© 2017 by the authors; licensee RonPub, Lübeck, Germany. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution license (http://creativecommons.org/licenses/by/4.0/).

Research online Publishing www.ronpub.com

Open Access

Open Journal of Big Data (OJBD)
Volume 3, Issue 1, 2017
http://www.ronpub.com/ojbd
ISSN 2365-029X

Combining Process Guidance and Industrial Feedback for Successfully Deploying Big Data Projects

Christophe Ponsard [A], Mounir Touzani [B], Annick Majchrowski [A]

[A] CETIC Research Centre, avenue Jean Mermoz 18, 6041 Gosselies, Belgium, {cp, am}@cetic.be
[B] Académie de Toulouse, rue Saint-Roch 75, 31400 Toulouse, France, mounir.touzani@ac-toulouse.fr

Abstract

Companies are faced with the challenge of handling increasing amounts of digital data to run or improve their business. Although a large set of technical solutions are available to manage such Big Data, many companies lack the maturity to manage that kind of projects, which results in a high failure rate. This paper aims at providing better process guidance for a successful deployment of Big Data projects. Our approach is based on the combination of a set of methodological bricks documented in the literature from early data mining projects to nowadays. It is complemented by learned lessons from pilots conducted in different areas (IT, health, space, food industry) with a focus on two pilots giving a concrete vision of how to drive the implementation with emphasis on the identification of values, the definition of a relevant strategy, the use of an Agile follow-up and a progressive rise in maturity.

Type of Paper and Keywords

Regular research paper: *big data, process model, agile, method adoption, pilot case studies*

1 Introduction

In today's world, there is an ever-increasing number of people and devices that are being connected together. This results in the production of information at an exponentially growing rate and opens the Big Data area. To give a few numbers, it is estimated that 90% of the current world's data has been produced in just the last two years, and that the amount of data created by businesses doubles every 1.2 years [45]. The total amount of data in the world reached one zettabyte (10^{21} bytes) around 2010, and by 2020 more than 40 zettabytes will be available. An important shift is that most of the data is now being generated by devices rather than people, due to the emergence of the Internet of Things.

Companies are facing the many challenges of processing such amounts of data. They typically view Big Data technologies as holding a lot of potential to improve their performance and create competitive advantages. The main challenges companies have to face with Big Data are often summarised by a series of "V" words. In addition to the *Volume* (i.e. the risk of information overload) already mentioned, other data dimensions are the *Variety* (i.e. the diversity of structured and non-structured formats), the required *Velocity* (i.e. highly reactive, possibly real-time, data processing), the *Visualization* need (in order to interpret them easily) and the related *Value* (in order to derive an income) [37].

The ease to collect and store data, combined with the availability of analysing technologies (such as NoSQL Databases, MapReduce, Hadoop) has encouraged many companies to launch Big Data projects. However, most

organisations are actually still failing to get business value out of their data. A 2013 report surveying 300 companies about Big Data revealed that 55% of Big Data projects fail and many others fall short of their objectives [31]. An on-line survey conducted in July 2016 by Gartner reported that many companies remain stuck at the pilot stage and that only 15% actually deployed their big data project to production[25].

Looking at the cause of such failures, it appears that the main factor is actually not the technical dimension, but rather the process and people dimensions, which are thus equally important [24]. Of course the technology selection for big data projects is important and needs to be kept up-to-date with the fast technological evolution to help selecting proper technologies [33]. However, much less is devoted to methods and tools that can help teams to achieve big data projects more effectively and efficiently [46]. There exists some recent work in that area, identifying key factors for a projects success [47], stressing management issues [12], insisting on the need for team process methodologies and making a critical analysis of analytical methods [46].

Our paper is aligned with those works and aims at helping companies engaging in a Big Data adoption process to be driven by questions such as:

- How can we be sure Big Data will help us?
- Which people with what skills should be involved?
- What steps should be done first?
- Is my project on the right track?

Our main contribution is composed of practical guidelines and lessons learned from a set of pilot projects covering various domains (life sciences, health, space, IT). Those pilots are spread over three years and are conducted within a large project carried out in Belgium. They are following a similar process which is incrementally enhanced. The reported work is based on the first four pilots while four others are in analysis phase. It significantly extends our first report published in [43] by:

- giving a more detailed overview of existing methodologies that form the building bricks of our approach,
- proving a detailed feedback over two industrial pilots respectively in the data centre maintenance and medical care domains,
- putting our work in the light of other work focusing on the successful adoption of Big data techniques. We also discuss in more detail some important issues like ethics and cybersecurity.

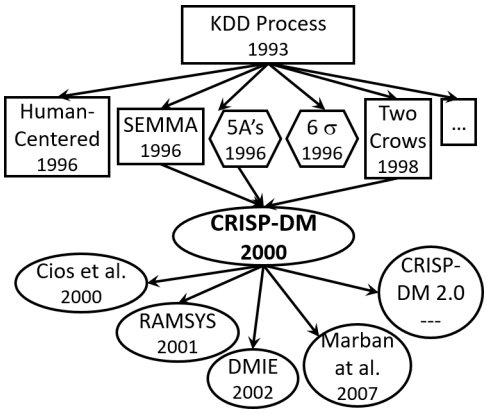

Figure 1: Evolution of data processing methodologies (source: [36])

This paper is structured as follows. Section 2 reviews the main available methodologies for dealing with Big Data deployment. Section 3 presents the process followed to come up with a method and validate it on our pilots. It stresses key requirements for successful deployment. Section 4 presents more detailed feedback and highlight specific guidelines. Section 5 discusses some related work. Finally Section 6 draws some conclusions and on-going extensions of our work.

2 Evolution of Methods and Processes for Data Oriented Projects

This section reviews existing methods and processes. It highlights some known strengths and limitations. First, methods inherited from the related data mining field are presented before considering approaches more specific to Big Data with a special attention to Agile methods.

2.1 Methods Inherited from Data Mining

Data mining was developed in the 1990's to extract data patterns in structured information (databases) and to discover business factors on a relatively small scale. In contrast, Big Data is also considering unstructured data and operates on a larger scale. A common point between them, from a process point of view, is that both require the close cooperation of data scientists and management in order to be successful. Many methodologies and process models have been developed for data mining and knowledge discovery [36]. Figure 1 give an overview of the evolution and parenthood of the main methodologies.

