only of an economic nature, closely linked to the birth of new products and services or to the improvement of the existing ones, but also of a more social nature, where, for example, there are improvements in the field of medicine and in the management of public affairs. However, a growing number of analysis highlight the emergence of problems related to **forms of discrimination**, including the disparity in data access and the tools used to analyse them.

In addition to the search for new business opportunities, companies use *big data* in order not only to predict future consumption, but also to direct them towards a product rather than another. These processes are now made possible by the increasingly massive use of **algorithms to address the choices of individuals**.[64] This happens, for example, when we communicate with friends and family, when we choose a road journey, when we use search engines, when we get informed on social networks. In this last aspect, according to a recent study, the majority of social network users, especially among those not having an IT background, is not aware that news appearing to the consumer (so-called news feed) are filtered by an algorithm.[65]

[61] STIGLER G.J., (1980), *An introduction to privacy in economics and politics*, Journal of Legal Studies 9, n.4; POSNER R.A., (1981), *The economics of privacy*, American Economic Review 71, n. 2.
[62] HIRSHLEIFER J., (1980), *Privacy: its origin, function and future*, The Journal of Legal Studies 9, n. 4.
[63] VARIAN H.R., (1996), *Economic aspects of Personal Privacy*, in Privacy and self-regulation in the information age, National Telecommunications and Information Administration.
[64] For an illustration of the role of algorithms in the information world, see Box 5.1.
[65] HAMILTON K., SANDVIG C., KARAHALIOS K., ESLAMI M., (2014), *A Path to Understanding the Effects of Algorithm Awareness*, ACM, Toronto.

Among the possible **discriminatory practices**, those **related to the price** are among the most widespread; the price, in fact, is one of the most important variables through which companies try to achieve the maximization of their profits when the market configuration differs from the competitive one; that is when companies can exercise their market power. Price discrimination is the practice allowing company to offer the same good or service at different prices (price discrimination) depending on the willingness to pay of individual consumers (or reserve price), as identified by the operators through techniques of *big data analytics*.[66]

Price discrimination is a very significant phenomenon, especially following the diffusion of online purchases made by consumers. There are different ways in which companies manage to sell the same good or service at different prices; one of the classifications most commonly used in the economic field, and which better comply with problems of *big data*, is based on the level of information held by companies regarding tastes and preferences of consumers, since these information are strongly correlated with their willingness to pay.

Depending on the type of information available to the company, it is possible to classify price discrimination in three types: 1st, 2nd and 3rd degree. The 3rd degree implies that only some characteristics of the buyers are well observable and, therefore, these information may be used to prepare different rates according them; these are strategies often used when it is possible to set different prices based on easily observable characteristics such as age, gender and sometimes employment (think of discounted prices for younger or mature consumers, free entry according to gender, and discounted tickets for categories such as students). In order to implement 2nd degree price discrimination strategies, instead, companies use information regarding the quantity of good or service that consumers use; noting the heterogeneity in consumption models, in fact, companies are able to offer a range of offers with respect to which consumers self-select themselves, as typically occurs in contracts for the provision of telecommunications services (two-stage rates), in the "pay two and get three" offers, or in all those marketing strategies aimed at reaching customer loyalty (quantity discounts, loyalty cards, etc.). Finally, 1st grade discrimination, also called "perfect discrimination", occurs when the seller is able to apply to each customer a price corresponding to the estimate of the maximum price he is willing to pay; this is a case that in economics has been considered as merely theoretical, considering the high information necessary to companies in order to implement such a strategy. **The huge availability of individual data connected to the advent of big data is making it increasingly possible for online operators to implement perfect price discrimination strategies.**

In general, price discrimination strategies are considered as a phenomenon able of increasing, or in any case of not decreasing, social welfare and, therefore, in some ways have been considered desirable in the past. Through these practices, in fact, on one hand, social welfare can be increased because it is possible to sell goods that otherwise would have never been sold in the absence of the identification by the seller of consumers willing to pay a price higher than others; on the other hand, it is possible to ensure a wider participation in the exchange by involving those consumers who, by presenting a low willingness to pay, without discrimination would not participate in the exchange as they would have to bear a higher effective price than that maximum that they are willing to pay.

The practice of price discrimination, however, was, in some cases, challenged by relying on a question of fairness rather than on purely economic grounds; in an intuitive way, in fact, offering the same good at

[66] It is useful to remember that in a perfectly competitive market the law of the single price is valid; this is due to the fact that if it were possible to discriminate, arbitrage mechanisms would be put into motion such that secondary markets would be generated. In order to make it possible for companies to implement price discrimination strategies, therefore, it is necessary the presence of market power, i.e. non-competitive market forms, and the absence of secondary markets because consumers are not aware of the fact that the asset is sold at different prices, or because resale is prohibited by law, as it happens for electricity services. Furthermore, we can talk of price discrimination only if the difference in price for the same good cannot be attributed to a difference in production costs.

different prices causes damage to those who show a willingness to pay higher. Furthermore, if the discrimination is effectively implemented, it can allow the company to achieve the maximum profit available in the market, leaving consumers without any gains from the transactions.[67] Finally, it is useful to highlight also the existence of some criticality related to the costs incurred by companies to put in place price discrimination, since such strategies need the distribution of resources that otherwise would be allocated to other activities.[68]

The advent of the *big data* made the application of strategies by companies offering the same good or service at different prices increasingly frequent; the possibility of implementing 1st level discrimination, as mentioned above, derives mainly from the possibility of using digital information regarding an individual, which represents a good information base, to read into her / his willingness to pay. Moreover, **thanks to the identification of general patterns through big data analytics techniques, little individual information are necessary to foresee the economic and social behaviour of each individual** (for example, few likes on a social network are enough to predict, with a very high probability of success, sensitive information such as those related to political orientation, religious belief, ethnicity, sexual orientation, sentimental situation and even drugs addiction; see also Box 1). This makes the strategies for the differentiation of prices of goods and services sold online according to the characteristics of individuals increasingly practiced.

One of the first cases of price discrimination, based on the traces left on the net by individuals, concerned, in September 2000, a customer of *Amazon.com*, who bought a DVD for $ 24.49; the following week, he noticed that the same product had risen by $ 1.75 with a cost of $ 26.24. The same consumer found how the simple elimination of the cookies and of electronic tags from his computer, elements helping the tracking of activities carried out by individuals on the web, was enough to bring back the price of the DVD below the initial level, that is to 22.74 $.[69] This form of price discrimination was one of the first cases in which a company put in place a strategy of "dynamic prices" based on the measurement of the buyer's desire.[70]

Another case involved, in 2009, the *Microsoft* company and its *Bing Cashback* service: although it was designed as a service aimed at consumers to save money on their online transactions, it appears that some third-party vendors 'websites, using the URL of consumers, would have practiced price discrimination to the detriment of visitors redirected by *Bing*. Similarly, in 2012 the travel agency website *Orbitz Worldwide Inc.* found out that *Mac* computer users on average had a 30% higher spending propensity for a hotel overnight stay; consequently, the online agency decided to proceed first with a different view of the offers, and then set prices for *Mac* users higher than those proposed to users of "windows-based" devices.[71]

More generally, the advent of e-commerce (network transactions for the acquisition of goods and services), together with the increase in the capacity for aggregation and analysis of a huge amount of data - often unstructured and related to the consumption of individuals - , has led to the multiplication of databases containing information about their consumption preferences. Given the technological evolution, the companies want to use these databases to get as closer as possible to the purchasing

[67] In economic terms, this situation refers to cases in which the loss of well-being is reduced, caused by the presence of market power by the companies (such as the monopoly), and therefore generates additional wealth that, however, is redistributed exclusively to companies.
[68] LEESON PT, SOBEL RS, (2007), *Costly price discrimination*, Economics Letters 9.
[69] STREITFELD D., (2000), *On the Web, Price Tags* Blur, The Washington Post.
[70] KRUGMAN P., (2000), *Reckonings; what price Fairness?,* New York Times.
[71] MATTIOLI D., (2012), *On Orbitz, Mac Users Steered to Pricier Hotels*, The Wall Street Journal., WHITE M.C., (2012), *Orbitz Shows Higher Prices to Mac Users*, Time.

inclinations of individuals and propose them customized offers, also to guide / direct consumption towards a product rather than another, always looking for the maximum profit.

Discriminatory practices can be the result of unconsciousness, but also of a specific choice. In this regard, a good example of this ambiguity is provided by the application of *big data* to the context of the labour market. In the labour market, it was thought that with the advent of the *big data* it was possible to minimize subjectivity in the recruitment of personnel, an activity bringing a certain level of discrimination; basing the recruitment choices of personnel on an algorithm using a very large set of information on individuals, it would have been possible to overthrow the bias linked to subjectivity in the choice. Furthermore, the use of data on the behaviour of job seekers from sources such as websites could also be useful to reduce a type of prejudice in recruitment practices called "social network segregation" deriving from the fact that a significant part of the jobs, in particular from the diffusion of social networks, takes place through word of mouth within a connections network. The chain effect, in an intuitive way, is that in organizations where minorities are already poorly represented, the minorities themselves will have less and less access to jobs and, therefore, will be increasingly discriminated.

However, these discriminatory phenomena do not seem to have loosened even with the use of *big data* predictive techniques. If the algorithms that are used to make decisions about the personnel to be hired are not optimally designed, there is a strong risk that such discrimination will not only continue over time, but will even be strengthened.[72]

In conclusion, access to increasingly large information sets makes it possible to implement ever more stringent price discrimination strategies, up to those of perfect discrimination. It is worth highlighting how, by means of the identification of specific behaviour patterns through large amounts of data, little information are needed to foresee, with a high degree of success, certain economic and social behaviours and specific individual characteristics, even starting from anonymised data. Price discrimination strategies have a certain effect of social redistribution in favour of online providers. Furthermore, such practices, even when efficient, **present very significant social risks**. It is indeed very easy to understand how such algorithmic strategies can automatically extend, even involuntarily, to differences in population based on ethnicity, race, sexual orientation, state of health, etc.

BOX 1 – PSYCHOMETRICS: THE PROFILING PROCESS

In the sixties, two psychologists, Ernest C. Tupes and Raymond Christal, advanced a theory according to which the personality of individuals could be described through the identification of some essential traits[1]. The initial model was later improved until, in the early nineties and thanks to the studies carried out by Mccrae & Costa[2] and John[3], it led to the identification of five essential personality qualitative factors, known as the "big five". Each of them was associated to and was counterbalanced by a specular trait: extroversion corresponded to introversion, pleasantness to unpleasantness, conscientiousness to negligence, neurosis to emotional stability, mental openness to mental closure. These five personality components can also be referred to as OCEAN (**Figure 1.1**): openness to experience, conscientiousness, extraversion, agreeableness, and neuroticism.

[72] MCILVAINE A.R., (2014), *The Power (and Peril) of Predictive Analytics*, Human Resource Executive Online.

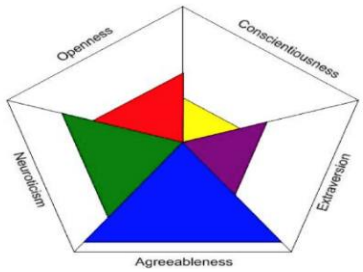


Figure 1.1 – The OCEAN model

Source: Jennifer Golbeck et al., "Predicting Personality from Twitter," in 2011 IEEE Third Int'l Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third Int'l Conference on Social Computing (IEEE, 2011)

Over the years, research has shown that the "big five" model, based on psycholinguistic measurement of personality,[4] is still valid: various tests, both in psycholinguistics and psychometrics, have not been able to question its validity.[5] The "big five" are therefore the basis of today's psychometrics: research shows how, for example, personality traits can be deduced from the choices that Facebook users make when they allow some people to enter their network of contacts (friendships).[6]

Through the use of machine learning and computational techniques from the Big Data world, it was Michael Kosinski, creator of the *Mypersonality app* - now at the centre of the Facebook and Cambridge Analytica scandal[7] - who argued that "*personality assessments are more accurate when made on a computer than when made by an individual*".[8] In particular, in his studies, the author compared the accuracy of the opinions expressed about the personality of individuals with the opinions created by machines. Recent results of this group of studies show that a few dozen "likes" are enough to identify, with a probability of 85%, the political orientation of a subject, and, with a probability of 82% (distinguishing between Christians and Muslims), their religious belief. Gender is correctly predicted in 93% of cases **(Figure 1.2)**.

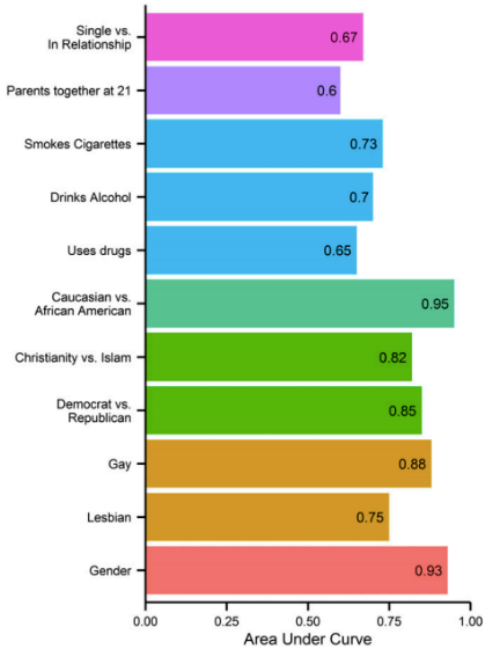


Figure 1.2 – Model predictions

Source: Michal Kosinski, David Stillwell, and Thore Graepel, "Private Traits and Attributes Are Predictable from Digital Records of Human Behavior"

Bibliografia

[1] Ernest C. Tupes and Raymond E. Christal, "Recurrent Personality Factors Based on Trait Ratings," Journal of Personality 60, no. 2 (June 1, 1992): 225–51, doi:10.1111/j.1467-6494.1992.tb00973.x.

[2] Robert R. McCrae and Oliver P. John, "An Introduction to the Five-Factor Model and Its Applications," Journal of Personality 60, no. 2 (June 1, 1992): 175–215, doi:10.1111/j.1467-6494.1992.tb00970.x. R. R. McCrae and Paul T. Costa, Personality in Adulthood: Emerging Lives, Enduring Dispositions (New York: Guilford, 1990).

[3] OP John, E Donahue, and R Kentle, "Big Five". Factor Taxonomy: Dimensions of Personality in the Natural Language and in Questionnaires," Handbook of Personality: Theory and Research, 1990, 66–100, http://thenetworktufh.org/wp-content/uploads/2015/10/Newsletter2004-02_0.pdf#page=25.

[4] Boele De Raad, "The Big Five Personality Factors: The Psycholexical Approach to Personality," Hogrefe & Huber Publishers, 2000, http://psycnet.apa.org/record/2001-17509-000.

[5] Jennifer Golbeck et al., "Predicting Personality from Twitter," in 2011 IEEE Third Int'l Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third Int'l Conference on Social Computing (IEEE, 2011), 149–56, doi:10.1109/PASSAT/SocialCom.2011.33.

[6] Maarten Selfhout et al., "Emerging Late Adolescent Friendship Networks and Big Five Personality Traits: A Social Network Approach," Journal of Personality 78, no. 2 (April 1, 2010): 509–38, doi:10.1111/j.1467-6494.2010.00625.x.

[7] Carole Cadwalladr and Emma Graham-Harrison, "Revealed: 50 Million *Facebook* Profiles Harvested for Cambridge Analytica in Major Data Breach," The Guardian, 2018, https://www.nytimes.com/2018/03/17/us/politics/cambridge-analytica-trump-campaign.html Matthew Rosenberg, Nicholas Confessore, and Carole Cadwalladr, "How Trump Consultants Exploited the *Facebook* Data of Millions," The New York Times, 2018.

[8] Wu Youyou, Michal Kosinski, and David Stillwell, "Computer-Based Personality Judgments Are More Accurate than Those Made by Humans.," Proceedings of the National Academy of Sciences of the United States of America 112, no. 4 (January 27, 2015): 1036–40, doi:10.1073/pnas.1418680112.

2.4. *The APP market*

One of the main trends that is making digital data management issues increasingly relevant is the rapid growth in **the use of online contents and services via mobile devices**.[73] These dynamics in consumption are made possible by the pervasive use of handsets that allow web browsing, especially smartphones. On a global level, as shown in **Figure 2.1**, for the first time in October 2016, the rate of internet access from a mobile device (smartphones and tablets above all) exceeded desktop internet access (fixed and portable PCs), thanks to the massive use of mobile devices recorded in Asian countries where, actually, access from a mobile device exceeded access from a fixed location as early as May 2014.

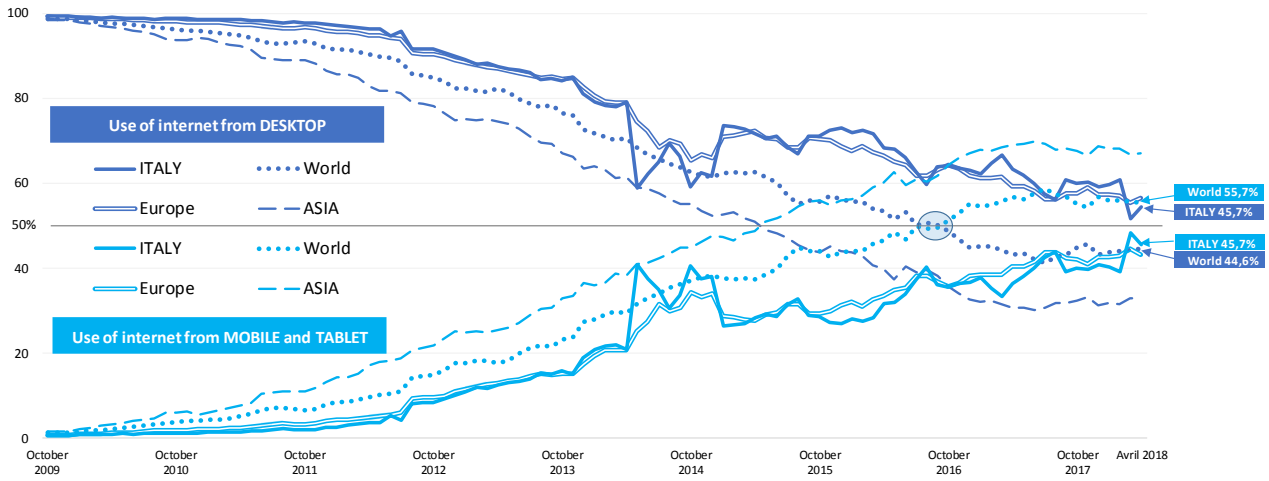


Figure 2.1 – Worldwide internet usage by device type (October 2009 – April 2018)
Source: AGCOM elaboration on monthly data from *StatCounter.com*

The implications of this trend are manifold and concern various areas; with reference to what is "the digital trail" (online footprint see section 2.2), of which individuals leave a trace as a result of their activities on the web, undoubtedly **the use of mobile devices has strongly contributed to the growth of the datasphere,** since mobile devices, much more than fixed devices, are used exclusively by the owner/holder. Generally speaking, in fact, mobile devices belong to a single individual (they are therefore personal), while fixed network workstations are more likely to be shared by a plurality of individuals (e.g. the family). In addition, **mobile devices are required to generate data that can be associated with the location of individual users**, the latter being increasingly important for the provision of targeted services for users (see section 1.4).

