Big data management: How Organizations Create and Implement Data Strategies

Author, Francesco Corea

A Data Science Foundation White Paper

September 2018

----

www.datascience.foundation

Data Science Foundation
Data Science Foundation, Atlantic Business Centre, Atlantic Street, Altrincham, WA14 5NQ
Tel: 0161 926 3641   Email: contact@datascience.foundation Web: www.datascience.foundation
Registered in England and Wales 4th June 2015, Registered Number 9624670

1. Introduction

There are many ways to define what big data is, and this is probably why it still remains a really difficult concept to grasp. Today, someone describes big data as dataset above a certain threshold, e.g., over a terabyte (Driscoll, 2010), others as data that crash conventional analytical tools like Microsoft Excel. More renowned works though identified big data as data that display features of *Variety*, *Velocity*, and *Volume* (Laney, 2001; McAfee and Brynjolfsson, 2012; IBM, 2013; Marr, 2015). Even though they are all partially true, there is a definition that seems to better capture this phenomenon (Dumbill, 2013; De Mauro et al., 2015; Corea, 2016): big data analytics is an innovative approach that consists of different technologies and processes to extract worthy insights from low-value data that do not fit, for any reason, the conventional database systems.

In the last few years the academic literature on big data has grown extensively (Lynch, 2008). It is possible to find specific applications of big data to almost any field of research (Chen et al., 2014). For example, big data applications can be found in medical-health care (Murdoch and Detsky, 2013; Li et al., 2011; Miller, 2012a; 2012b); biology (Howe et al., 2008); governmental projects and public goods (Kim et al., 2014; Morabito, 2015); financial markets (Corea, 2015; Corea and Cervellati, 2015). In other more specific examples, big data have been used for energy control (Moeng and Melhem, 2010), anomaly detection (Baah et al., 2006), crime prediction (Mayer-Schönberger and Cukier, 2013), and risk management (Veldhoen and De Prins, 2014).

No matter what business is considered, big data are having a strong impact on every sector: Brynjolfsson et al. (2011) proved indeed that a data-driven business performs between 5% and 6% better than its competitors. Other authors instead focused their attention on organizational and implementation issues (Wielki, 2013; Mach-Król et al., 2015). Marchand and Peppard (2013) indicated five guidelines for a successful big data strategy: i) placing people at the heart of Big Data initiatives; ii) highlighting information utilization to unlock value; iii) adding behavioral scientists to the team; iv) focusing on learning; and v) focusing more on business problems than technological ones. Barton and Court (2012) on the other hand identified three different key features for exploiting big data potential: choosing the right data, focusing on biggest driver of performance to optimize the business, and transforming the company's capabilities.