The seminal approach is KDD (Knowledge Discovery in Database) [22]. It was refined into many other approaches (like SEMMA [48], Two Crows [53]). It was then standardised under CRISP-DM (Cross Industry

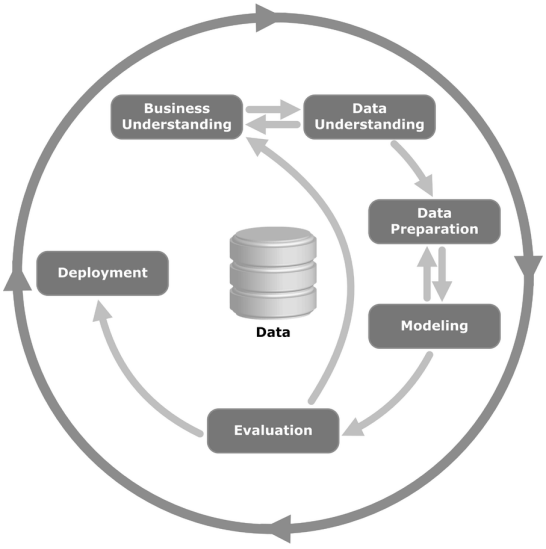

Figure 2: CRISP-DM method (source: [30])

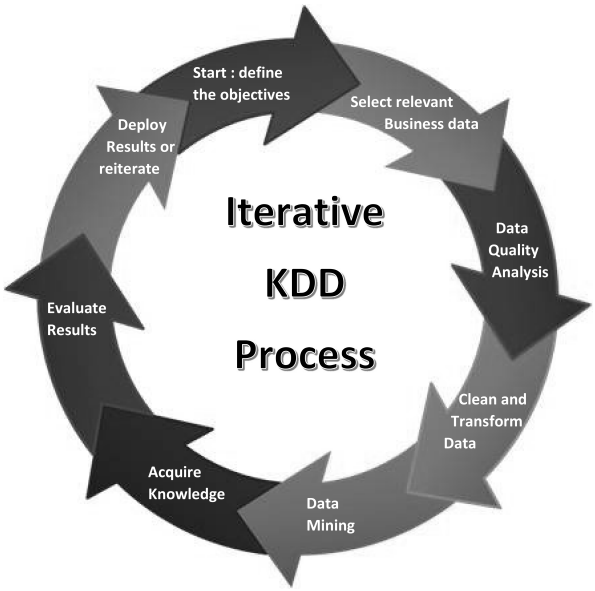

Figure 3: Agile KDD method (source: [11])

Standard Process for Data Mining) [50] which is depicted in Figure 2.

CRISP-DM is composed of six main phases, each one is decomposed in sub-steps. The process is not linear but rather organised as a global cycle with usually a lot of back and forth within and between phases. CRISP-DM has been widely used for the past 20 years, not only for data mining but also for predictive analytics and big data projects.

However, CRISP-DM and alike methods suffer from the following issues:

- they fail to provide a good management view on communication, knowledge and project aspects,
- they lack some form of maturity model enabling to highlight more important steps and milestones that can be progressively raised,
- despite the standardisation, they are not always known by the wider business community, hence difficult to adopt for managing the data value aspect,
- the proposed iterative model is limited: the planned iterations are little used in practice because they do not loop back to the business level but rather stay in the internal IT context. In addition to the lack of control on the added value, those iterations are very often postponed. This is the reason why more agile models were introduced.

2.2 Methods Adopting Agile Principles

Agile methods, initially developed for software development, can also be applied to data analysis in order to provide a better process guidance and value orientation. An agile evolution of KDD and CRISP-DM is AgileKDD [15] depicted in Figure 3. It is based on the OpenUP life cycle which supports the statement in the Agile Manifesto [1]. Projects are divided in planned "sprints" with fixed deadlines, usually a few weeks. Each sprint needs to deliver incremental value to stakeholders in a predictable and demonstrable manner.

For example, IBM has developed ASUM-DM, an extension and refinement of CRISP-DM combining traditional project management with agility principles [26]. Figure 4 illustrates its main blocks and its iterative principle driven by specific activities at the level of the last columns. These include governance and community alignment. However, it does not cover the infrastructure/operations side of implementing a data mining/predictive analytics project. It is more focused on activities and tasks in the deployment phase and has no templates nor guidelines.

Although it looks quite adequate, deploying an Agile approach for Big Data may still face resistance, just as it is the case for software development, typically in a more rigid kind of organisation. A survey was conducted to validate this acceptance [23]. It revealed that quite similarly as for software development, companies tend to accept Agile methods for projects with smaller scope, lesser complexity, fewer security issues and inside organisation with more freedom. Otherwise, a more traditional plan-managed approach is preferred.

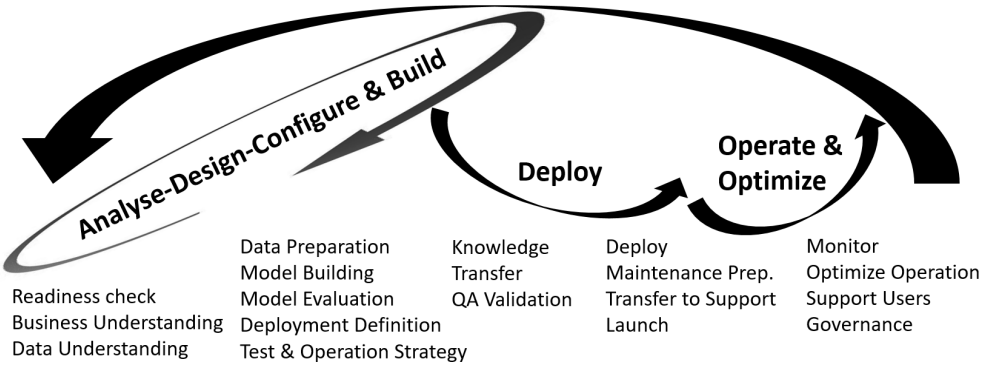

Figure 4: ASUM-DM method (source: [29])

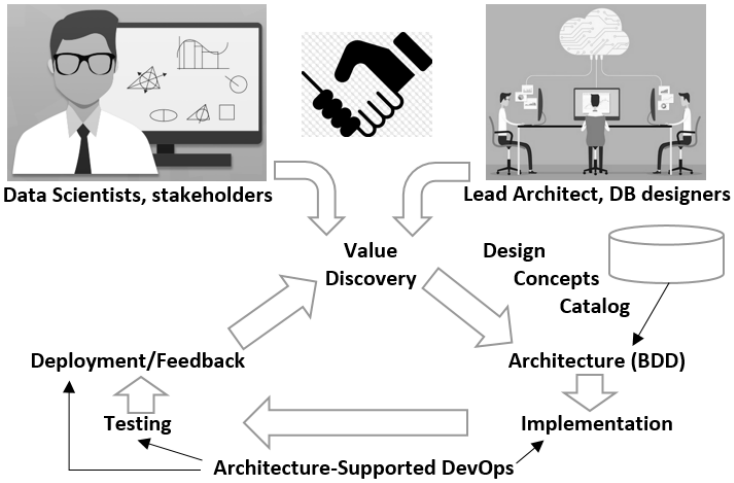

Figure 5: AABA Method (source: [9])

2.3 Methods Developped for Big Data Projects

Architecture-centric Agile Big data Analytics (AABA) addresses technical and organizational challenges of Big Data [9]. Figure 5 shows it supports an Agile delivery. It also integrates the Big Data system Design (BDD) method and Architecture-centric Agile Analytics with architecture-supported DevOps (AAA) model for effective value discovery and continuous delivery of value.

The method was validated on 11 case studies across various domains (marketing, telecommunications, healthcare) with the following recommendations:

1. Data Analysts/Scientists should be involved early in the process, i.e. already at business analysis phase.

2. Continuous architecture support is required for big data analytics.

3. Agile bursts of effort help to cope with rapid technology changes and new requirements.

4. The availability of a reference architecture and a technology catalog ease the definition and evolution of the data processing.