Through mobile devices, individuals can use communication services "anywhere and anytime", a condition that has partly replaced internet access from a fixed location, that appears to be subject to greater constraints. In addition, fixed networks increasingly connect through mobile devices that work wirelessly (often via Wi-Fi) in order to use services, which, using wired connectivity, need increasing bandwidths (audio-video streaming, videoconferencing, etc.).

The growth of mobile connection to the internet represents the structural basis for the development and subsequent success of APPs,[74] i.e. those programs, or packages of programs -

[73] According to data collected by the Communication market monitoring system, through which, among other things, AGCOM monitors telecommunications markets, during 2017 data traffic from mobile phones increased by 56% and, in the same period, unit consumption rose from 1.84 to 2.76 Giga/month per user, with a growth of 49.5% (data as of December 2017). The number of SIM cards with internet access was 52.2 million, equal to 63.9% of the entire customer base, almost double compared to 2012, when this type of SIM cards represented 27.8% of the total SIM cards.

[74] Compared to the APPs available on a PC-Desktop, these programs are characterized by greater compactness and ease of use that are well suited to the limited hardware resources of mobile devices. APPs must therefore be able to address and overcome specific problems related to the nature of the handset, such as *i)* the limited resources (memory, CPU), *ii)* the lack

application software - that each user installs, or finds already installed, on his/her handset to carry out specific activities (e.g. word processing programs, photo editing programs, games, etc.). In this respect, the term **"APP economy"** was coined to cover both the set of activities that includes the design, production and distribution of mobile applications, and the different players involved in the market.[75] Basically, APPs are designed to make the use of mobile devices easier and more enjoyable by allowing individual users to access content and services easily, anytime, anywhere.

In a few years, the world of APPs has shown exponential growth; so much so that, falling within the category of general purpose technologies, it is difficult to precisely define their scope and the extent of their effects on the entire economic and social system.[76] This phenomenon is quite recent, since the first APP store, created by the company *Apple*, was launched in 2008, a year after the release of the first *iPhone* model. Only three years after the creation of the first APP store, about one million applications were already available to users, distributed over four virtual stores. The turnover generated varies depending on the type of studies and on the specific element on which they focus; according to calculations by the analysis *App Annie*,[77] the world of APPs (including pay applications, advertising and mobile purchases) was worth more than $ 1,300 billion in 2016, while for 2021 the expected turnover will reach over $ 6,300 billion, an increase of 385%. This forecast is mainly driven by two forces: the spread of mobile devices among the population but, above all, the growth of the average time spent using APPs, with an average of 3 hours per day per user at the end of 2017.[78]

Some APPs are already installed (built-in) in the mobile device, since it is difficult to make the device work with only the operating system; among these, it is should be underlined that other APPs are available on APP stores, through which users have access to the virtual store where they can download applications. It is difficult to determine which of these pre-installed APPs are actually necessary for the operation of the mobile device, as much depends on the needs of individual users.[79] In some cases, these are unwanted APPs, but it is very difficult to disable them from the device because they are pre-set as system applications.[80] Decisions on the pre-installation of APPs are made by handset manufacturers and/or operating system providers.

Most applications are available on virtual shops called APP stores: these are distribution platforms through which APPs are made available to the public. Once in the virtual store, unlike pre-installed APPs, users may choose which application to download and it is therefore possible, despite the restrictive

---

of external power, *iii)* the different data transfer protocols for internet access (WiFi, GPRS, etc..) and *iv)* the small size of the display, which make the development of operating systems much more critical.

[75] In the Prospectus relating to the listing of shares on NASDAQ in 2012, the company ZYNGA, specialized in the production of video games, included the emergence of the APP economy among the opportunities of growth of its business, defining it as follows: "*Emergence of the App Economy. In order to provide users with a wider range of engaging experiences, social networks and mobile operating systems have opened their platforms to developers, transforming the creation, distribution and consumption of digital content. We refer to this as the "App Economy." In the App Economy, developers can create applications accessing unique features of the platforms, distribute applications digitally to a broad audience and regularly update existing applications.*" See https://www.sec.gov/Archives/edgar/data/1439404/000119312511180285/ds1.htm, page 64.

[76] T. BRESNAHAN, J.P. DAVIS, PAI-LING YIN, (2014): *Economic Value Creation in Mobile Applications* in: The Changing Frontier: Rethinking Science and Innovation Policy, pp 233-286, National Bureau of Economic Research, Inc.

[77] Source: https://www.appannie.com/en/insights/market-data/app-economy-forecast-6-trillion-market-making/

[78] Source: https://www.appannie.com/en/insights/market-data/apps-used-2017/

[79] For example, the APP *Watch APP* is pre-installed on *Apple* devices, but it can be used only with the specific watch produced by Apple (*iWatch).*

[80] With the *iOS* 10 operating system, for example, *Apple* allows you to delete 23 pre-installed APPs; in fact, this is not a real deletion, as it will not free memory, although *Apple* points out that these are applications that altogether occupy only 150 Mb of memory. Users who would like to have an app available on their device after uninstalling it should simply access *Apple's* virtual store and download the APP again. https://support.apple.com/en-gb/HT204221

policies adopted by the platforms, to install products that could cause vulnerability problems to the operating systems with consequent damages to the device.[81]

For reasons related to vertical integration, virtual shops may present incompatibilities with operating systems: for example, the *Google APP* store works on *Android* but not on *iOS* terminals and, conversely, the *Apple store* operates on its own operating system but not on *Android*. This is significantly important, as it is likely to have an impact on the market and its competitive structure: in particular, the market position of the main vertically integrated players (in particular *Google* and *Apple*) is effectively strengthened. Moreover, the position held by the latter in both the operating systems and the APP stores constitutes, in fact, a situation of great advantage in the big data market, especially in the data collection phase.

The number of applications available in virtual shops is one of the factors that determine the choice of users relating to the type of device, and connected operating system, to buy, triggering a **feedback effect** (so-called direct and indirect network externalities) where the more companies develop applications for a specific operating system and its virtual store, the greater will be the number of users using such APP store, and this in turn will lead to greater use of the store by developers, and so on.

The type and number of APPs available in a store have grown exponentially since 2008: in March 2017, 2.2 million applications were available on Apple's APP store and 2.8 million on Google Play, while far fewer APPs are present in other virtual shops. It should be remembered that these "figures" are quickly exceeded, due to the fact that every day APP stores include new applications.[82]

As regards the players involved, both on the demand side (end users) and on the supply side (APP developers), for the purposes of this Report, an analysis of market shares in terms of volumes developed by stores seems to be extremely interesting. In terms of revenues, the *Apple* store generates a turnover higher than *Google Play*;[83] however, in terms of volumes, the situation is precisely the other way around (**Figure 2.2**): being part of an open system (see below), the *Google* store is leader in terms of the number of APPs available and uploaded by developers as well as of downloads made by end users.

The APP store segment, therefore, appears to be characterized by a high level of concentration, with two operators (*Google* and *Apple*) in a market-leading position, which is also due to the vertical integration and the share held in the related mobile operating systems market.

[81] Currently, numerous APP virtual stores are available on various platforms including Microsoft, Google, Apple and Amazon. It is interesting to note that, over time, these virtual shops have considerably expanded the range of products, selling not only APPs but also content (films, videos, books, etc.).

[82] According to a study by pocketgamer.biz, an average of 1,000 APPs are submitted to the Apple store in one day, 90% of which are approved by Apple in 7 days. http://www.pocketgamer.biz/metrics/app-store/.

[83] According to the analyses developed by the platform *Statista.com*, and in view of the minor importance of alternative stores to Apple store and Google Play, the former holds about 60% of the joint profits, the latter 40% and, more significantly, this division is expected to continue in the coming years. https://www.statista.com/statistics/259510/revenue-distribution-between-the-apple-app-store-and-google-play/

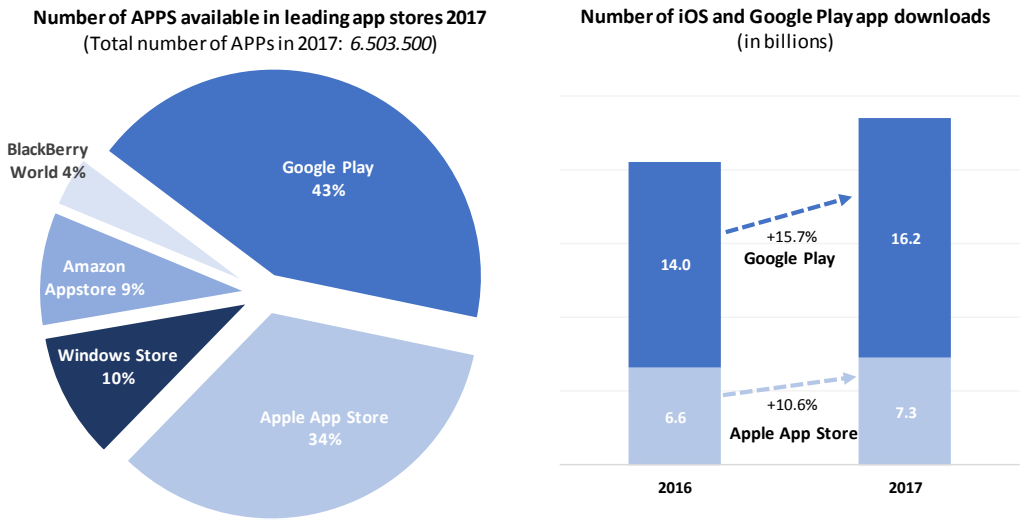


Figure 2.2 – Market shares in volume (2017)
Source: AGCOM elaboration on data from *Statista.com*

Again in terms of structural characteristics, it must be considered that this is a market in which intense **network externalities** operate; for the individual user, the benefit deriving from the purchase and use of a good or service increases with the increase in the number of other users who use the same good or service (direct externalities), or compatible goods and services (indirect externalities). The greater benefit for the individual user (utility) increases their willingness to pay, which has a considerable impact on the functioning of the market, leading to a greater degree of concentration.

Moreover, from the point of view of the economic theory, the APP sector can be classified as a two-sided or multi-sided market, given that there are more than two players involved.[84] As a result, cross-side network externalities also emerge when decisions made by individuals on one side of the market (users) affect agents on the other side (developers) and vice versa; in the case of APPs, both sides of the market benefit from the number of downloads made through the APP store. The interest of users towards a specific platform, in fact, increases as the number of developers who operate there increases and, conversely, developers will be led to produce APPs for a store as more users access it. The intensity of the cross-side elasticity becomes a relevant factor in strategic choices related to the price of the APPs set by the companies. In many cases, in fact, consumption on one side of the market is supported on the other side, by setting low prices, often even below the cost of production. For instance, a significant number of APPs are available to users for free, even though they have a no null production cost.[85]

Network externalities emerge mainly through the achievement of what is defined as a "critical mass of users": once this quantity is reached, in fact, the subsequent growth of the network follows a self-fueling process ("snowball effect" or "bandwagon effect"). The achievement of critical mass may partly explain why companies, strategically, in a first phase can consider it useful to offer their product at

[84] AGCOM has extensively dealt with the subject of two-sided markets in previous surveys, among which it is worth mentioning the *Survey on Advertising Collection*, Chapter 1 (Annex A to Resolution no. 551/12/CONS) and the *Survey on the Internet Services sector and Online Advertising*, Chapter 1 (Annex A to Resolution no. 19/14/CONS) and lastly in the *Survey on the development of digital platforms and electronic communication services,* Part II - Digital Platforms. For a literature review relating to two-sided markets, see M. RYSMAN, (2009): *The Economics of Two-Sided Markets*, Journal of Economic Perspectives 23. In the case of APPs, a store acts as an intermediary between application developers and users who use the services.

[85] In general, when the intensity of network effects are not different in both sides, then the possibility of subsidising one side of the market is not entirely possible. On the contrary, as asymmetries intensity increase, it may be useful to subsidise access to that side of the market that generates most value, leading to a situation where consumer welfare increases. However, the extent of these effects depends on the level of competition in which a platform operates, since competition imposes constraints which do not allow the company to set prices in a monopolistic market configuration that is optimal for it. Competition would lead to cost-oriented prices on both sides of the market and would not allow for subsidy practices.

a low price or, as often happens, free of charge; in fact, low prices attract users more quickly and therefore allow to reach the critical threshold earlier, compared to a high price policy that in many cases could discourage the purchase of goods or services by consumers.

Economic literature points out that the presence of network externalities determines some important effects in terms of market balance; a first effect, of a general nature, is that in markets with network externalities there is a limited number of "technological standards" (with, often, a single standard that operates in a monopolistic regime), be they operating systems, APP stores, or single categories of APPs (such as social networking APPs).[86] A further important effect is linked to the fact that it is not always the best available technology that prevails on the market; when the critical mass threshold, and therefore the starting point of the "snowball effect", is reached, the boost given by initial consumers to the diffusion of the specific type of technology becomes crucial and, consequently, the technology that will prevail is strongly influenced by the initial choices of a certain number of consumers (lock-in effect).

All these aspects related to two (or more) sided markets can lead to the formation of monopolistic markets or oligopolies ("the winner takes all"), where the advantage of the first players that overcome a critical mass of users (not necessarily the first operator to enter the market) becomes increasingly relevant over time, offsetting the initial low market entry costs and raising barriers to entry. In this sense, the APPs developed by technological giants have evolved, increasingly representing real platforms through which they deliver content to users (content consumption platform). The Facebook APP, in this sense, is an emblematic case (see Chapter 3); thanks to the enormous quantity and variety of data released by users and collected and stored at an ever-increasing speed, *Facebook* has introduced, over time, numerous innovations in the use of content, in particular video content, and access to other services (such as, for example, APPs concerning games) so as to make its APP increasingly similar to a real distribution platform.

In addition, given the increasing time that users are devoting to the use of mobile devices, in particular online services, the APP sector plays a crucial role in acquiring advertising resources, as shown in **Figure 2.3**. Due to the characteristics of the market described above (presence of economies and network externalities), this increase has mainly benefited technological giants such as *Google* and *Facebook*.[87]

[86] The effect of network externalities also depends on how problems linked to consumer choice and business. As far as the demand side is concerned, it should be stressed that the consumption of goods and services subject to network externalities depends on the size of the network itself and above all on the expectations of users as to the size that the network can reach; this consideration makes it possible to state that it will be the users themselves, in particular the initial users, who will direct the technology that will be most widespread among other users (interdependence in consumer choices) bearing in mind the need to overcome problems of coordination in choices resulting from the inertia of consumers to adopt new technologies or, on the contrary, their excessive mobility that leads them to "try" different technologies. The way in which these coordination problems are resolved determines, among the many possible balances, what will actually happen. On the supply side, however, issues concern the way in which the predominant technology will be chosen or promoted; for example, in the presence of network externalities, in many cases it is governments that promote certain standards. Where choices about standards are left to the free play of the market, companies may decide not to make their products compatible, thereby reducing the network itself, or to make their technology compatible. Lack of compatibility, however, may be a deliberate choice to trigger lock-in (or "consumer capture") processes such that users' transaction costs for switching technology become so high that it does not make it easy for them to move.

[87] According to data provided by *eMarketer*, in the US online advertising market, the joint market share of *Google* and *Facebook* in 2018 was 56.7% (37.2% Google and 19.6% Facebook). According to the forecasts, in the coming years a slight reduction in *Google's* market share to the benefit of another technological giant, *Amazon*, is expected.
https://www.recode.net/2018/3/19/17139184/google-facebooks-share-digital-advertising-ad-market-could-decline-amazon-snapchat

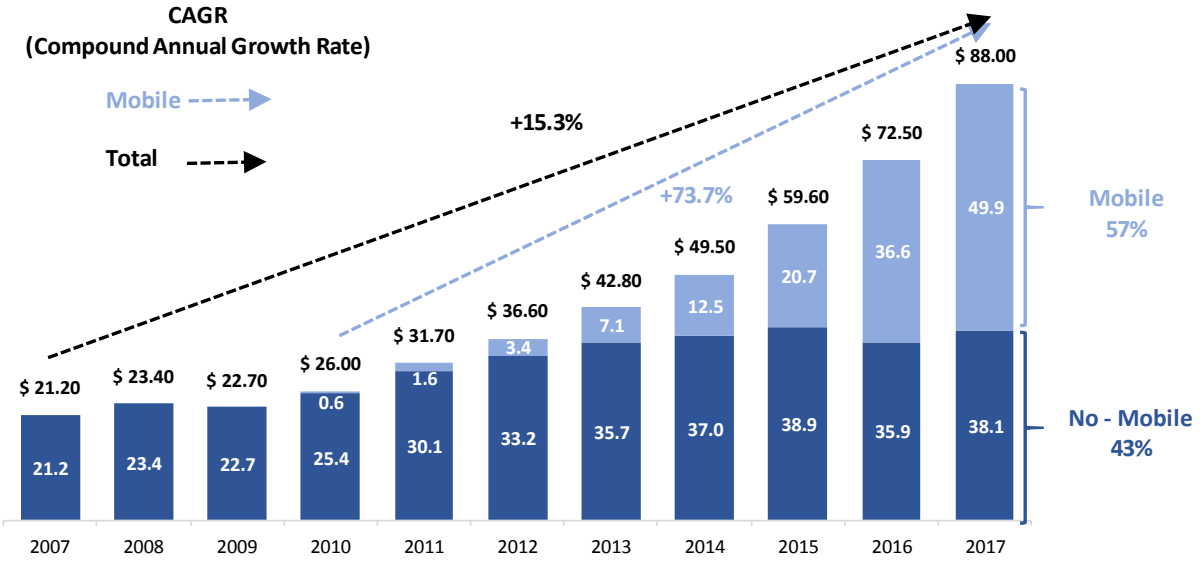


Figure 2.3 – Online advertising revenues worldwide (2007 - 2017)
Source: AGCOM elaboration on data from IAB

The "APP world" does not seem easy to define; in addition to aspects closely linked to the communications sector, in fact, it is necessary to consider the influences exerted by many other industrial sectors, including entertainment, device, hardware and software manufacturers. The result is a very complex system in which diversity appears to be a relevant element; depending on the positioning along the value chain, in fact, those who operate in the APP system can adopt a different business model, just as they can meet a different competitive level, to the point that some companies are at the same time competing and collaborating with each other.