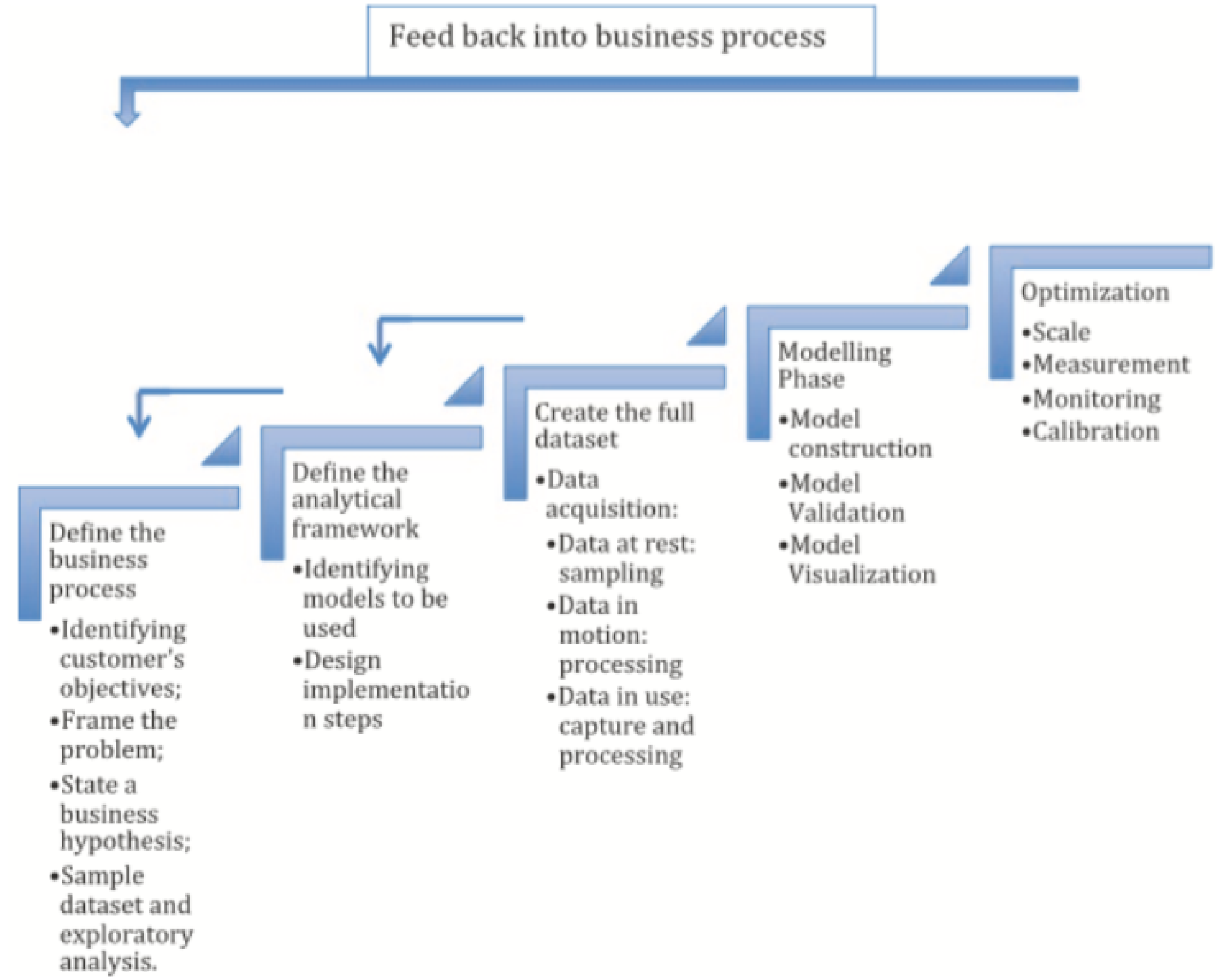
Given then the importance of big data analytics in an organizational environment, this paper proposes a few insights on how to implement internally a data strategy. The rest of the work is pided as follows: in the first section, it will be described a lean approach to data problems. The following section will show a data maturity map, useful to organizations to rationalize on what has already been done and understand how to move forward. The last section will deal with an organizational model for data science teams within a company and the paper will be concluded with some considerations on data projects failure.

2. A Data Lean Approach

Data are quickly becoming a new form of capital, a different coin, and an innovative source of value. It has been mentioned above how relevant it is to channel the power of the big data into an effective strategy to manage and grow the business. However, a consensus on how and what to implement is difficult to be achieved and what is then proposed is only one possible approach to the problem.

Following the guidelines given by Doornik and Hendry (2015), we find a lean approach to data problem to be not only useful but above all efficient. It actually reduces time, effort and costs associated with data collection, analysis, technological improvements and ex-post measuring. The relevance of the framework lies in avoiding the extreme opposite situations, namely collecting all or no data at all. The next figure illustrates key steps towards this lean approach to big data: first of all, business processes have to be identified, followed by the analytical framework that has to be used. These two consecutive stages have feedback loops, as well as the definition of the analytical framework and the dataset construction, which has to consider all the types of data, namely data at rest (static and inactively stored in a database), at motion (inconstantly stored in temporary memory), and in use (constantly updated and store in database). The modeling phase is crucial, and it embeds the validation as well, while the process ends with the scalability implementation and the measurement. A feedback mechanism should prevent an internal stasis, feeding the business process with the outcomes of the analysis instead of improving continuously the model without any business response.

Figure 1. Big data lean deployment approach


Feed back into business process
Define the business process
•Identifying customer's objectives;
•Frame the problem;
•State a business hypothesis;
•Sample dataset and exploratory analysis.
Define the analytical framework
•Identifying models to be used
•Design implementatio n steps
Create the full dataset
•Data acquisition:
•Data at rest: sampling
•Data in motion: processing
•Data in use: capture and processing
Modelling Phase
•Model construction
•Model Validation
•Model Visualization
Optimization
•Scale
•Measurement
•Monitoring
•Calibration