5. Feedback loops need to be open, e.g. about non-functional requirements such as performance, availability and security, but also for business feedback about emerging requirements.

Stampede is another method proposed by IBM to its customers. Expert resources are provided at cost to help companies to get started with Big Data in the scope of a well-defined pilot project [28]. Its main goal is to educate companies and help them get started more quickly, in order to drive value from Big Data. A key tool of the method is a half day workshop to share definitions, identify scope/big data/infrastructure, establish a plan and most importantly establish the business value. The pilot execution is typically spread over 12 weeks and carried out in an Agile way with a major milestone at about 9 weeks as depicted in Figure 6.

Figure 6: The IBM Stampede method (source: [28])

Table 1: Maturity Model from Nott and Betteridge (IBM) (source: [39])

Level	Ad hoc	Foundational	Competitive	Differentiating	Breakaway
Business strategy	Use of standard reporting. Big Data is just mentioned	Data-related ROI identified	Data processing encouraged	Competitive advantage achieved	Business innovation is driven by data processing
Analytics	Limited to the past	Event detection	Prediction of the likelihood of specific evolution	Optimisation of decision support	Optimisation and process automation
IT Alignment	No coherent architecture of the information system	Define architecture but not oriented towards analytics	Definition of Big Data architectural patterns for Big Data	Defined and standardised Big Data oriented architecture	Architecture fully aligned with Big Data requirements
Culture and governance	Largely based on key people	Rambling artefact management, resistance to change	Policy and procedure well defined, partial adoption	Large adoption, daily use	Generalised Adoption

2.4 Some Complementary Approaches

2.4.1 Introducing Maturity Models

Some attempts are being made to develop some "Capability Maturity Model" (CMM) for scientific data management processes in order to support the evaluation and improvement of these processes [13, 39]. Such a model describes the main types of processes and the practices required for an effective management. Classic CMM characterizes organizations using different maturity levels that represent their ability to reliably perform processes of growing complexity and scope. A 5-level scale is typical and proposed both by [13] and [39]. The first uses standard levels ranging from "defined" to "optimized" while the latter uses a more specific nomenclature ranging from "ad hoc" to "breakaway". Table 1 details the main criteria relating to the place of the data in the business strategy, the type of data analysis used, the alignment of the IT infrastructure, as well as aspects of culture and governance.

2.4.2 Cognitive "Sensemaking"

The Sensemaking approach has also an iterative nature. There are actually two internal cycles in the method: a more classical "Foraging" loop trying to dig into the data to find out relations among them and a second approach "Sensemaking" loop trying to build sense out of the data by reproducing the cognitive process followed by humans in order to build up a representation of an information space for achieving his/her goal. It focuses on challenges for modelling and analysis by bringing cognitive models into requirements engineering, in order to analyse the features of data and the details of user activities [32].

2.4.3 Critical Success Factors

In complement to processes, many key success factors, best practices and risk checklists have been published, mostly in blogs for Chief Information Officers, e.g. [4]. A systematic classification of Critical Success Factors has been proposed by [24] using three key dimensions: people, process and technology. It has been further extended by [46] with tooling and governance dimensions. A few key factors are the following:

- Data: quality, security, level of structure in data.
- Governance: management support, well-defined organisation, data-driven culture.
- Objectives: business value identified (KPI), business case-driven, realistic project size.
- Process: agility, change management, maturity, coping with data growth.
- Team: data science skills, multidisciplinarity.
- Tools: IT infrastructure, storage, data visualization capabilities, performance monitoring.

3 Method Development and Validation Process

The global aim of our project is to come up with a systematic method to help companies facing big data challenges to validate the potential benefits of a big data solution. The global process is depicted in Figure 7.

The process is driven by eight successive pilots which are used to tune the method and make more technical bricks available through the proposed common infrastructure. The final expected result is to provide a commercial service to companies having such needs.

The selected method is strongly inspired by what we learned from the available methods and processes described in Section 2:

- the starting point was Stampede because of some initial training and the underlying IBM platform. Key aspects kept from the methods are the initial workshop with all stakeholders, the realistic focus and a constant business value driver,
- however, to cope with the lack of reference material, we defined a process model based on CRISP-DM which is extensively documented,
- the pilots are executed in an Agile way, given the expert availabilities (university researchers), the pilots are planned over longer periods than in Stampede: 3-6 months instead of 12-16 weeks. The popular SCRUM approach was used as it emphasizes collaboration, functioning software, team-self management and flexibility to adapt to business realities [49].

The global methodology is composed of three successive phases detailed hereafter:

1. *Big Data Context and Awareness*. In this introductory phase, one or more meetings take place with the target organisation. A general introduction is given on Big Data concepts, the available platform, a few representative applications in different domains (possibly with already a focus on the organisation domain), the main challenges and main steps. The maturity of the client and a few risk factors can be checked (e.g. management support, internal expertise, business motivation).
2. *Business and Use Case Understanding*. This is also the first phase of CRISP-DM. Its goals are to collect the business needs/problems that must be addressed using Big Data and also to identify one or some business use cases generating the most value out of the collected data. A checklist supporting this phase is shown in Table 2.

Table 2: Workshop checklist (source: [29])

Business Understanding	Use Case Understanding
Strategy & Positioning	**Assess Business Maturity**
Global Strategy	
Product/Services Positioning	**Determine**
Scorecards - KPI's	**Use Case Objectives**
Digital Strategy	Value to the Client
Touchpoints for	Business Success Criteria
Customers/Prospects	
Search, e-Commerce,	**Assess Situation**
Social Media, Websites,...	Resource Requirements
Direct Competitors	Assumptions/Constraints
Disruptive Elements	Risks and Contingencies
Disruptive Models	Terminology
Disruptive Technologies	Costs and Benefits
Disruptive Behaviour	
Select Potential Use Cases	**Refine Data Mining Goals**
Objectives	Data Mining Goals
Priorities	Data Mining KPIs
Costs, ROI, Constraints	
Value to the Client	**Produce Project Plan**
Scope	Approach
New Data Source	Deliverables
	Schedule
High Level Feasibility	Risk Mitigation (Privacy,...)
Data Mining Goals & KPIs	Stakeholders to involve
Resources Availability	Initial Assessment of
Time to Deliver	Tools and Techniques

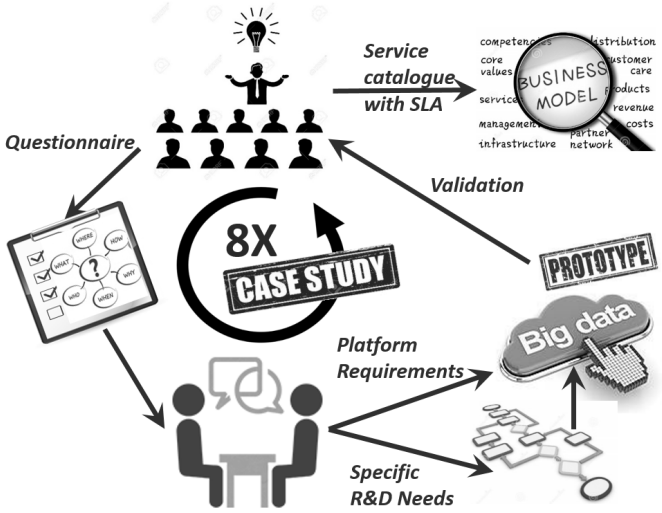

Figure 7: Iterative development of the platform and method

This phase is organised based on one or a few workshops, involving the Business Manager, Data Analyst, IT architect and optionally selected specialists, such as the IT security manager if there are specific security/privacy issues that need to be checked at this early stage. Both the as-is and to-be are considered. Specific tools to support the efficient organisation of those workshops are described in Section 4. At the end of this step, a project planning is also defined.

