



***Data science
and the transformation of the financial
industry***

Design and Layout

Marketing and Communication Department
Management Solutions

Photographs

Photographic archive of Management Solutions
iStock

© Management Solutions 2015

All rights reserved. Cannot be reproduced, distributed, publicly disclosed, converted, totally or partially, freely or with a charge, in any way or procedure, without the express written authorization of Management Solutions. The information contained in this publication is merely to be used as a guideline. Management Solutions shall not be held responsible for the use which could be made of this information by third parties. Nobody is entitled to use this material except by express authorization of Management Solutions.

Content



Introduction

4



Executive summary

6



A financial industry undergoing transformation

12



Data science: an emerging discipline

24



Case study: social networks and credit scoring

38



Bibliography

42



Glossary

44

Introduction

Without data you're just another person with an opinion.
W. Edwards Deming¹

The world is changing, and doing so at high speed. We are witnessing a technological revolution of a magnitude never observed before.

This is not a transitory event. The paradigm shift rate (the rate at which new ideas are adopted) doubles every decade; while it took nearly half a century for the telephone to be adopted, and while acceptance of television and radio took several decades, it took under 10 years² for computers, Internet and mobile phones to catch on. In 2014, the number of mobile phones already equaled the number of people on the planet (7 billion) and a third of them were smartphones, while the number of Internet users reached almost 3 billion³.

Every year information technology doubles its capacity and price/performance ratio, as predicted by Moore's Law⁴, which to date has proven to be true (Fig. 1). The result is exponential growth in the technology available and an equivalent reduction in its cost, regardless of the crises experienced over the past few years, and this trend is expected to continue in the coming decades.

But this technological revolution has taken on a new dimension in recent years. Along with increased technical performance has come increased capacity to generate, store and process information, and at an exponential rate too, a situation that has been called the «big data» phenomenon. Some evidence of this is:

- ▶ The total volume of data in the world doubles every 18 months^{5,6}.
- ▶ Over 90% of the data that exist today were created in the last two years⁶.
- ▶ The per capita capacity to store data has doubled every 40 months since 1980 and its cost has decreased by more than 90%⁶.
- ▶ Processing capacity has increased 300-fold since the year 2000, making it possible to process millions of transactions per minute⁶.

The impact of this technological transformation is particularly significant in the financial industry because it adds to four other major trends that are shaping this industry:

1. A macroeconomic environment characterized by weak growth, low inflation and low interest rates, which has penalized the banking industry's profit margins in mature economies for a long period of time; and uneven performance in emerging countries, with a trend towards slower growth and a rise in default levels.
2. A more demanding and intrusive regulatory environment, where regulation is becoming global in terms of corporate governance, solvency, liquidity, bail-out limitation, consumer protection, fraud prevention and data and reporting requirements, among other areas.
3. A profound change in customer behavior, as consumers' financial culture has improved and customers expect and demand excellence in service while manifesting growing confusion at the complexity and disparity of the products and services offered, which makes them more dependent on opinion leaders.
4. New competitors entering the financial market, some with new business models that impact the status quo.

¹William Edwards Deming (1900-1993). American statistician, university professor, author, consultant and promoter of the concept of total quality, known for his work on Japan's development and growth after World War II.

²Kurzweil [Director of Engineering at Google] (2014).

³International Telecommunications Union (2014).

⁴Observation by Gordon Moore, Intel co-founder in 1965, that technology develops at a rate such that the number of transistors on an integrated circuit doubles approximately every two years. Moore (1965).

⁵It is estimated that each day of 2012 there were 2.5 exabytes of data, an information volume equivalent to 12 times all printed books in the world.

⁶Federal Big Data Commission (2014).

The combined effect of these four factors, together with technological transformation among other reasons, is causing industry players to put the focus on the efficient use of information, thus giving rise to a financial industry discipline which so far has been more focused on the IT industry: data science.

Data science is the study of the generalizable extraction of knowledge from data using a combination of automated learning techniques, artificial intelligence, mathematics, statistics, databases and optimization, together with a deep understanding of the business context⁷.

All the above disciplines were already used in the financial industry to varying degrees, but data science has features that make this specialist subject essential to successfully navigate the industry transformation already underway.

More specifically, addressing all elements in the complex financial industry environment previously mentioned requires large data volumes and elaborate analytical techniques, which is exactly the specialty field of data science. Also, data science as a discipline has become more prominent as a result of the big data phenomenon, and therefore data scientists are professionals that are qualified to handle massive quantities of unstructured data (such as those from social networks) which are increasingly significant for financial institutions.

This dramatic increase in data creation, access, processing and storage, as well as in data-based decision making, together with other circumstantial factors already described, has not gone unnoticed by regulators. Indeed, there is a global trend, substantiated among others by the Basel Committee on Banking Supervision (through BCBS 239), towards a requirement for a robust data governance framework that will ensure data quality, integrity, traceability, consistency and replicability for decision-making purposes, especially (but not only) in the field of Risk.

This trend has been complemented by the US Federal Reserve and OCC regulations⁸ requiring entities to implement robust model governance frameworks to control and mitigate the risk arising from the use of models, known as «model risk»⁹.

Financial institutions are taking decisive steps to develop these governance frameworks (data and models), which together constitute the governance of the data science capabilities.

In this changing environment, the transformation of financial institutions is not a possibility: it is a necessity to ensure survival. This transformation is closely linked to intelligence, which is ultimately the ability to receive, process and store information to solve problems.

Against this backdrop, the present study aims to provide a practical description of the role played by data science, particularly in the financial industry. The document is divided into three sections which respond to three objectives:

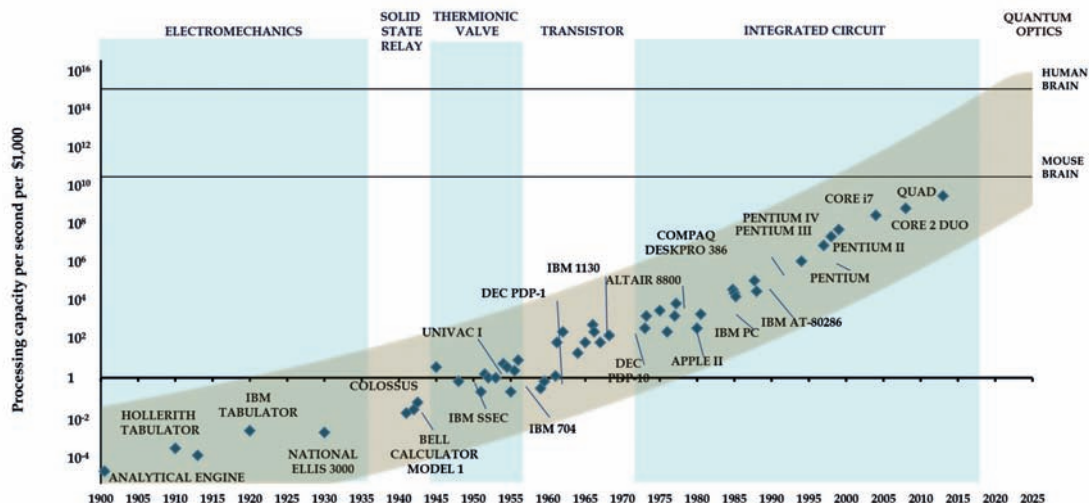
- ▶ Describing the IT revolution in which the financial industry is immersed and its consequences.
- ▶ Introducing the data science discipline, describing the characteristics of data scientists and analyzing the trends observed in this respect, as well as their impact on governance frameworks for data and data models in financial institutions.
- ▶ Providing a case study to illustrate how data science is used in the financial industry, consisting of the development of a credit scoring model for individuals using data from social networks.

⁷Dhar [Center for Data Science, New York University] (2013).

⁸OCC/Fed (2011).

⁹This trend has been analyzed in depth in Management Solutions (2014).

Fig. 1. Moore's Law: exponential growth of processing capacity per second and 1,000 dollars.

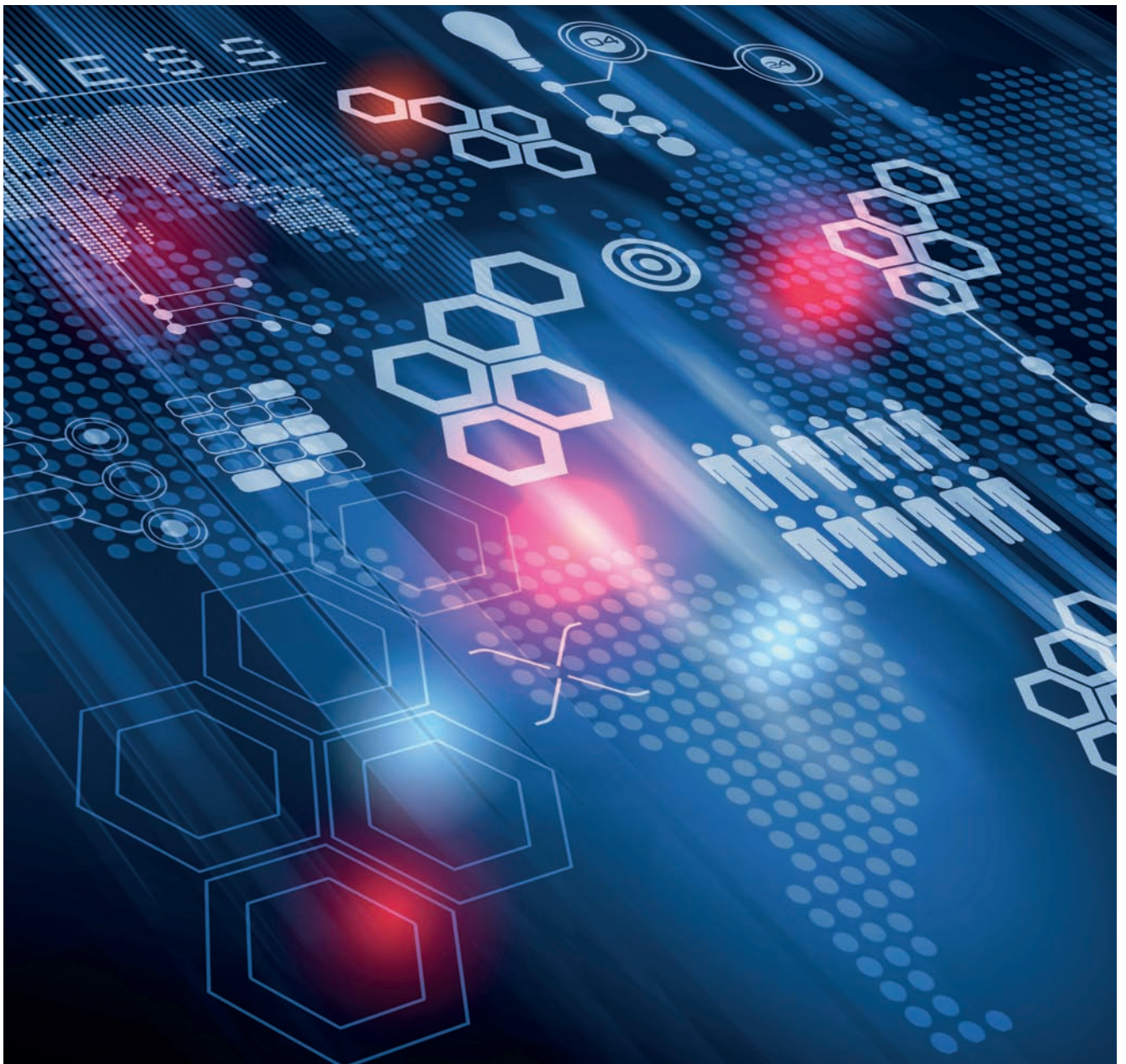


Source: Kurzweil (2014).

Executive summary

*If you can't explain it simply,
you don't understand it well enough.*

Albert Einstein¹⁰



A financial area under transformation

1. Financial institutions are facing an unprecedented technological revolution that is transforming the world in terms of what can be done and the cost at which it can be done, and that is consequently having a substantial impact on their activity. This revolution is evident in the creation of and access to information as well as in how data are stored, processed and modeled.
 - ▶ The speed at which information is generated is increasing at a breathtaking pace: it is estimated that the total volume of data in the world doubles every 18 months¹¹. For years now, these data have for the most part been digital, and 50% of them are already accessible via the Internet, with 80% being unstructured (videos, images, emails)¹¹. Moreover, much of these data come from new sources: social networks, activity logs, etc. As a result, the big data phenomenon is being characterized by a surge in the «three Vs»: volume, variety of sources and velocity in data generation.
 - ▶ Also skyrocketing is access to information through mobile devices. The global cell phone market is close to saturation point, and smartphones will reach a 51% share in 2016¹². Although there still is room for growth, smartphone ownership is growing at a slower rate, which suggests there will be a new replacement technology in the near future, possibly having to do with the «Internet of things»¹³ and with wearable¹⁴ devices such as glasses, watches, etc. that will integrate mobile technology.
 - ▶ Storage capacity is also growing exponentially, and its unit cost is decreasing at the same rate: storing 1 GB of data cost \$10 million in 1980, whereas today it barely costs ten cents of a dollar¹⁵. This has led to the amount of stored information in the world becoming massive:

in 2015 there are 8 ZB of data, three times that of 2012^{16,17}.

- ▶ The same phenomenon can be seen in data processing: the capacity to execute instructions per second for each thousand dollars-worth of processor time has increased almost 300-fold since the year 2000¹⁸. Furthermore, new developments in distributed computing mean operations can be run in parallel across multiple nodes and, backed by technology giants and retailers like Google or Amazon, is emerging as the future of processing. In the financial industry, digital banking and information system requirements mean institutions need higher processing capabilities, and are therefore already adopting high performance, distributed computing equipment.
- ▶ Modeling capabilities are rapidly evolving driven by new technologies and data availability, opening a horizon of possibilities that were previously unthinkable¹⁹. The number of decisions made automatically using models in financial institutions is

¹⁰Albert Einstein (1879-1955). German physicist (later acquired Swiss and US citizenship), who is the author of the theory of relativity, one of the two pillars of modern Physics.

¹¹Federal Big Data Commission (2014).

¹²International Telecommunication Union (2014).

¹³Digital interconnection of everyday objects with the Internet. An estimated 26 billion devices will be connected to the Internet of Things in 2020. (Gartner, 2013)

¹⁴Clothing or accessories that incorporate advanced electronic technology.

¹⁵McCallum (2014).

¹⁶SiliconAngle (2014).

¹⁷The financial industry also reflects this upward trend: each institution handles 1.9 PB on average in their information systems (DeZyre, 2014), which is promoting the use of distributed storage platforms.

¹⁸Kurzweil [Director of Engineering at Google] (2014).

¹⁹E.g. the rapid increase of «machine learning», which allows for more agile modeling (5,000 models/year with 4 analysts vs. 75 models at present).

multiplied each year, which undoubtedly results in benefits (efficiency, objectivity, automation), but also carries risks.

2. All this is turning data into a new commodity: data are produced, stored and processed at a very low cost, are consumable (become obsolete) and, suitably processed, have the potential to contribute enormous value. All this requires professional and specialized tools, in other words: data science.
3. Financial services should be one of the industries benefitting most from data science, especially since it is an area that handles the largest volume of data and requires the greatest quality of information from its (current and potential) customers in order to extract knowledge and incorporate it into its value proposition (an understanding of the customer's needs, a customized offer, a suitably adjusted multi-channel relationship model, etc.).
4. However, in the financial industry this technological revolution takes place together with a particularly complex environment that combines cyclical elements with strong regulatory pressures and changes in customer behavior.
5. In the developed economies, the current macroeconomic environment is characterized by a prolonged period of low interest rates, weak growth and low inflation, which penalizes banks' profit margins. Emerging economies, highly dependent on public investment and expansionary fiscal policies, are seeing a slight slowdown in bank lending activity, slower growth and rising default levels. These factors make it necessary to increase the focus on managing the income statement.
6. In terms of regulations, there is a «regulatory tsunami» marked by the proliferation, harmonization, and tightening of standards, as well as their transnational and more intrusive nature, in several areas: (1) capital and liquidity:

capital buffers, liquidity and leverage ratios, review of requirements for credit, market and operational risk; (2) prudential supervision: strengthening of SREP, ICAAP and ILAAP²⁰, supervisory stress tests; (3) limitation of public support: recovery and resolution plans, TLAC and MREL, ring-fencing; (4) corporate governance: increased demands on the Board and Senior Management, new roles (CRO, CDO, CCO²¹, etc.); (5) consumer protection: Compliance function, quality control, complaints management, conduct risk; (6) the fight against fraud and tax havens: FATCA, tougher penalties for money laundering; and (7) cybersecurity and information security: FISMA, Budapest Cybercrime Convention, EU Network and Information Security Directive, ISO 27032; and (8) information and reporting: RDA&RRF, COREP, FINREP, FR Y-14, FR Y-9C, etc.

7. This costly regulatory compliance process is however a differential element for bank customers as it gives them access to safer and more supervised regulated processes, something that financial institutions will eventually leverage against new competitors.
8. Bank customers have become more demanding, are permanently connected (using their cell phones 110 times a day²²) and consult the social media before buying. In addition, they no longer perceive banks as the sole providers of financial services or branches as the basic relationship channel, and have become used to being offered personalized services. At the same time, customers

²⁰SREP: Supervisory Review and Evaluation Process; ICAAP: Internal Capital Adequacy Assessment Process; ILAAP: Internal Liquidity Adequacy Assessment Process.