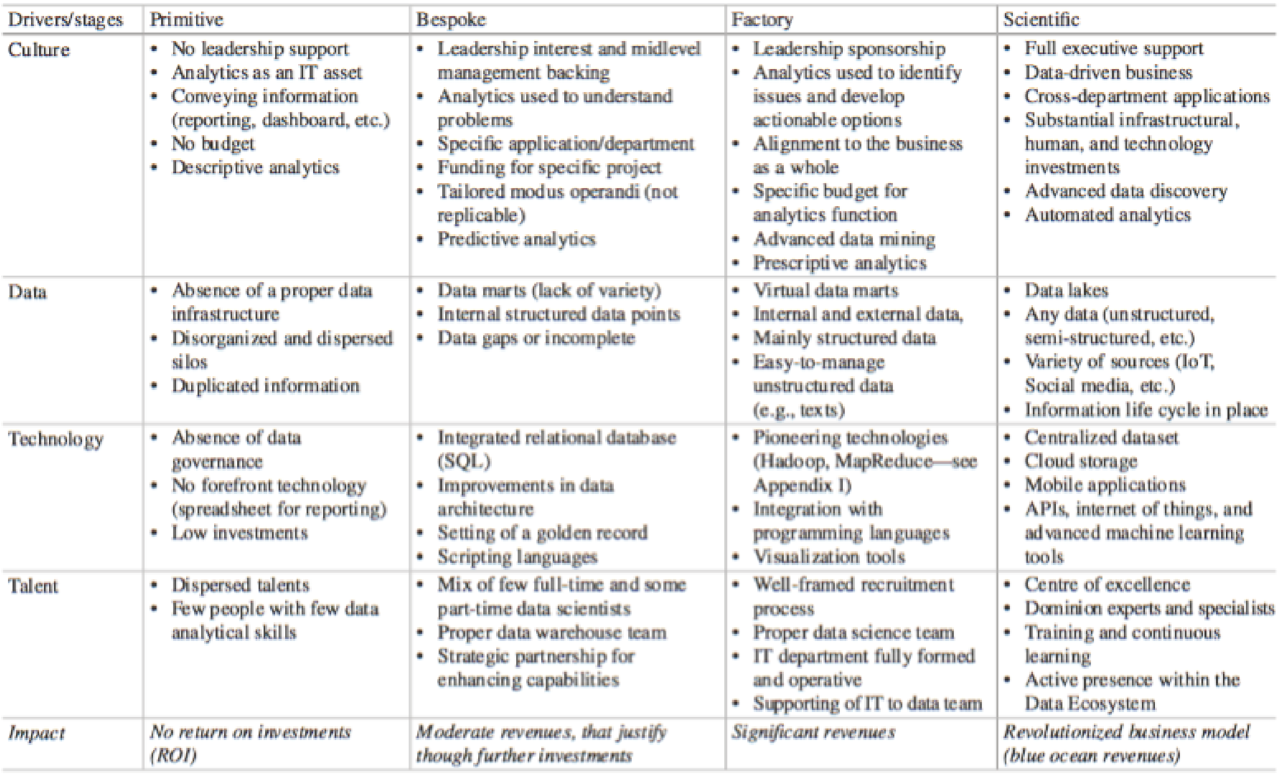
Data need to be consistently aggregated from different sources of information, and integrated with other systems and platforms; common reporting standards should be created – the master copy - and any information should need to be eventually validated to assess accuracy and completeness. Finally, assessing the skills and profiles required to extract value from data, as well as to design efficient data value chains and set the right processes, are two other essential aspects. Having a solid internal data management, jointly with a well-designed golden record, helps to solve the huge issue of *stratified entrance*: dysfunctional datasets resulting from different people augmenting the dataset at different moments or across different layers.

3. A Maturity Map

Even if a data lean approach is used, companies may incur many problems. It is essential then to develop a framework to track internal developments and obstacles, as well as to draw the next steps in the analytics journey. A *Data Stage of Development Structure* (DS2) is a maturity model built for this purpose, a roadmap developed to implement a revenue-generating and impactful data strategy. It can be used to assess a company’s current situation, and to understand the future steps to undertake to enhance internal big data capabilities.

Table 1 provides a four by four matrix where the increasing stages of evolution are indicated as *Primitive*, *Bespoke*, *Factory*, and *Scientific*, while the metrics they are considered through are *Culture*, *Data*, *Technology*, and *Talent*. The final considerations are drawn in the last row, the one that concerns the financial impact on the business of a well-set data strategy.

Table 1. Data Stage of Development Structure.