3. *Pilot Implementation of Service or Product*. In this phase, the following implementation activities are carried out in an Agile way:

   - Data Understanding: analyse data sets to detect interesting subset(s) for reaching the business objective(s) and make sure about data quality.
   - Data Preparation: select the right data and clean/extend/format them as required.
   - Modelling: select specific modelling techniques (e.g. decision-tree or neural networks). The model is then built and tested for its accuracy and generality. Possible modelling assumptions are also checked. The model parameters can be reviewed or other/complementary techniques can be selected.
   - Evaluation: assess the degree to which the model meets business objectives, using realistic or even real data.
   - Deployment: transfer the validated solution to production environment, make sure user can use it (e.g. right visualization, dashboard) and start monitoring (performance, accuracy).

Our pilots are kept confidential. However, Table 3 presents the main features of the first four pilots based on the three first "V" of Big Data [37].

4 Lessons and Recommendations Learned from our Pilot Cases

In this section, we present some lessons learned and related guidelines that are useful to support the whole process and increase the chances of success. We also illustrate our feedback based on some highlights from two pilot cases used as running examples: the IT maintenance pilot case and the clinical pathway pilot case.

4.1 Defining Progressive and Measurable Objectives.

Through the deployment of a Big Data solution, a company expects to gain value out of its data. The business goals should be clearly expressed. There exists different methods to capture goals. In our pilots we integrated goal oriented requirements engineering techniques to elicit and structure business goals and connect them with data processing processes and components [55, 56]. Such methods also include specific techniques to verify that goals are not too idealised by helping in the discovery of obstacles and their resolution in order to define achievable goals.

Another way to connect the goals with the (business) reality is to define how to measure the resulting value which should be defined right from the business

Table 3: Main characteristics of first pilot wave

#	Domain	Volume	Velocity	Variety	Main challenge
1	Life science	20 Go/analysis 2 To/week	High (requires parallel processing)	Business data and traceability (food, pharmaceutical, cosmetic industry)	Product quality
2	Space	Galileo ground segment maintenance (12 EU sites, 16 remote sites)	Medium	High: messages, logs	Predictive maintenance of costly equipment. High dependability (99.8%)
3	Health	900 beds on 3 sites	Real-time	Several sources and formats	Reduce morbidity and mortality, guarantee confidentiality
4	IT Maintenance	About 3000 servers	High (databases, events, logs...)	Real-time	Predictive maintenance, cost optimisation

understanding phase, typically by relying on KPIs (Key Performance Indicators). Companies should already have defined their KPIs and be able to measure them. If this is not the case, they should start improving on this: in other words, Business Intelligence should already be present in companies.

Based on this, different improvement strategies can be identified and discussed to select a good business case. In this process, the gap with the current situation should also be considered, it is safer to keep a first project with quite modest objectives than risking to fail by trying a too complex project that could bring more value. Once a pilot is successful, further improvements can be planned in order to add more value.

Computer Maintenance Area Case Study. The large IT provider considered here manages more than 3000 servers that are hosting many web sites, running applications and storing large amount of related customer data. No matter what efforts are taken, servers are still likely to go off-line, networks to become unavailable or disks to crash and generally at times that are not expected, less convenient and more costly to manage, like during the night or weekends. The considered company is currently applying standard incident management and preventive maintenance procedures based on a complete monitoring infrastructure covering both the hardware (network appliances, servers, disks) and the application level (service monitoring).

In order to reduce the number of costly reactive events and optimise preventive maintenance, the company is willing to develop more predictive maintenance by trying to anticipate the unavailability of the servers in such a way they can react preventively and, ultimately, prevent such unavailability. In the process, the client wants to diagnose the root causes of incidents and resolve them in order to avoid possible further incidents which can turn into a nightmare when occurring in a reactive mode. The ultimate goal is to increase the service availability, the customer satisfaction and also reduce the operating costs.

The resulting KPI is called Total Cost of Ownership (TCO) and typical breakdown costs to be can be considered are:

- maintenance on hardware and software that could be reduced through a better prediction,
- personnel working on these incidents,
- any penalties related to customer Service Level Agreements (SLAs),
- indirect effects on the client's business and its brand image.

Clinical Pathway Case Study. Hospitals are increasingly deploying clinical pathways, defined as a multidisciplinary vision of the treatment process required by a group of patients with the same pathology with predictable clinical follow-up [6]. The reason is not only to reduce the variability of clinical processes but also to improve care quality and have a better cost control [54]. It also enables richer analysis of the data produced and thus the profiling of patients with higher risks (for example due to multi-pathology or intolerances).

A typical workflow (e.g. for chemotherapy) is shown in Figure 8. It is a sequence of drugs deliveries or cures, generally administered in day hospital. Each cure is followed by a resting period at home that lasts for a few days to a few weeks. A minimal interval between cures is required because chemotherapy drugs are toxic

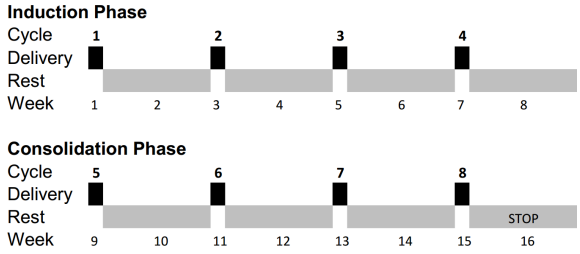

Figure 8: A typical chemotherapy workflow

and the body needs some time to recover between two drugs deliveries. When following the ideal treatment protocol, the number of cancerous cells are progressively reduced, hopefully to reach a full healing or cancer remission. If for some reason, chemotherapy cures do not closely follow the intended periodicity or if doses are significantly reduced, the treatment efficiency may be suboptimal. In such conditions, cancerous cells may multiply again, which can result in a cancer relapse.

Figure 9 shows the high level goals for the optimal organisation of care pathways. Goals and obstacles are respectively depicted using blue and red parallelograms. Agents (either people or processing components) are pictured using yellow hexagons. Some expectations on human agents are also captured using yellow parallelograms. The adequate workflow should be enforced for all patients within the recommended deadlines given the possible impact on patient relapse. Ethical principles also require a fair allocation of resources, i.e. every patient deserves optimal care regardless of his medical condition or prognosis. The workload should also be balanced to avoid the staff having to manage unnecessary peak periods.