²¹CRO: Chief Risk Officer; CDO: Chief Data Officer; CCO: Chief Compliance Officer.

²²KPCB (2014).



show signs of confusion at the huge variety and complexity of the products and services offered, which favors the influence of opinion leaders. This change is forcing companies to rethink their offering and channels as well as to adopt a more customer-focused approach, with a significant impact on all areas.

9. In addition, new competitors with new business models are entering the financial sector. These new entrants come from sectors that are not subject to strict banking regulations but have a brand image that is perceived very favorably by consumers.

Data science: an emerging discipline

10. The environment that has been described is causing the financial industry to adopt an emerging discipline, largely inherited from the IT industry, which is necessary to tackle the transformation institutions are facing: data science.
11. What essentially characterizes data science²³ is its scientific nature, as it approaches value extraction from data through a scientific method, the «data science process»: formulating a question or hypothesis; obtaining information from various massive and possibly unstructured data sources to answer it; exploring data using descriptive statistics; modeling the phenomenon with the available data; and visualizing and communicating the findings, which will either confirm or refute the hypothesis or question asked.

12. Data science therefore entails the development of traditional modeling, largely driven by the big data environment, and uses innovative tools and techniques²⁴ that allow for data self-service, mobile access, data merging from various sources, non-relational connectivity, use of the cloud and interactive data visualization²⁵.
13. Thanks to these capabilities, adopting data science allows institutions to ask and answer questions regarding customers and their environment, and even regarding the organization itself, which were previously unthinkable in all areas (risk, marketing, finance, operations, etc.).
14. By way of illustration, it is already possible to enhance credit scoring models with information from social networks and the digital footprint, improve models for estimating income by cross-matching publicly available data on the network with geolocation, prevent customer turnover by analyzing recordings from call centers through natural language processing or detect fraud and money laundering by identifying behavior patterns in activity logs, among many other possibilities.
15. The transition towards these capabilities, however, is not without challenges. The cost and difficulty of handling massive amounts of data, issues related to privacy, ethics and safety in data management, recruiting and training talent in data science, the risk of relying on automatic models to make many important decisions, and the governance of data and models.



²³Beyond the data science definition in the introduction, most studies analyze the skills and knowledge needed by data scientists: (1) training in Mathematics, Physics, Statistics, etc., and machine learning algorithms, optimization, simulation or time series, among others; (2) technological skills, proficiency in languages such as SAS, R or Python, relational and non-relational database management and use of platforms such as Hadoop; and (3) in-depth knowledge of the business, which is key to successful models.

²⁴New devices and customer relationship channels, new payment methods, Business Process Management (BPM) solutions, social networks as a channel for contracting, brand monitoring and handling complaints, distributed and horizontally scalable systems, infrastructure as a service, new databases (NoSQL and in-memory), real-time data capture and processing, new ETL and query engines for unstructured data, data discovery tools and new programming languages, and new on-line data visualization and exploitation tools.

²⁵To gauge the importance of this, it should be noted that, in February 2015, President Barack Obama created the role of Chief Data Scientist and personally appointed Dhanurjay «DJ» Patil.



16. Financial institutions have gradually adapted to the phenomenon described by transforming their data generation and reporting processes, though in many cases in an unstructured manner and rather a result of incremental requests from supervisors and regulators, of unplanned management needs or entity integration processes.
17. Regulators have identified information gaps as one of the causes of the financial crisis, which has led to the publication of specific regulations with strong requirements for data quality, consistency, integrity, traceability and reproducibility (especially for risks²⁶). This leads to the need to review institutions' data governance frameworks.
18. The data governance scheme should be developed into a management framework describing the principles, the participants (with new roles such as the CDO²⁷), committee structure, critical processes related to data and information, tools (data dictionary, data warehouse architecture, exploitation solutions, etc.), and data quality control.
19. Data governance presents several challenges, including the involvement of Senior Management, defining the scope of the data to be governed by the framework, aspects of privacy and security in the use of data and cybersecurity (including protection against «hactivism», financial cybercrime, espionage and theft of information), or adapting to novel storage architectures such as data lakes²⁸. But the relevance such governance has acquired in the

²⁶Risk Data Aggregation and Risk Reporting Framework; see BCBS (2013).

²⁷Chief Data Officer.

²⁸Massive repositories of unprocessed data (in their original format).

management of institutions is now a fact, having become a necessary condition for the proper provisioning of available data and, in short, a strategic pillar for institutions.

20. Regulations also insist on the need to have a framework in place to govern models²⁹. The elements of this framework were discussed in detail in previous Management Solutions publications³⁰.
21. More advanced institutions in this area have already developed model risk management frameworks to govern model inventory and classification as well as documentation and a monitoring scheme.
22. The governance of models also presents challenges, including Senior Management involvement, considerations on perimeter (what is a model and which models should be governed), segregation of functions (ownership, control and compliance³¹), effective challenge or the use of model inventory and workflow tools, among others. But it is unquestionable that model governance is an issue that requires involvement at the highest level because institutions depend on it to make the right decisions.
23. In short, data and model governance is a strategic element for financial institutions, driven by regulations and in response to the big data phenomenon. This issue impacts different areas of an institution, from organization, to policies and procedures, to information tools and systems, and will be a key area of action in the coming years.

Case study: social networks and credit scoring

24. To illustrate some of the concepts that have been described through a case study, a credit scoring model that uses data from social networks is presented, integrated with a traditional model. An analysis is also offered on how incorporating such data improves the traditional model's predictive power.
25. Some of the findings are: (1) data from social networks are of much lower quality than internal data; less than half of customers have data, of which few are complete; (2) there is a disambiguation³² problem to uniquely identify each customer; (3) despite this, the predictive power of the model that is based on social networks is comparable to the traditional model, and when combined the resulting predictive power increases³³; (4) for this, variables related to the customer's formal and non-formal qualifications, professional experience, geographic location and other information on hobbies and interests have been used.
26. The study shows the potential of applying data science to the field of risk, and can be extended to other types of models (valuation of collateral, customer loyalty programs, income, attrition, propensity to buy, etc.) and information sources (internal logs, public databases, web information, etc.).
27. Data science is emerging as a multidisciplinary field that opens up new possibilities for the financial industry by applying a scientific approach to the big data phenomenon, taking advantage of the surge in information and technology capabilities to increase intelligence in organizations³⁴.



²⁹OCC/Fed (2011-12).

³⁰See Management Solutions (2014): Model Risk Management: quantitative and qualitative aspects of model risk management.

³¹The model owner defines the model requirements and is usually the end user. Control includes measuring model risk, setting limits and follow-up as well as independent validation. Compliance includes processes that ensure the model owner and control roles are performed according to established policies.

³²A technique that makes it possible to uniquely identify a customer from among several who share the same name and other characteristics (location, age, etc.).

³³Measured through ROC (receiver operating characteristic), a predictive power metric for a binary classification model.

³⁴Capacity to receive, store and process information to solve problems.

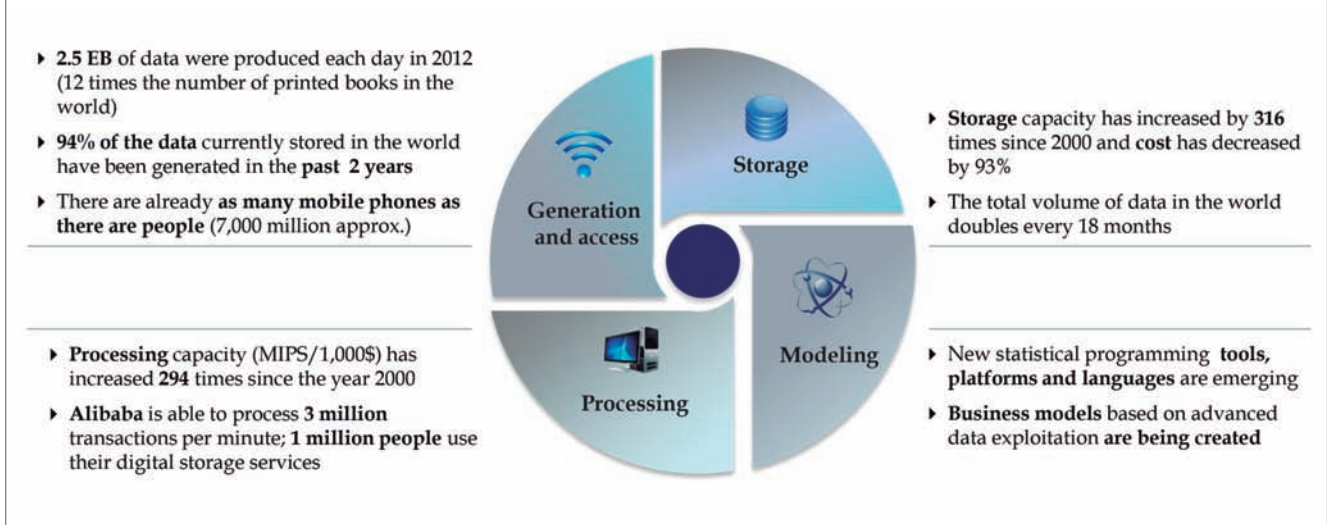
A financial industry undergoing transformation

The technology and infrastructure needed to connect people is unprecedented, and we believe this is the most important problem we can focus on.

Mark Zuckerberg³⁵



Fig. 2. The IT revolution: a snapshot of the current landscape.



The technology revolution

The financial industry is facing a technological revolution of unprecedented magnitude and reach that is substantially transforming its activity.

This revolution is above all marked by its rapid pace of development. As predicted by Moore's Law, technology power, measured by different criteria, is growing exponentially, and predictions suggest that this phenomenon will continue.

To understand this phenomenon and its implications, it is necessary to observe it in four areas: data generation and access to information, storage, processing and modeling (Fig. 2).

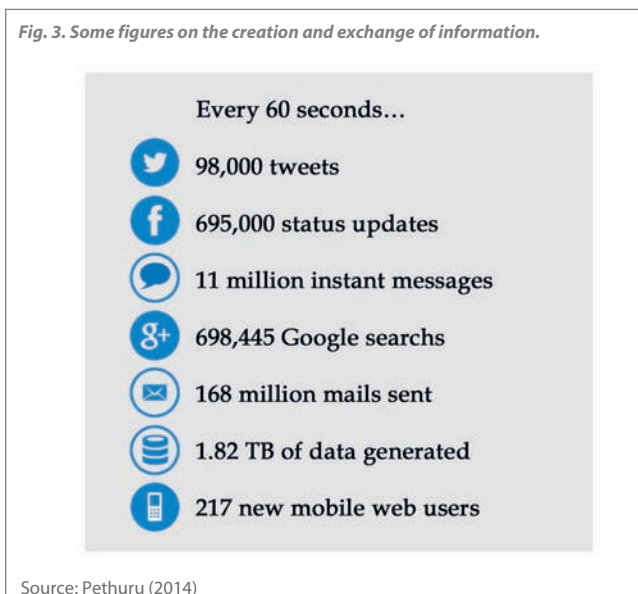
The first three areas will be discussed in this section, while the detail on modeling and data science will be addressed in the next section.

Data generation and access to information

The first aspect of this phenomenon is the increased speed at which digital data are generated. Estimates of the velocity and acceleration of digital data generation vary depending on the analyst, but they all agree that this acceleration is exponential in all areas; to cite some examples^{36,37} (Fig. 3):

- ▶ In 2012, 2.5 exabytes of data were generated every day; this rate has continued to rise.
- ▶ Over 90% of all data that exist today were created in the last two years.
- ▶ 12 hours of video are uploaded to YouTube every minute.
- ▶ 12 terabytes of tweets are created in Twitter every day.
- ▶ 5 billion financial transactions are conducted every day.
- ▶ In 2012 there were 2.4 billion Internet users in the world, almost half of them in Asia ...
- ▶ ... they exchanged 144 billion emails a day, of which 69% were spam.
- ▶ Also in 2012, Facebook surpassed the 1 billion user mark, and in 2016 it is expected to have more users than the population of China.

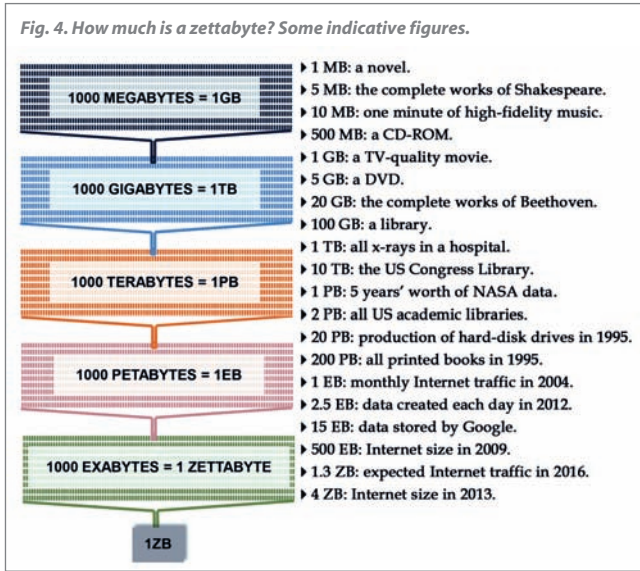
Fig. 3. Some figures on the creation and exchange of information.



³⁵Mark Elliot Zuckerberg (b. 1984). Co-founder and CEO of Facebook.

³⁶IBM (2014a).

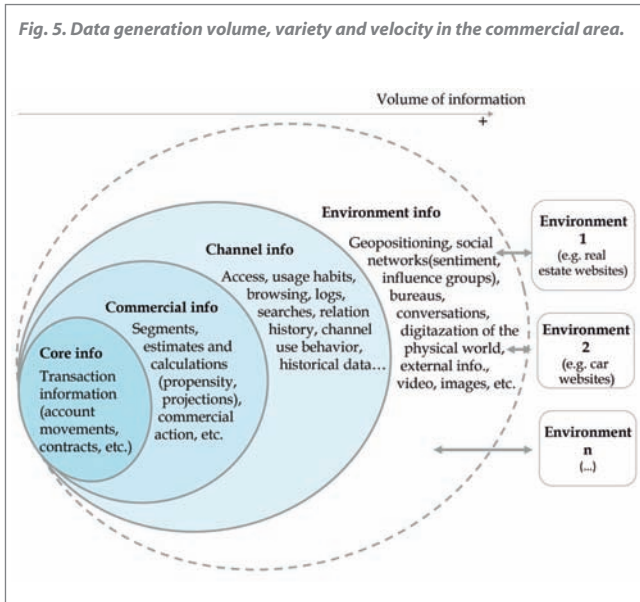
³⁷Pingdom (2015).



The Federal Big Data Commission, to which the US government has entrusted the task of understanding the big data phenomenon in government agencies³⁸, describes it like this:

In recent years, federal, state and local government agencies have struggled to navigate the tidal wave of sheer volume, variety, and velocity of data that is created within their own enterprise and across the government ecosystem. [...]

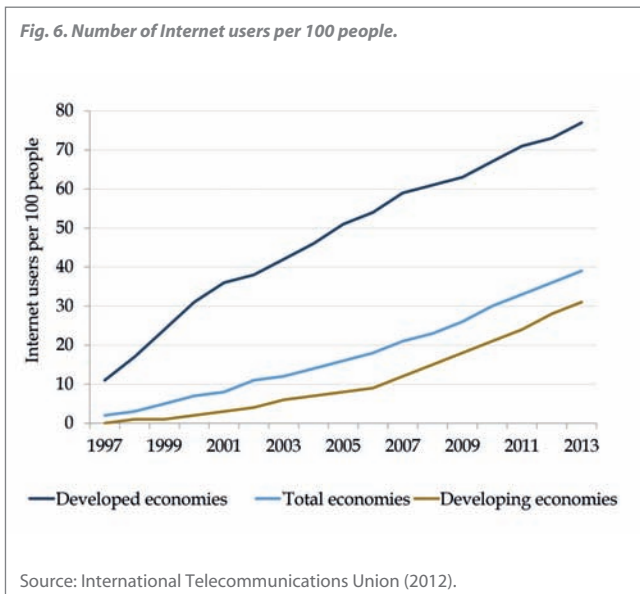
Since 2000, the amount of information the federal government captures has increased exponentially. In 2009, the U.S. Government produced 848 petabytes of data and U.S. healthcare data alone reached 150 exabytes. Five exabytes (10¹⁸ gigabytes) of data would contain all words ever spoken by human beings on earth. At this rate, Big Data for U.S. healthcare will soon reach zettabyte (10²¹ gigabytes) scale and soon yottabytes (10²⁴ gigabytes).



These dizzying figures are hard to imagine; see Fig. 4 for an indicative reference on each of these magnitudes.

These data are no longer generated only in a structured way; on the contrary, 80% of the data being generated every day are unstructured: videos, pictures, emails, etc., and come from a wide variety of new sources: social networks, sensors, internet browsing records, activity logs, call logs, transactions, etc.

In other words, the big data phenomenon is an explosion in the volume, variety and velocity of data generation, which have been called «the three V's of big data» (some authors add a fourth V: «veracity»). As an example, the commercial side of financial institutions is experiencing a surge in these three areas, as it has developed from core transaction data, to commercial data from channels, to customer environment data, which is very rich, varied and heterogeneous (Fig. 5).



Regarding access to information, although differences remain between developed and emerging economies, and some countries have little access to the Internet, the trend is clear. It is estimated that, in a few years' time, full use of the Internet will be widespread in most parts of the world (Fig. 6).