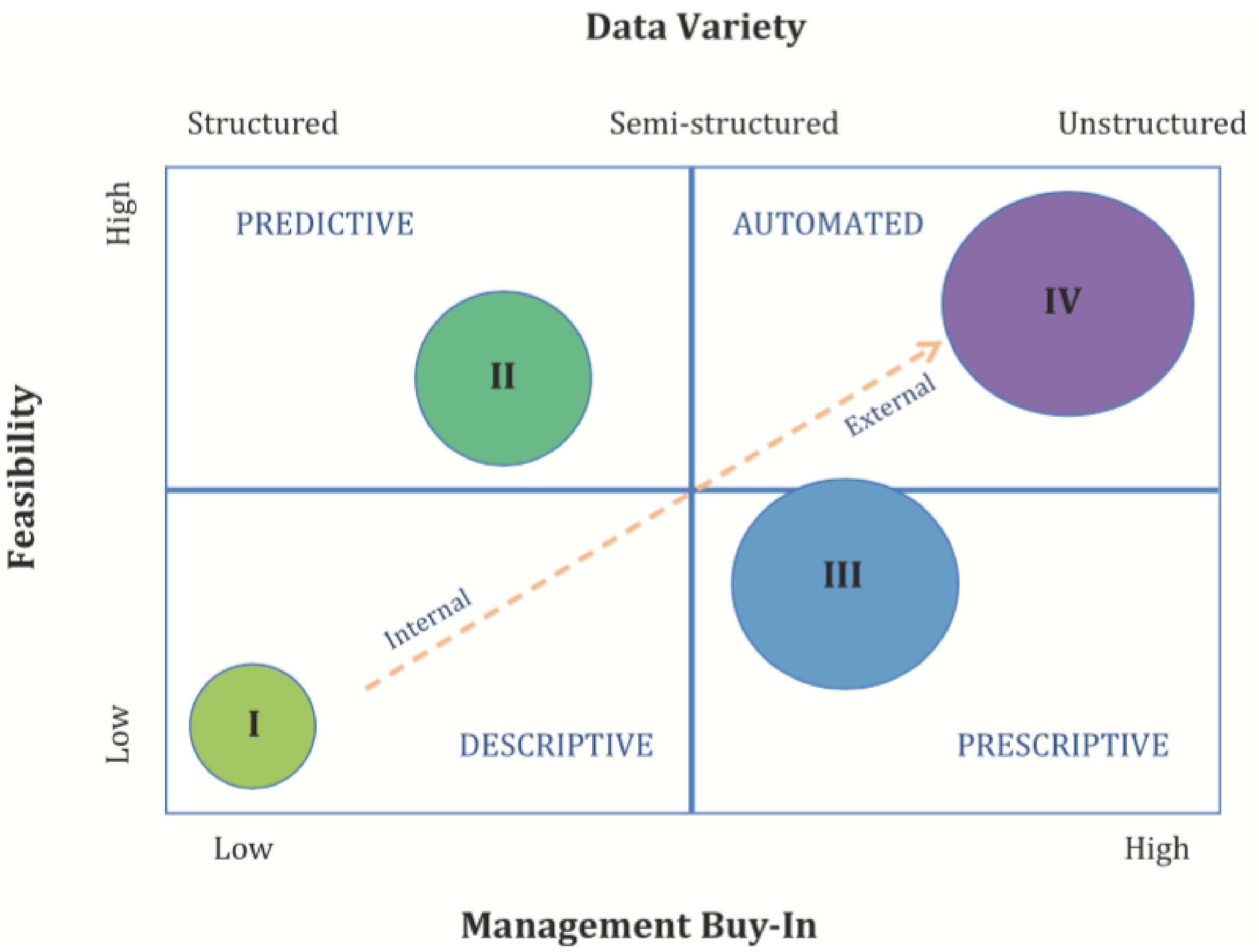
Drivers/stages	Primitive	Bespoke	Factory	Scientific
Culture	• No leadership support • Analytics as an IT asset • Conveying information (reporting, dashboard, etc.) • No budget • Descriptive analytics	• Leadership interest and midlevel management backing • Analytics used to understand problems • Specific application/department • Funding for specific project • Tailored modus operandi (not replicable) • Predictive analytics	• Leadership sponsorship • Analytics used to identify issues and develop actionable options • Alignment to the business as a whole • Specific budget for analytics function • Advanced data mining • Prescriptive analytics	• Full executive support • Data-driven business • Cross-department applications • Substantial infrastructural, human, and technology investments • Advanced data discovery • Automated analytics
Data	• Absence of a proper data infrastructure • Disorganized and dispersed silos • Duplicated information	• Data marts (lack of variety) • Internal structured data points • Data gaps or incomplete	• Virtual data marts • Internal and external data, • Mainly structured data • Easy-to-manage unstructured data (e.g., texts)	• Data lakes • Any data (unstructured, semi-structured, etc.) • Variety of sources (IoT, Social media, etc.) • Information life cycle in place
Technology	• Absence of data governance • No forefront technology (spreadsheet for reporting) • Low investments	• Integrated relational database (SQL) • Improvements in data architecture • Setting of a golden record • Scripting languages	• Pioneering technologies (Hadoop, MapReduce—see Appendix I) • Integration with programming languages • Visualization tools	• Centralized dataset • Cloud storage • Mobile applications • APIs, internet of things, and advanced machine learning tools
Talent	• Dispersed talents • Few people with few data analytical skills	• Mix of few full-time and some part-time data scientists • Proper data warehouse team • Strategic partnership for enhancing capabilities	• Well-framed recruitment process • Proper data science team • IT department fully formed and operative • Supporting of IT to data team	• Centre of excellence • Dominion experts and specialists • Training and continuous learning • Active presence within the Data Ecosystem
Impact	*No return on investments (ROI)*	*Moderate revenues, that justify though further investments*	*Significant revenues*	*Revolutionized business model (blue ocean revenues)*