Reaching those goals together, of course, requires enough resources to be available and a number of related obstacles (in red) like monitoring the flow of patients joining and leaving the pathway is important. The available workforce can also be influenced by staff availability and some public holidays reducing the available slot for delivering care. A number of mitigation actions are then identified to better control that the workforce is adapted. An agent with a key responsibility in the system is the scheduler, which must manage every appointment. Human agents are not very good at this task because the problem is very large and it is difficult to find a solution that simultaneously meets all patients and service constraints. Moreover, the planning must constantly be reconsidered to deal with unexpected events and the flow of incoming/outgoing patients. In contrast, **a combined predictive and prescriptive solution is very interesting because it has the capability to ensuring optimal care and service operation by also taking into account risks that some patient could be delayed**.

In order to measure the quality of chemotherapeutic cares, a quantifiable indicator called the "Relative Dose Intensity" (RDI) was defined [35]. It captures both the fact that the required dose is administered and the timing of the delivery, on a scale from 0% (no treatment) to 100 % (total conformance).

$$RDI = \frac{\text{planned dose}}{\text{delivered dose}} \times \frac{\text{real duration}}{\text{planned duration}}$$

Medical literature has shown, for a number of cancers, that the relapse-free survival is strongly correlated with the RDI. For instance, for breast cancer, a key threshold value is 85% [41]. Hence this indicator can be seen has a gauge that should be carefully managed across the whole clinical pathway.

4.2 From Descriptive to Predictive and then Prescriptive Data Analysis.

Analytics is a multidisciplinary concept that can be defined as the means to acquire data from diverse sources, process them to elicit meaningful patterns and insights, and distribute the results to proper stakeholders [10, 44]. Business Analytics is the application of such techniques by companies and organisations in order to get a better understanding of their level of performance of their business and drive improvements. Three complementary categories of analytics can be distinguished and combined in order to reach the goal of creating insights and helping to make better decisions. Those analytics consider different time focus, questions and techniques as illustrated in Table 4 [38, 51].

In a number of domains, it is interesting to consider an evolution scheme starting from immediate reaction raised by analysing data to more intelligence in anticipating undesirable situations, or even considering how to prevent them as much as possible.

Computer Maintenance Area Case Study. In terms of maintenance, starting from the identified KPI of total cost of ownership (TCO) including the cost of purchase, maintenance and repair in the event of a breakdown. Different strategies can be envisaged:

- *react* to problems only after the occurrence of a breakdown. This translates into a generally high cost because quick reaction is required to minimize downtime. Moreover, any unavailability has a negative impact in terms of image or even penalty if a Service Level Agreement (SLA) has been violated. This should of course be minimised

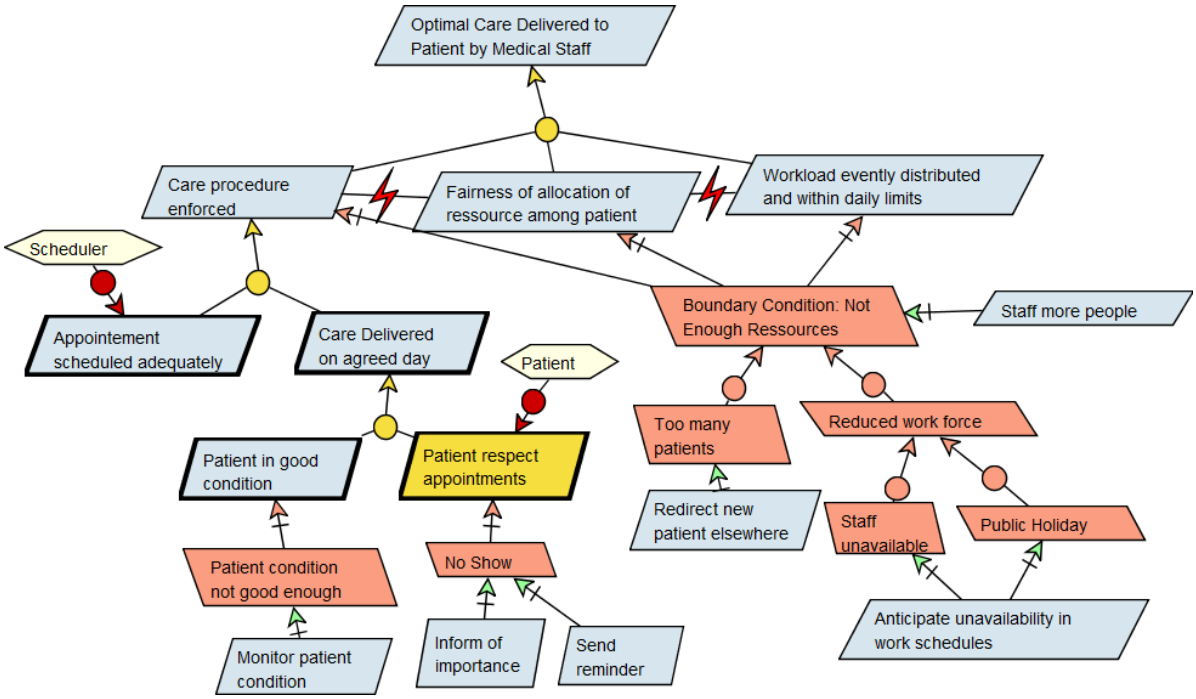

Figure 9: Goal analysis for clinical pathways: strategic goals and main obstacles

through the use of next level strategies that can benefit from data analytics,

- *anticipating* occurrence of breakdown based on system monitoring. Simple strategies can be implemented. For example, an alert can be triggered when a storage approaches a threshold close to the maximum capacity. However, this does enable the prediction of failures resulting from complex sequence of events. Mostly descriptive techniques are used at this level,

- *try to predict* problems based on known history and observation of the system. At this level, predictive data analysis techniques can discover cause-effect relationships between parts of the system which, in cascade, can cause unavailability. For example, applying a badly validated patch may affect a service that can itself paralyse a business process,

- *improving the system* is the ultimate step. It is necessary to ensure that the system operates under optimum conditions by eliminating the root causes that could trigger some failure. Prescriptive techniques are used at this level.

The predictive solution was the best option, but it should only be considered if the preventive step is carried out. Similarly, the most common time patterns should be identified and processed first. For example, a storage is more likely to be saturated on days when backups are performed, usually predictably (weekend or month end). An anticipation would avoid expensive interventions, especially during weekends.

Clinical Pathway Case Study. The operation of clinical pathways is characterised by the occurrence of many events which may be expected or not and thus impacting the scheduled behaviour. An important concern is to detect such events and decide about how to manage possible deviations to minimise their impact, especially on the quality of care KPI. Different possible strategies can be explored for this purpose:

- *reactive strategies* should be avoided as much as possible because the impact on patient is irreversible. Some case of reactive can be related to a patient no-show or last minute medical no-go. The action is to reschedule an new appointment as soon as possible.