In the APP sector, a **primary role is played by stores, which represent platforms that meet the demand of users of mobile internet services provided through applications offered by developers**. Stores are part of a broader system of integrated markets (linked together through complementary relationships), ranging from operating systems to mobile device manufacturers and telecommunications service providers. This set of services, products and markets forms a so-called mobile platform.

APP stores are virtual stores where users can find applications for the specific operating system installed on their device. As mentioned above, the first successful APP store was the *Apple store* created in 2008. Although there were some previous attempts to create virtual stores for software exchange, it is thanks to the idea and service offered by *Apple* that the market has developed.[88] At that time, in fact, the first *iPhone* was launched, a tool that has profoundly changed internet consumption habits, triggering a developing process that has pushed more and more towards mobile connectivity.

The Apple store was followed by the *Google* APP store, which has distinctive features. In fact, *Apple* has adopted a closed model, with a greater profit margin, while i is the virtual store from which the greatest number of APPs have been downloaded and where the greatest number of applications are uploaded and made available to users (leadership in volumes; see **Figure 2.2**).

Consumers' assessment of a mobile platform highly depends on what they find on the "shelves" of virtual shops, in particular, the number of downloadable APPs and their variety (categories). At the same time, the other side of the market, the developer side, will be attracted to stores where it is possible to reach as many potential customers as possible. This triggers a phenomenon of positive

[88] J. WEST e M. MACE, (2010), *Browsing as the killer app: Explaining the rapid success of Apple iPhone*, Telecommunications Policy, 34 (5-6).

feedback between the two sides of the market, which leads, as explained above, to a high level of concentration.

Figure 2.4 shows the rapid growth of the number of APPs downloaded by users worldwide on all existing APP stores: in just 8 years, the number of downloads has increased by about 10,500% compared to 2009, while last year there was a 19.5% growth.

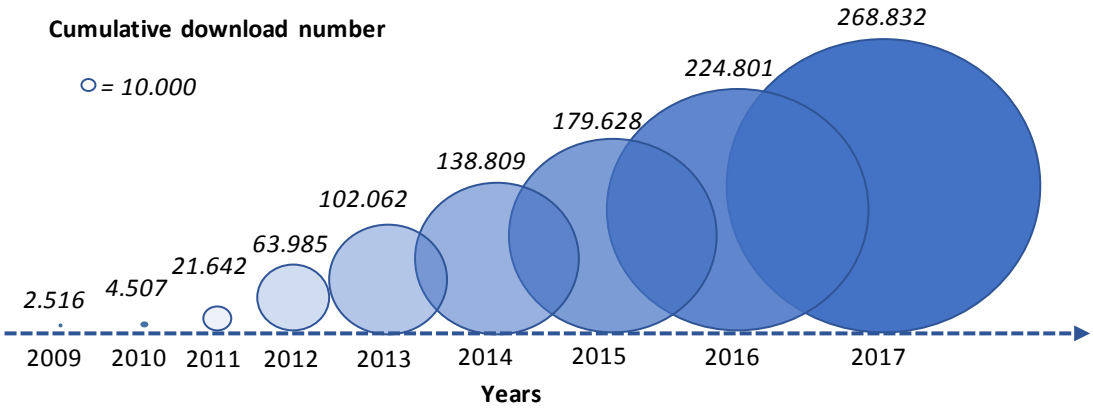


Figure 2.4 – Number of mobile apps downloaded worldwide from 2009 to 2017 (millions)
Source: AGCOM elaboration on annual data from *Statista.com*

According to the *AppAnnie* website, specialised in the field, the average smartphone user has 80 apps on their device (including the numerous pre-installed apps) and uses an average of 40 apps per month.[89]

With the spread of smartphones and the increase in the number of downloads, stores also need to adapt their offer, not only in terms of numbers but also in terms of the variety of products and services offered. Moreover, when two-sided markets, such as the APP market, are successful, a significant number of transactions are generated, so that the player acting as an intermediary may find it useful to apply a handling fee on each transaction. In this regard, it should be noted that the two main stores provide, among the provisions that developers must accept in order to use the services of the platform, a similar fee of 30% on each transaction carried out.[90]

Since APPs are designed for specific operating systems and since there are two most popular operating systems in the world, it is not surprising that the two largest stores are owned by the two most popular operating systems (vertical integration: see also above): *iOS* (Apple's App Store) and *Android* (Google Play).

The *Google Play* APP store is by far the largest in terms of the number of available APPs: in 2017, more than 2.8 million applications were available, about 94% allow free downloads and forms of in-APP purchase (i.e. additional sales after downloading), while about 70% are totally free. This generated total revenues equal to $ 9.8 billion.[91]

At the end of 2017, more than 2.2 million APPs were available on the *iOS* operating system shop, while in 2013 only 300.000 APPs were available. Approximately 1 million APPs are defined as "native" in the sense that they are developed to work specifically on devices produced by Apple. Turnover at the end of 2017 generated profits equal to 11.4 billion dollars (about 5% of *Apple's* total turnover compared to 1.5%

[89] https://www.appannie.com/en/insights/market-data/app-annie-2017-retrospective/
[90] Per *Google* Play si veda https://support.google.com/googleplay/android-developer/answer/112622?hl=it, che prevede: *"Per le applicazioni e i prodotti in-app che vendi su Google Play, la commissione sulle transazioni è pari al 30% del prezzo. Ricevi il 70% del pagamento. Il restante 30% è destinato al partner di distribuzione e alle commissioni dell'operazione.* Per l'*APP store* di Apple si veda https://developer.apple.com/programs/whats-included/ , dove si indica testualmente: "*You get 70% of sales revenue. 85% for qualifying subscriptions*"
[91] Source: *statista.com* https://www.statista.com/statistics/444476/google-play-annual-revenue/

in 2016); in just 7 days, from Christmas Eve to the beginning of 2018, an estimated 890 million dollars were spent in the store.[92]

Despite such a considerable number of applications, when looking at the most downloaded ones, it is evident how the market is concentrated in the hands of a few companies, which correspond to the most important companies operating in the digital economy, such as *Google* and *Facebook* (**Figure 2.5**). This again highlights a high degree of integration of players throughout the online services chain

APP	Impresa	Data di rilascio
Facebook	Facebook	dic-11
WhatsAPP Messenger	Facebook	gen-12
Facebook Messenger	Facebbok	nov-11
Instagram	Facebook	mar-12
Clean Master	Cheetah Mobile	set-12
Skype	Microsoft	nov-11
Line	Line	gen-12
Viber	Rakuten	ott-11
Twitter	Twitter	nov-11
Flashlight by Surpax	Surpax	gen-12

APP	Impresa	Data di rilascio
Facebook	Facebook	lug-08
Facebook Messenger	Facebook	ago-11
You Tube	Google	set-12
Instagram	Facebook	ott-10
Skype	Microsoft	mar-09
WhatsApp Messanger	Facebook	mag-09
Find My iPhone	Apple	giu-10
Google Maps	Google	dic-12
Twitter	Twitter	ott-09
iTunes	Apple	gen-12

Figure 2.5 – Top 10 APPs for download in the two main stores (2017)
Source: AGCOM elaboration on data from *AppAnnie.com* and *AppBrain.com*

In this regard, the effects deriving from the vertical integration with the operating systems are not evident from the data produced in **Figure 2.5** since, as said, the majority of the APPs developed by the integrated operators (*Google* and *Apple*) are pre-installed on the devices. According to statistics produced by *appbrain.com*, in fact, the *Google Play services* APP, pre-installed on devices with *Android* operating system and allowing access to the virtual store, seems to have been installed more than 5 billion times since its appearance until today.

That said, APPs with a high diffusion created by "independent" developers, i.e. not integrated in any other stage of the value chain, are very rare (in the case of the *Apple store*, *Twitter* is the only one among the leading ten APPs; right side of **Figure 2.5**). This highlights **the importance of integration at the various stages of the supply chain in terms of the position acquired by operators in each of them. Evidently, this also affects data acquisition processes and competitive structures in these areas.**

With respect to this, it is useful to remember that about 94% of apps are for free, i.e. they are made available to users free of charge.[93] In many cases, it is a monetization model that may be defined *freemium*; that is, the combination of the word "free" (free of charge) with the word "premium" (surcharge). This implies the free transfer of the so-called "core" of the product and involves the sale of additional products (called premium products). This pricing policy is not only aimed at exploiting the positive feedback effect deriving from network externalities (see above), but is also **aimed at acquiring**

[92] Source: *forbes.com* https://www.forbes.com/sites/chuckjones/2018/01/06/apples-app-store-generated-over-11-billion-in-revenue-for-the-company-last-year/#7ceb836f6613
[93] As of 19 February 2017, a total of 2.734.073 APPs were available on *GooglePlay*, of which 2.520.462 were free, equal to 92%, and 213.611 were paid APPs, equal to 8% (source: Appbrain.com).

the greatest number of users' data, which are then monetized by the operators, directly and/or indirectly (i.e. by selling them to third parties), through various uses.

Figure 2.6 shows the apps distribution in the two main virtual shops (*Apple Store* and *Google Play*) by category type. While there is no perfect correspondence between the categories, as each platform classifies APPs according to its own criteria, a very similar distribution emerges. In fact, it should be remembered that many APPs (especially the most popular ones) are developed in the two versions in order to be sold in both virtual shops, thus reaching a wider audience of potential users.

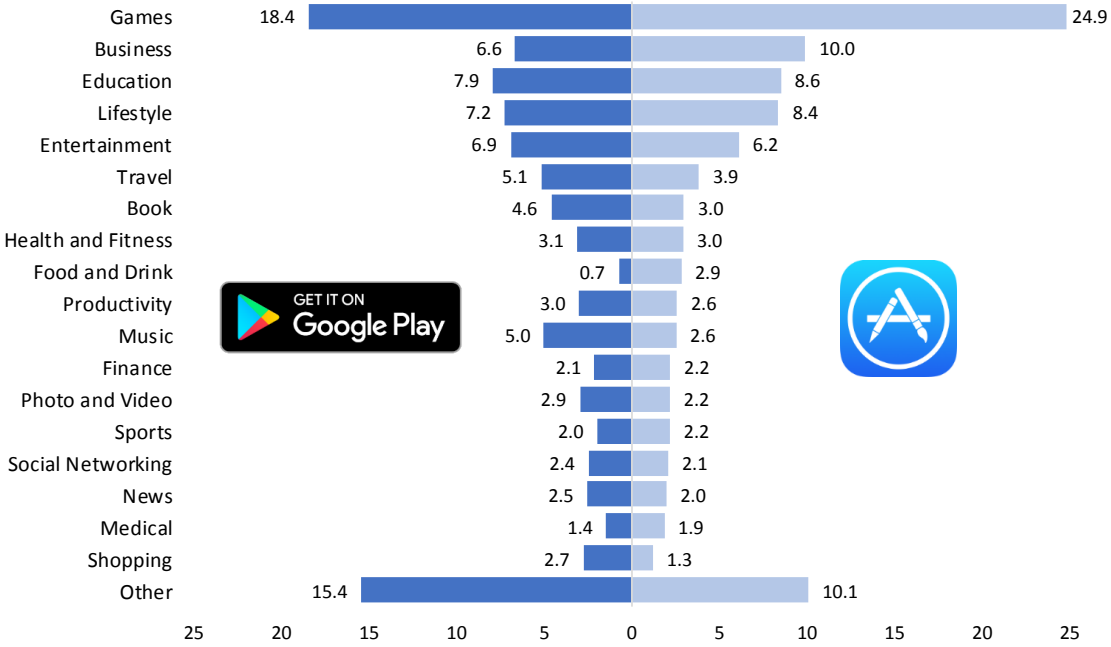


Figure 2.6 – APPs by most popular categories in 2017 (%)
Source: AGCOM elaboration on data from *Statista.com* and *AppBrain.com*

First of all, the important role of the applications developed for games emerges; this category is the one that, in addition to involving the user for a longer time, produces the greatest flow of direct revenues (i.e. revenues generated directly through the sale of the service to end users). For example, according to *Statista.com*, in 2017 the games *Arena of Valor* and *Fantasy Westward Journey* generated more than $ 3.4 million in revenues.[94]

Secondly, it is interesting to notice that the lifestyle and entertainment categories are those that reach the majority of users, being by their nature transversal APPs: about 78% of consumers who own an iOS operating system have downloaded at least one APP belonging to one of these two categories.

As the owners of the APP stores play a key role in the intermediary process between developers and users, they are able to guide consumers' choices, orienting them in various ways towards different applications. They, in fact, decide the conditions for acceptance, publication and dissemination of the APPs in the store.

2.5. A market solution to data transactions: permissions

The size of the market described above involves a huge and growing flow of data whose origin is the individual; the information collected is used to improve the products offered and make the consumer experience more enjoyable and, at the same time, to create databases from which it is possible

[94] Source: https://www.statista.com/statistics/505625/leading-mobile-games-by-global-revenue/

to extract values in many ways (so-called primary and secondary uses). This data can be collected directly through a user interface, such as an APP, but also indirectly, i.e. without any specific activity on the part of the user, through telephone numbers that, given the personal character of the device, represent real unique identifiers (UDID - Unique Device Identifier).

All information available on a mobile device and related to all the activities carried out by the user, such as data storage, contact list, videos, photos, messages, e-mail and the various passwords used to access specific services, such as financial ones, can be potentially collected. Increasing importance is given to information concerning the location of the user, thanks to which users may be reached anywhere and provided with appropriate responses to their needs, such as finding places for lunch or for overnight stays.

Technological development has allowed the increasing diffusion of "datization" processes, i.e. the transformation of any type of content (films, books, voice messages, body movements, etc.) into digital format. By way of example, it is possible to say that Facebook, LinkedIn and Twitter have transformed, respectively, social relations, working relations and opinions into data; the same may be said for APPs that offer health services.

The degree of users' awareness on the transfer of their information is a central aspect: according to a study conducted by the European Union in 2015 (Special Eurobarometer 431 - Data Protection), 69% of Italian users consider the transfer of individual information a natural consequence of current lifestyles, and 54% connect it with access to digital services. At the same time, 52% point out that providing individual data entails a certain degree of risk.

Among the most perceived risks, there are the use by internet companies, directly or through the transfer to third parties, of their information without their consent and the possibility of being victims of fraud, identity theft and receiving unsolicited advertising. As a result, citizens expect rules that allow better protection of private information; in particular, when incidents of information misuse (breach) occur, 51% of users hope that the company should no longer be allowed to use individual data and 39% claim compensation for damages.

On the other hand, AGCOM holds that, where possible, the analysis of issues related to the world of big data should be carried out more profitably through methodologies consistent with the system under observation. Therefore, the subsequent study (see in particular section 2.7) was conducted on **millions of observations regarding network behaviour actually adopted by users** (and not through surveys carried out, by means of answers to questionnaires, on limited samples of citizens).

In particular, the analysis focuses on virtual shops. In fact, APP stores are an important example of a service/method through which digital data is exchanged through online platforms. Users are informed about how the data will be processed prior to downloading an APP through an End-User License Agreement (EULA). In some cases, it is strictly necessary for the proper functioning of an APP to allow access to certain hardware and software components of the smartphone as well as to some sensitive information: for example, if the user is interested in an application that reports weather forecasts, it is necessary for the APP to be able to access the user's location information in order to provide weather news relating to the user's location at that time.

For the above reasons, it seems logical to analyse APP stores, and in particular *Google's*, as an example of market outcome to the management of online data. In fact, the above shows:

a) the growing trend (and now most common among users) to connect from mobile devices (see **Figure 2.1**);
b) the use of mobile platforms (devices, operating system, APP stores) as personal tools not only for communication but also for the management of a range of daily activities carried out by citizens (see section 2.2);
c) the emergence of APPs, and consequently of virtual shops from which they can be downloaded, as intermediaries facilitating these activities (see **Figure 2.4** and **Figure 2.6**);
d) the emergence of *Google Play* as the virtual store holding the leading position both in terms of the number of downloads of APPs by users and in terms of distribution of mobile applications by developers (see **Figure 2.2**).

In this context, the APP stores have adopted specific data processing policies, in order to comply with the data protection regulations in force in the individual State and to create a more trustworthy relationship with the users of their services.

Privacy Permission is the term used by *Google Play* to indicate the fact that users, through the granting of permissions, transfer a series of information in order to install and use an APP and use its services. **The system of permissions, therefore, represents the mechanism adopted by this market component to regulate the transfer of data from the user to the developer of the APP. This system is defined and monitored by the store manager.**

In order to distribute their products on the *Google Play* store, developers must, in addition to using a series of technical tools made available on the platform expressly created by Android,[95] sign a distribution agreement that explicitly provides that the developer agrees, "*to protect the privacy and legal rights of users, if You make Your Products available through Google Play. If the users provide You with, or Your Product accesses or uses, usernames, passwords or other login information or personal information, You agree to make the users aware that the information will be available to Your Product, and You agree to provide legally adequate privacy notice and protection for those users.*" (Article 4.8, Google Play Developer Distribution Agreement as of February 2018).[96]

In order to be downloaded by the user and work properly, applications may require access to certain device features and information contained therein. Precisely because of the multiple features of smartphones and mobile devices, these devices produce and store a considerable amount of information and data related to the individual, from geo-positioning to call and message logging. Such information may be crucial for the functioning of the APPs, but users must grant their consent for its use. It is no coincidence that when downloading the APP, the **operating system asks the user to check the "accept" box after providing all the information the APP should access**. The mechanisms through which developers reveal how their APPs interact with users' devices and with the individual information conveyed by the devices themselves are, as already mentioned above, called "permissions".

Users can find detailed information on the permissions that an APP requires in blogs, specialized sites, the *Android* page, etc. The most obvious case is the message that appears on the screen of your device when you choose to download an application.

[95] https://developer.android.com/index.html.
[96] https://play.google.com/about/developer-distribution-agreement.html.

Usually, on *Android*-based smartphones (or tablets), once the user clicks on the "install" icon, a screen like the one shown in **Figure 2.7** appears.

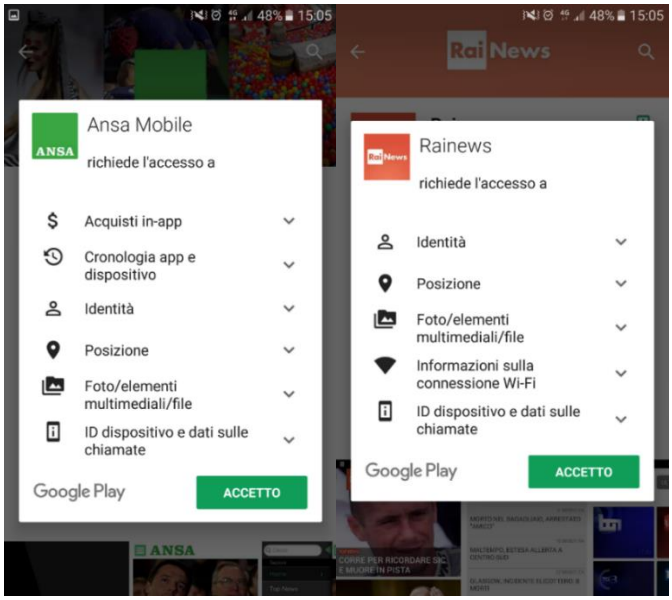


Figure 2.7 – Screenshots of the permissions requested by two information APPs
Source: *Google Play Store*

Furthermore, before the APP is installed, users can check the required permissions by accessing the "*authorization details*" option of Google Play, as shown in **Figure 2.8**. The list of permissions will be updated each time the APP is updated.