There has also been a dramatic increase in the level of information accessed through mobile devices. The number of mobile phones already equals the number of people in the world, and in developed economies it exceeds it by 21%³⁹, while developing economies are close to parity with nine mobile lines for every ten people.

³⁸Federal Big Data Commission (2014).

³⁹International Telecommunication Union (2014).

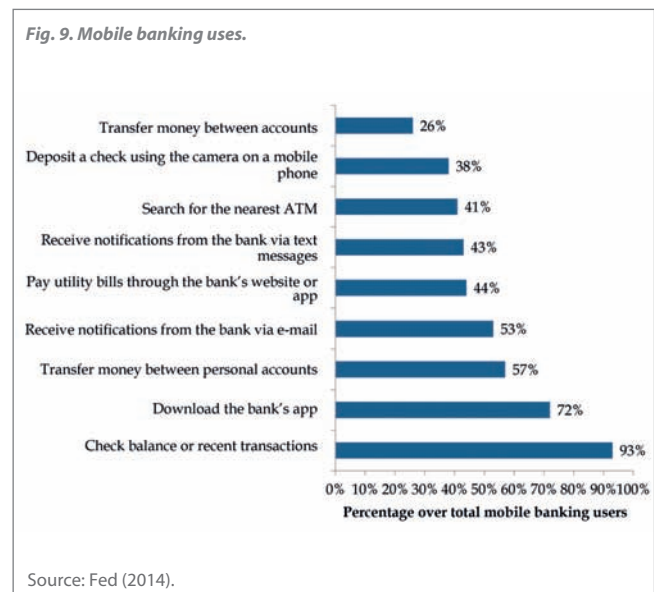
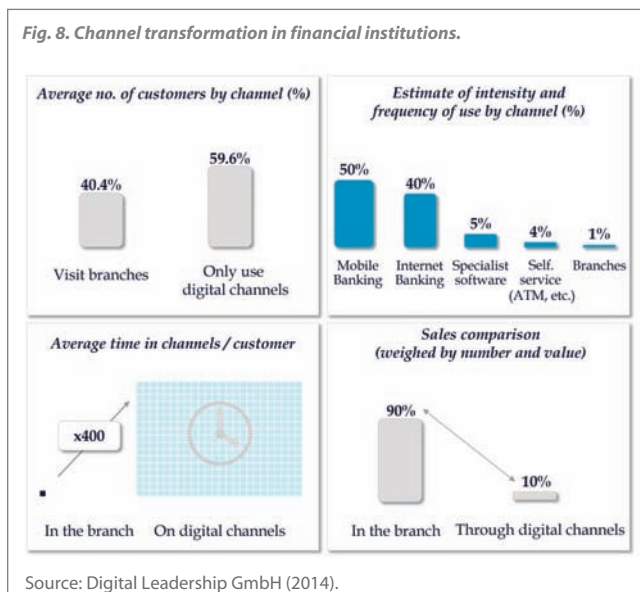
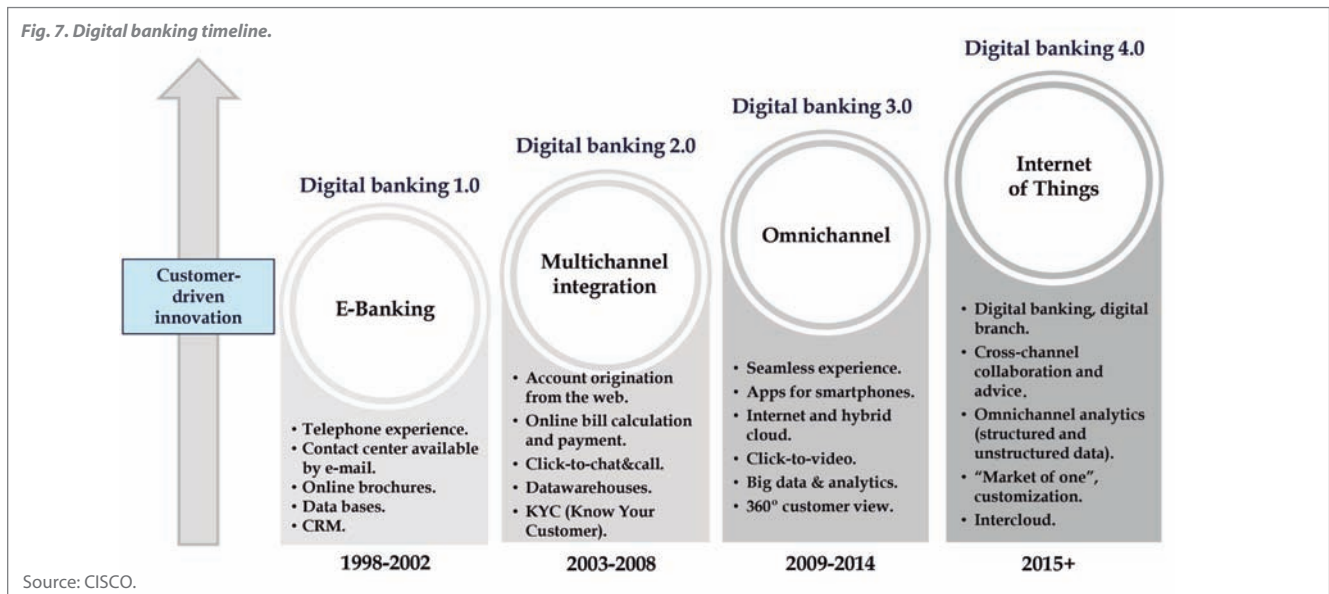
Financial institutions have not been exempt from this proliferation of devices, and in developed countries digital banking, which encompasses mobile banking and beyond, is in constant evolution, largely driven by customers themselves (Fig. 7). From the e-banking and multichannel integration that took place in the first decade of the twenty-first century, through the omni-channel phenomenon that has occurred in recent years, the trend going forward is marked by the so-called «Internet of things»: the prevalence and ubiquity of smart devices, capable of generating digital information on their own.

In the US, nearly 60% of customers already operate via digital channels only, spend 400 times longer on these channels than they do in branch offices and only 1% of transactions are carried out from the branch. However, 90% of sales continue to take place in the branch (Fig. 8).

Development is therefore uneven, and many institutions are still in the process of adapting to mobile banking; most have developed applications to provide mobile access to key services. According to a study conducted by the US⁴⁰ Federal Reserve, more than 50% of smartphone users in the country have used mobile banking in the last 12 months and the most common financial uses (Fig. 9) are checking account balances and recent transactions (93%) and transferring money between own accounts (57%).

The trend is clear: digital banking is gaining market share and, since customers under 29 years use digital channels almost four times as much as they use traditional channels, growth is likely to continue along this path in the coming years. However, monetization of this phenomenon can still be improved.

⁴⁰Fed (2014).



Data storage

Along with data generation and access to information, storage capacity is also growing exponentially in line with Moore's Law, and unit costs continue to decline at the same rate (Fig. 10). While in 1980 storing a gigabyte required devices worth 10 million dollars, today it only costs ten cents of a dollar to store it in a minute fraction of an SSD hard drive.

Storage devices have rapidly evolved from magnetic tapes in 1920, to CRT in 1940, the first hard disk in 1956 (Fig. 11), cassettes in 1963, DRAM memory in 1966, floppy disks in the 1970's, CD's in the 1980's, zips and DVD's in 1994 and 1995, flash cards in 1995, MMC cards in 1997, pendrives in 1999, SD cards in 2000, Blu-Ray in 2003, to solid memory and cloud storage in the 2010's. Note how the number of new formats has increased exponentially with each decade, as has the capacity of the devices.

This has caused the amount of information stored in the world to grow massively in recent years. There was an estimated total of 2.75 zettabytes of digitally stored data in 2012, and this figure is estimated to have reached 8 zettabytes in 2015 and continues to grow⁴¹.

In the financial industry, storage systems have evolved in parallel with the need to collect and manage large amounts of information. In the late 1970s, banks began implementing host servers, technological environments whose purpose was to receive, process and store all the information generated through transaction management. Later, the incorporation of information systems made it possible to decouple massive data querying from operational processes, opening the door to much larger amounts of data and to data archiving.

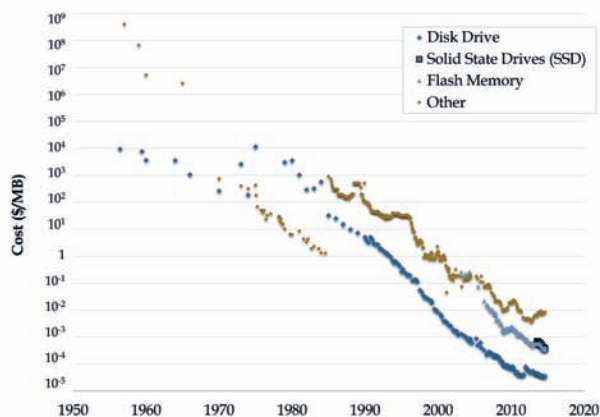
This trend was reflected in the creation of data warehouses, structured information repositories with four main characteristics:

- ▶ Focused: stored data are structured and grouped by theme.
- ▶ Updated: they make it possible for new information to be added over time.
- ▶ Historical: they keep a record of historical data, it is not necessary to remove them.
- ▶ Integrated: they ensure consistency of data registered by different systems.

At the beginning of the twenty-first century, the financial industry added new technology to modernize and optimize data storage, with large dedicated systems that used software and hardware designed specifically for data exploitation. At the same time, more sophisticated query software began to be implemented, allowing the user greater freedom (OLAP tools, query & reporting tools, etc.).

Currently, financial institutions already handle around 1.9 petabytes in their information systems⁴², which challenges the implemented architecture. As a result, organizations are experiencing a new revolution in information storage systems: distributed storage platforms. This technology allows unstructured data to be stored on a massive scale and to be managed through nodular architecture.

Fig. 10. Storage costs, in dollars per megabyte (note that the Y axis is logarithmic).



Source: McCallum (2014).

Fig. 11. IBM 350 hard-drive, from 1956, with a capacity of 5 megabytes.



⁴¹SiliconAngle (2014).

⁴²DeZyre (2014).

Some major financial institutions, and in the United States at least three of the five largest banks⁴³ have already adopted distributed storage platforms and are beginning to exploit their storage and data processing potential, albeit still in a limited way.

Data processing

Like data generation and information storage, processing power is also experiencing rapid growth⁴⁴. Because of lower costs, the capacity to process one million instructions per second per thousand dollars of CPU has increased almost 300-fold since year 2000. This means some retailers are able to process millions of commercial transactions per minute, which is at the core of their business model.

Moreover, the emergence of distributed computing has made it possible to combine the capabilities of a large number of processors (nodes) to execute operations in parallel. It is estimated⁴⁵ that in 2012 Google had some 7.2 million nodes in more than 1.8 million machines, and their combined power was able to execute some 43 trillion operations per second, about four times the capacity of the world's most powerful machine (Fujitsu K). The same study estimated that Amazon could reach 24 trillion operations per second.

The main technology players, aware that their distributed processing capacity is itself a valuable service, sell access to distributed computing in their respective clouds; surplus capacity allows them to offer it for rent under one dollar per hour⁴⁶.

In the financial industry, processing requirements for traditional banking activities (through branches or ATMs) is supported at an acceptable level by the technology currently in place. However, the new channels such as digital banking require an increase in parallel transaction processing capacity.

The sophisticated modeling techniques in today's information world mean institutions require increased computing and mass data processing power. By using parallel or distributed computing techniques, which are closely linked to the storage structure itself, institutions are able to take full advantage of information while reducing processing time.

A new commodity: data

This surge in the ability to generate, store and process information, and to access it anytime, anywhere, through mobile devices is causing a new phenomenon: data have become a new commodity. Indeed, data are generated, stored and processed at a very low cost, are consumable since they quickly become outdated, and are a raw material which, when transformed, result in services of all kinds.

However, this new commodity has two particular features: like energy, it has become essential for the operation of most businesses, including financial services; and, like all commodities, it requires professionals and specialized tools in order to be processed. This is precisely the field of data science.



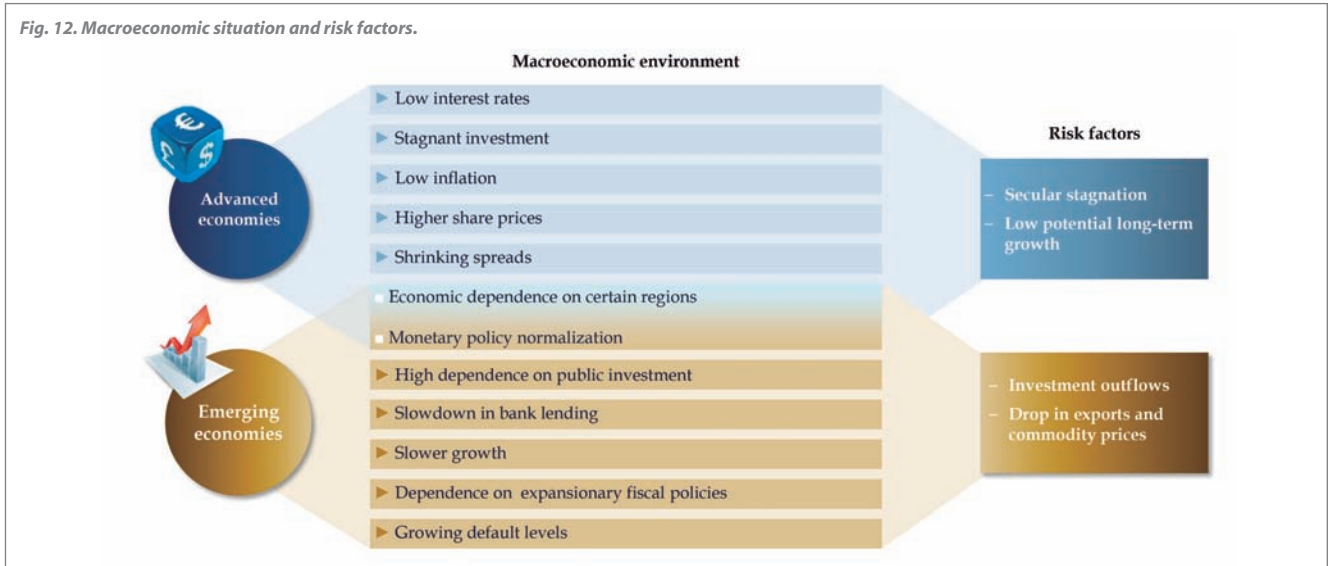
⁴³Goldman (2014).

⁴⁴See graph in the Introduction.

⁴⁵Pearn (2012).

⁴⁶For 8 nodes with 30 GB of RAM in Amazon, Google and Windows, for instance.

Fig. 12. Macroeconomic situation and risk factors.



Financial industry environment

While this technological revolution significantly impacts all industries, the financial industry is expected to be among those that will benefit most from the adoption of data science as a strategic pillar. This is the industry that handles the greatest amount of data and quality information from customers and targets. Therefore there is huge potential for the industry to extract knowledge and incorporate it in its value proposition, which distinguishes it from other industries.

However, in the financial industry this technological revolution is taking place in a singular context, characterized by a difficult macroeconomic situation, a demanding regulatory environment and a change in customer behavior patterns, a number of factors that are not affecting other industries in the same way.

Macroeconomic situation

From a macroeconomic perspective (Fig. 12), the dual nature of the world economy (developed vs. emerging countries) continues to be the case in terms of growth, inflationary pressures and investment flows. As a result, some macro magnitude patterns (Fig. 13) are maintained which affect banking business trends in terms of both sources of funding and investment and financial margins.

In advanced economies, a prolonged scenario of low interest rates has led to a relative increase in share prices, tighter spreads and a general decline in volatility to pre-crisis levels. However, this has not led to a surge in investment, unlike savings levels, which have risen and resulted in weaker private demand. Thus, in advanced economies this stagnation (which has been described as «secular stagnation»⁴⁷) is expected to continue for several years, though unevenly across countries.

There is also ever increasing evidence that potential growth in advanced economies began to decline before the outbreak of the financial crisis due to the impact of an aging population on the labor force and weak growth in total factor productivity⁴⁸.

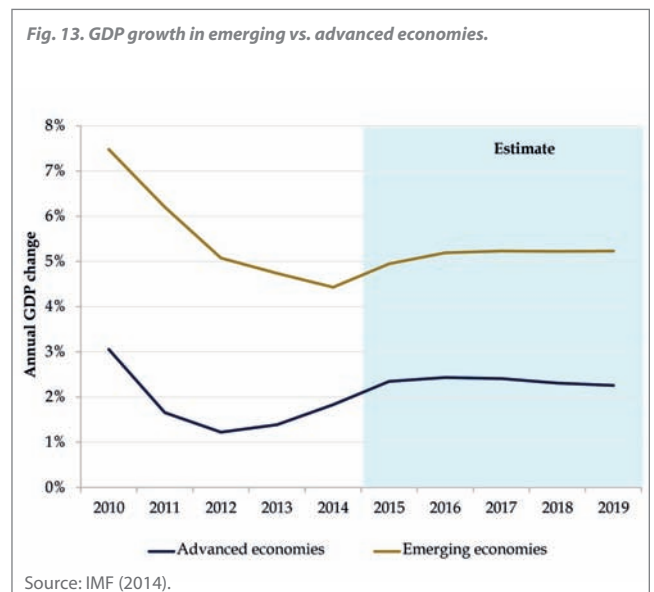
Moreover, in some advanced economies inflation is low or there are signs of deflation, and there is no margin for reduction of reference rates. All this has a negative impact on credit growth and on the margins of financial institutions.