- *preventive strategies* can be used to reduce the risk of now-show, for example by sending a reminder (e.g. text message, phone call) one or two days before the appointment. Descriptive data analytics are enough at this level.

- *predictive strategies* relying on predictive analytics can be used to learn risk factors for specific patients which could result in more careful planning or

Table 4: Overview of analytics in terms of questions, techniques and outputs (source: [51])

	Business Analytics		
	Descriptive Analytics	**Predictive Analytics**	**Prescriptive Analytics**
Questions	What has happened? Why did it happen?	What will happen? Why will it happen?	What should be done? Why should it be done?
Techniques	Statistical Analytics Data Integration Data Augmentation Data Reduction	Data Mining Machine Learning ...	Optimisation Simulation Operation Research Management Science
Outputs	Reports on historical data Insight from raw data ...	Future opportunities Future risks ...	Recommended business decisions Optimal courses of actions ...

Table 5: Some workshop questions about data

Q.UD.1	What are the data sources and data types used in your current business processes?
Q.UD.2	What tools/applications are used to deal with your current business processes?
Q.UD.3	Are your present business processes performing complex processing on data?
Q.UD.4	How available is your data? What happens if data is not available?
Q.UD.5	Do different users have different access rights on your data?
Q.UD.6	Does your data contain sensitive information (e.g. personal or company confidential data)?
Q.UD.7	What are the consequence of data alteration? Do you know the quality level of your data?

guide the drug selection. For example, the possible intolerance or interaction with another pathology could be anticipated and solved by selecting an alternative drug cocktail.

- *prescriptive strategies* will deploy an globally optimising scheduler able to solve the planning problem by taking into account the constraints resulting for the treatment plan of each patient and the service availabilities. Such a system was successfully prototyped and is reported in [42].

4.3 Using Questionnaires for Workshops

Conducting a workshop requires to pay attention to many issues while also focusing the discussion on the most relevant ones. A questionnaire can provide an efficient support both as possible preparation before the workshop and as checklist during the workshop. Table 5 shows a few questions about the data to process.

Table 6: Evaluation readiness checklist (partial)

R.EV.1	Are you able to understand/use the results of the models?
R.EV.2	Do the model results make sense to you from a purely logical perspective?
R.EV.3	Are there apparent inconsistencies that need further exploration?
R.EV.4	From your initial glance, do the results seem to address your organizations business questions?

4.4 Using Modelling Notations

Modelling using standard modelling notations is useful to support business and data understanding. During workshops, a whiteboard can be used to sketch models together with the audience. Note this should not to be confused with the data modelling step in CRISP-DM which is related but actually occurs later in the process.

In our experience, data-flow and workflow models help to understand which process is generating, updating or retrieving data. UML class diagrams also help to capture the domain structure [40].

On the other hand, use cases should be avoided because they only focus on a specific function and cannot provide a good global picture of the problem. Those should be used later in the process when considering the implementation.

4.5 Defining Activity Checkpoints

An Agile approach allows the process to be quite flexible and enable to go back and forth across activities. To make sure an activity can be started with enough input, we defined some readiness checklist shown in Table 6.

5 Related Work and Discussion

5.1 Methodologies Focusing on Adoption

Section 2 gives an exhaustive historical perspective of the evolution of relevant methodologies. While the first proposed approaches based on data mining were already iterative in nature [50], their evolution over time is clearly about paying a growing attention on how to ease the adoption of such methodologies by the companies. The Agile culture has been a key milestone to better include the customer in the process and to drive that process towards the production of business value [23]. Commercial methods like IBM Stampeded are also strongly inspired by this trend [28]. In complement to those methods, the need to identify barriers and adoption factors has also been addressed by recent work also discussed earlier, such as critical success factors [24, 46].

Consolidated Big Data Methodologies are also being published under more practical and simplified presentation forms in order to be attractive for companies. The DISTINCT method is based on only four steps (acquire, process, analyse, visualise) and considers the use of feedback loops to enable repeated refinements of the data processing [18]. Although the analysis phase is not explicitly mentioned, this iterative approach can be used to set up a feedback channel between IT and business people. After each feedback cycle, the system can then be refined by enhancing the data preparation or data analysis steps.

The well-known "for Dummies" series has also a dedicated book on Big Data [27]. It contains a section about how to create an implementation roadmap based on factors like business urgency, budget, skills, and risks level; an agile management approach is also recommended. The availability of Business Intelligence is also identified as an easing factor.

Work carried out in related fields about how to address organisational challenges is also worth being investigated. For example, Cloud Computing Business Framework (CCBF) helps organisations achieve good Cloud design, deployment and services. Similar to our approach, CCBF is a conceptual and an architectural framework relying on modelling, simulation, experiments and hybrid case studies [7, 8].

Given the variety and multidisciplinary nature of complex system being analysed (e.g. supply chains, IT system, health systems), it is useful to consider a Multi Disciplinary Engineering Environment (MDDE) approach. A very good and comprehensive survey on the approaches of data integration based on ontologies is described in [17]. It also gives guidelines for the selection of technologies of data integration for industrial production systems.

5.2 Ethical Concerns about Data Privacy

The interaction with companies also raised some ethical concerns and questions like: "Are we sufficiently cautious about the Big Data phenomenon?" It is certainly a great technological revolution of our time to collect large amounts of data and to value them to improve the health and living conditions of humans. Nevertheless, we are faced with a problem of ethics when using predictive algorithms. Regulatory intensification is therefore necessary to find a good compromise between the use of personal data and the protection of privacy.

For example, in the field of health, we can wonder about the way governments intend to exploit the data collected. Should those data be made available (by open data) or should a solution be found to support the exploitation of private data?

By the use of massive data in the medical community, the legal and economic aspects change at great speed. This challenges ethical principles and rules in the relationship between a doctor and a patient. This also disturbs the balance between confidentiality and transparency and creates a feeling of declining confidence in the health environment around the management and exploitation of Big Data. The ethics of this type of data requires a well supervised control of the use of medical information [5, 16].

Studies have also demonstrated the segmentation power of predictive modelling and resulting business benefits for life insurance companies [3]. While some customers with "lower risk" could enjoy better conditions for their insurance, customers with higher anticipated risks could be excluded due to unaffordable rates, thus reducing the solidarity effect of life insurances.

Data is also increasingly carrying location information due to the large development of mobile applications and the emergence of the Internet of Things. A specific branch of Big Data called location analytics is specifically focusing on this area and can endanger privacy if applied without safeguards. Specific guidelines and techniques are being developed for this purpose. Some guidelines are issued e.g. by the European Commission for public administration [2]. Specific data processing techniques and algorithms are also being developed for privacy preserving location

based services [34, 52].

At a more general level, in order to better control the huge quantities of data processed every day and to ensure that every single person is respected, the European Commission has issued the General Data Protection Regulation in 2016 that will come into force in May 2018 [19]. An EU portal with extensive resources is available to give some starting points to companies [20].

Our recommendation, based on our pilots, is to investigate this issue early in the process. This can already be envisioned at the business and data understanding phases and involve relevant people like Chief Information Security Officer or even a more specific Data Protection Officer if this role is defined. Actually this happened quite naturally in most of our pilot cases because the data had to be processed outside of the owning organisation. However, the focus was more on confidentiality than on the purpose of the data processing itself.