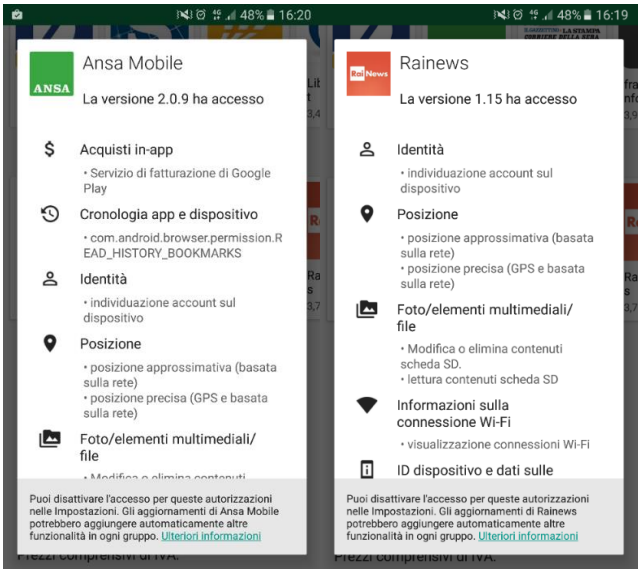


Figure 2.8 – Screenshot of the authorization details required by two information APPs
Source: *Google Play Store*

Permissions may allow apps to access a wide range of user data, such as recreational and travel data, browsing and shopping habits, media consumption, photos and videos taken and shared.

A survey conducted by the *Pew Research Center* on U.S. users showed a high level of awareness of the information required by APPs, with 90% of respondents pointing out that this aspect is (very or fairly) important when choosing whether or not to download an application, and 60% of respondents preferring

not to download an APP after discovering that the information requested was, in many cases, not necessary for the successful operation of the APP.[97]

In the *Android* operating system, APPs require users to agree to the terms of use at the time of their installation and, at the same time, users may view a brief description of the permission. The methods used to inform users about how their data will be used by an APP represent a connection between the user, *Google* (the provider and developer of the operating system and store) and third party developers of mobile apps.

Permissions, as already mentioned, are the way *Goiogle* asks developers to reveal how their APPs will interact with the user's devices and what information they need (or are required). On the pages dedicated to developers on the Android system, it is also possible to find a more specific classification made by the company based on the functionalities that the APP requests access to, as shown in **Table 2.1** below.

Table 2.1: Permission groups[98]

Permission groups	Description
Calendar	Used for runtime permissions related to user's calendar
Camera	Used for permissions that are associated with accessing camera or capturing images/video from the device
Contacts	Permessi *run-time* relativi ai contatti e ai profili sul *device*
Location	Used for permissions that allow accessing the device location
Microphone	Used for permissions that are associated with accessing microphone audio from the device
Phone	Used for permissions that are associated telephony features
Sensor	Used for permissions that are associated with accessing body or environmental sensors
SMS	Used for runtime permissions related to user's SMS messages
Storage	Used for runtime permissions related to the shared external storage

Source: developer.android.com

Two elements are significant in the Google Play permissions system: **Permission Group** membership and **Protection Level**. The first element is used to group the permissions needed and then present them to the user during the installation process, while the Protection Level specifies how the operating system should behave at the time of installation of the APP and, therefore, when necessary, users will be asked for their consent.

The main function of permissions groups, therefore, is to facilitate public understanding. Such system should allow for greater comprehensibility of permissions, thus making users more conscious when deciding whether to give their consent. Each group of permits may also include a number of individual permits that meet the criteria established by the *Android* platform (**Figure 2.9**).

Therefore, when the APP requests access to a specific Permit, the end user must grant the permit to the entire category. For example, suppose that an APP needs permission to monitor, modify or interrupt an outgoing call (*process outgoing calls*), as happens with VOIP (*Voice Over Internet Protocol*) applications; the

[97] Pew Research Center (2015), *Apps Permissions in the Google Play Store*, www.pewinternet.org. The survey was conducted on a sub-sample of 461 adults (over 18 years) from the GFK Group Knowledge Panel.
[98] https://developer.android.com/reference/android/Manifest.permission_group.html.

developer must indicate this to the platform operator, declaring it in the *permission manifest*; however, the user will only be asked to give consent to the entire group of permissions to which it belongs (i.e. the "*phone*" group of permissions).

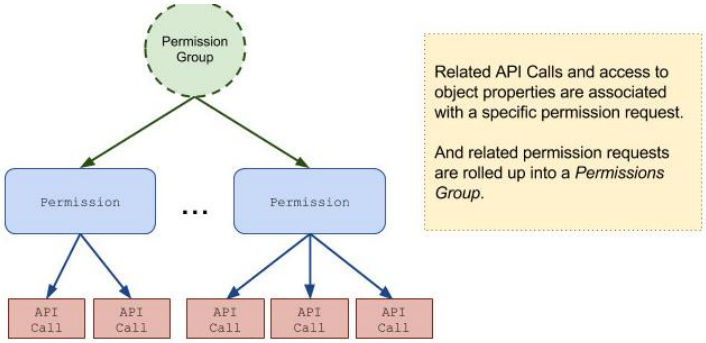


Figure 2.9 – The structure of permits in the Android system
Source: developer.android.com

Some permissions are not intended mainly to collect information, although they do, but to ensure interaction between the APP and the device's operating system in order to ensure the proper operation of the application while ensuring a minimum level of security.

In this sense, **Android makes a basic distinction based on the levels of protection to be attributed to individual permissions, identifying three levels of protection that developers must consider: normal, dangerous and signature**. This distinction is made not by considering the hazards related to the processing of users' data only, but mainly taking into account the level of risk associated with the proper functioning of the hardware.

Normal permissions are those that pose a low risk to other applications, the operating system, and users. The system automatically allows access to this type of permission, without even asking for the explicit consent of the user who, however, can always decide not to install the APP. **Dangerous permissions** are more risky because they allow access to user data or require control on the device. Because these APPs are potentially risky, explicit consent is required from the user, or the user should receive an explicit notification. The **signature feature**, on the other hand, refers to permissions whose dangerousness is similar to that of normal permissions, but which concern applications that have the same acronym (or certificate) used to sign the application that first certified the permission. If the two certificates match, then permission is granted automatically, without the explicit consent of the user. An additional layer of protection for signature permissions is the **signatureOrsystem feature**, which allows users to connect a signature permission also to other applications in the operating system; this feature is often used by application developers working for large enterprises to grant an advantage to their own applications

It should be remembered that APP developers can still identify new specific permissions; in addition, individual permissions, groups and risk levels defined by the store manager (in this case *Google*) are subject to continuous review in response to the increasingly sophisticated functionality of the devices, the privacy needs of users and, of course, the security of the operating system.

For the purposes of this study, in light of a more rigorous identification of permissions that makes it possible to go beyond the mere technical and IT perspective, more precise, additional classifications will be outlined in the next section.

In conclusion, it is worth observing how the management of data relating to a single individual in the digital world makes use of tools such as permissions; thanks to this system, no application,

by default, is allowed to perform any function that could have a negative impact on the functioning of other applications, on the operating system and, clearly, on the user.

It may be therefore stated that permissions rule the relationships that exist between users and developers of APPs related to the data flow and their treatment. An analysis of these tools makes it possible to understand the real behaviour of consumers and businesses in such exchange and to assess their economic and social efficiency.

In this context, it should be noted, finally, that the structure of permissions, also on the basis of national regulations on privacy, is defined and classified by the **owners of mobile platforms (i.e. operating systems and APP stores) which, therefore, are in a privileged position and are therefore able to guide the data market.**

2.6. The existence of an implicit exchange between users and web operators

Recent research has shown that there is a causal link between the number of permissions required by an APP and users' behaviour. The OECD[99] underlines that the permissions policy is strategically relevant to the success of an APP, as users seem to be aware of the amount and type of data that are transmitted when deciding to download and use an APP.

However, empirical studies show that there are a number of conflicting results with regard to the users' assessment of data collection carried out by APPs. The work of Grossklags & Acquisti (2007), for example, shows an unpredictable outcome between users' maximum willingness to pay for reducing the request of data and the minimum willingness to transfer data; according to the authors, the processes determining users' decisions differ according to whether to protect data or provide data; users' willingness to pay for protection is low.[100]

The work of Savarge & Waldman (2014), based on a sample of individuals who were asked to pay in order to avoid the transfer of their data, shows that users are willing to pay an estimated one-time fee of $2,28 to delete search data and a fee of $ 4.05 to hide data related to one's contact agenda.[101]

Kummer & Schulte (2016), who conducted an analysis similar to the one carried out in this Report, successfully test the hypothesis that developers offer APPs at a lower price in exchange for more and/or better quality digital data produced by users. They find that both APP demand and supply are significantly influenced by the number of permits required, underlining the existing trade-off between the demand and supply sides.[102]

The analysis carried out by this work also shows that on average, **paid APPs have a lower number of permits than free APPs. Therefore, implicitly, the market seems to attribute an economic value to users' digital data: the release of less data, in fact, implies a higher price to be paid for the user.**

[99] OECD, (2013); *The App economy*.
[100] J. GROSSKLAGS e A. ACQUISTI, (2007), *When 25 Cents is too much: An Experiment on Willingness-To-Sell and Willingness-To-Protect Personal Information*, WEISS.
[101] S. J. SAVAGE e D. M. WALDMAN, (2014), *The Value of Online Privacy: Evidence from Smartphone Applications*, Technical Report.
[102] KUMMER M. E., SCHULTE P., (2016), *When private information settles the bill: money and privacy in Google's market for smartphone applications*, ECONSTOR.

2.6.1. The study on (millions of) APPs and permissions

The present work, carried out by AGCOM's Department of Economics and Statistics in cooperation with the Department of Automatic Information and Management Engineering of the "La Sapienza" University of Rome,[103] will continue along this line of research, taking into account the information obtained from a rich dataset of APPs and the relative permissions required.

The dataset includes information on 1.135.700 GooglePlay apps, i.e. about 80% of the total number of apps available on the store. Such information has been collected on the basis of a process called "crawling". The remaining 20% is part of a residual share of applications, belonging to the "long tail", that were not frequently downloaded by users.

For each APP, the information gathered concerns:

- the category to which it belongs;
- the price;
- the minimum *in-app purchase* threshold;
- the maximum in-app purchase threshold;
- the rating assigned by users;
- the number of reviews written by users;
- the number of downloads;
- the types of permissions required.

As far as **categories** are concerned, they represent a fundamental tool, since their classification allows the user to navigate more easily through a store that contains about 3 million applications. In fact, with the exponential growth in the number of APPs, it has become important for the platform operator to introduce shortcuts that make users' browsing simple and easy. As an example of this, in July 2016, some existing categories were modified and others added.[104]

Table 2.2 shows the distribution of APPs by categories; it is important to note that the category of games includes 30 subcategories. The data collected are in line with what is already shown in **Figure 2.6**.

[103] The collection of data and information on APPs from the Google store was carried out by the Automatic Information and Management Engineering Department of the "La Sapienza" University of Rome, in cooperation with AGCOM's Department of Economics and Statistics. Special thanks go to Prof. A. Vitaletti and Mr. A. De Carolis.

[104] For example, from 27 July 2016, on Google Play there will be 8 new categories; Art and Design, Cars and Vehicles, Beauty, Dating, Events, Eating and Drinking, Home and Furnishing and Parents. Two other categories have been renamed; Transport in Maps and Navigators, and Media and Video in Video Tools http://www.androidauthority.com/google-play-store-new-app-categories-706028/

Table 2.2: Apps distribution by category

Categories	# of APPs	%
Games	228,823	20.16
Education	100,293	8.83
Tools	82,799	7.29
Entertainment	80,915	7.12
Lifestyle	79,158	6.97
Personalization	72,457	6.38
Business	64,551	5.68
Books & Reference	59,990	5.28
Travel & Local	49,093	4.32
Music & Audio	41,793	3.68
Productivity	33,821	2.98
News & Magazines	33,002	2.91
Health & Fitness	32,844	2.89
Finance	25,793	2.27
Communication	25,462	2.24
Social	22,246	1.96
Shopping	19,039	1.68
Transportation	17,618	1.55
Photography	16,859	1.48
Medical	16,646	1.47
Media & Video	14,843	1.31
Family	6,313	0.56
Weather	4,632	0.41
Comics	3,460	0.3
Libraries & Demo	3,250	0.29
Total	**1,135,700**	**100**

Source: AGCOM elaboration on data from *Google Play*

As described above, an APP shall require access to a set of information regarding both the hardware component of the device and the user's data, in order to function properly. Permissions are, therefore, a complex system, which has within it a series of classification problems. **Some permissions perform a technical function necessary to interact with the operating system installed on the devices; others are necessary in relation to the service offered; finally, others are technically redundant.**

The APPs analysed in the study contain, in total, 266 single permissions; the platform for developing APPs in *Android* also allows developers to predict new types of permissions, which means that such permissions are subject to continuous change; as a result, some of them are replaced over time. It should be noted that a considerable number of permits concern technical aspects related to the proper functioning of an APP, e.g. if you are interested in an application that measures road journeys, you need to allow the APP access to use the device's position sensors.

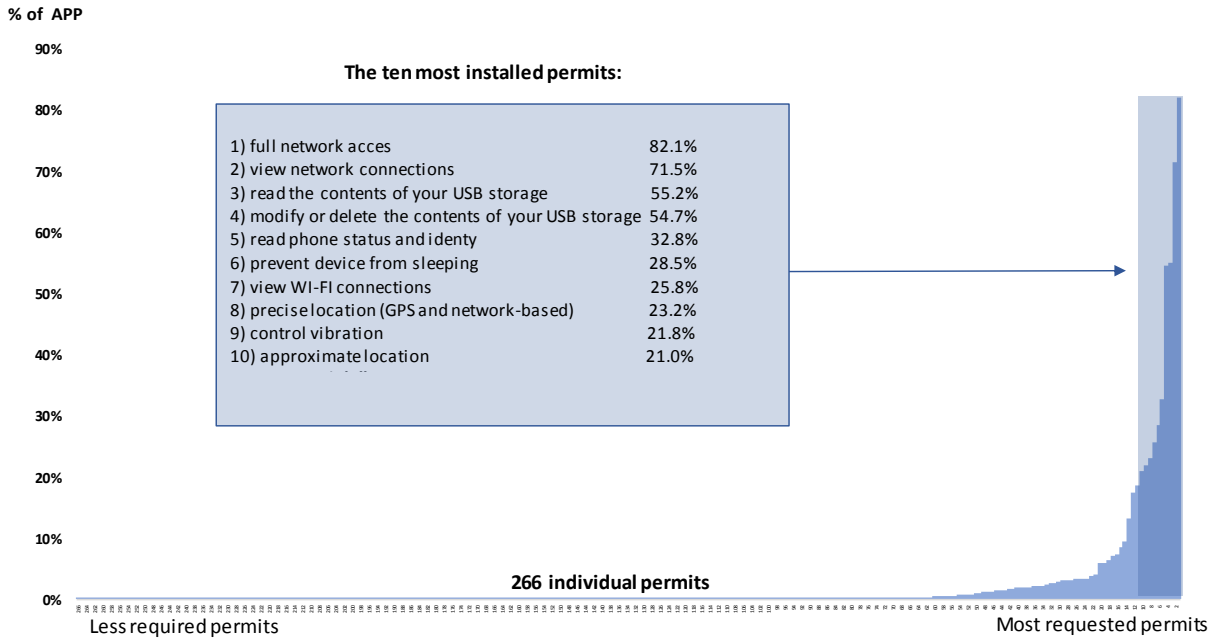


Figure 2.10 – Distribution of permissions
Source: AGCOM elaboration on data from *Google Play*

Figure 2.10 shows the distribution of permissions among APPs: only 10 out of 266 permissions types are used by more than 20% of APPs, while a considerable number of permissions are used by very few applications. As many as 20 single permissions, for example, are needed for by a single APP.

Obviously, the greatest interest is for the permissions used by the largest number of applications. Among these, it is important to identify a criterion **to distinguish between those that are most likely required to collect individual data and those that are technically necessary for the proper functioning of APPs.** The most important classifications of permits in the technical and economic literature are used in this study.

First, the *Pew Research Center*, in a search conducted on a sample of APPs available on the *Google Play Store*, distinguishes **permissions into two categories: those regarding hardware and those regarding user data** (so-called user info).[105] The authors of the research point out that the definition of "access to information of users" allows to distinguish between permissions that can interfere with any information of the user and those that do not require access to any of them.

A second important classification is the one developed by the researchers Kummer and Schulte (2016). The two scholars, on the basis of a previous classification carried out by Sarma et al. (2012), identify **a limited number of permits that may be critical in terms of access to sensitive data**.[106] They therefore construct an indicator that includes a series of permissions linked to the collection of user data; this variable can in turn be divided into four further categories, depending on the type of information acquired: device status, location, communication activities and profile.

The identification of some criteria to classify the most sensitive permissions with respect to the data of individuals, however, has no relevance; it is necessary, in fact, to consider how many APPs use these

[105] Pew Research Center (2015), *Apps Permissions in the Google Play Store*, www.pewinternet.org
[106] KUMMER M. E., SCHULTE P., (2016); *When private information settles the bill: money and privacy in Google's market for smartphone applications*, ECONSTOR.
SARMA B.P., et al., (2012); *Android permissions: a perspective combining risk and benefits*, in *Proceedings of the 17th ACM symposium on Access Control Models and Technologies.*

specific permissions and how many downloads have been made to understand their dissemination among users.

Referring to the 10 most common permissions, **Table 2.3** indicates, in addition to their description, which permissions can be considered "sensitive" according to the classifications of the Pew Research Center and Kummar & Schulte, or whether they can be considered "dangerous" or "normal" based on the categorization provided by the store manager itself, that is to say, *Google*.