In the case of emerging economies, growth has slowed but still remains relatively high compared to advanced economies. Private consumption contributes significantly to this growth, although there is also a high dependency on public investment and expansionary fiscal policies.

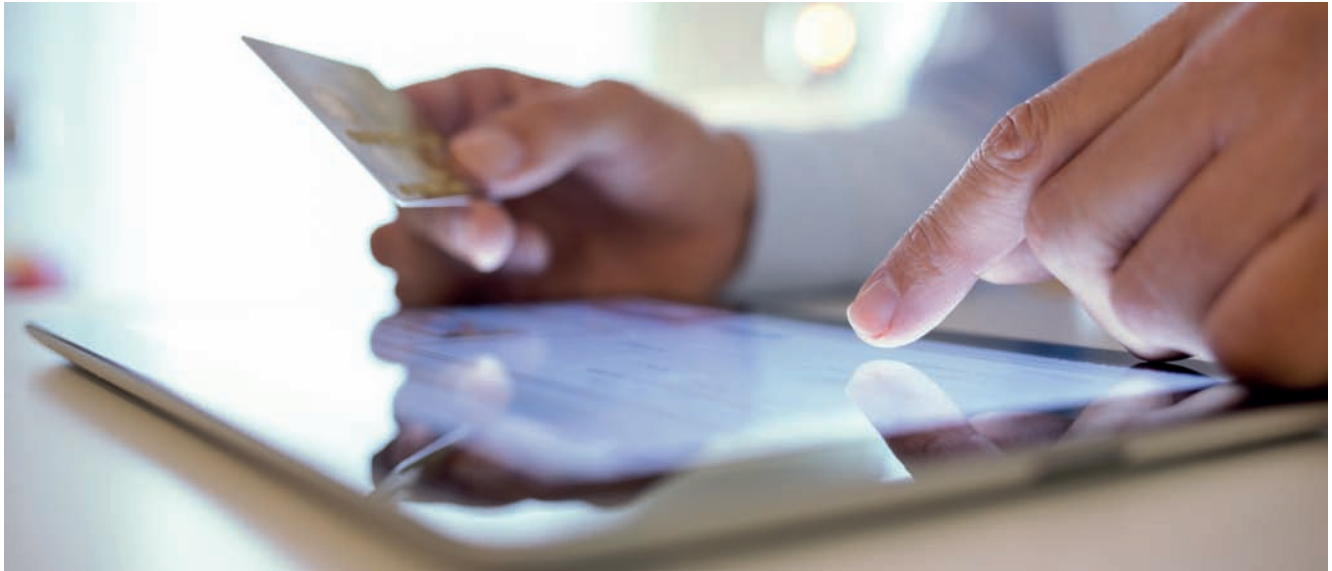
⁴⁷IMF (2014).

⁴⁸For the US, see for instance: Fernald (2014), Gordon (2014) and Hall (2014).

Fig. 13. GDP growth in emerging vs. advanced economies.



Source: IMF (2014).



Although bank credit expansion is slowing down in some emerging markets (Brazil, India, Russia), two-digit growth rates (Fig. 14) are being maintained. Moreover, the default rate is generally increasing in emerging economies due to several causes, which include the problems faced by certain areas (such as the mining industry in Peru or the public sector in Chile and Brazil⁴⁹) and the incorporation of new customers previously without banking services and with a worse credit profile.

Finally, there are two elements that create uncertainty in this environment. On the one hand, economic dependence on China, whose slowdown could cause a massive contraction in exports in the rest of the world, lower the price of raw materials and result in a decrease in consumer and business confidence indices. In this respect, growth-related risks are seen in China due to excess production capacity and an overabundance of credit, which are the main drivers of its growth.

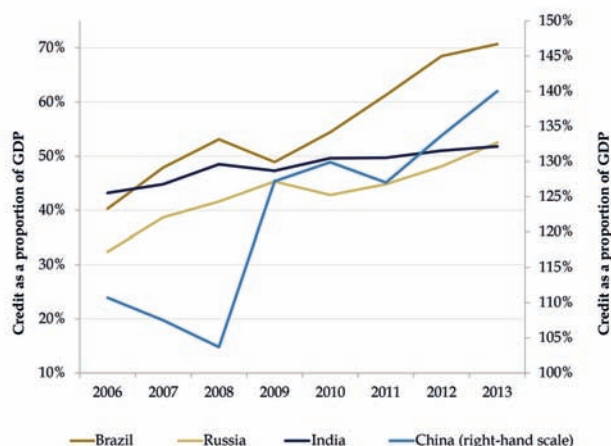
On the other hand, monetary policy normalization in the United States, Japan and the European Union, which in recent years has expanded to countries like Chile, Mexico and Peru, poses a risk for the possible deflationary effect and for attracting investment flows from emerging countries.

This macroeconomic situation is causing financial industry margins to narrow, but is also introducing pressure on capital and liquidity. The consequence has been that banks have increased efforts towards profitability, capital and balance-sheet structure management, providing more analytical intelligence and risk insight while paying attention to the economic outlook and its potential impact.

Regulatory environment

The financial sector is experiencing a remarkable surge, which could be described as a «tsunami», in both supranational and local regulation in the financial areas that have had the most influence on the crisis that began in 2007: accounting, prudential supervision, conduct and compliance, corporate governance, consumer protection and risk, in a broad sense.

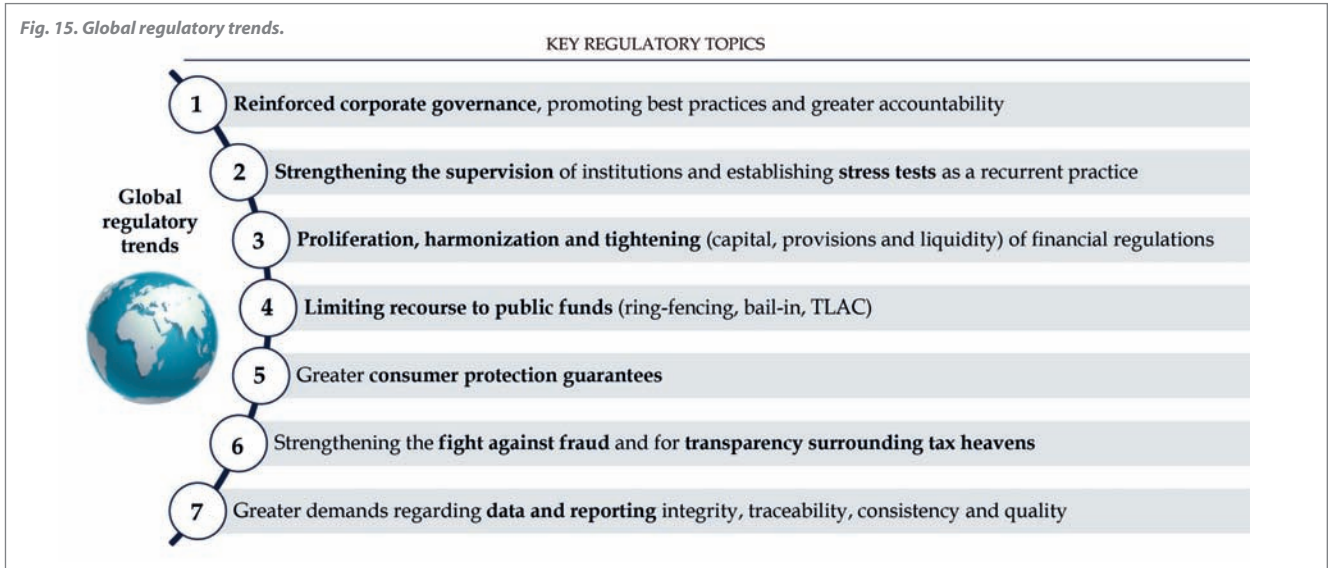
Fig. 14. Growth in credit and credit relative to GDP in emerging economies.



Source: IMF (2014).

⁴⁹BBVA Research (2014).

Fig. 15. Global regulatory trends.



There is a trend towards harmonization of standards across countries, for which the creation of supranational regulators and supervisors such as the European Banking Authority (EBA), the European Central Bank’s Single Supervisory Mechanism (SSM) or the Financial Stability Board (FSB) is a contributing factor.

At the same time, rules are becoming more intrusive and prescriptive, and leave less room for adaptation and interpretation. As an example, in the European Union Basel III has been adopted through a combination of a Regulation (therefore immediately applicable in all EU countries) and a Directive (which must be transposed into local legislation), while Basel II was adopted only as a Directive.

Specifically, this proliferation and harmonization of financial regulations is resulting in more restrictive rules in several areas, including most notably (Fig. 15):

- ▶ Capital and liquidity: Basel III resulted in greater capital requirements (in terms of both quantity and quality), a new leverage ratio and two liquidity ratios (short and long term⁵⁰). Capital requirements for credit, market and operational risk⁵¹ were also reviewed and simplified.
- ▶ Reinforced prudential supervision: common guidelines⁵² for the supervision of financial institutions were established, thus reinforcing the SREP, ICAAP and ILAAP processes⁵³ (especially in Europe with the entry into force of the Single Supervisory Mechanism in November 2014). Also, supervisory stress tests⁵⁴ were strengthened and established as a recurrent practice.

⁵⁰Management Solutions (2012).

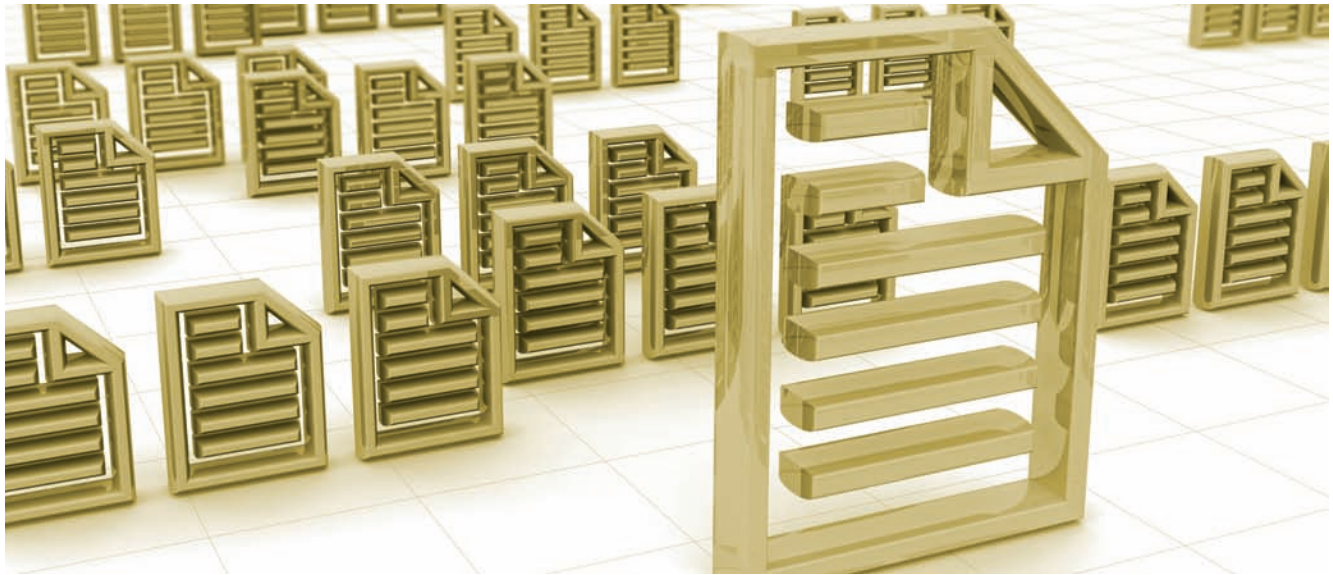
⁵¹In this regard, in 2014 the Basel Committee on Banking Supervision issued several documents that review the Standardized approach to credit risk, simplify the calculation of capital for operational risk, set capital floors and review the calculation of capital requirements for portfolio risk trading, among others.

⁵²EBA (2014) and ECB (2014).

⁵³SREP: Supervisory Review and Evaluation Process; ICAAP: Internal Capital Adequacy Assessment Process; ILAAP: Internal Liquidity Adequacy Assessment Process.

⁵⁴Management Solutions (2013).





- ▶ Limited public support: especially in advanced economies, the aim is for the State not to have to rescue financial institutions with public funds because they are systemic, which is known as «the end of the too-big-to-fail». For this, institutions are required to have recovery and resolution plans⁵⁵ in place, as well as a sufficient amount of liabilities capable of absorbing losses (TLAC, MREL); and, in the European Union, an authority was created to manage the resolution of unviable financial institutions (Single Resolution Board). The US and the UK, and later other European Union countries, encouraged regulations on ring fencing, which requires legal separation between wholesale and traditional banking activities.
- ▶ Strengthened governance: greater demands were placed on the Board of Directors and Senior Management concerning the approval and compliance supervision of the business strategy, risk appetite and risk management framework, and new key roles (CRO, CDO, CCO⁵⁶, Risk MI corporate functions, etc.) were created.
- ▶ Consumer protection: financial sector scandals concerning products, distribution channels, payment technology, market abuse and money laundering gave rise to more intensive and prescriptive regulations⁵⁷ requiring the reinforcement of the Compliance function (resources, assets, capabilities and reporting lines), quality control (mystery shopping) and complaints management policy, in addition to a stronger focus on conduct risk measurement, management, control, monitoring and reporting vis-à-vis markets and customers. This trend was championed by the UK with the creation of a specific body, the Financial Conduct Authority, which only between 2013 and 2014 issued almost 2,000 million pounds in fines for conduct matters⁵⁸.
- ▶ Fight against fraud and tax havens: due to increased fraud as a result of intensive use of electronic channels and constant changes in organizations, steps were taken to regulate the need for intensive control of internal and

external fraud. Some countries brought forward aggressive policies to avoid tax evasion by citizens (e.g. FATCA⁵⁹). Requirements were increased regarding the fight against money laundering (high penalties⁶⁰), which requires financial institutions to make significant changes to their processes and systems.

- ▶ Cybersecurity: specific regulation is being issued to combat the increasing attacks to the security of financial institutions («hactivism», financial cybercrime, espionage, theft of information, etc.): in the USA, the Federal Information Security Management Act (FISMA), among others; in Europe, the Budapest Cybercrime Convention or the Network and Information Security Directive; and in the global domain, the ISO 27032, which provides specific guidelines on cybersecurity.

⁵⁵In the EU, through the Bank Recovery and Resolution Directive (BRRD), summarized by the European Commission (2014).

⁵⁶CRO: Chief Risk Officer; CDO: Chief Data Officer; CCO: Chief Compliance Officer; MI: Management Information.

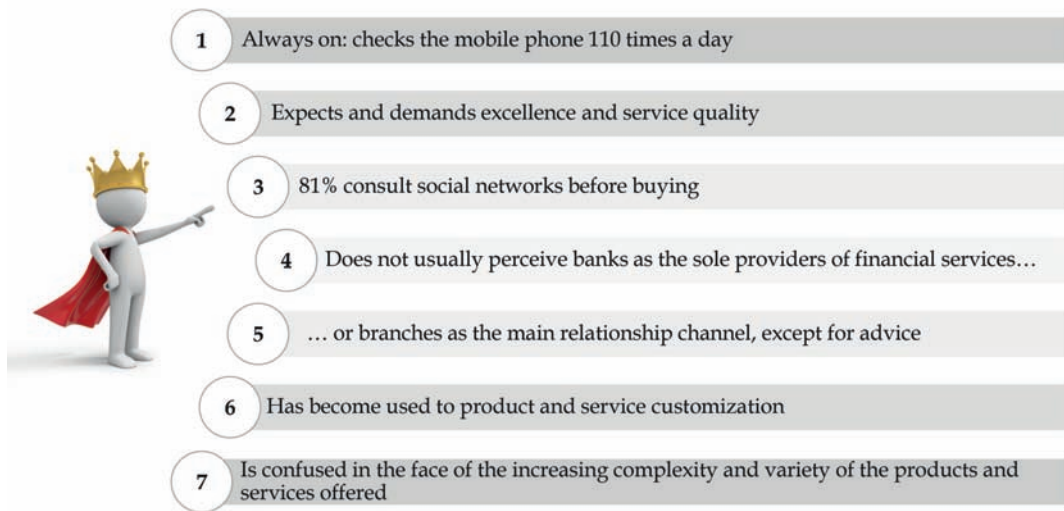
⁵⁷Mortgage Market Review and Retail Distribution Review in the UK, Market Abuse and Money Laundering Directives and Report on Consumer Trends in the European Union, among others.

⁵⁸FCA (2015).

⁵⁹Foreign Account Tax Compliance, federal law requiring financial institutions worldwide to report foreign accounts held by American citizens to the US tax agency.

⁶⁰Such as the \$1,900 million fine to HSBC for failure to implement money laundering controls; see Bloomberg (2013).

Fig. 16. Today's customer.



► Information and reporting: information generation and risk reporting processes have gradually lost effectiveness for various reasons, which has been identified as one of the causes of the crisis, and in response to which regulators have issued rules⁶¹ requiring financial institutions to comprehensively review Risk data and reporting to ensure quality, integrity, traceability and consistency, a measure known as «RDA & RRF»⁶². The aim is to strengthen risk data aggregation capabilities and reporting practices to improve management and decision making. Also, the criteria for reporting capital, liquidity and financial information (COREP, FINREP) were unified. All this requires financial institutions to thoroughly review their data generation and reporting systems and processes.