Stage one is about raising awareness: the realization that data science could be relevant to the company business. In this phase, there are neither any governance structures in place nor any pre-existing technology, and above all no organization-wide buy-in. Yet, tangible projects are still the result of inpidual’s data enthusiasm being channeled into something actionable. The set of skills owned is still rudimental, and the actual use of data is quite rough. Data are used only to convey basic information to the management, so it does not really have any impact on the business. Being at this stage does not mean being inevitably unsuccessful, but it simply shows that the projects performance and output are highly variable, contingent, and not sustainable. The second Phase is the reinforcing: it is actually an exploration period. The pilot has proved big data to have a value, but new competencies, technologies, and infrastructures are required – and especially a new data governance, in order to also take track of possible data contagion and different actors who enter the data analytics process at different stages. Since management’s contribution is still very limited, the potential applications are relegated to a single department or a specific function. The methods used although more advanced than in Phase one are still highly customized and not replicable. By contrast, Phase three adopts a more standardized, optimized, and replicable process: access to the data is much broader, the tools are at the forefront, and a proper recruitment process has been set to gather talents and resources. The projects benefit from regular budget allocation, thanks to the high-level commitment of the leadership team. Step four deals with the business transformation: every function is now data-driven, it is lead by agile methodologies (i.e., deliver value incrementally instead of at the end of the production cycle), and the full-support from executives is translated into a series of relevant actions. These may encompass the creation of a Centre of Excellence (i.e., a facility made by top-tier scientists, with the goal of leveraging and fostering research, training and technology development in the field), high budget and levels of freedom in choosing the scope, or optimized cutting-edge technological and architectural infrastructures, and all these bring a real impact on the revenues' flow. A particular attention has to be especially put on data lakes, repositories that store data in native formats: they are low costs storage alternatives, which supports manifold languages. Highly scalable and centralized stored, they allow the company to switch without extra costs between different platforms, as well as guarantee a lower data loss likelihood. Nevertheless, they require a metadata management that contextualizes the data, and strict policies have to be established in order to safeguard the data quality, analysis, and security. Data have to be correctly stored, studied through the most suitable means, and to be breach-proof. An information lifecycle has to be established and followed, and it has to take particular care of timely efficient archiving, data retention, and testing data for the production environment.

A final consideration has to be spared about cross-stage dimension "culture". Each stage has associated a different kind of analytics, as explained in Davenport (2015). Descriptive analytics concerned what happened, predictive analytics is about future scenarios (sometimes augmented by diagnostic analytics, which investigates also the causes of a certain phenomenon), prescriptive analytics suggests recommendations, and finally, automated analytics are the ones that take action automatically based on the analysis' results.

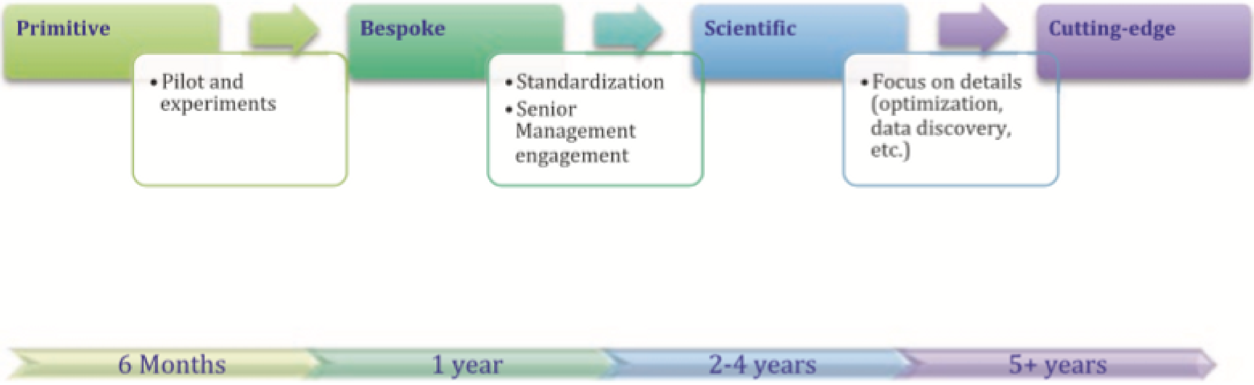
Some of the outcomes presented so far are summarized in Figure 2. The following chart shows indeed the relationship between management's support for the analytics function and the complexity and skills required to excel into data- driven businesses. The horizontal axis shows the level of commitment by the management (high vs. low), while the vertical axis takes into account the feasibility of the project undertaken – where feasibility is here intended as the ratio of the project's complexity and the capabilities needed to complete it. The intersection between feasibility of big data analytics and management involvement pides the matrix into four quarters, corresponding to the four types of analytics. Each circle identifies one of the four stages (numbered in sequence, from I – *Primitive*, to IV – *Scientific*). The size of each circle indicates its impact on the business (i.e., the larger the circle, the higher the ROI). Finally, the second horizontal axis keeps track of the increasing data variety used in the different stages, meaning structure, semi-structured, or unstructured data (i.e., IoT, sensors, etc.). The orange diagonal represents what kind of data are used: from closed systems of internal private networks in the bottom left quadrant to market/public and external data in the top right corner.

