# **Business-IT Alignment in Fraud Analytics:**

# **Fostering Collaboration between Domain Experts and Data Scientists**

A. Simonofski ([anthony.simonofski@unamur.be](mailto:anthony.simonofski@unamur.be) , Faculty of Computer Science, Namur Digital Institute, UNamur / Faculty of Economics and Business, KU Leuven), B. Vanderose (Faculty of Computer Science, Namur Digital Institute, UNamur), B. Frenay (Faculty of Computer Science, Namur Digital Institute, UNamur)

## Abstract

Fraud Analytics refers to the use of Big Data Analytics to detect fraud. In government, two groups of stakeholders are usually involved in this process: domain experts knowledgeable in fraud detection and data scientists knowledgeable in analytics. However, data scientists in government are rarely knowledgeable in the business domain (fraud detection) and domain experts do not always have an Information Technology (IT) profile. Thus, ensuring collaboration between the Business and IT sides of fraud analytics is a key challenge for governments. Alignment is increasingly important to detect frauds efficiently as the complexity of frauds, as well as the techniques used to detect them, keeps increasing and makes collaboration needed.

The goal of this chapter is to formalize the Fraud Analytics process and to illustrate this key alignment challenge. For this, we examine two case studies from the Belgian Federal government: the detection of tax frauds and of social security infringements. Data from these two cases has been collected through 21 interviews. As a result, we infer two fraud analytics processes and identify three crucial moments where alignment between Business and IT is needed: the identification of requirements of the business team before performing the analytics, the presentation of the output of the analytics to the business team and the feedback of the business to the data scientists. In order to foster this alignment in Fraud Analytics, we suggest a methodology drawing from agile methods, participation methods and design thinking literatures.

Keywords: big data, artificial intelligence, agile, participation, design thinking.

## Introduction

The concept of Big Data can be defined in several ways and still evolves. Daniell et al. (2016) and Klievink et al. (2017) talk about massive, complex collections of data that are constantly changing. Big data is characterized by three basic features: large volumes, high velocity (meaning rapidly generated data), large variety (meaning large heterogeneity) (Daniell et al., 2016; Kim et al., 2014; Klievink et al., 2017). Desouza & Jacob (2017) add a factor “complexity” to these “three V’s”, the others claim that big data is complex because of these three V’s. The benefits of Big Data for government reveal themselves when applying analytics to support policy-making in diverse domains (Kim et al., 2014).

Fraud Analytics is one of these domains and refers to the use of advanced analytics to detect fraud, with a focus on big data analytics due to the large amount of data necessary to ensure detection. Numerous techniques, ranging from data mining to social network analysis, are developed and applied to detect various types of fraud. While Fraud Analytics offers the promise of more efficiency in fighting fraud, it also raises legal, technical and organizational challenges for governments. A key organizational challenge relates to the necessary alignment between the business and Information Technology (IT) side of Fraud Analytics. In their seminal paper, Henderson and Venkatraman (1999) define business-IT alignment as a way through which organizations can optimally use IT to achieve their business objectives on strategic and operational levels. This challenge is reinforced with the introduction of emerging digital technologies such as Big Data or Artificial Intelligence (AI) that impact, even more, the way traditional organizations work. This challenge is particularly important for governments. Indeed, governments constitute a diverse ecosystem with multiple internal stakeholders, each of which has their own objectives. Among these stakeholders, there are different business teams and IT teams that do not always communicate with each other, leading to organizational silos. In the case of Fraud Analytics, the business teams are composed of jurists, tax specialists and inspectors whereas the IT teams are composed of data scientists and data miners (Baesens et al., 2015). Organizational silos may hamper the alignment of development projects in the organization.

The goal of this chapter is thus to formalize the Fraud Analytics process, to illustrate the Business-IT alignment challenge and to provide leads for solutions to foster this alignment. For this, we examined two case studies from the Belgian Federal government: the detection of tax frauds and of social security infringements. We chose these two case studies as they both apply fraud analytics in their daily work, performed by diverse business and data analysis teams. Furthermore, organizational silos are particularly present in large governmental organizations such as the Federal Public Services in Belgium (Chantillon et al., 2020), which makes them potential good cases to understand the problem of alignment. Data from these two cases have been collected through 21 interviews and relevant documents. As a way forward, this chapter then suggests a methodology to foster business-IT alignment based on three leads for solution: agile methods, participation methods and design thinking.