Dealing with this «regulatory tsunami» represents an enormous cost to financial institutions and is forcing them to embark on ambitious transformation processes. However, this transformation is clearly a differentiating element for entities, as it allows them to offer customers the security of knowing they have the safest, most regulated and supervised processes in all digital industries. This is a key aspect that institutions will ultimately leverage against new competitors entering the industry.

Customer behavior

Within a few years, the financial sector has seen a transformation in customer behavior: customers are more informed, more connected, more financially literate and demand more customized products (Fig 16). They require services that provide comfort, speed, personalization and fair treatment, in addition to access from mobile devices.

Customers have high expectations, compare the quality of service provided by their financial institution with that of

providers from other sectors (technology, retail, etc.) and expect a similar level of performance and real-time response.

Also, customers are competent and active in the use of social networks, which they use both to compare information (81% consults social networks before buying) and to express their disappointment with a poor experience.

Studies⁶³ show that customer experience is positively correlated with retention levels. In spite of this, they also reveal that, although the quality of the customer experience with financial institutions is gradually improving, it is still insufficient: over 50% of customers express their intention to switch banks within six months.

Moreover, customers no longer regard banks as the sole providers of financial services or branches as the basic connection channel (except for seeking advice). All this is forcing financial institutions to rethink their overall service and channel offering, in short, to adopt a «customer-centric» or «360°» view, which impacts all areas, from processes and systems, to organization, to risk management and business planning.

There is, however, a perception that customers are increasingly confused at the complexity and diversity of the products and services offered. This is causing financial institutions to adopt a commercial approach aimed at simplifying their offer, adapting it to the customer's needs (and therefore to review their product and services catalogue).

⁶¹BCBS 239 (2013).

⁶²Risk Data Aggregation and Risk Reporting Framework.

⁶³EFMA (2013).



A disruptive new phenomenon appears to be occurring simultaneously and in close association with the changing customer profile: the entry in the financial industry of new competitors, some of them from other sectors, (Fig. 17) who satisfy needs not entirely covered by traditional banking.

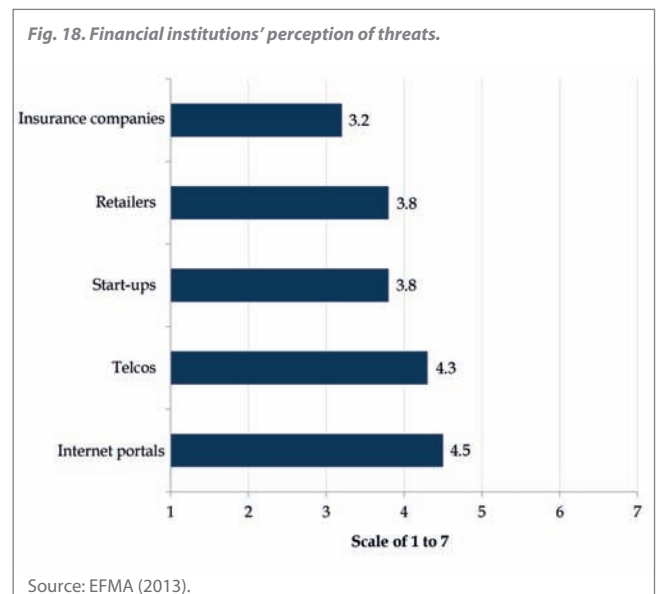
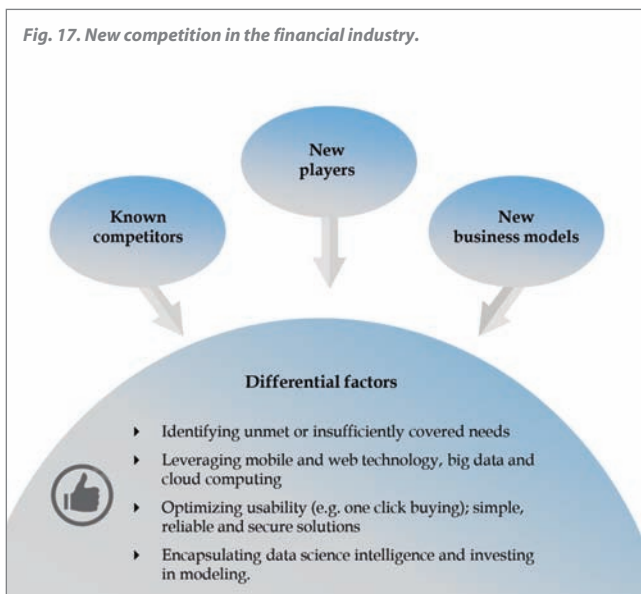
This new competition can be classified into three groups:

- ▶ Known competition offering new services (such as banks that are 100% mobile).
- ▶ New financial players that are new in the market and cover underserved niches.
- ▶ New business models from other industries, particularly technology, retail sales and telecommunications.

In the case of competition from other sectors, this threat is particularly harmful for several reasons: firstly, new competitors

are not subject to strict banking regulations; secondly, they have highly efficient, very low cost business models; thirdly, they have «ecosystems» combining many customer needs: physical devices, work environment, music, movies, books, magazines, etc., where it becomes natural to integrate financial services; and finally, they have brands that are perceived very positively by customers.

Banks still perceive these competitors as a moderate threat (Fig. 18), and it is true that at present they are concentrating on specific niches, such as payment methods (e.g. PayPal, Google Wallet or Apple Pay), and they face high regulatory entry barriers to access the core deposit and credit services. However, because of their size and influence, these competitors have the potential to overcome barriers and alter the market significantly in the near future.



Data science: an emerging discipline

Every company has big data in its future and every company will eventually be in the data business.

Thomas H. Davenport⁶⁴



What is data science?

As with any commodity, data commoditization and the resulting data and model governance entails the appearance of new tools and techniques to process these data. All these tools and techniques together constitute a discipline which, while not new, is emergent in nature and is increasingly receiving attention in the financial industry as well as in other areas. This discipline is data science.

Definition

The recent interest in this discipline, together with its innovative nature and links to big data technology, means that there is no formal, commonly accepted definition of data science. The Center for Data Science at New York University approaches this term as follows⁶⁵:

Data science is the study of the generalizable extraction of knowledge from data [using] mathematics, machine learning, artificial intelligence, statistics, databases, and optimization, along with a deep understanding of the craft of problem formulation to engineer effective solutions.

However, most approaches to the data science definition describe the skills and knowledge a professional needs in order to be considered a data scientist:

*A high-ranking professional with the training and curiosity to make discoveries in the world of big data. [...] More than anything, what data scientists do is make discoveries while swimming in data [...] bring structure to large quantities of formless data and make analysis possible. They identify rich data sources, join them with other, potentially incomplete data sources, and clean the resulting set.*⁶⁶

*A professional with a deep understanding of data who can work effectively with data in a scalable manner.*⁶⁷

*A data scientist represents an evolution from the business or data analyst role [...] with a solid foundation typically in computer science and applications, modeling, statistics, analytics and math. What sets the data scientist apart is strong business acumen, coupled with the ability to communicate findings to both business and IT leaders.*⁶⁸

As can be seen, a data scientist is a professional with a multidisciplinary profile, specifically combining at least three characteristics:

- ▶ Basic training in any science or quantitative discipline that includes knowledge of machine learning, algorithms, optimization, simulation, time series and association and classification models, among others.
- ▶ Advanced technological skills, including proficiency in statistical programming languages, but also technical expertise for efficient data extraction, use and storage, handling of relational and non-relational databases and the ability to extract data from the Internet and to process large amounts of information.

⁶⁴Thomas Hayes Davenport (n. 1954). US academic specializing in knowledge management and business process innovation. He was named one of the top three business analysts in the world in 2005 by Optimize magazine.

⁶⁵Dhar [Center for Data Science, New York University] (2013).

⁶⁶Harvard Business Review (2012).

⁶⁷Berkeley (2015).

⁶⁸IBM (2014b).

- And, what possibly marks a major difference with other similar profiles: a deep understanding of the business in which they carry out their work as data scientists.

The third aspect, business acumen is particularly relevant because it matches analytical capabilities with - in the case of banking, financial knowledge, which is key for the full integration of models in the management process, and a prerequisite for their success and good use (Fig. 19).

An indication of the importance of this role is the fact that US President Barack Obama created the position of Chief Data Scientist in February 2015, and personally appointed Dhanurjay 'DJ' Patil⁶⁹ to it, with the mission to promote new big data applications in all Government areas⁷⁰.

Data science process

As pointed out by some authors⁷¹, the most important feature of data science is precisely its science status: faced with the massive amount of data a financial institution has to deal with, a data scientist's task is to formulate a theory (a question or hypothesis) based on the reality of the business, and to apply knowledge and skills to either verify or discard the data. This is known as the «data science process», which consists of five stages (Fig. 20):

- 1. Formulation:** a relevant question for the business is asked which must be answered using the data and techniques available. One of the key changes brought about by the data science discipline is precisely the formulation of questions or hypotheses that were previously impossible to

verify. Today, however, the current abundance of data, tools and techniques opens new possibilities. For example, «judging from their comments in recent calls to the call center, what is the probability that each of my customers will change banks in the next six months, and what should I do about it?».

- 2. Data collection:** all available sources are localized, including structured sources (data warehouses, data marts, etc.) and unstructured sources (activity logs, social networks, etc.). The massive amount of data and sometimes their unstructured nature are at the core of the computational challenge the whole process represents. Legal aspects such as data protection, confidentiality or usage restriction clauses are also dealt with in this phase.
- 3. Data exploration:** descriptive statistical techniques are applied to conduct a first exploratory analysis. At this point, data science brings new exploration techniques that make the work easier and, given the potential for parallelization in these tasks, benefits are obtained from distributed computing platforms.
- 4. Modeling:** traditional model construction and validation are enriched by high performance algorithms developed on an ad hoc basis for large data sets, as well as by model types that are an alternative to classic models as they make them to improve in terms of stability, robustness and exploitation of the wealth of information, e.g. random forests and support vector machines, among others (Fig. 21). Because of this, traditional vendors of analytical tools are expanding their product suites, and new statistical programming languages, many of them open source, are emerging.

⁶⁹Reputed data scientist who has worked at LinkedIn, eBay, PayPal and Skype, among others, and who is credited with having created the term «data scientist».

⁷⁰Wired (2015).

⁷¹O'Neil and Schutt (2013).

Fig. 19. Data scientist profile.

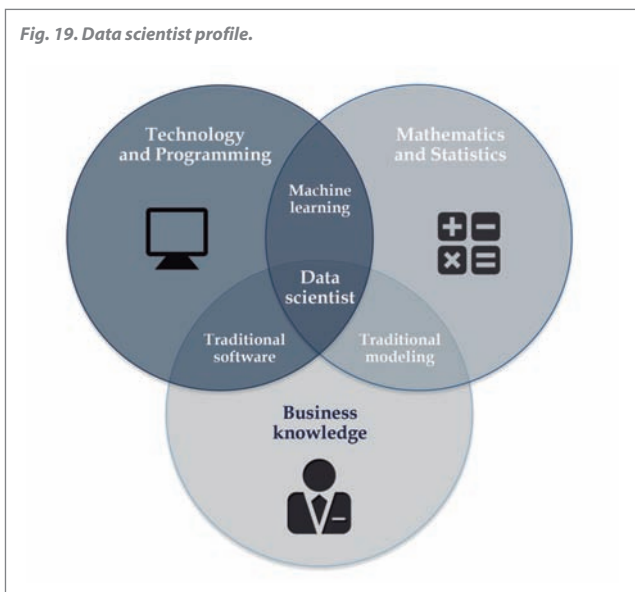
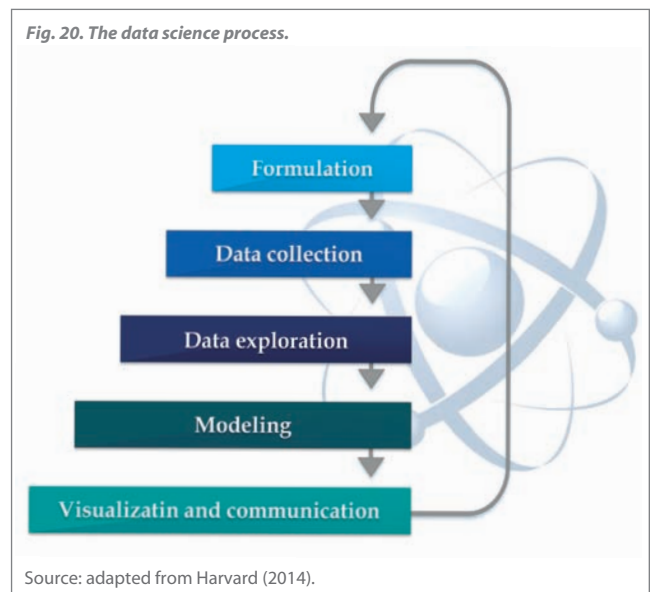
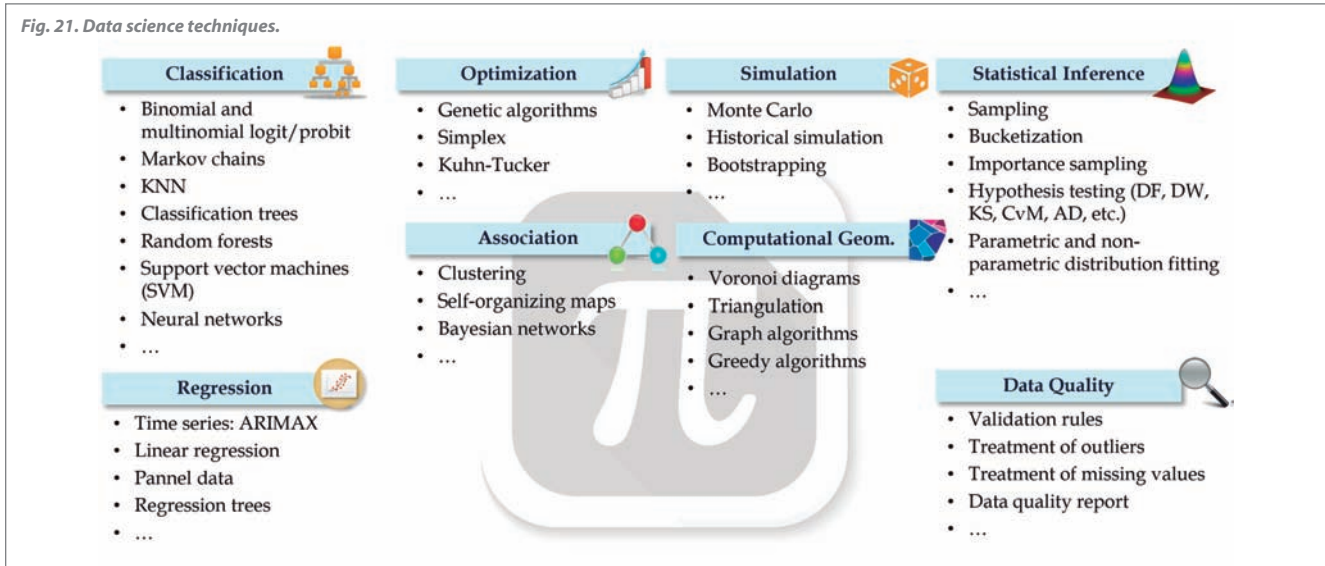


Fig. 20. The data science process.



Source: adapted from Harvard (2014).

Fig. 21. Data science techniques.



Data lake: a new information architecture

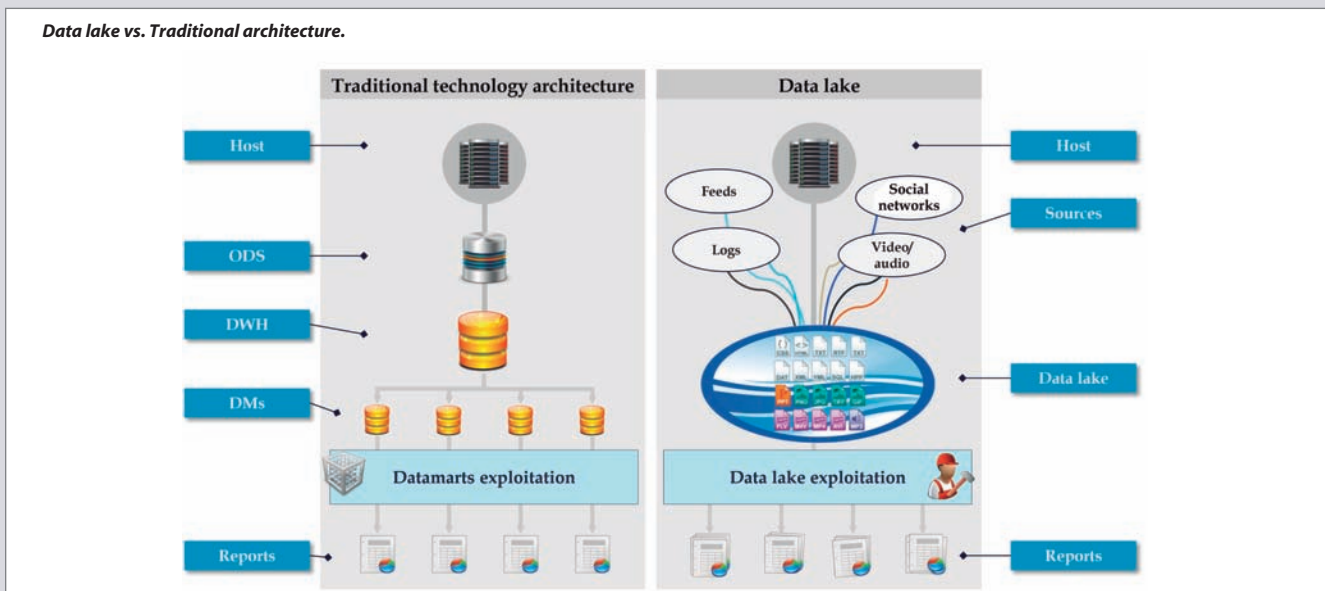
The growing volume of information and varied nature of data sources calls for new techniques and technologies capable of storing information in an optimized way. It is this scenario that gives rise to the concept of «data lake» or «data pool», a huge repository of data in their original format with the following characteristics:

- ▶ **Integrity.** A data lake is a single repository where all data are stored. This ensures the information can be traced.
- ▶ **Flexibility.** Data lakes store all information of interest, regardless of format, removing any specific standards on data capture and storage.