Figure 2. Big Data Maturity Map



Once the different possibilities and measurements have been identified (see Corea, 2016 for the full details on the framework), they can be used to understand what stage a firm belongs to. It is also useful to see how a company could transition from one level to the next and in the following figure some recommended procedures have been indicated to foster this transition.

Figure 3. Maturity stage transitions.



In order to smoothly move from the *Primitive* stage to the *Bespoke*, it is necessary to proceed by experiments run from single inpiduals, who aim to create proof of concepts or pilots to answer a single small question using internal data. If these questions have a good/high-value impact on the business, they could be acknowledged faster. Try to keep the monetary costs low as possible (cloud, open source, etc.), since the project will be already expensive in terms of time and manual effort. On a company level, the problem of data duplication should be addressed. The transition from *Bespoke* to *Factory* instead demands the creation of standard procedures and golden records, and a robust project management support. The technologies, tools, and architecture have to be experimented, and thought as they are implemented or developed to stay. The vision should be medium/long term then. It is essential to foster the engagement of the higher- senior management level. At a higher level, new sources and type of data have to be promoted, data gaps have to be addressed, and a strategy for platforms resiliency should be developed. In particular, it has to be drawn down the acceptable data loss (*Recovery Point Objective*), and the expected recovered time for disrupted units (*Recovery Time Objective*). Finally, to become data experts and leaders and shifting to the *Scientific* level, it is indispensable to focus on details, optimize models and datasets, improve the data discovery process, increase the data quality and transferability, and identifying a blue ocean strategy to pursue. Data security and privacy are essential, and additional transparency on the data approach should be released to the shareholders. A Centre of Excellence (CoE) and a talent recruitment value chain play a crucial role as well, with the final goal to put the data science team in charge of driving the business. The CoE is indeed fundamental in order to mitigate the short-term performance goals that managers have, but it has to be reintegrated at some point for the sake of scalability. It would be possible now to start documenting and keeping track of improvements and ROI. From the final step on, a process of continuous learning and forefront experimentations is required to maintain a leadership and attain respectability in the data community.

In Figure 3 it has also been indicated a suggested timeline for each step, respectively up to six months for assessing the current situation, doing some research and starting a pilot; up to one year for exploiting a specific project to understand the skills gap, justify a higher budget allocations, and plan the team expansion; two to four years to verify the complete support from every function and level within the firm, and finally at least five years to achieving a fully- operationally data-driven business. Of course, the time needed by each company varies due to several factors, so it should be highly customizable.

4. The Organization Model

A few more words should be spent regarding the organizational home for data analytics (Pearson and Wegener, 2013). We claimed that the Centre of Excellence is the cutting-edge structure to incorporate and supervise the data functions within a company. Its main task is to coordinate cross-units activities, which embeds: maintaining and upgrading the technological infrastructures; deciding what data have to be gathered and from which department; helping with the talents recruitment; planning the insights generation phase, and stating the privacy, compliance, and ethics policies. However, other forms may exist, and it is essential to know them since sometimes they may fit better into the preexisting business model.