5.3 Cyber Security Issues

Among the challenges of the Big Data, data security is paramount against piracy and requires the development of systems to secure trade by ensuring strict control of access to the Big Data platform and thus guarantee the confidentiality of data. Securing a Big Data platform is nevertheless a domain in its own right because the very principle of this system is that it can be based on a heterogeneous architecture spread over several nodes. The ENISA has produced a landscape of Big Data threats and a guide of good practices [14]. This document lists typical Big Data assets, identifies related threats, vulnerabilities and risks. Based on these points, it suggests emerging good practices and active areas for research.

Storing sensitive data on the Cloud, for example, is not without consequences, because the regulations are not the same in all countries. A sensitive aspect is the management of the data storage and processing locations, e.g. the need to process data in a given country. However, as this situation is also hindering European competitiveness in a global market, the EU is currently working on a framework for the free flow of non-personal data in the EU [21].

6 Conclusions

In this paper, we described how we addressed the challenges and risks of deploying a Big Data solution within companies willing to adopt this technology in order to support their business development. We first looked at different methods reported over time in the literature. Rather than building yet another method, we realised the key when considering the adoption of Big Data in an organisation, is the process followed to come up with a method that fits the context, needs and will maximize the chance of success. Based on this idea, we defined a generic guidance process relying on available methods as building bricks. To be meaningful our approach is also strongly relying on lessons learned from industrial cases which on one hand helped in validating our process guidance and on the other hand can also be used as concrete supporting illustration.

Moving forward, we plan to consolidate our work based on what we will learn in the next series of pilot projects. This includes investigating challenges from other domains. We plan to address life sciences which requires a sustained processing of high volume of data and the space domain with highly distributed infrastructures. Considering the global development process, until now we have mainly focused on the discovery and data understanding phases. So our plan is to provide more guidance on the project execution phase using our most advanced pilots that are now reaching full deployment. In our guidance process, we also had to face a number of problems which sometimes blocked all further progress. In some cases the reason was a lack of business value or maturity, for which the recommended action was to postpone the process. In other cases, some blocking issues could not be overcome or were delaying the project a lot longer than expected, e.g. to set up a non-disclosure agreement about data access, to get actual data access, to configure proprietary equipment, etc. Guidance about how to detect and avoid such cases is also part of our work as it helps to increase the chance of successful deployment.

Acknowledgements

This research was partly funded by the Walloon Region through the "PIT Big Data" project (grant nr. 7481). We thank our industrial partners for sharing their cases and providing rich feedback.

References

[1] R. Balduino, "Introduction to OpenUP," https://www.eclipse.org/epf/general/OpenUP.pdf, 2007.

[2] L. Bargiotti, I. Gielis, B. Verdegem, P. Breyne, F. Pignatelli, P. Smits, and R. Boguslawski, "European Union Location Framework Guidelines for public administrations on location privacy. JRC Technical Reports," 2016.

[3] M. Batty, "Predictive Modeling for Life Insurance Ways Life Insurers Can Participate in the Business

Analytics Revolution," Deloitte Consulting LLP, April 2010.

[4] T. Bedos, "5 key things to make big data analytics work in any business," http://www.cio.com.au, 2015.

[5] J. Béranger, *Big Data and Ethics: The Medical Datasphere*. Elsevier Science, 2016.

[6] H. Campbell, R. Hotchkiss, N. Bradshaw, and M. Porteous, "Integrated care pathways," *British Medical Journal*, pp. 133–137, 1998.

[7] V. Chang, *A Proposed Cloud Computing Business Framework*. Commack, NY, USA: Nova Science Publishers, Inc., 2015.

[8] V. Chang, R. J. Walters, and G. Wills, "The development that leads to the cloud computing business framework," *International Journal of Information Management*, vol. 33, no. 3, pp. 524 – 538, 2013.

[9] H.-M. Chen, R. Kazman, and S. Haziyev, "Agile big data analytics development: An architecture-centric approach," in *Proceedings HICSS'16, Hawaii, USA*. Washington, DC, USA: IEEE Computer Society, 2016, pp. 5378–5387.

[10] H. Chen, R. H. L. Chiang, and V. C. Storey, "Business intelligence and analytics: From big data to big impact," *MIS Q.*, vol. 36, no. 4, pp. 1165–1188, Dec. 2012.

[11] K. Collier, B. Carey, E. Grusy, C. Marjaniemi, and D. Sautter, "A Perspective on Data Mining," Northern Arizona University, 1998.

[12] F. Corea, *Big Data Analytics: A Management Perspective*. Springer Publishing Company, Inc., 2016.

[13] K. Crowston, "A capability maturity model for scientific data management," *BibSonomy*, 2010.

[14] E. Damiani *et al.*, "Big data threat landscape and good practice guide," https://www.enisa.europa.eu/publications/bigdata-threat-landscape, 2016.

[15] G. S. do Nascimento and A. A. de Oliveira, *An Agile Knowledge Discovery in Databases Software Process*. Springer Berlin Heidelberg, 2012, pp. 56–64.

[16] EESC, "The ethics of Big Data: Balancing economic benefits and ethical questions of Big Data in the EU policy contex," European Economic and Social Committee, Visits and Publications Unit, 2017.

[17] F. J. Ekaputra, M. Sabou, E. Serral, E. Kiesling, and S. Biffl, "Ontology-based data integration in multi-disciplinary engineering environments: A review," *Open Journal of Information Systems (OJIS)*, vol. 4, no. 1, pp. 1–26, 2017. [Online]. Available: http://nbn-resolving.de/urn:nbn:de:101:1-201711266863

[18] T. Erl, W. Khattak, and P. Buhler, *Big Data Fundamentals: Concepts, Drivers & Techniques*. Prentice Hall, 2016.

[19] European Commission, "General Data Protection Regulation 2016/679," http://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32016R0679, 2016.

[20] European Commission, "The EU General Data Protection Regulation (GDPR)," http://www.eugdpr.org, 2016.

[21] European Commission, "A framework for the free flow of non-personal data in the EU," http://europa.eu/rapid/press-release_MEMO-17-3191_en.htm, 2017.

[22] U. Fayyad, G. Piatetsky-shapiro, and P. Smyth, "From data mining to knowledge discovery in databases," *AI Magazine*, vol. 17, pp. 37–54, 1996.

[23] P. Frankov, M. Drahoov, and P. Balco, "Agile project management approach and its use in big data management," *Procedia Computer Science*, vol. 83, pp. 576 – 583, 2016.

[24] J. Gao, A. Koronios, and S. Selle, "Towards A Process View on Critical Success Factors in Big Data Analytics Projects," in *AMCIS*, 2015.

[25] Gartner, "Gartner survey reveals investment in big data is up but fewer organizations plan to invest," http://www.gartner.com/newsroom/id/3466117, 2016.

[26] J. Haffar, "Have you seen asum-dm?" ftp://ftp.software.ibm.com/software/data/sw-library/services/ASUM.pdf, 2015.