- ▶ **Independence.** They largely eliminate dependencies on IT departments as data dumping is flexible and the user can perform the necessary extractions directly from the data lake.

The use and exploitation of data lakes and related data science techniques enables data scientists to work with unprocessed information in its original format. It allows for scale-out and removes limits on handling large volumes of unstructured information. This architecture does not necessarily replace the traditional one; on the contrary, it generally supplements data warehouses and data marts.

Data lake vs. Traditional architecture.



5. Visualization and communication: two aspects that have received much attention in data science. Visualizing results and communicating them to third parties in an intelligible form are two of the qualities expected from a data scientist. They are also enhanced by new tools that naturally and intuitively integrate code with documentation.

Although this might appear to be common sense, approaching data through a scientific method involves changing the work methodology. Analysts often approach the problem in the opposite way (running random models on a massive amount of data in search of hidden relationships), which may use up a large amount of resources without a clearly stated objective or a hypothesis to test.

In short, data science represents the development of traditional modeling in a big data environment and opens new, previously unthinkable possibilities that may even transform established business models. The adoption of data science as a strategic instrument for development is a priority for the technology industry and, as we shall see, is beginning to be so for the financial industry too.

Data science tools

Data commoditization also favors the development and emergence of new technological data science tools that facilitate processing, analysis and visualization. All traditional vendors are promoting analytical ecosystems, and start-ups are continuously emerging with innovative proposals as well as open source tools and languages (Fig. 22), which makes this market a focus of competition and accelerated development.









These tools make it possible to overcome the limitations of traditional systems, which were insufficient due to data heterogeneity (they could not analyze structured and unstructured information together), data defragmentation (the information was distributed into different silos under imprecise modeling), dependence on IT (business users had to delegate to Systems areas the task of collecting the information and organizing it in data warehouses, which entailed excessive data preparation time) and, in general, lack of adaptation to current data sources (traditional systems were not integrated with social networks, call centers, sensors, geographical location, etc., nor were they adequate to deal with the volume of information they generated).

Among the main contributions that these tools have incorporated are⁷²:

- ▶ **Self-service:** under the traditional scheme, only a few highly specialized professionals in the organization had access to data and analytical tools, whereas under the data science scheme tools are simpler, enabling more professionals to explore, analyze and visualize the data. This phenomenon has occurred in all industries and has contributed to enhancing the analytical capabilities of professionals.
- ▶ **Data fusion:** data fusion refers to combining information from different sources and in different formats. In the traditional system, this is done through ETL processes and the deployment of data models that can be highly complex. Under the more advanced data science system, data are dumped to a data lake, well documented with a data dictionary, and the tools are able to take the files and merge them in a short space of time.

⁷²Adapted from Gigaom Research (2014).

Fig. 22. Who is who in data science? Some of the main open source tools and languages.

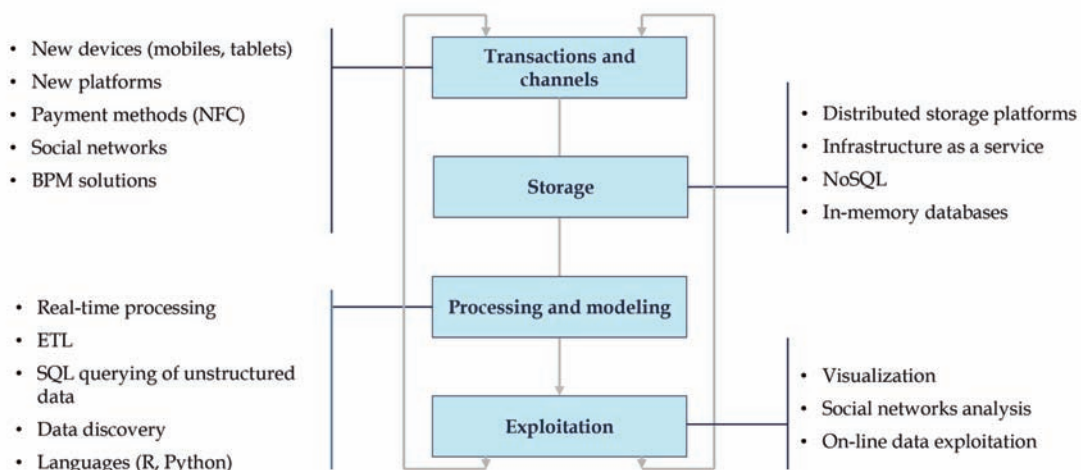
	Hadoop: open code programming infrastructure for parallel storage, processing and analysis of large data sets across large clusters of servers.
	Hive: Hadoop-based data warehouse system developed by Facebook. It is used for querying and handling large data sets in a distributed storage environment. It uses a SQL-like language called HiveQL (HQL), which does not require users to use Java or Hadoop APIs.
	Pig: open source application created by Yahoo and built on Hadoop, focusing on processing large data sets (structured and semi-structured) in batch mode.
	Impala: open source platform which provides real-time SQL querying functionality (processing and data analysis) in real time. It uses Hadoop components.
	Lucene: open source API that serves as a search engine. It is used for indexing and searching of data on limited datasets. It is implemented in the Twitter, LinkedIn, Apple, AOL and Eclipse web apps, among others.
	Mahout: open source project to implement scalable, distributed machine-learning algorithms. It is a Java framework that runs on top of Hadoop. It is used by Adobe, AOL, Intel, LinkedIn, Twitter and Yahoo, among others.
	R: statistical programming language and environment which is highly versatile due to its modularity (advanced functionality packages already created by other programmers can be installed) and open-source nature.
	Python: general-purpose programming language that emphasizes readability and intuitive coding. In data science, it is particularly suitable for capturing data from online sources.

New technological capabilities

The big data phenomenon brings new technological capabilities, which are structured in four layers:

- ▶ In the **transactional** layer, there are new devices and customer relationship channels (mobile, tablet, etc.), new platforms for the development of applications and provision of services (CRM, mobile apps to cover new financial and operational needs of customers, etc.), new payment technologies such as mobile payment (e.g. NFC) and Business Process Management (BPM) solutions for platform integration and process automation, such as on-line contracting or document management. Also, there is increased focus on social networks as a new customer relationship channel, as they are emerging as a potential contracting channel and are being used to analyze customer brand sentiment and to deal with complaints and claims.
- ▶ In the **storage** layer, there are new storage systems designed to run on low-cost hardware, providing high availability and tolerance to failure (data replicated on multiple nodes), as well as scale-out and mass data processing capabilities. Infrastructure as a service has emerged in the public cloud, private cloud or hybrid cloud modalities. There are new data bases (NoSQL) focused on the batch processing of large volumes of information and new data structures: columnar and documentary data bases as well as new in-memory data bases that rely on main memory to process data, providing high speed query response.
- ▶ In the **processing and modeling** layer, there are tools for capturing and processing information in real time, as well as new ETL for processing unstructured data, such as Pig, and new engines to query unstructured data using SQL language. There are also tools to implement data governance mechanisms: cataloging, transformation, traceability, quality, consistency and access control, and data discovery tools for extracting knowledge freely from various sources, both structured and unstructured. And finally, there are new techniques, mathematical algorithms and languages for the recognition of patterns in data, predictive analytics, implementation models and machine learning.
- ▶ Finally, in the **exploitation** layer there are new multidimensional analysis and reporting tools with capabilities for accessing large volumes of information in memory, specific solutions for the analysis of information from social networks and the exploitation of data streams on-line for decision-making and for real time triggering of events such as fraud detection, identifying and launching commercial events and risk scoring, among many other uses.

New capabilities by IT architecture layer.



- ▶ **Non-relational connectivity:** unlike traditional tools, which only provided for connection with databases, data science tools make it possible to connect with other sources of information NoSQL, distributed computing platforms, and information from social networks, on the cloud or on software as a service systems, are increasingly important for organizations.
- ▶ **The cloud:** among the most important new developments is the use of the cloud in the analytical field, as this provides data storage functionality in its most basic version, allowing the data scientist's work to be decoupled from a specific location or server and making work from different geographies easier. Also integrated in some cases are ETL services, data visualization and deployment to mobile devices, creating a complete analytical ecosystem that simplifies the analysis effort.
- ▶ **Data visualization:** one of the distinguishing features of the data science discipline, linked to business knowledge, is data visualization. Some tools go well beyond graphics generation with annotations, and are now able to automatically produce dashboards and interactive presentations that allow users to dynamically look further into the analysis.

Data science in the financial industry

The financial industry is experiencing the same explosion in data generation and storage requirements as other sectors. It has the potential to draw much insight from both customers and the environment (competition, industry-specific economic activity, geolocation, etc.) which could not be accessed before.

With this in mind, financial institutions are developing a series of technology and methodology skills that are opening new possibilities in the industry, in spite of the challenges.

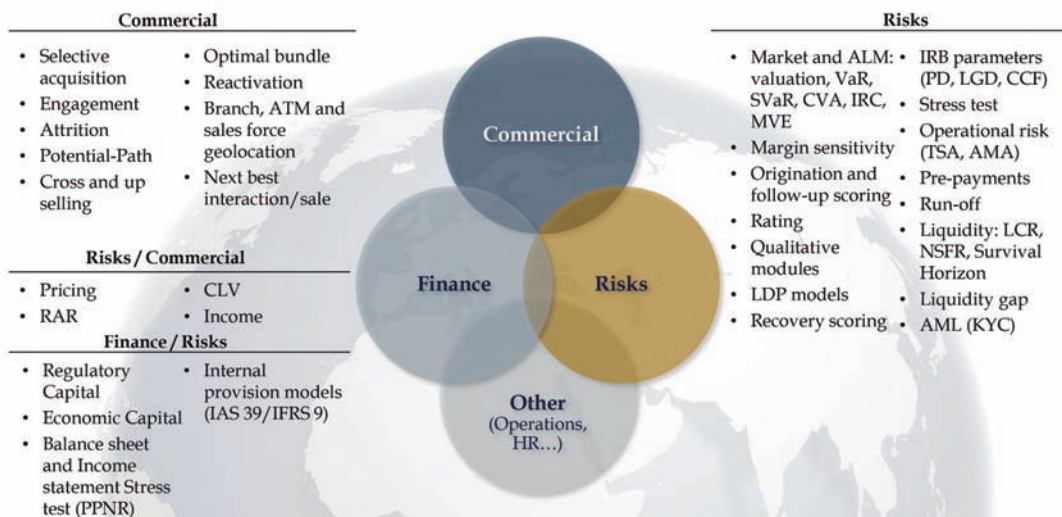
New opportunities in the financial industry

Based on these new capabilities, traditional models are expanded and enriched in all areas of activity in financial institutions, including risks, marketing, finance, operations, etc. (Fig. 23). This helps to better exploit all available information to improve decision-making in these areas and, in some cases, to automate some processes.

By way of example, we will draw attention to some emerging data science applications in the industry, based on data from social networks, geolocation, multimedia or logs, etc., that received no attention before:

- ▶ **Credit scoring with digital footprint:** the credit scoring of individuals, usually based on a few variables (between 5 and 20, depending on the portfolio and the information available). This is enriched and expanded from the information available in social networks and on the Internet in general, in what is known as the «digital footprint». Based on this information, models are built which substantially improve predictive power and therefore delinquency control, especially in the non-customer population, who are traditionally rated worse due to data availability constraints.
- ▶ **Customer churn prevention through natural language processing (NLP):** call center recordings, which were previously almost exclusively used for internal quality control purposes, are revealed as a valuable source for preventing customer churn. Using speech recognition

Fig. 23. Main models in financial institutions.



models, all conversations with customers are automatically transcribed, and text mining and computational linguistics techniques are applied on the resulting texts to identify the probability that a particular customer will decide to change banks in the coming weeks. This begins with a lexical analysis (detection of certain words that are associated with the intention to change). On an experimental basis work is also being done at a semantic level, where the model looks at more complex meaning patterns in the customer's speech.

- ▶ Income and propensity models based on social networks cross-matched with geolocation: customer information available on social networks is cross-matched with census, housing and Google Maps data as well as information from other sources, and this is used to improve estimates of the customer's income level, savings capacity, financial product needs, the value of the property where the customer lives (used also to assess the collateral in a securitization), etc. This supplements the available information and helps to improve commercial action on customers.
- ▶ Customized promotions to reduce customer acquisition costs: all available information on all customers' transactions through all channels and social network data are collected and cross-matched, obtaining a 360° view of the customer. Since this narrows down the target niche for each campaign, the customer capture ratio increases and acquisition costs are reduced.
- ▶ Customer reward campaigns through card transaction analysis: card movements show the behavior of card holders: when they buy, travel, shop, etc., and make it possible to launch customer award campaigns at the right time, with the greatest likelihood of success.

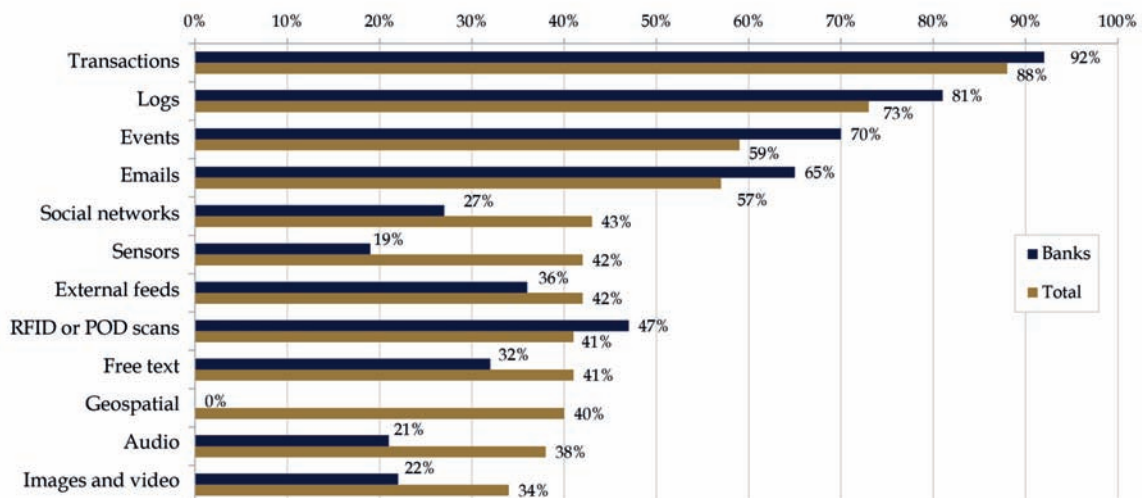
- ▶ Fraud and money laundering detection, and improved quality control through logs: activity logs are large, loosely structured files where all actions performed by customers or employees on a digital platform (PC, mobile device, ATM, etc.) are recorded. Identifying behavior patterns in logs is complex because it requires a particularly massive data processing effort, but it can serve to identify fraud attempts (both internal and external) and money laundering. It is also the basis for a new form of quality control, potentially very far-reaching, which can cover anything from response times in an office to the difficulty in using a new application, to customer channel preferences for each type of transaction, among many other uses.

These are just a few examples; opportunities are as many as questions can be formulated, given the proliferation of data sources in financial institutions and the growing capabilities in data science (talent and tools) of which institutions are already availing themselves.

In this regard, process automation and the improvement of models used in the financial industry are closely linked to financial institutions' capabilities for capturing relevant information on customers, processes, products, etc., and for storing, processing and exploiting this information using data science tools at a later stage.

The sources of information currently available to be exploited by financial institutions are virtually limitless. This emphasizes the fact that any information, regardless of its source (internal or external) and nature (structured or unstructured), is potentially relevant for decision-making. In fact, it is reckoned that most mass data in banks come from transactions, logs, events and emails - well above any other sources (Fig. 24).

Fig. 24. Big data sources in banking and other industries.



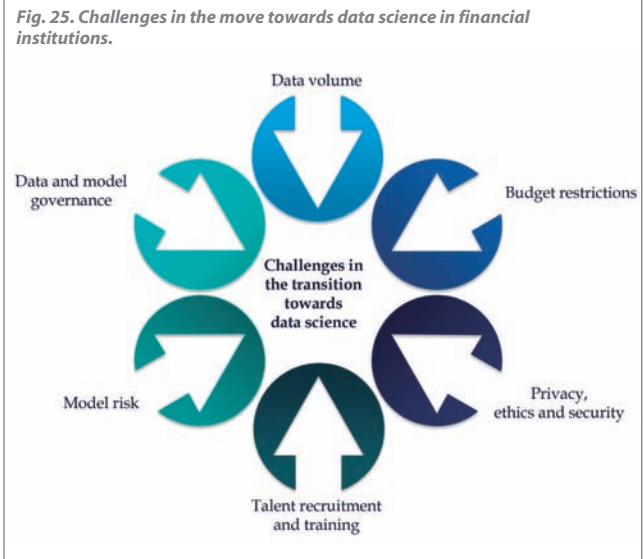
Source: IBM & University of Oxford (2015).