Figure 4. Data analytics organizational models

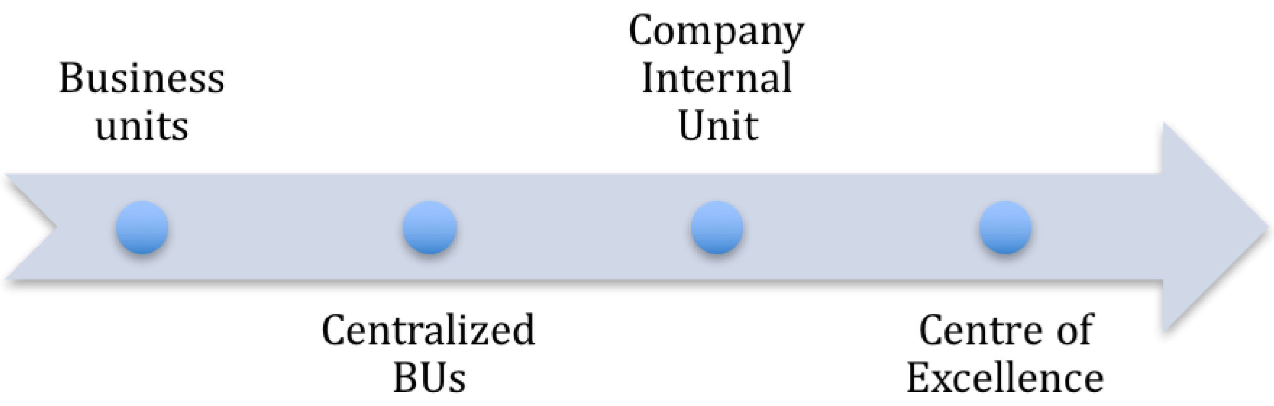


Figure 4 shows different combinations of data analytics independence and business models. It ranges between business units (BUs) that are completely independent one from the other, to independent BUs that join the efforts in some specific projects, to an internal (corporate center) or external (center of excellence) center that coordinates different initiatives.

In spite of everything, all the considerations made so far mean different things and provide singular insights depending on the firm's peculiarities. In particular, the different business life cycle phase in which the company is operating deeply influences the type of strategy to be followed, and it is completely unrelated to the maturity data stage to which they belong (i.e., a few months old company could be a *Scientific* firm, while a big investment bank only a *Primitive* one).

5. Conclusion

It is not clear when a company should start worrying about switching or going for a big data strategy. Of course, there is not a unique standard answer, because the solution is tightly related to business specificities, but broadly speaking it is necessary to start thinking about big data when every source of competitive advantage is fading away or slowing down, i.e., when the growth of revenues, clients acquisitions, etc., reaches a plateau. Big data are drivers of innovation, and this approach could be the keystone to regain a competitive advantage and to give new nourishment to the business. However, it should be clear by now that this is not something that may happen overnight, but it is rather a gradual cultural mind-shift that requires many small steps to be undertaken.

The insights proposed in this work do not guarantee the success of the strategy but for sure lower the likelihood of incurring in common failure scenarios. It happens often indeed that the scope is inaccurate because of lacking proper objectives or too high ambitions. On the other hand, the excessive costs and time employed in developing efficient project result from high expectations as well as an absence of scalability. Managing correctly expectations and metrics to measure the impact of big data on the business is essential to succeed in the long term.

References

Baah, G. K., Gray, A., Harrold, M. J. (2006). “Online anomaly detection of deployed software: a statistical machine learning approach”. In Proceedings of the 3rd international workshop on Software quality assurance: 70-77.

Barton, D., Court, D. (2012). “Making Advanced Analytics Work for You”. Harvard business review, 90 (10): 78-83.

Brynjolfsson, E., Hitt, L. M., and Kim, H. H. (2011). “Strength in Numbers: How Does Data-Driven Decisionmaking Affect Firm Performance?”. Available at SSRN: http://ssrn.com/abstract=1819486.

Chen, M., Mao, S., Zhang, Y., Leung, V.C. (2014). “Big Data: Related Technologies, Challenges and Future Prospects, SpringerBriefs in Computer Science, 59.

Corea, F. (2015). “Why social media matters: the use of Twitter in portfolio strategies”. International Journal of Computer Applications 128 (6), 25- 30.

Corea, F. (2016). “Big Data Analytics: A Management Perspective”. Studies Series in Big Data, 21. Springer International Publishing.

Corea, F., Cervellati, E. M. (2015). “The Power of Micro-Blogging: How to Use Twitter for Predicting the Stock Market”. Eurasian Journal of Economics and Finance 3 (4), 1-6.

Davenport, T. H. (2015). “The rise of automated analytics”. The Wall Street Journal, January 14th 2015. Accessed on October 30th 2015 (available at http://www.tomdavenport.com/wp-content/uploads/The-Rise-of-Automated- Analytics.pdf)

De Mauro, A., Greco, M., Grimaldi, M. (2015). “What is big data? A consensual definition and a review of key research topics”. AIP Conference Proceedings, 1644: 97-104.