Table 2.3: Main permissions by dissemination and relevance to the processing of sensitive data

Permission	The permission allow the APP to…	Pew Center	Kummer & Schulte	*Google*
full network access	…create network sockets and use custom network protocols.	Yes	No	Dangerous
view network connections	…view infomration about network connections (which network exist and are connected).	No	No	Normal
read the contents of your USB storage	…to read from external storage (SD cards).	Yes	No	Normal
modify or delete the contents of your USB storage	…write to the USB storage (SD cards).	Yes	No	Dangerous
read phone status and identity	…access to your device's identifiers (IMEI/IMSI, SIM ID, voice mailbox number, your phone number and, if a call is in progress, the remote number).	Yes	Yes	Dangerous
prevent device from sleeping	….prevent processor from sleeping or screen from dimming.	No	No	Normal
view WI-FI connections	…. check the state of your connection before trying to access the internet.	Yes	No	Dangerous
precise location GPS and network-based	…access precise location from location sources such as GPS, cell towers, and Wi-Fi.	Yes	Yes	Dangerous
control vibration	…access to the vibrator.	No	No	Normal

Fonte: elaboration on data from *Google Play*

Table 2.3 therefore refers to the most common permissions among APPs. The work of Kummer and Schulte also identifies many other permissions that can be considered relevant to the digital data of individuals, such as those concerning communication activities (e.g. allowing an APP to read SMS or MMS, to record audio or to control outgoing calls), and user profiling (allowing an APP to read the calendar, address book and search history).

2.6.2. The value of individual data for businesses and consumers

The following analysis concerns the **price of applications**, i.e. the value attributed to them by businesses and consumers, **behind which the value attributed to the data transferred after the purchase of the APP through the permissions system described above is hidden**.

As regards the **price** of APPs (**Table 2.4**), there is a rather asymmetric distribution: 86% of applications can be downloaded for free, while only 0.5% (i.e. 5.171 applications) have a price above €10.

Table 2.4: Distribution of APPs by price range

Price (€)	Number of APP	%
0	977,244	86.0
0-0.99	65,676	5.8
1-1.99	46,882	4.1
2-4.99	33,415	2.9
5-9.99	7,312	0.6
≥10	5,171	0.5
Total	**1,135,700**	**100.0**

Source: AGCOM elaboration on data from *Google Play*

Although some APPs are downloaded for free, users may, at a later stage, decide to use the *in-app purchase* service to buy additional services (*fremium* modality). However, only 3% of free APPs have a minimum level of in-app purchase, that is to say, between 0.40 and 0.99 cents.

The **rating** and the number of **reviews** are two very important variables in users' decision-making processes; when users download an APP, in fact, they are able to increase their information about the technical specifications and quality of the APP by drawing useful information from the comments and the vote that the users of the service have issued. The APP stores themselves recommend that you read this information in order to minimize the download of APPs that could be harmful to the proper operation of the device.

As mentioned above, **the search for an economic value to be attributed to the exchange of digital data and use of a service (downloading an APP) involves the study of permissions; each permission, in fact, can be associated with the release of a specific set of data related to the individual.**

A first interesting result emerges from **Table 2.5**; **free APPs require a significantly higher number of permissions than paid ones** (on average, about twice as many, or 6.4 compared to 3.8). 50% of paid APPs require up to 3 permissions, while 50% of free APPs require up to 5 permissions.

Table 2.5: Average number of permissions

	Number of APP	% on the total	Average no. of permissions
Free APPs	977.244	86%	6,4
Paid APPs	158.456	14%	3,8
Totale	1.135.700	100%	6,0

Source: AGCOM elaboration on data from *Google Play*

When considering the more "sensitive" permissions, the results are confirmed (**Table 2.6**): both in the case of APPs that require at least one sensitive permission to process individual data (**Panel B of Table 2.6**), and, more in general, for all APPs, i.e. including those that do not require sensitive data (**Panel A**

of Table 2.6), the average number of permits requested is decidedly higher when the APPs can be downloaded for free.

Table 2.6: Average number of "sensitive" permissions

Panel A: all APPs

	Pew Center	*Google*	Kummer & Schulte
Free APPs	3.3	3.6	1.1
Paid APPs	1.9	2.1	0.6

Panel B: APPs that require at least one "sensitive" permission

	Pew Center	*Google*	Kummer & Schulte
Free APPs	8.1	7.1	9.7
Paid APPs	6.1	5.4	7.2

Source: AGCOM elaboration on data from *Google Play*

A first conclusion of the study is that the free provision of an APP implies the provision, through the permissions system, of a greater number of digital data, in general, and of individual data, in particular. It can be therefore inferred that there exists an implicit exchange between users and internet providers that affects the primary commercial relationship concerning the purchase and sale of APPs.

Another interesting aspect concerns the correlation between **downloads** and permissions. The average number of downloads represents, in fact, a valid approximation of users' demand for APPs, and, therefore, of the potential quantity (**volume**) of data collected by developers and brokerage platforms (APP stores). **Figure 2.11** shows the distribution of APPs by number of downloads divided into paid and free APPs. The two values differ considerably, since, as it was easy to imagine, free APPs have a higher number of downloads. More than 80% of paid APPs, on the other hand, are downloaded 1 to 100 times, whereas such values falls to 45% for free APPs. Conversely, nearly one-third of free applications are downloaded more than 1.000 times, a value which stands below 4% for paid applications.

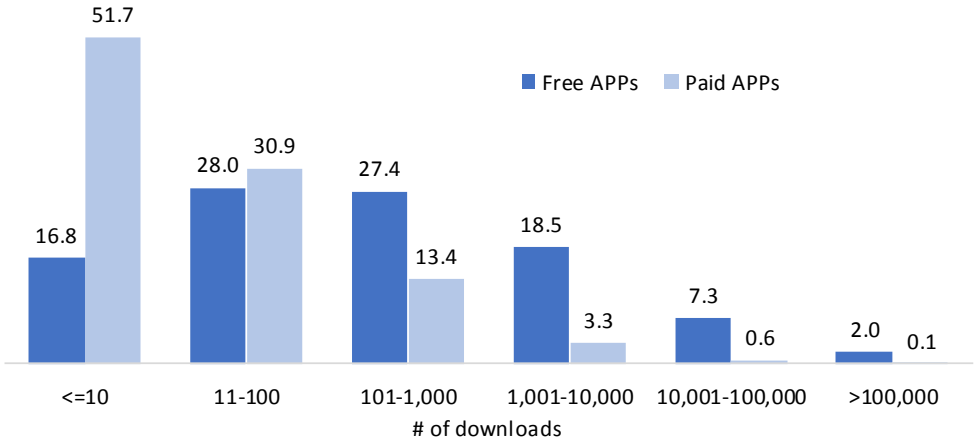


Figure 2.11 – Distribution of APPs by number of downloads
Source: AGCOM elaboration on data from *Google Play*

The download trend shows a clear phenomenon of "long tail" affecting also the APP market, as the rest of the web: in fact, considering both paid and free APPs, about 50% of APPs are downloaded less than 100 times and about 98% less than 100.000 times. **This determines that only a handful of APPs – that is to say, 2% – are installed by a considerable number of users, in line with what has already been shown above. Worldwide, only six APPs are installed more than 1 billion times: Facebook, Google Gmail, Youtube, Google Maps, Google Search and Google Play Services. This figure shows once again how, with a very high number of applications and operators, the market is actually concentrated in few large platforms, which reach distribution numbers that cannot be replicated by other operators.**

From this analysis, although descriptive, **two important trends emerge: *i)* the price of APPs decreases with the increase in the average number of permissions required, even though only those related to individual data are considered; and *ii)* the most frequently downloaded APPs are characterized by a greater presence of permissions related to individual data.**

A further step in the analysis is to verify the existence of a causal relationship in order to prove the possible existence of a relationship between users' demand for APPs (downloads) and the number of permissions required by the applications themselves, as well as between the price of the APPs set by the developers and the quantity and quality of the digital data processed.

In general (see Box 2), **there is a clear connection between the number of downloads and the permissions concerning sensitive data.** In particular, an increase in the permissions defined user info according to the *Pew Research* classification, leads to a reduction of 5% of downloads.

Taking into account a more detailed classification, such as that proposed by Kummer & Schulte, the effects of the permissions are divergent; on the one hand, in fact, those that require full access to the network and the device, which are characterized by a more technical nature, have a positive impact on the downloads; on the other hand, the permissions regarding communication activities and user location impact negatively on the number of downloads.

As far as price is concerned, **the results confirm what has already emerged from the analysis of the descriptive statistics. A business model that provides for the request of a price greater than zero, actually, is associated with a lower demand for permissions.** In this sense, with the exception of permissions that require access to the status of the user's device, the effect of permissions, when significant, considerably reduces the probability of paying to download an APP.

From the empirical study carried out on millions of applications, a significant effect of the system of permissions connected to the functioning of APPs emerges, both in terms of consumer choices (downloads) and of the business models that companies intend to adopt. In particular, study reports highlight how the permissions system is the tool through which data is exchanged between businesses and consumers.

However, **this exchange does not take place in the context of a definite contractual transaction where, inter alia, the price of the product is set, but takes place as an implicit exchange which is part of the buying and selling of other services** (APPs). This clearly poses enormous problems regarding the efficient functioning of markets and their regulation.

BOX 2 –APP SUPPLY AND DEMAND: AN ECONOMETRIC MODEL ON DATA EXCHANGE

With regard to the **analysis of APP demand**, a linear regression connected to demand of APPs was estimated, in terms of downloads, as a function of the permissions that it requires, as well as a series of control variables such as price, category, average rating and number of reviews that the application has.

The estimated model is as follows:

$$DEMAND_i = \alpha + \beta D_i + \theta X_i + \epsilon_i$$

where *Demand* is represented by the logarithm of the total number of downloads of the generic APP *i*, while X includes a series of control variables (such as the price, the total number of permissions requested, the average rating, the number of reviews and category, the developer of the APP).

The β parameter is the most interesting since it is associated with a dummy variable (i.e. binary) that takes into account whether or not an APP requires consent to a permission relating to individual sensitive data. Consequently, if the estimated parameter is negative, this implies that the presence of permissions concerning individual sensitive information leads to a reduction in demand.

A simple OLS has been used to estimate the model, the results of which are shown in the table below, while the classification of permissions has been based on the one shown in **Table 2.1**.

OLS estimations: demand models

Dependent variable: log. of installations	Model A: *Pew Research classification*	Model B: *Google classification*	Model C: *Kummer and Schulte*
User information permissions	-0.05***		
	(0.00)		
Dangeorus permissions		-0.01***	
		(0.00)	
Full Internet access permissions			0.07***
			(0.00)
View network state permissions			-0.13***
			(0.00)
Phone state permissions (read phone state and ID)			0.07***
			(0.00)
Location permissions (Gps)			-0.05***
			(0.00)
Communication permissions (read sms, intercept outgoing calls, ecc.)			-0.10***
			(0.00)
Users profile permissions			-0.02***
			(0.00)
Other permissions			-0.06***
			(0.01)
Constant	1.95***	1.92***	2.00***
	(0.02)	(0.02)	(0.03)
Controls	Yes	Yes	Yes
Categories	Yes	Yes	Yes
Adjusted R^2	0.84	0.84	0.84
# of observations	1,135,700	1,135,700	1,135,700

Heteroskedasticity-robust standar errors in brackets. ***, **, * significantly different from 0 at 1%, 5% and 10% levels, respectively

With regard to the **analysis of supply**, a probabilistic model was estimated in which, by means of a dichotomous variable equal to 1 for paid APPs and equal to 0 in the case of free APPs, the choices concerning the business model adopted by developers was analysed.

The estimated model appears as follows:

$$\Pr(PRICE_i = 1) = \Lambda[\alpha + \beta D_i + \theta X_i + \epsilon_i]$$

Again, the main parameter of interest is represented by the β parameter combined with the dummy variable that identifies whether or not a permission is to be considered relevant with respect to individual data: it is expected that as the number of permissions required by the APP increases, the probability that the application will have a price greater than 0 will be reduced. The following table shows the results of the probabilistic model estimates, in which the previous permit classifications are reported

Probit estimations: supply models

Dependent variable: APPs price	**Model A:** ***Pew Research classification***	**Model B:** ***Google classification***	**Model C:** ***Kummer and Schulte***
User information permissions	-0.26***		
	(0.00)		
Dangeorus permissions		-0.67***	
		(0.00)	
Full Internet access permissions			-0.41***
			(0.01)
View network state permissions			-0.55***
			(0.00)
Phone state permissions (read phone state and ID)			0.09***
			(0.00)
Location permissions (Gps)			-0.30***
			(0.01)
Communication permissions (read sms, intercept outgoing calls, ecc.)			0.01
			(0.01)
Users profile permissions			-0.00
			(0.01)
Other permissions			-0.03
			(0.02)
Constant	-0.33***	0.08**	0.29***
	(0.03)	(0.03)	(0.04)
Controls	Yes	Yes	Yes
Categories	Yes	Yes	Yes
Adjusted R^2	0.14	0.16	0.21
# of observations	1,135,700	1,135,700	1,135,700

Heteroskedasticity-robust standar errors in brackets. ***, **, * significantly different from 0 at 1%, 5% and 10% levels, respectively

2.6.3. The inefficiency of the data exchange system

The use of *big data* has become pervasive, extending to a growing number of sectors of the economic system. Data are collected and used for different purposes (so-called primary and secondary uses) and through increasingly complex and innovative processes and technologies. In particular, mobile platforms and their main components (device, operating system, APPs and APP stores), being personal internet access tools, now play a primary role both in the daily life of citizens and, consequently, in the online collection of data carried out by companies.

However, the data sector suffers from market failures. As AGCOM has already pointed out on several occasions, the markets under examination show, due to factors such as the presence of strong network externalities, a natural tendency towards concentration (with, at the very least, monopolistic situations such as "the winner takes all").

In this chapter **the, implicit and explicit, commercial relationship between users and internet service providers in the exchange of services (APP) and data was investigated**. In particular, the characteristics of the permissions that regulate the transfer of data from users to APP developers in the context of virtual stores (APP store) were analysed, thanks to a research project developed in collaboration with the Department of Automatic and Management Computer Engineering of the University "La Sapienza" of Rome, which allowed AGCOM to carry out a study on over a millions of **APPs available in the Google virtual store** (*Google Play*).

In a context in which users show, at the same time, a certain degree of awareness in the transfer of their data when surfing the net and a strong scepticism about the risks associated with the management of such information by web operators, the relationships that regulate this transfer assume a central importance. However, **the exchange of individual data against the offer of web services, often free of charge, or almost always available at a price lower than the underlying costs, is implicit, i.e. it is not clearly and legally established, so that the market is not able to set a price at the transaction.**

The fact that data are used not only for **primary purposes makes the structure of the digital ecosystem profoundly different from that of other media**, which are also partially financed by advertising sales. In this case, in fact, **data are used not only for the sale of (personalized) contacts to advertisers, but also for further uses, often unknown at the time of collection**.

In this context, **data represent the main bargaining chip of economic agents (users, operators, traders, etc.); however, the structure of the big data sector is incomplete to the extent that there is no mechanism, at least on the users' side, that regulates price formation.**

The AGCOM analysis has studied in detail the implicit commercial relationship, which takes the form of a transfer by users of rights on their data, in exchange not for an economic consideration, but for the offer of web services for free, or at low prices, close to marginal costs.

A rigorous quantitative analysis, involving millions of APPs, has shown that **all economic agents - both on the consumer demand side and on the developer supply side - are suffering, in their behaviour, from the absence of an institutional mechanism that regulates data trade**. On the one hand, consumers are willing to transfer their data at lower prices, on the other hand, web operators offer their APPs at lower, if any, prices, only in exchange for users' detailed information.

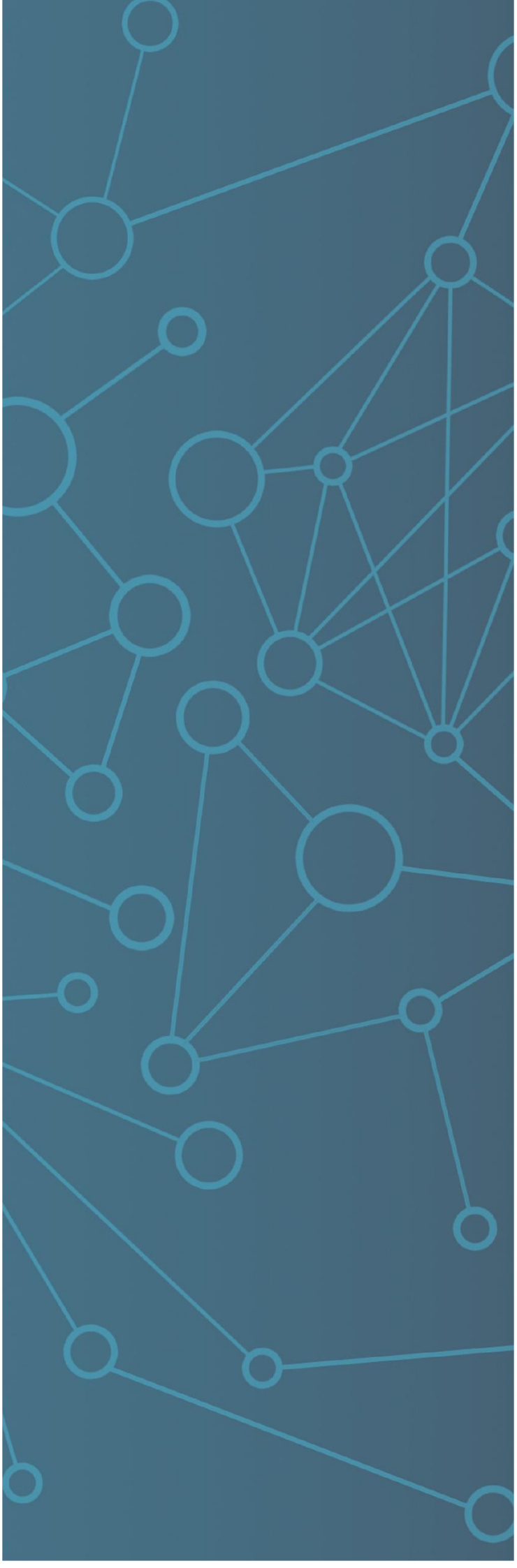
In the absence of a transparent market and institutional mechanisms that can guarantee a stable framework of rules for the actors involved in the purchase and sale of data, the digital system has self-regulated itself, the incompleteness of such transaction affecting the price of the services through which data are acquired by the operators and transferred by the users.

The absence of a real market mechanism makes these reports incomplete and inefficient. First of all, although permissions have started to distinguish the different types of information acquired by the operator, by categorizing them, **the consumer does not have a clear perception of which data are transferred and how they are processed, both for primary uses and, above all, for secondary uses**. It is, in fact, a **one-off transaction concerning other services (APPs) in view of the progressive and continuous use of users' information**. Therefore, it may be said that the market structure and the relevant transactions are distorted, leading to incomplete markets that inevitably fail.

What is missing is the main instrument that regulates, statically and dynamically, all (efficient) markets: the price of data. Moreover, since data are extremely diverse, prices should also be highly differentiated, depending on the type of information exchanged.