Challenges facing the adoption of data science

It appears to be evident that the possibilities opened by the data science discipline in the financial industry are numerous and have the potential to substantially improve performance metrics in all areas including customer experience, efficiency, risk control, commercial action efficiency, etc.

However, progress towards these capabilities is neither simple nor immediate. Institutions face a number of challenges (Fig 25), which have been linked to the challenges of big data in industry surveys (Fig 26). These challenges include the following:

- ▶ **Data volume and sharing:** the same massive amounts of data that have given rise to data science represent a challenge in terms of database sizing, storage architecture, processing capacity costs, computation time and need for optimal algorithms. It is solved in part with new tools and platforms, but requires careful technology planning. Also, data sharing between different areas in financial institutions poses a technological and organizational challenge which is complex to solve.
- ▶ **Budgetary constraints:** in relation to the above, the move towards data science requires investment in infrastructure and talent, which in an environment of pressured margins and abundant regulation needs to be coordinated with budgetary constraints.
- ▶ **Privacy, ethics and security:** the use of abundant customer information raises questions about data privacy and ethical use. Service is more personalized the more information is used and, since regulatory activity on many of the nuances of privacy is still incipient, and it is difficult for it to move

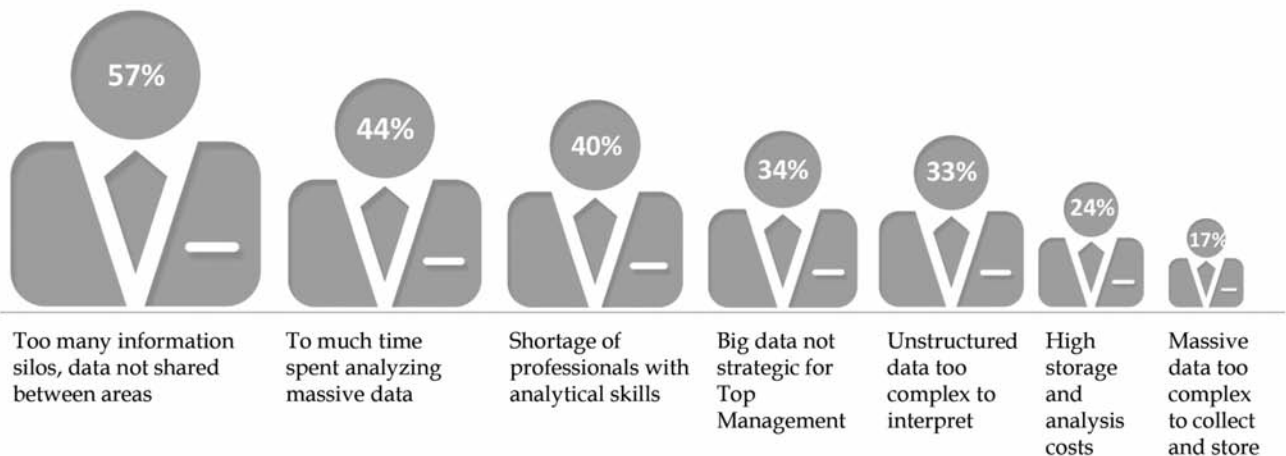


forward at the speed at which the big data phenomenon moves. To find a balance between privacy and customer experience it is necessary to involve the legal areas and to listen to the customer. This point also covers aspects related to data security and to ensuring that the data collected from various sources are safe from hacking, identity theft and improper sale to third parties.

- ▶ **Talent recruitment and training:** data scientists are still a relatively scarce profile, and demand for these professionals far exceeds supply. Therefore, attracting data scientists on the market and providing traditional methodologists with new skills is a key challenge. Some studies⁷³ point to a shortage of over 140,000 data scientists only in the United States in 2018.

⁷³McKinsey (2011).

Fig. 26. Results of a survey on «major barriers to success with big data in banks» (percentages over total responses).



Source: adapted from Inetco (2015)

- ▶ Model risk: using models entails risks, which arise from the data used in these models but also from the model's estimation itself and from potential model misuse. Model risk management and control are increasingly a focus of attention with the emergence of data science⁷⁴.
- ▶ Data and model governance: finally, the organization and governance of internal structures necessary to properly manage a financial institution's data and models is a key element for a successful move towards data science in the organization, as detailed below.

Impact on the governance of data and models

Although the growing speed in data generation and access, and the possibilities this offers, are a relative novelty in the financial industry, the reality is that neither financial institutions nor regulators and supervisors are oblivious to the phenomenon described.

On the contrary, financial institutions have been making changes to their data generation and information reporting systems and processes, especially in the risk, financial and commercial areas. However, in many cases these changes have been made in an unstructured form and with a limited perspective. Due to incremental requests from regulators and supervisors, as well as data migrations resulting from mergers and acquisitions management needs have not been planned using a global approach. As a result, the data generation and information reporting processes have gradually become less effective, and data consistency is sometimes not ensured.

Moreover, regulators have pointed out data shortcomings as one of the causes of the financial crisis that began in 2007:

*One of the key lessons from the global financial crisis that began in 2007 was that the inadequacy of banks' information technology (IT) and data architectures prevented the integrated management of financial risks. [...] In some banks, the inability to properly manage risks was caused by shortcomings in risk data aggregation and reporting practices. This had serious consequences for the banks themselves and for the stability of the financial system as a whole.*⁷⁵

The use of models for decision-making is also a rapidly increasing phenomenon, which brings significant benefits such as improving objectivity, automation and efficiency. However, their use also carries «model risk», understood as the potential damage (financial, reputational, etc.) caused by decisions based on erroneous models or inappropriately used models⁷⁶.

Both the regulation and the management issues derived from the wealth of information available and its use in models for decision-making call for the need to establish a new framework to properly govern data and models in each financial institution. This section will look into industry practices in relation to these governance frameworks.

Data governance

Establishing data governance mechanisms is a complex strategic need in financial institutions which becomes particularly pressing as the big data phenomenon develops and it becomes essential to make the most of the information available.



⁷⁴Management Solutions (2014).

⁷⁵BCBS (2013).

⁷⁶Management Solutions (2014).

This has not gone unnoticed by supervisors, who have in some areas, issued specific regulations on the governance of data and reporting. Particularly noteworthy are the *Principles for effective risk data aggregation and risk reporting* (BCBS, 2013), known as «RDA & RRF»⁷⁷ with requirements on data quality, consistency, integrity, traceability and reproducibility. This legislation is binding for globally systemic banks and their subsidiaries, and will also apply to locally systemic institutions in the future.

The importance of these aspects is shown by the fact that some institutions have already articulated maximum involvement of the Board of Directors and Senior Management on information and reporting issues. In some cases, data governance and reporting is substantiated by the creation of Board-delegated committees for data management. Banks are also broadly addressing the creation of organizational roles such as Chief Data Officer (CDO) or Head of Risk Management Information (RMI), and are launching strategic initiatives to strengthen the infrastructure supporting the information generation process.

The benefits of robust data governance are clear: It helps to achieve uniform and consistent reporting, using standard concepts throughout the organization and aligning the parent and subsidiaries in the case of international groups; it ensures consistency between regulatory and management reporting; achieves greater efficiency (increased automation, fewer redundancies), improves time-to-market and makes report generation more flexible; and in the case of risk, it enables accurate knowledge of risks by Senior Management and ultimately contributes to improving risk management and control.

Elements of a data governance and reporting framework

The above therefore implies that institutions that are most advanced in this field have a data reporting governance framework that details the basic principles, participants and their roles, governance structure and support elements (processes and tools) in relation to data management and information generation.

With regard to the principles, the framework should identify the basic guidelines for data and reporting governance, including the scope of the framework and its area of implementation, data ownership, consistency between the different areas, and mechanisms deployed to ensure data quality.

As for organization and governance, the framework identifies those involved in the process and their roles, specifically including those responsible for producing the information, for ensuring its quality and for the data repositories. Some key roles are the Chief Data Officer (CDO), responsible for quality assurance and end-to-end data traceability in reports to Senior Management and for data consistency (for which this role relies on tools such as the data dictionary). Also key are those responsible for data (by area) or, in the field of Risk, the Head of Risk Management Information (RMI), among others.

This framework also provides the definition of the governance bodies responsible for data and reporting, whose responsibilities include promoting the development and effective implementation of the data governance model, the review and approval of relevant changes in the data generation process, the adoption of data quality objectives and, in general, the definition of data management strategy. These committees

⁷⁷Risk Data Aggregation and Risk Reporting Framework.





should include representation from all data users, which often includes the Business, Risk and Finance divisions as well as those responsible for systems, the CDO and those responsible for generating information.

Committees include several levels, including technical committees responsible for supporting senior committees and resolving potential conflicts on data involving several areas or geographies. Additionally, for international organizations it needs to be ensured that data governance is consistently aligned across geographies. This requires setting up relevant committees in the different countries and establishing appropriate mechanisms for reporting and escalation to corporate level committees.

Data governance requires a number of elements for proper coordination, which will facilitate compliance with the principles of quality, traceability, consistency and granularity required by regulators⁷⁸. Some of which are as follows:

- ▶ Data dictionary: a unified inventory of metrics, dimensions and components related to reports, with clear functional definitions that are unified for the entire organization.
- ▶ Metadata: specific information on data, contained in the data dictionary, which allows cataloging and conditions its use. The trend is to enrich existing metadata (business and technical), supplementing them with additional metadata about the origin of data their transformation, and their quality, which influence the use that can be assigned to each type of data.
- ▶ Datawarehouses and data lakes: databases and other information sources having sufficient quality to build metrics which are eligible for inclusion in reports.
- ▶ Exploitation tools: analytical data processing and visualization tools, among which those with big data processing capabilities play a key role.

Finally, it is critical to ensure the quality of the data used. Best practice considers the following:

- ▶ Establishing a data control model that includes monitoring the production of reports for Senior Management and defining quality tolerance levels to be applied to the data to be reported.
- ▶ Identifying, defining and implementing KPIs to measure the degree of quality of information at multiple levels in the data lifecycle, and defining tools (dashboards) for the aggregation and monitoring of data quality levels in reports to Senior Management.
- ▶ Implementing data quality plans at various levels (operational systems, information repositories and reports), which take the form of initiatives to cleanse historical data (often addressed through crash programs), and to improve new data production (through process changes).

⁷⁸E.g. the Basel Committee on Banking Supervision, the Fed and the OCC.

Data governance challenges

Developing sound data governance entails, a number of challenges that institutions need to face, some of which are:

- ▶ Ensuring the involvement of Senior Management in data governance, reporting and quality.
- ▶ Defining the scope of the data governed by the model (especially considering the exponential growth in terms of diversity and volume), and seeing that the model is operational and ensures adequate levels of data quality, traceability and consistency without impairing the organization's ability to optimize its use.
- ▶ Resolving the issues relating to data privacy and security and ensuring that data are safe from fraudulent use.
- ▶ Strengthen cybersecurity, which includes protection against «hactivism» (attacks against financial institutions for ideological reasons through virus, malware, etc.), fraudulent data use, financial cybercrime, espionage and information theft. (It is worth mentioning that the risk of cyberattacks was incorporated in 2014 to the top 5 global risks list of the World Economic Forum).
- ▶ Identifying and implementing tools that facilitate data governance and adapting data governance mechanisms to new architectures such as data lakes.
- ▶ Implementing a single dictionary of concepts for uniform understanding throughout the organization and, where appropriate, subsidiaries.

- ▶ Involving the various subsidiaries, in the case of a financial group, in joint data governance.

Model governance

According to the Fed and the OCC, the term «model» refers to «a quantitative method, system or strategy that applies statistical, economic, financial or mathematical theories, techniques and hypotheses to process data and obtain quantitative estimates»⁷⁹.

To date, there is little regulation specifically governing model risk, and it tends to be nonspecific both in its definition and the expected treatment. The exception is the *Supervisory Guidance on Model Risk Management* published in 2011-12 by the OCC and the Fed (US banking regulators).

This publication defines model risk for the first time as «the potential for adverse consequences from decisions based on incorrect or misused model outputs and reports». It establishes, through a set of guidelines, the need for institutions to have a Board-approved framework in place to identify and manage this risk.

These guidelines cover all stages of the model's life cycle: development and implementation, use, validation, governance, policies, control and documentation from all participants. Among the key aspects required is the need to address model risk with the same rigor as any other risk, with the particularity that it cannot be eliminated, but only mitigated through effective challenge.

⁷⁹OCC/Fed (2011-12).



Elements of an objective MRM model

Financial institutions that are most advanced in this area have a model risk management (MRM) framework which is embodied in a Board-approved document and details aspects relating to organization and governance, model management, etc.

With regard to organization and governance, the MRM framework is characterized by its cross-functional nature (covering several areas, such as business lines, Risk, Internal Audit, Technology, Finance, etc.), the explicit definition of the three roles required by the regulator (ownership, control and compliance⁸⁰), their assignment to specific functions in the organization and, above all, the creation of a Model Risk Management function, whose responsibility is to create and maintain the MRM framework.

Regarding the management of models, the MRM framework includes aspects such as: (a) a model inventory which catalogues all models in all areas of the institution (risk, commercial, financial, etc.), usually supported by an appropriate technological tool which tracks all changes and versions; (b) a system to classify models into tiers according to the risk they pose for the institution, which determines the level of thoroughness in the monitoring, validation and documentation of models; (c) complete and detailed documentation on each model, allowing replication by a third party as well as the transfer to a new modeler without loss of knowledge; and (d) a model follow-up methodology allowing for early detection of deviations in model performance relative to expectations as well as of model misuse, in order to take appropriate corrective action.

⁸⁰The model owner defines model requirements and tends to be its final user. Control includes model risk measurement, setting of limits and monitoring, as well as independent validation. Compliance comprises the processes needed to ensure that the roles of model owner and control are performed in agreement with the policies.

Model validation is central to model risk management, with the fundamental principle being to critically, effectively and independently challenge all decisions made in the development, monitoring and use of a model. The frequency and intensity of validation for each model should be proportionate to the risk it poses, measured through its tier, with the validation process and outcome fully documented in turn.

Model governance challenges

All this leads to the need to define a sound and stable model governance approach, which poses a number of challenges to financial institutions, including:

- ▶ Questioning what constitutes a model and which models should undergo these procedures (possibly depending on the model and its tier) including how this need for model governance can be reconciled with greater use of models for multiple purposes.
- ▶ Solving the difficulties posed by the existence of increasing volumes and types of data (not all subject to the same quality controls) used in the modeling process.
- ▶ Ensuring the involvement of Senior Management in the governance of models and more specifically, defining and approving the model risk framework at the highest level.
- ▶ Defining the organizational structure of the data science function (or functions) in terms of centralization or decentralization, both geographically and between areas of the organization, and delimiting responsibilities between corporate and local areas in the case of international groups.
- ▶ Building or strengthening governance mechanisms around each of the processes associated with the analysis function.

In short, governing the data and their transformation into knowledge, which in turn implies governing the models that articulate this transformation, has become a strategic priority for any organization, particularly for financial institutions. Consequently, the undeniable trend in the coming years will see financial institutions decisively boosting their respective governance frameworks.



Case study: social networks and credit scoring

*It is a capital mistake to theorize before one has data.
Insensibly one begins to twist facts to suit theories,
instead of theories to suit facts.*

Sir Arthur Conan Doyle⁸¹



Purpose

In order to directly illustrate how the data science discipline is implemented in the financial industry, it was considered of interest to conduct a quantitative exercise applying some of the previously described tools for a specific use in a financial institution.

Specifically, the analysis aims to develop a credit scoring model for individuals using data from social networks, integrate it with a traditional consumer loan model, and see to what extent predictive power is improved.

Analysis details

The analysis was carried out using the following data and models:

- ▶ A real construction sample from a consumer loan scoring model, consisting of some 75,000 records, with a default rate of around 12%.
- ▶ Additional variables on the sampled customers, allowing these customers to be searched on social networks.
- ▶ A scoring model built on the aforementioned sample, using 12 variables with average predictive power (ROC⁸²) of around 73%.

Main conclusion

The main conclusions from the analysis are as follows:

- ▶ The quantity and quality of the information available on social networks is significantly lower than that of the bank's internal data: only 24% of clients have data, and of these, only 19% have complete or nearly complete information.
- ▶ In addition, there are disambiguation issues in the extraction of data from social networks: customers do not identify themselves unequivocally through the use of ID on the network, so there is a probability of error in the matching of customers with their respective social network profiles. For the purpose of this analysis, customers for whom this probability was estimated to be above 25% were discarded.
- ▶ The variables drawn from social networks also are mostly qualitative and can include a large number of values, which complicates their treatment, though it is possible to build very rich variables.
- ▶ The scoring model based on social networks uses nine numeric and categorical variables (some discretized) for predictive power. These variables cover various aspects of the customer's professional profile (especially their employment history, but also their industry, academic background and languages), and with a ROC of 72% achieves predictive power comparable to that of the original model.
- ▶ The combination of both models, however, substantially increases predictive power, reaching 79%.