Doornik, J. A., Hendry, D. F. (2015). “Statistical model selection with Big Data”. Cogent Economics & Finance, 3: 1045216.

Driscoll, M. E. (2010, Dec. 20). “How much data is "Big Data"?”, [Msg 2]. Message posted to https://www.quora.com/How-much-data-is-Big-Data.

Dumbill, E (2013). “Making Sense of Big Data”. Big Data, 1 (1): 1-2.

Howe, A. D., Costanzo, M., Fey, P., Gojobori, T., Hannick, L., Hide, W., ... Rhee, S. Y. (2008). “Big data: The future of biocuration”. Nature, 455 (7209): 47–50.

IBM (2013). “The Four V's of Big Data”. Retrieved from http://www.ibmbigdatahub.com/infographic/four-vs-big-data .

Kim, G. H., Trimi, S., Chung, J. H. (2014). “Big-data applications in the government sector”. Communications of the ACM, 57 (3): 78-85.

Laney D. (2001). “3D Data Management: Controlling Data Volume, Velocity, and Variety”. META group Inc., 2001. http://blogs.gartner.com/doug- laney/files/ 2012/01/ad949-3D-Data-Management-Controlling-Data-V olume- Velocity-and-Variety.pdf. Accessed on Oct 27, 2015.

Li, Y., Hu, X., Lin, H., Yang, Z. (2011). “A framework for semisupervised feature generation and its applications in biomedical literature mining”. IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB), 8 (2): 294–307.

Lynch, C. (2008). “Big data: How do your data grow?”. Nature 455: 28- 29.

Mach-Król, M., Olszak, C. M., Pełech-Pilichowski, T. (2015). Advances in ICT for Business, Industry and Public Sector. Studies in Computational Intelligence, Springer, 200 pages.

Marchand, D., Peppard, J. (2013). “Why IT fumbles analytics”. Harvard Business Review, 91 (1/2): 104-113.

Marr, B. (2015). Big Data: Using SMART Big Data, Analytics and Metrics To Make Better Decisions and Improve Performance. Wiley, 256 pages.

Mayer-Schönberger, V., Cukier, K. (2013). Big Data: A Revolution that Will Transform How We Live, Work, and Think. Eamon Dolan/Houghton Mifflin Harcourt.

McAfee, A., Brynjolfsson, E. (2012). “Big Data: The Management Revolution”. Harvard business review, 90

(10): 60-6.

Miller, K. (2012a). “Leveraging Social Media for Biomedical Research: How Social Media Sites Are Rapidly Doing Unique Research on Large Cohorts”. Biomedical Computation Review (available at http://biomedicalcomputationreview.org/content/leveraging-social-media- biomedical-research; accessed October 27, 2015).

Miller, K. (2012b). “Big Data Analytics in Biomedical Research,” Biomedical Computation Review (available at http:// biomedicalcomputationreview.org/content/big-data-analytics- biomedical- research; accessed October 27, 2015).

Moeng, M., Melhem, R. (2010). “Applying statistical machine learning to multicore voltage and frequency scaling”. In Proceedings of the 7th ACM international conference on Computing frontiers: 277–286.

Morabito, V. (2015). Big Data and Analytics: Strategic and Organizational Impacts. Springer International Publishing, 183 pages.

Murdoch, T. B., Detsky, A. S. (2013). “The Inevitable Application of Big Data to Health Care”. JAMA, 309 (13): 1351-1352.

Pearson, T., Wegener, R. (2013). “Big data: the organizational challenge”. Bain & Company White paper.

Veldhoen, A., De Prins, S. (2014). “Applying Big Data to Risk Management”. Avantage Reply Report: 1-14.

Wielki, J. (2013). “Implementation of the Big Data concept in organizations – possibilities, impediments, and challenges”. Proceedings of the 2013 Federated Conference on Computer Science and Information Systems: 985- 989.

This is an excerpt from my forthcoming book “Introduction to Data” edited by Springer (2019).

About the Data Science Foundation

The Data Science Foundation is a professional body representing the interests of the Data Science Industry. Its membership consists of suppliers who offer a range of big data analytical and technical services and companies and individuals with an interest in the commercial advantages that can be gained from big data. The organisation aims to raise the profile of this developing industry, to educate people about the benefits of knowledge based decision making and to encourage firms to start using big data techniques.

Contact Data Science Foundation

Email:contact@datascience.foundation
Telephone: 0161 926 3641
Atlantic Business Centre
Atlantic Street
Altrincham
WA14 5NQ
web: www.datascience.foundation