The absence, moreover, of correct market mechanisms tends to create a situation of **"overproduction" of data**: the allocation of resources is socially inefficient, not only for the absence of specific contracts that delineate the trading, but, given the structure of implicit prices, the quantity of data, considered as "goods", the amount of data gathered appear to be very different from their ideal quantity, from the economic and social point of view.

.



3

BIG DATA IN THE INFORMATION SYSTEM

3.1. Big data, online platforms and online news

The use of big data, as already pointed out in previous chapters, has become pervasive, extending to a growing number of sectors, from energy to finance, from insurance to automotive, from e-commerce to medicine, from entertainment to news and so on. Large amounts of data, differently structured, are collected and used for different purposes (primary and secondary uses), through increasingly sophisticated processes and technologies; this phenomenon concerns, in particular, online operators, whether they are providers of horizontal (satisfying a plurality of user needs) or vertical services (satisfying specific needs).

In this context, individuals are a primary source of data (see Chapter 3); this appears to be the consequence of the endless trail of pieces of information and traces that each user leaves, more or less consciously, while communicating or carrying out actions online, via fixed and mobile devices. This generates data (images, videos, typed words, emails, ...) concerning a multitude of characteristics and aspects of an individual's life, related to his or her personal details, location, spending, habits, interests, research carried out, social information (including contacts, reactions, shared content, opinions expressed).

The market operators that acquire, directly and by means of cookies or other tracking systems, the greatest amount of data generated by individuals when surfing the net through browsers or applications are certainly online platforms. In fact, as they are privileged tools for accessing the internet, they now play a primary role in citizens' daily life and, consequently, in the collection of data. Often, in fact, there is an implicit exchange between online platforms and users (see Chapters 2) whereby, in view of the information released and provided by individuals, online platforms, thanks to the management and processing of acquired data, offer services for free or at reasonable prices and customize the user experience on the network, with the aim of maximizing their profits.

More generally, the ability to access data regarding its users and use them as a strategic asset according to the logic of multi-sided markets is an almost constant element in the business model adopted by online platforms.[107] In recent years, commercial uses of consumers' data have become increasingly important both for platform activities and for the development of online services and the news media system.

Search engines and social networks are, in this sense, the most obvious example of how many internet services are provided for free to end users who, however, must provide the platform with a series of personal data, whether consciously or not, in order to use the service.

Moreover, the use of big data by search engines and social networks represents a crucial aspect because of the increasingly important role played by these platforms in the news media system, both at the international and national level. On the one hand, they became world leaders in the field of online advertising - which still represents the main source of funding for online news publishing and media[108] – thanks to the amount of individuals' and personal data that they own and which allow for an accurate user profiling; on the other hand, they now represent the main distribution vehicle for online news, since internet users increasingly access news through search engines and social networks.

As a consequence, big data also plays a crucial role in the protection of information pluralism, given that the availability and subsequent use of data on actual or potential consumers of information services have

[107] See AGCOM (2014), Indagine conoscitiva sui servizi internet e la pubblicità online.
[108] See. AGCOM (2017), 2017 Annual Report on the activity performed and work programmes pages 125-132; AGCOM (2015), Indagine conoscitiva su Informazione e internet in Italia. Modelli di business, consumi, professioni.

become a key competitive lever for internet companies, on both sides of the market on the users' side and on the advertising sales side.[109]

The acquisition and use of individuals' data, in fact, is at the core of the operating mechanisms (such as crawling, classification, association, prioritization and filtering) of online platforms that distribute online contents. In this regard, search engines and social networks are often defined as "algorithmic sources" of information to recall algorithmic customization, made possible by the quantity and quality of data collected on individuals, which characterizes the processes of creation and dissemination of informative content on the same platforms.

More specifically, the distribution of information on the internet is carried out through a process of disaggregation and disintermediation of the traditional information offer and subsequent regrouping and re-intermediation by algorithmic sources. Therefore, the algorithms underlying their operation have become decisive in determining how consumers use information, significantly orienting the success or not, in terms of audience, of one piece of news (or a publisher) compared to another (on the mechanisms that govern the operation of the algorithms, see also Box 5.1).

From the perspective of the online news demand, given that the internet in Italy is by now the second most widely used means of information (after television), algorithmic sources are constantly present in the news media consumption of most citizens. The most recent data available to AGCOM[110] (see **Figure 3.1**) show that, in 2017, 54.5% of Italians acquired news via means governed by algorithms, while 39.4% of the population obtained news directly from websites and publishers' applications (daily press and periodicals, radio and television, and online-only publications). Specifically, as regards the acquisition of information, search engines and social networks receive similar preferences among the population, both reaching a share of 36.5%.

Algorithmic sources, therefore, seem to have a "gatekeeper" function in accessing information; they represent the "places of transit" preferred by users/citizens, and, as a consequence, a necessary tool for publishers that wish to reach consumers. This clearly affects the strategies of informative content distribution implemented by the latter, which, on the other hand, in the medium to long term, risk losing direct contact with the public; this may lead to users ceasing to consider publishers as providers of such news media contents, in favour of the intermediary through which online content is delivered.

[109] With regard to the role played by online platforms in the information system, AGCOM is also conducting the specific survey "*Digital platforms and information system",* aimed at analysing the structure and functioning of the platforms in disseminating information on the Internet, highlighting any critical issues from the point of view of information pluralism. In addition, AGCOM has recently set up the *"Technical Committee for ensuring pluralism and fairness of information on digital platforms*", which aims to promote self-regulation of platforms and the exchange of good practices for identifying and fighting online misinformation phenomena (see section 3.4).

[110] See AGCOM (2018), Rapporto sul consumo di informazione

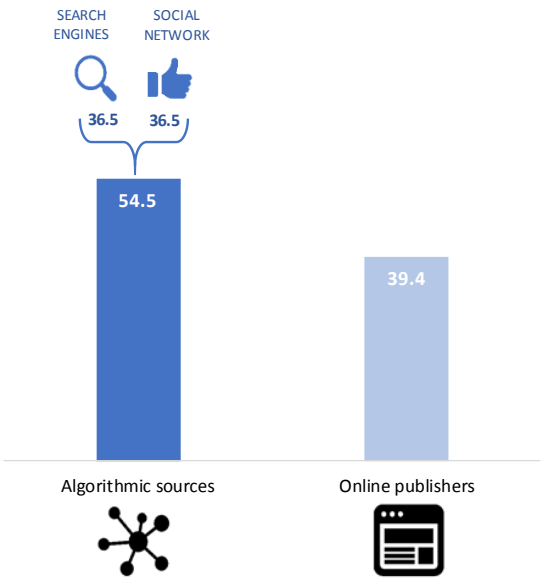


Figure 3.1 – Sources of access to online information used by Italian citizens (2017; % of population)
Source: AGCOM elaboration on data from *GfK Italia*

The fact that users prefer to acquire information through algorithmic sources rather than through online editorial sources is a further proof of the importance that individuals give to the former, which, since this may be taken as an indication of the greater attention paid to the use of disclosed information, can be considered as an sign of the actual consumption that users make of these sources with the purpose of getting informed.

In this regard, **Figure 3.2** shows that 19.4% of the population believes that an algorithmic source is the most important source to get information. It is worth underlining that great importance is given to search engines (9.7% of citizens) and social networks (6.8% of Italians), which, following national free-to-air television channels, national newspapers and digital daily newspapers, represent respectively the third and fourth source of news, often considered the most important information source, considering all classic and online media.

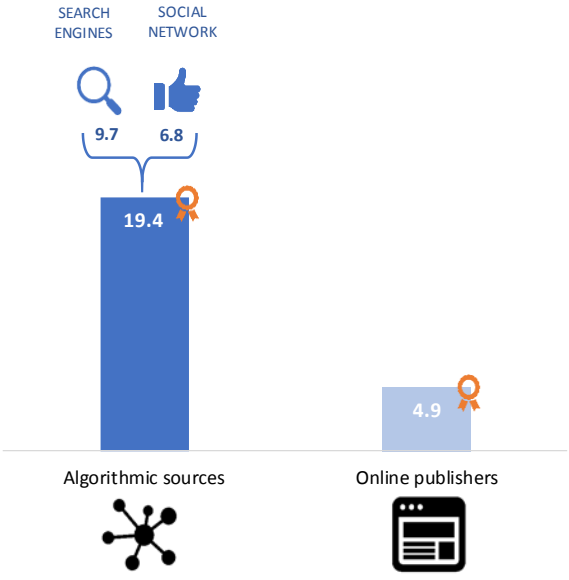


Figure 3.2 – Source of information considered most important by Italian citizens (2017; % of population)
Source: AGCOM elaboration on data from *GfK Italia*

In short, the diffusion of big data is also changing the structure of the global information system. As AGCOM has repeatedly pointed out in recent years, on the one hand, the emergence of data analytics has revolutionized the advertising sector, which finances (in part) the main traditional information sources and (predominantly) online sources. In this field, and particularly in online advertising, **platforms, which have a very wide range of data and analysis technologies, enjoy a huge and growing competitive advantage over other players, such as publishers, who are active on the internet**. On the other hand, **big data platforms such as social networks and search engines have become the preferred gateway for citizens to access online news**. This not only affects users, who consume news and information, but also publishers, who wish to reach end users with their informative products. Search engines and social networks, with their algorithms (indexing, content presentation, newsfeed, news classification, recommendation, etc..; see Box 3), are a fundamental instrument for those who produce and aim at distributing information on the network. The next section will therefore focus on the role played by these platforms and in particular by social networks in the online news media system.

BOX 3 – ALGORITHMS FOR ONLINE PLATFORMS AND INFORMATION

Algorithms are logical subsystems based on mathematical functions present in various types of software. In the online news media system, algorithms are powerful tools used mainly, but not only, to filter the available news and deliver it to users in an order, often customized, resulting from the application of certain criteria. First, it is possible to distinguish five categories of algorithms used by online platforms with reference to the information system:

- *web search*: these are algorithms for indexing and presenting open content on web pages in HTML format, with reference to the responses to queries made by users (typical of Google - PageRank - and other search engines);
- creation of news *feeds* on social networks: this second type of algorithm, which includes the algorithm used by Facebook (EdgeRank), determines the visibility of each content within the social networking platform and the customization of the News page (the news feed) of each user, based on certain characteristics of the post and the user/page that publishes it;
- recommendation systems: these are selection algorithms that suggest certain customized contents to users;
- control and removal of contents: these are algorithms that intervene in automatic mode, performing a classification of the post on the basis of the subject dealt with and the opinions expressed;
- automatic classification of news: these are algorithms on which the operation of platforms that deliver news are usually based (e.g. Google News).

Algorithms are therefore crucial in defining the online news consumption patterns by users and also have a significant value on the supply side in guiding the success of certain news (and publishers) compared to others and in determining the choices of publishers and journalists. As a result, changes in the main algorithms can profoundly change the online information system, both on the demand side and on the supply side.

3.2. The role of social networks in the online news system

Among all online platforms, **social networks** - given the time spent by users using them, the many actions that individuals perform and reactions that they express through their profiles/pages/accounts, as well as social relationships that establish - **are certainly among the internet intermediaries with major capabilities of acquiring the greatest variety and volume of data (i.e. big data, see Chapter 1) on individuals, including those related to ideological and political preferences and informative content read, displayed, enjoyed, commented and shared.**

In addition, within the online information and communication system, social networks are distinguished by their location and peculiar characteristics compared to all other types of online intermediaries, for at least three main reasons.

1) As argued in the previous section, the operation of social networks, also in terms of offer of informative contents, is governed by algorithms that constantly filter, according to predetermined criteria, the available news and provide it to users in a generally customized order, i.e. one that takes into account the type of user.

2) In a context characterized by the "unpacking" of news products and by a fragmented use of the contents (articles, comments, videos, posts, etc.), social networks act as intermediaries for citizens that access information; access which, very often, is the result of a casual and random discovery of news by users.

3) Finally, social networks allow the entry into the information system of sources that usually stand outside the classic information circuit, through profiles of common users, non-professional information pages/accounts, satirical pages/accounts, etc.. In most cases, however, on social networks informative contents of a journalistic nature, as well as contents generated by users and other non-professional figures, assume the same importance, since the selection and prioritization of the information to be shown to users is done with automatic updating mechanisms, based on data such as the similarity of the content, the posts of friends, reactions, sharing and comments, rather than on the credibility and journalistic quality or relevance in terms of public interest of the contents/news.

Social networks, therefore, are the virtual space where users can directly communicate with all the other different players operating in the information system: from journalists to publishers, from politicians to influencers and to non-professional figures who spread news to other users (friends and friends of friends).

On social networks, each individual is not a mere user of news but potentially becomes an active part in the dissemination of information, opinions and points of view, being more or less involved in this process, depending on the actions that he or she choices to perform. In this sense, users can not only click on the link to a piece of news, but can also express their reaction to it, share it, comment on it, participate in a discussion on such piece of news and post their own images, photos and videos on the subject. All these actions contribute to the dissemination of news that may "go viral".

At an international level, **social networks have become an integral part of the daily news consumption by citizens. In large parts of the world, in fact, they are considered among the sources most frequently used to find all kinds of news, as they have a significant and increasing influence in forming the public opinion, and, therefore, on media pluralism** (in this regard, see also the next section).

A recent study by the *Pew Research Center*, carried out in 38 countries, shows that, in 2017, daily access (i.e., once or more times a day) to social networks for the purpose of getting informed involves on average

more than a third of the adult population (see **Figure 3.3**), and this figure reaches 48% when taking into account those who get informed through social networks less often than every day.[111] **Figure 3.3** shows, however, that on the whole there are no significant differences between developed countries (where the average daily use of social networks to acquire information stands at 36%) and developing countries (where an average 33% of the adult population get informed daily through social networks). Focusing on single nations, in 3 out of the 38 countries analysed (South Korea, Lebanon and Argentina) more than half of the adults get informed every day on social networks while 7 countries (Canada, Australia, Sweden, Vietnam, Turkey, Chile and Brazil) show usage rates higher than 40%. Only in 4 countries (Indonesia, Senegal, India and Tanzania), on the other hand, the rate of daily use of social networks for information purposes is less than 20%.

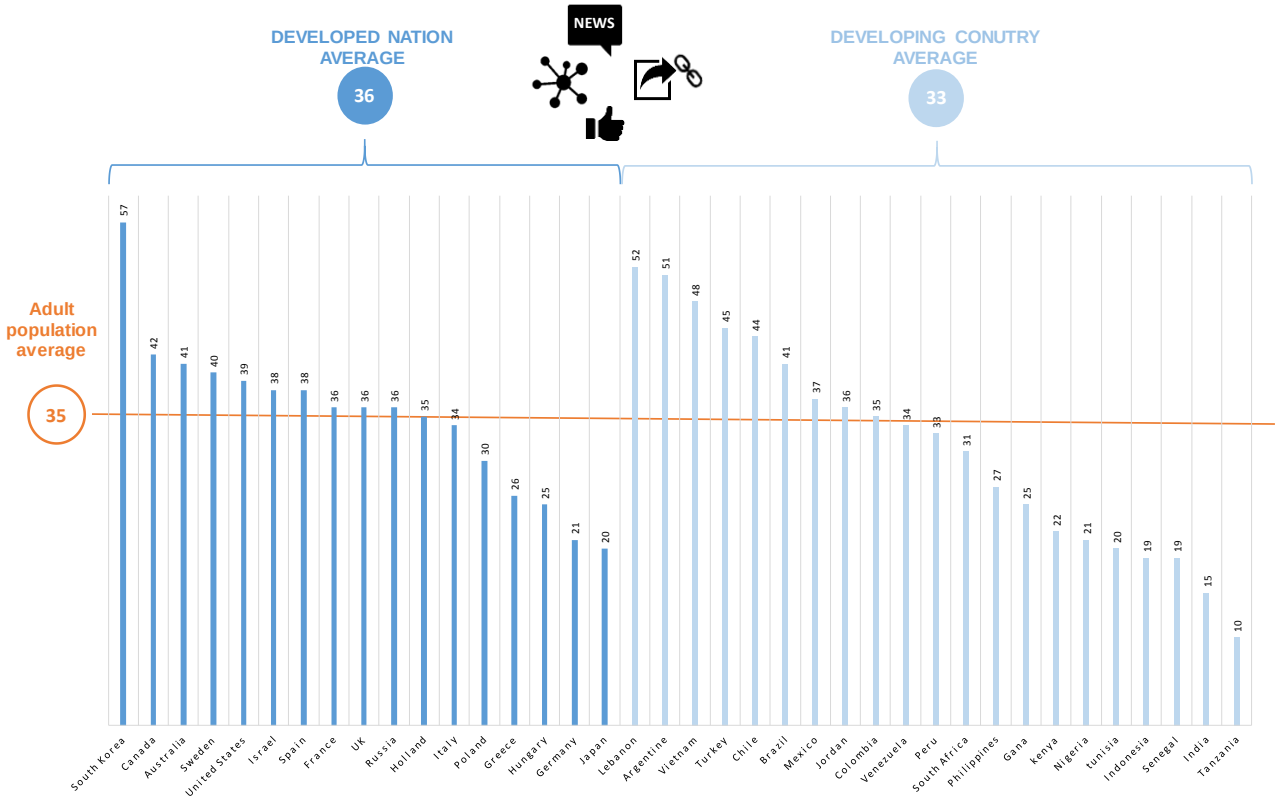


Figure 3.3 – Using social networks to get informed on a daily basis (2017; % of population aged 18 and over)
Source: Pew Research Center, Global Attitudes Survey 2017

As mentioned before (see **Figure 3.2**), the role of social networks in meeting the demand for news by the Italian population is further supported by the overriding importance that individuals attribute to them, despite the fact that acquiring information is only one of the many purposes that can determine their use. The recent study carried out by AGCOM on the "Online news consumption in Italy",[112] which has devoted extensive analysis to the access to and consumption of political information contents, also shows that 14.9% of Italian voters entrust their search for information to social networks in order to make their political and electoral choices (see **Figure 3.4**). This value, which places social networks, that stand first among online news media sources, in third place (after television channels and national newspapers) in the overall ranking of the means used to get informed on politics, corresponds to 43.8% of the total number of voters who use the internet to acquire information on politics.