⁸¹Sir Arthur Ignatius Conan Doyle (1859-1930). Scottish writer and physician, famous for his Sherlock Holmes character.

⁸²Receiver Operating Characteristic, a measure of a binary response model's predictive power.

In summary, this study reveals that information from social networks contributes substantially different information, which complements and significantly enriches the traditional scoring model. However, there are certain difficulties inherent in the use of this information, which are expected to be largely solved in the future through structured data capture by financial institutions as part of their credit origination and monitoring process.

Description of analysis

The analysis follows four stages: data extraction from social networks, data cleansing and processing, construction of a «social module» and integration with the «traditional module».

For the extraction of data from social networks, we used a combination of a tool specifically designed for Python, which is connected via the social networks' own APIs, and a VBA module that extracts and sorts the information into an accessible format.

The information from social networks is very irregular in terms of completeness and quality, and is generally composed of qualitative variables which do not match predetermined lists of values. In any case, it is not extracted under a classical relational scheme, but using records requiring an analysis process known as parsing for conversion into usable data.

For this analysis, data were found in approximately 18,000 out of 75,000 customers. Of these, the information was reasonably complete in approximately 4,000 customers. However, since missing values were thoroughly treated, the number of ultimately usable records was greater.

For those customers, 30 original variables were extracted with different levels of missing information. The variables covered several areas of the customers' personal and professional

profiles: formal and non-formal education, professional experience, geographic location and other information related to hobbies, interests, etc.

The processing of data required the construction of combined variables from the original variables, in addition to the transformation needed to manage the information. Thus, from the 30 variables extracted more than 120 fields were created, which are mostly categorizations based on univariate and bivariate analysis, as well as a temporal analysis of each customer's professional history.

The model construction phase is very similar to that of any scoring model; expert analysis was combined with a stepwise variable selection process, variables that were redundant from a business perspective were removed, and a 95% confidence level acceptance criterion (p-value less than 0.05) was applied.

The algorithm used was a combination of a decision tree (reinforced with a pruning algorithm to reduce entropy and therefore improve robustness and stability) and a binomial logistic regression:

$$P(Y = 1) = \frac{1}{1 + \exp(-\beta_0 - \beta_1 x_1 - \dots - \beta_n x_n)}$$

The result was a model of nine variables, summarized in Fig. 27. The ROC curve, which measures its predictive power, is shown in Fig. 28.

As can be seen, this model has average predictive power, comparable to that of the traditional model, despite using only nine variables; this is attributable to some extent to the presence of qualitative variables. Other statistics show that the model is robust and has appropriate statistical properties.



The last phase is the integration of the social module with the traditional scoring module. To do this, a new logistic regression was built, taking the scores for each module as the independent variables:

$$P(Y = 1) = \frac{1}{1 + \exp(-\delta_0 - \delta_1 \text{score}_{\text{trad}} - \delta_2 \text{score}_{\text{social}})}$$

Both scores are significant at a 95% confidence level, and as shown in Fig. 29, the final model has an area under the ROC curve of 79%, which substantially improves the predictive power of each separate model.

The present analysis is focused on the case of a credit scoring model with the incorporation of data from social networks, but this exercise can be extended to other types of models (collateral valuation, loyalty, income, customer churn, propensity to buy, etc.) and information sources (internal logs, public databases, web information, etc.).

In conclusion, as has been shown, incorporating variables from other sources has the potential to significantly increase the discriminatory power of traditional models.

Fig. 27. Social module variables in the credit scoring model.

Variable	Description	Relative weight
Time in current position	Number of months in current position	18%
Length of service	Number of months in service	15%
Minimum position length	Minimum length of time in a position during professional career	15%
Maximum position length	Maximum length of time in a position during professional career	13%
Activity sector	Official classification of professional activity	12%
Number of jobs	Number of current and past jobs	9%
Time without studying	Time elapsed since last studies	7%
Languages	Number of languages spoken	7%
Positions/years ratio	Number of positions/number of years of professional career	4%

Fig. 28. ROC curve of the social module in the credit scoring model.

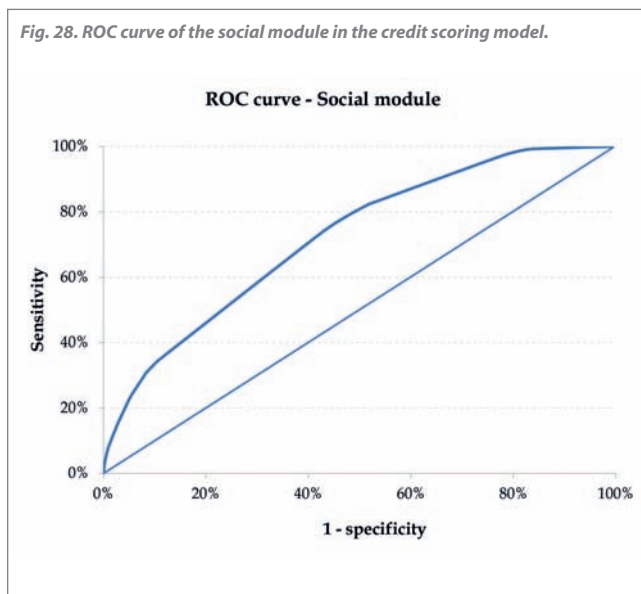
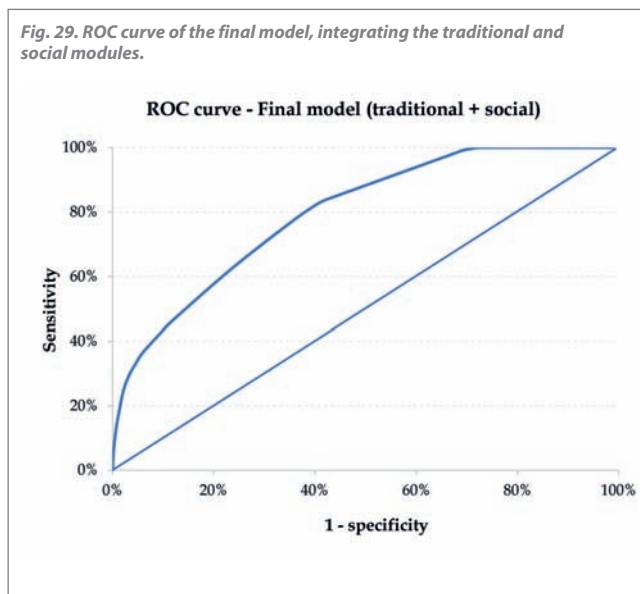


Fig. 29. ROC curve of the final model, integrating the traditional and social modules.



Bibliography

Basel Committee on Banking Supervision 239 (2013). *Principles for effective risk data aggregation and risk reporting*.

BBVA Research (2014). *Situación Latinoamérica. Cuarto trimestre de 2014*.

Berkeley (2015). <http://datascience.berkeley.edu/about/what-is-data-science/>

Bloomberg (2013). *HSBC Judge Approves \$1.9B Drug-Money Laundering Accord*.

DeZyre (2014). *Hadoop in Financial Sector*.

Dhar, V. (2013). *Data Science and Prediction, Association for Computer Machinery*.

Digital Leadership GmbH (2014). *What banks will have to work on over the next couple of years*.

EFMA (2013). *World Banking Report 2013*.

European Banking Authority (2014). *Guidelines on common procedures and methodologies for the supervisory review and evaluation process (SREP)*.

European Central Bank (2014). *ECB Banking Supervision*.

European Commission (2014). *EU Bank Recovery and Resolution Directive (BRRD): Frequently Asked Questions*.

Evans, P. (2014). *How data will transform business. TED*.

Federal Big Data Commission (2014). *Demystifying big data, a practical guide to transforming the business of Government*.

Federal Reserve (2014). *Consumer and Mobile Financial Services 2014*.

Fernald, J. (2014). *Productivity and Potential Output before, during, and after the Great Recession. NBER Working Paper 20248*, National Bureau of Economic Research, Cambridge, Massachusetts.

Financial Conduct Authority (2015). fca.org.uk/firms/being-regulated/enforcement/fines

Gartner (2013). *Gartner Says the Internet of Things Installed Base Will Grow to 26 Billion Units By 2020*.

Gigaom Research (2014). *Sector RoadMap™: data discovery in 2014*.

Goldman, R. (2014). *Big Data, Risk Management, and Full-Fidelity Analytic*. Cloudera.

Gordon, R. (2014). *The Demise of U.S. Economic Growth: Restatement, Rebuttal, and Reflections*. NBER Working Paper 19895, National Bureau of Economic Research, Cambridge, Massachusetts.

Hall, R. (2014). *Quantifying the Lasting Harm to the U.S. Economy from the Financial Crisis*. NBER Working Paper 20183, National Bureau of Economic Research, Cambridge, Massachusetts.

Harvard Business Review (2012). *Data Scientist: The Sexiest Job of the 21st Century*.

Harvard (2014). *CS109 Data Science*.

KPCB (2014). *Internet trends 2014*.

IBM (2014a). *Demystifying Big Data: Decoding The Big Data Commission Report*.

IBM (2014b). *What is a Data Scientist*.

IBM & University of Oxford (2015). *The real world of Big Data*.

Inetco (2015). *Driving Banking Engagement with Customer Analytics*.

International Monetary Fund (2014). *World Economic Outlook, oct. 2014*.

International Telecommunication Union (2014). *The World in 2014, facts and figures*.

Kurzweil, R. (2014). *The accelerating power of technology*.

Management Solutions (2012). *Liquidity risk: regulatory framework and impact on management*.

Management Solutions (2013). *Impact analysis of stress tests in the financial system*.

Management Solutions (2014). *Model Risk Management: quantitative and qualitative aspects*.

McCallum (2014). *Disk Drive Prices (1955-2014)*, jcmmit.com.

McKinsey (2011). *Big data: The next frontier for innovation, competition and productivity*.

Moore, G. (1965). *Cramming more components onto integrated circuits*. Electronics Magazine. p. 4.

Office of the Comptroller of the Currency and Board of Governors of the Federal Reserve System (2011-12). *Supervisory Guidance on Model Risk Management*.

Office of the Comptroller of the Currency (2014). *OCC Guidelines Establishing Heightened Standards for Certain Large Insured National Banks, Insured Federal Savings Associations, and Insured Federal Branches; Integration of Regulations*.

O'Neil, C. y Schutt, R. (2013). *Doing Data Science*. O'Reilly.

Pearn, J. (2012). *What is Google's total computational capacity?*

Pethuru, R. (2014). *Handbook of Research on Cloud Infrastructures for Big Data Analytics*. IGI Global.

Pingdom: royal.pingdom.com

Portio Research (2013). *Portio Research Mobile Factbook 2013*.

SiliconAngle (2014). *When Will the World Reach 8 Zettabytes of Stored Data?*

Wired (2015). *White House Names DJ Patil as the First US Chief Data Scientist*.

World Bank, Sabbata, S. and Graham, M. (2013). *Internet Population 2011 – DeSabbata Graham OII*.

Glossary

Bail-in: rescuing an institution with funds from shareholders and creditors.

Bail-out: rescuing an institution with public funds.

Basel Committee on Banking Supervision (BCBS): supranational body for the prudential regulation of banks, aiming to improve the quality of banking supervision and promote unified supervision standards.

Buffer capital: excess Capital with the purpose of ensuring that an entity is able to absorb losses arising from its activity during periods of stress.

COREP (Common Reporting): regulatory reporting framework issued by the EBA to standardize solvency reporting.

Credit scoring model (credit scoring): a system for the automatic rating of credit risk levels. It is used, inter alia, to calculate the probability of default as well as for automated decisions on the granting of credit.

EBA (European Banking Authority): independent European Union authority whose main objective is to maintain financial stability in the Union and safeguard the integrity, efficiency and functioning of the banking sector. It was established on January 1, 2011 as part of the European System of Financial Supervision (ESFS) and absorbed the former Committee of European Banking Supervisors (CEBS).

FATCA (Foreign Account Tax Compliance Act): federal law requiring financial institutions worldwide to report bank accounts held abroad by US citizens to the United States Internal Revenue Service (IRS). Its aim is to promote fiscal transparency.

Fed (Federal Reserve System): US central bank founded in 1913 with the goal of providing the nation with a safer, more flexible and stable monetary and financial system. Since inception, the Fed's role in the banking and financial sector has

expanded to include activities such as conducting national monetary policy, supervising and regulating banking institutions as well as providing financial services to depository institutions.

Federal Big Data Commission: federal commission whose aim is to provide advice to the US government on how to use data to increase efficiency and reduce costs.

Financial Stability Board (FSB): supranational organization which aims to increase the stability of the global financial system through better coordination between national financial authorities.

FINREP (Financial Reporting): regulatory reporting framework issued by the EBA to standardize the presentation of financial statements.

IAS 39 and IFRS 9: accounting standards for financial instruments: Among their requirements is the calculation of provisions through internal models.

ICAAP (Internal Capital Adequacy Assessment Process): internal process for self- assessment of capital adequacy in the banking industry.

LAAP (Internal Liquidity Adequacy Assessment Process): internal process for self-assessment of liquidity adequacy in the banking industry.

Internet of things: interconnection of everyday objects via the Internet. According to Gartner, by 2020 there will be 26,000 million connected objects in the world.

IRB (Internal Rating Based): advanced method for estimating regulatory capital based on internal rating models. To access this method, financial institutions must meet a set of requirements and obtain permission from the appropriate supervisor.

KYC (Know Your Customer): relevant customer information obtained for various purposes, such as regulatory compliance with respect to fraud, money laundering, terrorist financing or corruption.

MREL (Minimum Requirement for Own Funds and Eligible Liabilities): minimum requirement for internal funds and eligible liabilities for bail-in.

NFC (Near Field Communication): wireless technology for sending and receiving data at high speed over short distances. It is used, among other purposes, to make payments through mobile phones.

NLP (Natural Language Processing): studies the interaction between machines and human language through the analysis of syntactic structures and lexical level, among other elements.

OCC (Office of the Comptroller of the Currency): US federal agency responsible for the regulation and supervision of national banks, federal branches and branches of foreign banks. The OCC's main aim is to ensure that they operate in a safe and sound manner while complying with regulatory requirements, including fair and impartial treatment of customers and customer access to financial markets.

PPNR (Pre-Provision Net Revenue): net revenue before adjusting for loss provisions.

Ring-fencing: financial division of a firm's assets usually carried out for tax, regulatory or security reasons. In the financial industry, it refers to the legal separation between wholesale and traditional banking activities, as a protection measure for depositors.

ROC (Receiver Operating Characteristic) curve: curve used to analyze the predictive power of a binary output model. It represents the relationship between type 1 error (incorrectly classifying adverse events) and type 2 error (incorrectly classifying favorable events).

Single Resolution Board: single resolution authority, operational since January 1, 2015, which is responsible for taking action when a credit institution is not viable.

Single Supervisory Mechanism (SSM): mechanism created in 2014 to assume the supervisory powers of European financial institutions. It is comprised of the European Central Bank and the national supervisory authorities of euro area countries, with the main goals begin to ensure the soundness of the European banking system and to increase financial integration and security in Europe. The SSM directly supervises the 120 most significant institutions and indirectly supervises 3,000 institutions of less significance.

SREP (Supervisory Review and Evaluation Process): process aiming to ensure that financial institutions have adequate processes, capital and liquidity to ensure sound risk management and adequate risk hedging.

Stress test: simulation technique used to determine an institution's resistance to an adverse financial situation. In a broader sense, it refers to any technique used to assess the ability to withstand extreme conditions, and can be applied to institutions, portfolios, models, etc.

TLAC (Total Loss Absorbing Capacity): a requirement aiming to ensure that global systemically important banks (G-SIBs) have, in the event of resolution, the capacity to maintain critical functions during and immediately after the wind-down process without risking taxpayer funds and financial stability.



Our aim is to exceed our clients' expectations, and become their trusted partners

Management Solutions is an international consulting services company focused on consulting for business, risks, organization and processes, in both their functional components and in the implementation of their related technologies.

With its multi-disciplinary team (functional, mathematicians, technicians, etc.) of over 1,400 professionals, Management Solutions operates through its 18 offices (9 in Europe, 8 in the Americas and 1 in Asia).

To cover its clients' needs, Management Solutions has structured its practices by sectors (Financial Institutions, Energy and Telecommunications) and by lines of activity (FCRC, RBC, NT), covering a broad range of skills -Strategy, Commercial Management and Marketing, Organization and Processes, Risk Management and Control, Management and Financial Information, and Applied Technologies.

In the financial sector, Management Solutions offers its services to all kinds of companies -banks, insurance companies, investment firms, financial companies, etc.- encompassing global organizations as well as local entities and public bodies.

Luis Lamas

Partner at Management Solutions
luis.lamas.naveira@msspain.com

Carlos Francisco

Partner at Management Solutions
carlos.francisco.gallejones@msnorthamerica.com

Alberto Rilo

Partner at Management Solutions
alberto.rilo@msunitedkingdom.com

Javier Calvo

Research and Development Director at Management Solutions
javier.calvo.martin@msspain.com



Design and Layout

Marketing and Communication Department
Management Solutions

© Management Solutions. 2015

All rights reserved

www.managementtsolutions.com



Madrid Barcelona Bilbao London Frankfurt Warszawa Zürich Milano Lisboa Beijing
New York San Juan de Puerto Rico México D.F. Bogotá São Paulo Lima Santiago de Chile Buenos Aires