[27] J. Hurwitz, A. Nugent, F. Halper, and M. Kaufman, *Big Data For Dummies*. John Wiley & Sons, 2013.

[28] IBM, "Stampede," http://www.ibmbigdatahub.com/tag/1252, 2013.

[29] IBM, "ASUM-DM," https://developer.ibm.com/predictiveanalytics/2015/10/16/have-you-seen-asum-dm, 2015.

[30] K. Jensen, "Crisp-dm process diagram," https://commons.wikimedia.org/wiki/File:CRISP-DM_Process_Diagram.png distributed under CC-SHA2, 2012.

[31] J. Kelly and J. Kaskade, "CIOs & Big Data: What Your IT Team Wants You to Know," http://blog.infochimps.com/2013/01/24/cios-big-data, 2013.

[32] L. Lau, F. Yang-Turner, and N. Karacapilidis, "Requirements for big data analytics supporting decision making: A sensemaking perspective," in *Mastering data-intensive collaboration and decision making*, N. Karacapilidis, Ed. Springer Science & Business Media, April 2014, vol. 5, pp. 49 – 70.

[33] D. Lehmann, D. Fekete, and G. Vossen, "Technology selection for big data and analytical applications," *Open Journal of Big Data (OJBD)*, vol. 3, no. 1, pp. 1–25, 2017. [Online]. Available: http://nbn-resolving.de/urn:nbn:de:101:1-201711266876

[34] L. Liu, "From data privacy to location privacy: Models and algorithms," in *Proceedings of the 33rd International Conference on Very Large Data Bases*, ser. VLDB '07. VLDB Endowment, 2007, pp. 1429–1430.

[35] G. Lyman, "Impact of chemotherapy dose intensity on cancer patient outcomes," *J Natl Compr Canc Netw*, pp. 99–108, Jul 2009.

[36] G. Mariscal, s. Marbn, and C. Fernndez, "A survey of data mining and knowledge discovery process models and methodologies," *Knowledge Eng. Review*, vol. 25, no. 2, pp. 137–166, 2010.

[37] A. D. Mauro, M. Greco, and M. Grimaldi, "A formal definition of big data based on its essential features," *Library Review*, vol. 65, no. 3, pp. 122–135, 04 2016.

[38] M. Minelli, M. Chambers, and A. Dhiraj, *Big Data, Big Analytics: Emerging Business Intelligence and Analytic Trends for Today's Businesses*, 1st ed. Wiley Publishing, 2013.

[39] C. Nott, "Big Data & Analytics Maturity Model," http://www.ibmbigdatahub.com/blog/big-data-analytics-maturity-model, 2014.

[40] OMG, "Unified Modeling Language (UML) - Version 2.X," http://www.omg.org/spec/UML, 2005.

[41] M. Piccart, L. Biganzoli, and A. Di Leo, "The impact of chemotherapy dose density and dose intensity on breast cancer outcome: what have we learned?" *Eur J Cancer.*, vol. 36, no. Suppl 1, April 2000.

[42] C. Ponsard, R. D. Landtsheer, Y. Guyot, F. Roucoux, and B. Lambeau, "Decision making support in the scheduling of chemotherapy coping with quality of care, resources and ethical constraints," in *ICEIS 2017 - Proc. of the 19th Int. Conf. on Enterprise Information Systems, Porto, Portugal, April 26-29*, 2017.

[43] C. Ponsard, A. Majchrowski, S. Mouton, and M. Touzani, "Process guidance for the successful deployment of a big data project: Lessons learned from industrial cases," in *Proc. of the 2nd Int. Conf. on Internet of Things, Big Data and Security, IoTBDS 2017, Porto, Portugal, April 24-26*, 2017.

[44] D. J. Power, "Using 'Big Data' for analytics and decision support," *Journal of Decision Systems*, vol. 23, no. 2, Mar. 2014.

[45] E. Rot, "How Much Data Will You Have in 3 Years?" http://www.sisense.com/blog/much-data-will-3-years, 2015.

[46] J. Saltz and I. Shamshurin, "Big Data Team Process Methodologies: A Literature Review and the Identification of Key Factors for a Projects Success," in *Proc. IEEE Int. Conf. on Big Data*, 2016.

[47] J. S. Saltz, "The need for new processes, methodologies and tools to support big data teams and improve big data project effectiveness," in *IEEE International Conference on Big Data, Big Data 2015, Santa Clara, CA, USA, October 29 - November 1*, 2015, pp. 2066–2071.

[48] SAS Institute, "SEMMA Data Mining Methodology," http://www.sas.com/technologies/analytics/datamining/miner/semma.html, 2005.

[49] Scrum Alliance, "What is scrum? an agile framework for completing complex projects," https://www.scrumalliance.org/why-scrum, 2016.

[50] C. Shearer, "The CRISP-DM Model: The New Blueprint for Data Mining," *Journal of Data Warehousing*, vol. 5, no. 4, 2000.

[51] R. Soltanpoor and T. Sellis, *Prescriptive Analytics for Big Data*. Cham: Springer International Publishing, 2016, pp. 245–256.

[52] G. Sun, D. Liao, H. Li, H. Yu, and V. Chang, "L2p2: A location-label based approach for privacy preserving in lbs," *Future Generation Computer Systems*, vol. 74, no. Supplement C, pp. 375 – 384, 2017.

[53] Two Crows Corporation, "Introduction to Data Mining and Knowledge Discovery," http://www.twocrows.com/intro-dm.pdf, 2005.

[54] P. A. van Dam, "A dynamic clinical pathway for the treatment of patients with early breast cancer is a tool for better cancer care: implementation and prospective analysis between 2002–2010," *World Journal of Surgical Oncology*, vol. 11, no. 1, 2013.

[55] A. van Lamsweerde, "Goal-oriented requirements engineering: a guided tour," in *Requirements*

Engineering, 2001. Proceedings. Fifth IEEE International Symposium on, 2001, pp. 249–262.

[56] A. van Lamsweerde, *Requirements Engineering - From System Goals to UML Models to Software Specifications*. Wiley, 2009.

AUTHOR BIOGRAPHIES

Ir. Christophe Ponsard holds a master in Electrical Engineering and Computer Science. He runs the Software and System Engineering department of CETIC focusing on requirements engineering, model-driven development and software quality. He is actively contributing to several applied research programs at European level and transfer activities with local companies to foster the adoption of emerging technologies like Big Data, Machine Learning and IoT.

Mounir Touzani holds a PhD from the University of Montpellier. His areas of expertise are requirements engineering, business process analysis, business rule systems, database engineering and Data Science. He is actively involved in database operation for large scale administrative processes and is working on the development of Big Data deployment methodologies, Machine Learning and Cloud computing.

Annick Majchrowski is a member of the CETIC Software and Systems Engineering department since 2007. She holds a BA in Mathematics and a BA in Computer Science. She leads the activities related to software process audit and deployment. She is actively involved in software process improvement in several organisation both in the public and private sectors. She also contributes to develop methodologies for the optimal adoption of emerging technologies.