[111] See Pew Research Center (2018), *Publics Globally Want Unbiased News Coverage, but Are Divided on Whether Their News Media Deliver*. This study is based on the results of the *Global Attitudes Survey 2017*.
[112] See AGCOM (2018), *Rapporto sul consumo di informazione*

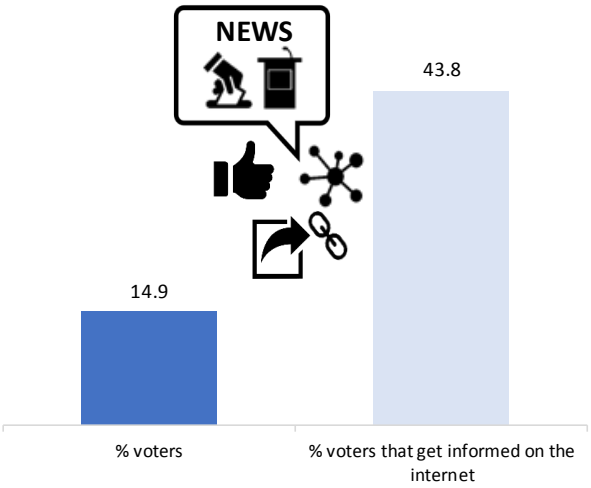


Figure 3.4 – Use of social networks with the purpose of making political and electoral choices in Italy (2017; %)
Source: AGCOM elaboration on data from *GfK Italia*

Despite the prominence given by citizens to social networks as information and communication tools and means to foster social relations, in recent years they are the subject of international discussions on the dissemination of pathological forms such as those related to polarization,[113] which triggers the formation of "ideological bubbles" (or *echo chambers*)[114] on the web and, more generally, of phenomena of misinformation.

In this regard, several scientific studies based on the analysis of millions of data deriving from the use of social networks examine the role played by ideological polarization and *confirmation bias* mechanisms (that is, the tendency to acquire information consistent with one's own ideological preferences at the expense of contrasting ones) in the dissemination of online misinformation. Among these are, in particular, the works that prove the importance assumed by forms of selective exposure to informative content on social networks in the creation and dissemination of "information disorders".[115] Research scholars go so far as to establish the existence of a direct link between the level of polarization of the topics dealt with and the tendency of the latter to become the object of misinformation.[116] Most scholars, however, also found that the trend to

[113] The analysis of the relationship between the ideological polarization of social networks' users and their activities on the web carried out in the aforementioned AGCOM's Report on Information Consumption (*Rapporto sul consumo di informazione*) has shown how polarization can have a significant effect on the greater engagement with the news disclosed by social networks. The link between the performance of all information actions on social networks (including those with a higher rate of user involvement) and polarization has clear repercussions on the occurrence of phenomena of diffusion of radicalized positions and the creation of ideological bubbles.

[114] *Echo chambers* are characterized by individuals who discuss only with people sharing the same ideological beliefs. They therefore tend to follow and intensify the problems of selective exposure and *confirmation bias*.

[115] See, among others, BESSI A., COLETTO M., DAVIDESCU G.A., SCALA A., CALDARELLI G., QUATTROCIOCCHI W. (2015), "Science vs Conspiracy: Collective Narratives in the Age of Misinformation", *PLoS ONE* 10(2); ZOLLO F., KRALJ NOVAK P., DEL VICARIO M., BESSI A., MOZETIČ I., SCALA A., CALDARELLI G., QUATTROCIOCCHI W., (2015), "Emotional Dynamics in the Age of Misinformation", *PLoS ONE* 10(9); BESSI A., PETRONI F., DEL VICARIO M., ZOLLO F., ANAGNOSTOPULOS A., SCALA A., CALDARELLI G., QUATTROCIOCCHI W. (2016), "Homophily and Polarization in the Age of Misinformation", *The European Physical Journal Special Topics*, 225(10); DEL VICARIO M., BESSI A., ZOLLO F., PETRONI F., SCALA A., CALDARELLI G., STANLEY H.E., QUATTROCIOCCHI W., (2016), "The Spreading of Misinformation Online", *Proceedings of the National Academy of Science* 113(3); ZOLLO F., BESSI A., DEL VICARIO M., SCALA A., CALDARELLI G., SHEKHTMAN L., HALVIN S., QUATTROCIOCCHI W., (2017), "Debunking in a World of Tribes", *PLoS ONE* 12(7); e SCHMIDT A.L., ZOLLO F., CALDARELLI G., SCALA A., QUATTROCIOCCHI W., et al., (2017), "Anatomy of News Consumption on *Facebook*", *Proceedings of the National Academy of Sciences*, 114(12), pp. 3035-3039.

[116] M. DEL VICARIO, W. QUATTROCIOCCHI, A. SCALA, F. ZOLLO (2018), "Polarization and Fake News: Early Warning of Potential Misinformation Targets", *arXiv preprint arXiv:1802.01400*.

polarization of users around the topics discussed (such as, for example, the Brexit[117] or the Italian constitutional referendum of 2016[118]) is shown in similar ways on different social networks (which reach different users), based on different algorithms.

Therefore, through social networks, the automatic personalization systems (operating on the basis of the *big data* acquired), on the one hand, and content sharing actions carried out by users, on the other, facilitate the proliferation of fake news and viral content propagation.

In this regard, Vosoughi, Roy and Sinan (2018)[119] have studied the ways and speed of dissemination of false news through social networks, comparing them with the propagation of real news.[120] Specifically, the research examines true and false news spread on Twitter from 2006 to 2017 and classified as "true" or "false" on the basis of the data acquired from 6 independent fact-checking organizations. The results produced by the analysis of 126,000 posts published on Twitter by 3 million people more than 4.5 million times reveal how, for all the categories of news considered (politics, economy, wars and terrorism, environmental disasters, science and technology, entertainment, curiosity), fake news are able to spread more quickly, more extensively and with greater pervasiveness than real news (see **Figure 3.5**).

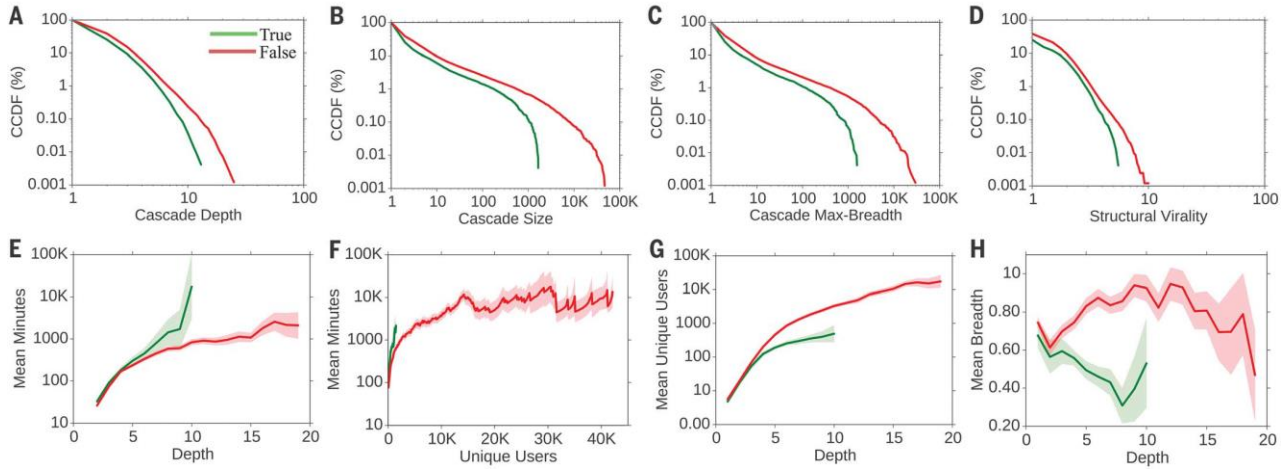


Figure 3.5 – Dissemination of real and false news on Twitter

Source: *Science*, 2018, Vol. 359, pp. 1146-1151

Moreover, these effects, as shown in **Figure 3.6**, become more pronounced when it comes to false news related to politics, where the most popular false news shows the largest (reaching the greatest number of people) and fastest dissemination (in terms of speed of propagation and popularity).

[117] See DEL VICARIO M., F. ZOLLO, CALDARELLI G., SCALA A., QUATTROCIOCCHI W., (2017), "Mapping Social Dynamics on *Facebook*: The Brexit Debate", *Social Networks*, Vol. 50, pp. 6-16. DEL VICARIO M., GAITO S., QUATTROCIOCCHI W., ZIGNANI M., ZOLLO F., (2017), "News Consumption during the Italian Referendum: A Cross-Platform Analysis on *Facebook* an Twitter", *IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, http://ieeexplore.ieee.org/document/8259827.

[118] See. DEL VICARIO M., GAITO S., QUATTROCIOCCHI W., ZIGNANI M., ZOLLO F., (2017), "News Consumption during the Italian Referendum: A Cross-Platform Analysis on *Facebook* an Twitter", *IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, http://ieeexplore.ieee.org/document/8259827.

[119] See VOSOUGHI S., ROY D., ARAL S., (2018), "The spread of true and false news online", *Science*, Vol. 359, pp. 1146-1151.

[120] Previous studies had focused on the ways in which certain types of false news spread on social networks. See, among others, BESSI A., COLETTO M., DAVIDESCU G.A., SCALA A., CALDARELLI G., QUATTROCIOCCHI W. (2015), "Science vs Conspiracy: Collective Narratives in the Age of Misinformation", cit.; e DEL VICARIO M., BESSI A., ZOLLO F., PETRONI F., SCALA A., CALDARELLI G., STANLEY H.E., QUATTROCIOCCHI W., (2016), "The Spreading of Misinformation Online", cit.

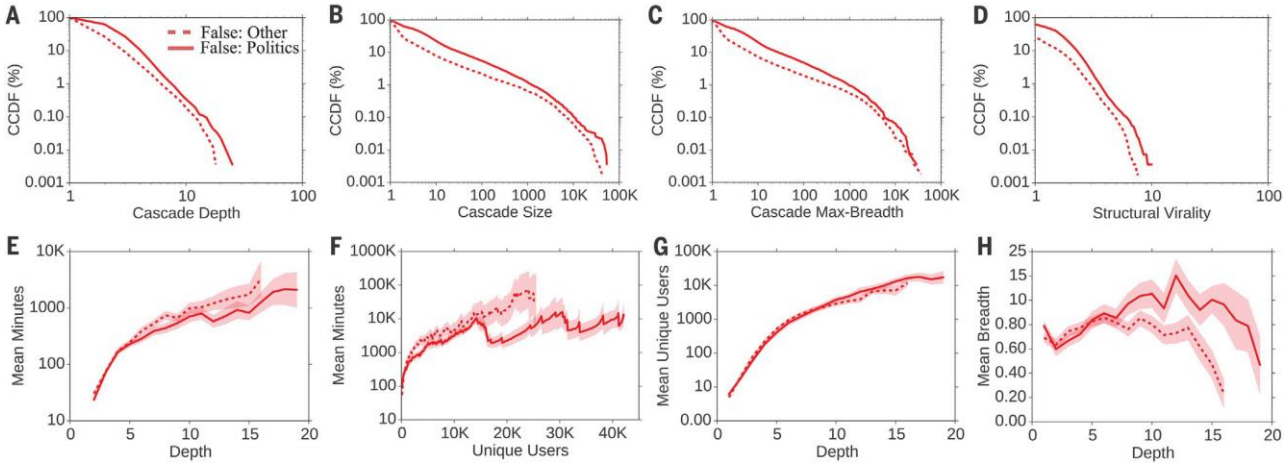


Figure 3.6 – Methods of disseminating false information on politics compared to other types of news on Twitter
Source: *Science*, 2018, Vol. 359, pp. 1146-1151

3.3. The influence of social networks on the forming of public opinion

In consideration of the growing importance of social networks for acquiring information and, at the same time, of the critical aspects in terms of protection of information pluralism that - according to the trends described above – may be connected to their use, various branches of literature have questioned the actual ability of social networks, which are virtual, to influence real (*offline*) behaviour. This issue is particularly important when taking into account the analysis on *big data*, given the crucial role that the acquisition and use of individual digital data plays in the functioning of social networks, including the distribution of news media content.

The production of scientific works on the subject, both theoretical and empirical, is already abundant and varied, although in continuous expansion, in the light of the sudden technological evolution and the analysable information generated by the digital traces left by users while browsing social networks.[121] For the purposes of the study carried out in this Report, it is considered appropriate, therefore, to review the significant evidence emerging from the most recent research on the influence exerted by social networks on the process by which individuals build their own vision of reality. On the basis of AGCOM's powers in this area, particular focus is placed on **the effects produced by social networks on the public opinion-forming process and political choices made by users, bearing in mind that the ability of social networks to reach large masses of individuals means that even small effects can generate behavioural changes involving thousands of people, affecting the election results. A subsequent Report by AGCOM will analyse, with the cooperation of prominent research scholars (and in particular of Prof. Walter Quattrociocchi),[122] all the processes preceding the electoral choice of the citizen, that is, those relative to the forming of public opinion on the web.**

A first aspect to consider, as it may trigger forms of "contagion" that could potentially concern the whole variety of human perceptions, including positive and negative attitudes towards a political topic/exponent, is **the ability of social networks to influence emotional states**. As early as 2014, Kramer, Guillory and Hancock pointed out that emotional states on social networks can be transferred

[121] In this regard, it is worth mentioning, among other things, the emergence of a new field of research, computational social science, which studies social phenomena associated with the consumption of social networks, adopting a quantitative (based on the use of large amounts of data) and multidisciplinary approach (involving areas such as mathematics, statistics, physics, sociology, information technology); see LAZER D., (2009), "Computational Social Science", Science, Vol. 323, pp. 721-723.
[122] See the outcome of the comparative procedure for the award of a research activity on 'Information and digital platforms', see https://www.agcom.it/documents/10179/8623861/Allegato+25-1-2018/e2200786-4963-4d19-a7e0-a437d24061c5?

to others through emotional contagion, leading different people to experience the same emotions without being aware of it.[123] In detail, the authors conducted an experiment on 689.003 Facebook users, changing the amount of emotional content shown in their news feed. The experiment shows that with the reduction of positive opinions, people produce less positive and more negative posts; conversely, with the reduction of negative expressions, the opposite occurs. These results suggest that the emotions expressed by others on social networks are likely to influence other users' emotions, constituting experimental evidence of a contagion on a large scale, even in the absence of personal interaction (exposure to the status of a friend who expresses an opinion is sufficient), underlining the extreme importance of the content selection by the online platforms.

Similarly, Coviello et al. (2014)[124], analysing data generated from millions of Facebook users, shows that the emotional content of users' status messages can influence friends' status messages, noting that social networks can amplify the intensity of emotions shared globally.

The sharing of emotions, sensations and, more generally, perceptions (including ideological ones) to social networks has also been studied by Jost et al. (2018).[125] The authors, reviewing a variety of works on protest movements in the United States, Spain, Turkey and Ukraine, observe that social networks can concretely play an important role in the exchange of information and coordination of collective action. This is not only due to the fact that the pieces of information relevant to the coordination of protest activities (such as news on transport, turnout, police presence, hazards, medical services, legal assistance) is likely to spread quickly and efficiently through social networks, but also because social platforms facilitate the transmission of emotional and motivational messages, both in support of and in opposition to the activity of protest, able to emphasize certain emotional states (such as moral indignation, social identification, group membership and concerns about equity, social justice, deprivation), as well as explicitly ideological themes, which, spreading on a large scale, can have a strong influence and determine the success or failure of the protest movements themselves.

In addition, Bond and Messing (2015),[126] through the analysis of data relating to over 6 million Facebook users, investigate the relationship between individuals' exposure to disagreement on the social network and their actual vote. In sum, the research shows that an increase in the ideological distance from friends is associated with lower voter turnout rates of the user considered.

The social influence on voter behaviour arising from the use of social networks was also the subject of a vast experimental study conducted on the occasion of the Congress elections held on 2 November2010 in the United States. Bond et al. (2012),[127] in order to verify the theory that political behaviour can spread through a social network, carried out an experiment on users aged 18 or over who had access to Facebook on election day. Users were randomly assigned to 3 different groups:

- a group, which was shown a **"social message"**, consisting of over 60 million people who, on 2 November 2010, viewed at the top of their Facebook news feed a message encouraging the user to

[123] KRAMER A.D.I., GUILLORY J.E., HANCOCK J.T., (2014), "Experimental evidence of massive-scale emotional contagion through social networks", *PNAS*, Vol. 111, N. 24, pp. 8788-8790.
[124] COVIELLO L., SOHN Y., KRAMER A.D.I., MARLOW C., FRANCESCHETTI M., CHRISTAKIS N.A., FOWLER J.H. (2014), "Detecting Emotional Contagion in Massive Social Networks", PLoS ONE 9(3).
[125] JOST J.T., BARBERP, BONNEAU R., LANGER M., METZGER M., NAGLER J., STERLING J., TUCKER J.A., (2018), "How Social Media Facilitates Political Protest: Information, Motivation, and Social Networks Advances", *Political Psychology*, Vol. 39, Suppl. 1.
[126] BOND R.M., MESSING S., (2015), "Quantifying Social Media's Political Space: Estimating Ideology from Publicly Revealed Preferences on *Facebook*", *American Political Science Review*, Vol. 109, N. 1, pp. 62-78.
[127] BOND R. M., FARISS C.J., JONES J. J., KRAMER A.D.I., MARLOW C., SETTLE J. E., FLOWER J. H., (2012), "A 61-million-person experiment in social influence and political mobilization", *Nature*, Vol. 489, pp. 295-298.

vote, which included a link for searching local polling places, a clickable button with the inscription "I voted", a counter that indicated how many other Facebook users had already voted, as well as showing up to 6 profile images of Facebook friends selected randomly among those who had already clicked on the button "I voted" (see **Figure 3.7**);

Figure 3.7 – Social message shown during the experiment carried out by Bond et al.
Source: *Nature*, 2012, Vol. 489, pp. 295-298

- a group, which was shown an **"information message"**, consisting of 611,044 users, who displayed a message encouraging them to vote that included a link for searching local polling places, contained the clickable button with the inscription "I Voted" and the counter indicating how many other Facebook users had already voted. Unlike the "social message", the "information message" did not show any faces of friends who had voted (see **Figure 3.8**);

Figure 3.8 – Information message shown during the Bond et al experiment.
Source: *Nature*, 2012, Vol. 489, pp. 295-298

- a control group, consisting of 613,096 users, who were not shown any messages at the top of their news feed.

At the end of the experiment, the authors examined both the direct effects that the messages generated on the behaviour of the users of each group, and the indirect effects produced on the friends of the friends of the users subjected to the experiment.

As far as direct effects are concerned (see **Figure 3.9**), Bond et al. found in users exposed to the "social message" a greater disposition to publish a self-declaration of vote, to search for information on local polling stations, as well as to participate in the vote (verified through the examination of public registers).

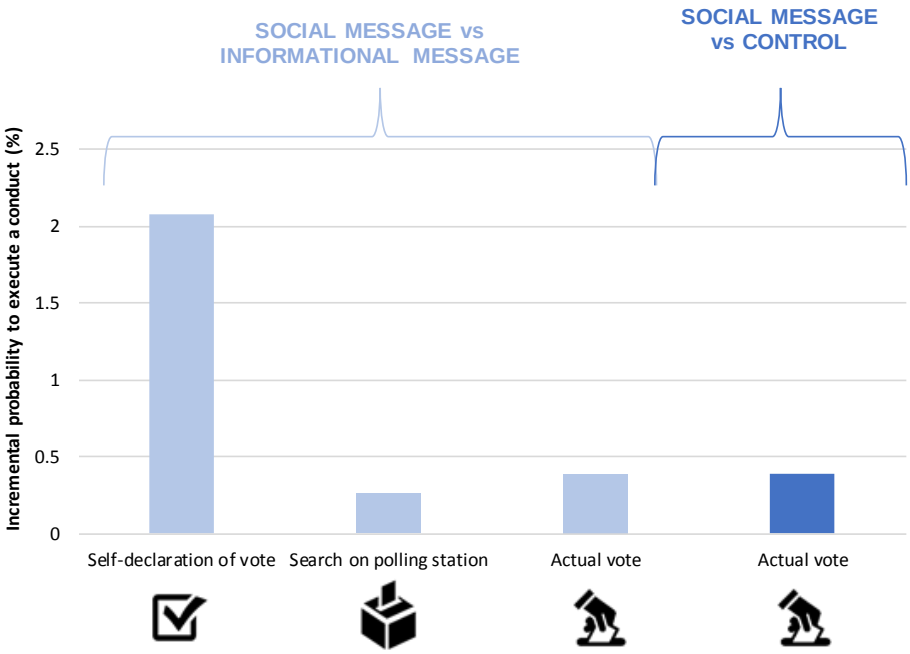


Figure 3.9 – Direct effects of message exposure on user policy actions

Source: *Nature*, 2012, Vol. 489, pp. 295-298

To study the indirect effects, the authors reconstructed the composition of the networks of friends of the sample users (made up of an average of 149 friends), distinguishing the relationships characterized by strong links from those characterized by weak links, based on the levels of interaction evaluated for each pair of subjects. The effect of the experiment on each friend was then measured by comparing the behaviour of friends connected to a user exposed to the "social message" with the behaviour of friends connected to a user who was part of the control group. The results obtained showed that the closer the bond with the user exposed to the message, the stronger the effects of the experiment on a friend. In particular, strong ties (between close friends) are important for spreading voting behaviour in the real world.

Overall, the authors estimated that the "social message" shown on Facebook has increased the direct turnout of about 60,000 voters and indirectly, through social contagion, has determined the participation in the vote of further 280,000 voters, for a total of 340,000 additional votes (corresponding to 0.14% of the electorate).

The Bond et al. experiment (2012) was successively replicated by Jones et al. (2017) during the 2012 U.S. presidential elections.[128] Once again, there was a significant increase in voter participation as a result of the direct and indirect effects of exposure to the "social banner". Also in this work, the authors register a significant increase in the turnout among close friends of those who received the message encouraging them to vote, and the total effect on friends appears even wider than the direct effect.

In conclusion, the results described in this and previous sections have demonstrated the importance of *big data*, which constitute the basis of the mechanisms through which online platforms, such as social networks, operate in the information system. Through the platforms that use them, in fact, big data have a crucial effect on information pluralism, on the demand side as well as on the supply side (in terms of

[128] JONES J.J., BOND R.M., BAKSHY E., ECKLES D., FOWLER J.H., (2017), "Social influence and political mobilization: Further evidence from a randomized experiment in the 2012 U.S. presidential election", *PLoS ONE* 12(4).

both economic resources deriving from advertising sales and algorithmic customisation, automatic selection and prioritisation of the informative contents displayed).

In this sense, when considering means of communications such as the internet - already used quite extensively by the most politically and ideologically active individuals, according to trends that lead to the formation of *eco chambers* characterized by self-referenced circles and polarized positions[129] - the effects of actions carried out on social networks by the most active users as well as by customization algorithms tend to make polarization go viral.[130]

More generally, exposure to informative messages rather than any other types of message on online platforms, where the big data collected are decisive, not only affects the perceptions of users, but is able to influence the forming of their opinions, translating them into real choices and actions, including those that prove to be decisive for the election results.

3.4. AGCOM's regulatory approach: the Technical Roundtable to safeguard pluralism and fairness of information within online platforms

For some time now, online misinformation, conveyed above all through *big data* digital platforms, and its repercussions on public opinion, including on political choices, have been at the centre of a regulatory, analysis and policy process promoted by AGCOM, which appears to be the result of the changes occurred due to the emergence of an (increasingly) *data-driven* society (see **Figure 3.10**).

In this sense, AGCOM, in the last five years, has carried out monitoring and studying activities concerning the specific technical and economic functioning of online platforms as means of access and distribution of news as well as the impact of these new online market players on the news media system and on pluralism. AGCOM was thus able to confirm the emergence of critical issues related to the increasing use of social networks and search engines in political elections and referendum campaigns, as well as the spread of misinformation strategies (based, among other things, on the acquisition of *big data*) through online platforms.

The evolution of the online news media system, from both the demand and supply sides, has been the subject of further analysis through the drafting of reports, the organisation of workshops and the conduct of study inquiries such as the one launched on "Digital platforms and the online news system" (Resolution no. 309/16/CONS) and the joint inquiry on "Big data" launched in cooperation with the Italian Competition Authority and the Authority for the Protection of Personal Data, which this Report is part of.

Moreover, current analyses have shown that **the examination of the online disinformation phenomena requires a multidisciplinary approach, as well as the adoption of cooperation and comparison initiatives with the subjects operating in the online information system, research institutions and sector associations, in order to acquire an adequate knowledge of complex phenomena, such as the impact of platforms on public opinion, and to encourage forms of self-regulation by those involved in the information system.**

[129] For a more detailed analysis, see AGCOM (2018), *Rapporto sul consumo di informazione*, cit
[130] See, inter alia, BESSI A., ZOLLO F., DEL VICARIO M., SCALA A., CALDARELLI G., QUATTROCIOCCHI W., (2015), "Trend of Narratives in the Age of Misinformation", *PLoS ONE* 10(8); QUATTROCIOCCHI W., A. Scala, C. R. SUNSTEIN (2016), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2795110; QUATTROCIOCCHI W., VICINI A., (2016), *Misinformation: Guida alla società dell'informazione e della credulità*, FrancoAngeli; QUATTROCIOCCHI W., VICINI A., (2018), *Liberi di crederci: Informazione, internet e post-verità*, Codice edizioni.

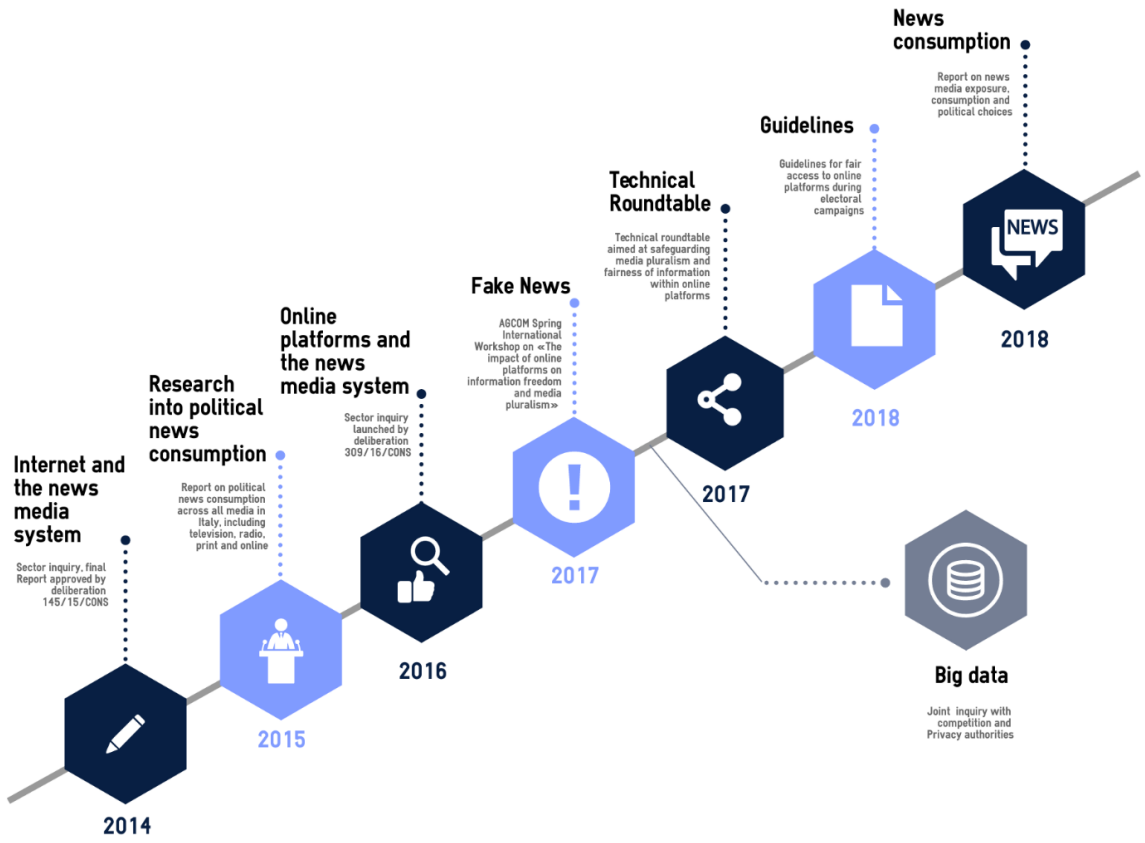


Figure 3.10 – AGCOM's regulatory approach for online information
Source: AGCOM

Based on the above and in view of the changes introduced by the use of big data, at the end of 2017 AGCOM set up the "***Technical Roundtable aimed at safeguarding media pluralism and fairness of information within online platforms***" (Resolution no. 423/17/CONS), with which it intended to pursue the objective of promoting self-regulation of platforms and the exchange of good practices for the identification and combating of online disinformation phenomena resulting from targeted strategies. In particular, the main purpose of the body is to encourage the sharing of information, the comparison and the emergence of suitable methods of detection, as well as the identification of transparency tools and the adoption of the most appropriate rules and actions to ensure, especially during election campaigns, equal treatment for all political actors on the platforms and fairness and impartiality of information for users.

The Roundtable promoted by AGCOM is an unprecedented experience, in Italy and abroad, of cooperation with online platforms to fight against disinformation in the digital environment. Although the adoption of similar instruments of self-regulation by platforms and publishers, for example in the field of fact-checking, has already been tested abroad, **the Italian one is the first case of coordination among the actors of the information system promoted by an independent regulator of communications that operates in order to facilitate dialogue between stakeholders.** At the current stage of the system development, in fact, AGCOM decided to support and monitor the self-regulatory initiatives implemented by the companies involved, also encouraging the confrontation with and contribution of international experts, universities, research centres and trade associations.

More specifically, as shown in **Figure 3.11**, **the Technical Panel involves the representatives of all the components of the information system**: online platforms; traditional publishers with online

information offers and those operating exclusively online; journalists; advertising and consumer associations.

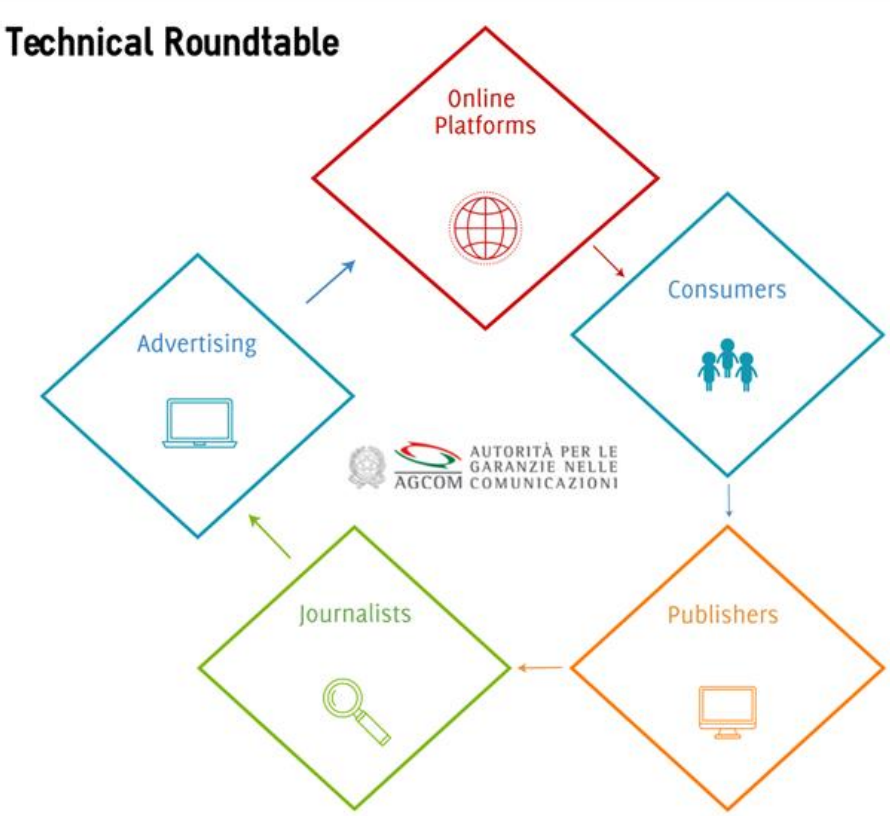


Figure 3.11 – The components of the Technical Roundtable
Source: AGCOM

The Technical Roundtable, whose activities are organized in meetings, plenary meetings and working groups (see **Figure 3.12**), had an initial operational phase - corresponding to the period prior to the start of the campaign for the general elections of 4 March 2018 - aimed at identifying, with respect for freedom of expression, self-regulatory instruments for:

i) preventing and countering online disinformation strategies in election campaigns;

ii) the creation of a free and conscious debate on the internet, including issues which are typically at the centre of the political and electoral debate;

iii) ensuring equal access for all competing political actors in an election campaign, including in an information context where big data platforms have a prominent role.

In this context, in February 2018, the **"Guidelines for equal access to online platforms during the 2018 election campaign"** were adopted, thanks to which the participating platforms (Google and Facebook) made available some specific tools, including the information campaign launched by Facebook on the pages of its Italian users for the detection of false news, and specific services for policymakers interested in making their program known to citizens (for example, Google Posts and Facebook Issues). The Guidelines also included activities to verify and remove illegal content, or content contrary to national legislation on equal conditions (*par condicio*) (for example, surveys published in the 15 days

preceding the elections), upon report, as well as fact-checking initiatives carried out with the support of independent organizations.

Figure 3.12 – Organisation and Structure of the AGCOM Technical Roundtable
Source: AGCOM

The first operational phase of the Panel is being followed by another, which has seen the establishment of five working groups related to:

a) **methods for classifying and detecting online misinformation**;

b) **definition of systems for monitoring the economic advertising flows**, from domestic and foreign sources, aimed at financing fake contents;

c) **fact-checking**, organisation, techniques, tools and effects;

d) **media literacy** and online disinformation;

e) planning and carrying out **information campaigns on misinformation aimed at consumers**.

Among the activities of the working groups, a first shared effort was made to outline the peculiar and identifying aspects of the problems related to disinformation on the internet. At the end of the survey, **the main elements to consider in order to classify the various disorders of online information (misinformation, "mal-information" and disinformation) were identified**.[131] These are elements

[131] In particular, the term "mis-information" identifies the category of information contents disclosed on the Internet that are untrue or inaccurately reported, which are likely to be perceived as real, but not created with malicious intent. "Mal-information" means the category of information contents based on real facts (including private facts) disseminated on the

that relate to the phases of production of information content (false contents and their ability to influence public opinion; malicious intent underlying their creation; political/ideological or economic motivation of those who create and then spread them), dissemination of the same (in a massive way) and impact on information pluralism (consequences on the forming of public opinion).

In identifying the characteristics of contents that could create online misinformation and prejudice information pluralism, the analysis focused once again on the importance of the use of big data for the implementation of techniques aimed at disseminating and making false information go viral through online platforms.

Specifically, the cross-cutting analysis carried out by the groups of the Technical Roundtable is currently focusing on the role of big data in the creation of online disinformation strategies and, more generally, in **the identification of false online information contents**, with the purpose of identifying the main activities, organizational methods, technologies and resources used (including big data) for the creation, production and distribution of fake contents, and the actual implementation of online disinformation strategies. These strategies, in fact, are characterized by the presence of an organized structure, which sets objectives, economic and otherwise, in the short, medium and long term. In particular, AGCOM had the opportunity to carry out a thorough examination of the above strategies, including those that are based on economic motivations and those that are based on ideological-political motivations. From the analyses carried out so far, a complex picture emerges, **in which a variety of real business models of online disinformation coexist, some based on the collection of advertising, others on the direct contribution of users through fraudulent actions and on the dissemination of contents that aim at damaging the brand and the image of the companies that have been targeted.**

In the course of the Roundtable's work, these models, resource flows and non-commercial strategies will be investigated in detail, also with the help of case studies, in order to highlight the critical issues to be faced in combating online disinformation and define the most appropriate combination of technical and market solutions and self-regulatory codes.

Each working group, moreover, is continuing in the specific activities related to its topic of reference, adopting an analytical approach that, focusing on the new paradigm introduced by the pervasive dissemination of big data in the information system, can only be based on knowledge and the most advanced interdisciplinary and data-driven methods, that require the use of large amounts of data.

In conclusion, the Technical Roundtable aimed at safeguarding media pluralism and fairness of information within online platforms is a fundamental tool in the monitoring and regulatory process on online information and disinformation that AGCOM is carrying out. More precisely, **the body is the expression of a new analysis, regulatory and policy approach, aimed at acquiring knowledge in order to be able to make adequate choices**. In other words, an approach based on the acquisition, preliminary and essential, of all the information and knowledge necessary to investigate complex phenomena, resulting not only from the technological development related to the emergence of big data, which is at the basis of online platforms operation (through algorithms), but also from the extreme importance of the various categories of rights (individual and collective) involved, including social and political rights that necessarily require timely and *ex ante* measures to be duly protected. **By using**

Internet and contextualized in such a way as to go also viral and convey a message with the precise intent of damaging a person, an organization or a country, or affirming/disregarding a thesis; while "dis-information" means the category of information content, including sponsored content, artfully created in such a way as to be plausible, characterized not only by the falsehood of the facts, but also by their "contagiousness", as well as by the intentional massive publication and dissemination.

instruments such as the Technical Roundtable, AGCOM - through all the tools at its disposal, the constant dialogue with the stakeholders and the continuous cooperation with the international scientific community – implement such new approach in the national and EU regulatory context of reference.

More generally, starting from the evidence presented above, the next phase of the Sector Inquiry on big data, concerning the aspects relating to AGCOM, will focus on issues connected to the relationship between big data and pluralism 2.0. In addition, policy suggestions and changes to the regulatory framework will be developed.