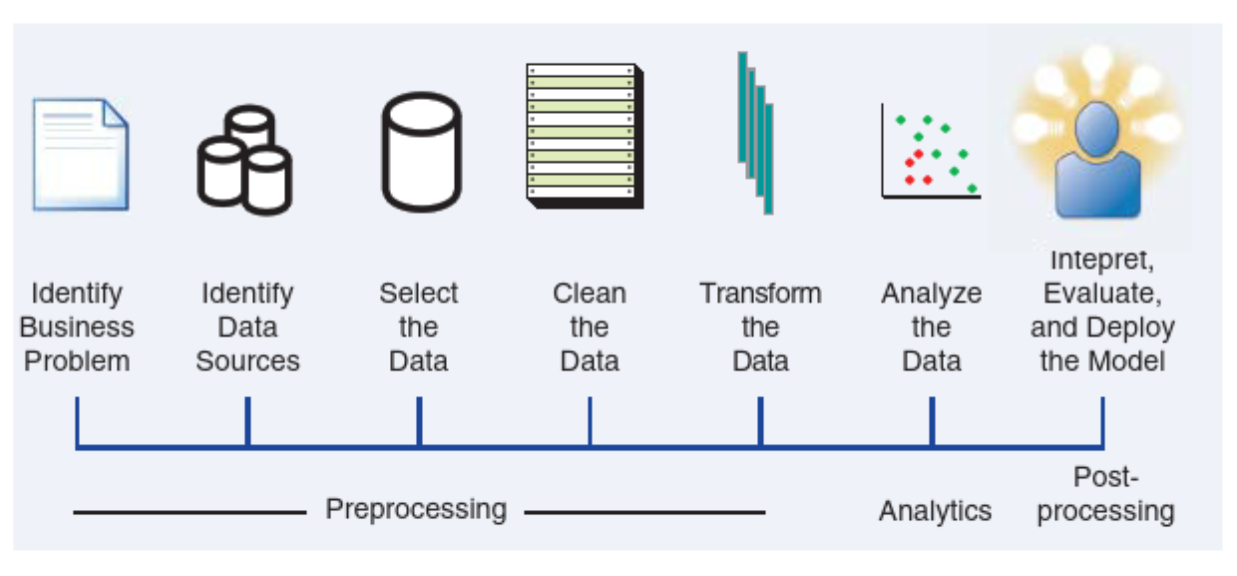
The chapter is structured as follows. In Section 2, we summarize the literature related to the use of Big Data analytics to detect frauds and highlight the research gaps to be tackled in this chapter. In Section 3, we present the methodology we followed to formalize the Fraud Analytics process and illustrate the Business-IT alignment challenge. In Section 4, we present this process in the Belgian use cases and identify the moments where the alignment challenge takes place. In Section 5, we suggest a methodology based on three concrete leads for solution to foster the alignment between business and IT profiles in Fraud Analytics. Finally, in Section 6, we summarize the contributions of this chapter as well as summarizing the limitations and further research leads.

## Background: Unpacking the Fraud Analytics Process

In this section, we unpack the Fraud Analytics process, defined as Big Data Analytics applied in a Fraud Detection context. Then, we detailed how the advanced techniques used in Fraud Analytics challenge the alignment between data scientists and fraud investigators and calls for the research performed in this chapter.

In order to detail the Big Data Analytics process, we rely on the three-step formalization suggested by two comprehensive models in the literature. On the one hand, Klievink et al. (2017) suggest a usage process for the analysis of big data and its relevance for policy-making that we follow in this section. First, there is the *Pre-Processing* stage where data is identified, collected from several sources, combined and cleansed (Section 2.1.). Second, there is the *Data Analytics* stage (Section 2.2.) where several techniques are applied to analyse the data. Third, there is the *Post-Processing* stage, where the output of the analysis is presented to the relevant stakeholders, interpreted and, in case of governments, has an impact on policy-making (Section 2.3.). On the other hand, Baesens et al. (2015) suggest a process of big data analytics that also follows this overall structure as shown in the figure below. This process is shown in Figure 1.

*Figure 1. Formalization of the Big Analytics Process (Baesens et al., 2015)*



The following sections detail the available literature for each of these three key stages.

### Pre-Processing: Data Collection and Combination

The first step of the pre-processing stage relates to the identification and understanding of the business problem to be tackled in the analytics project. In a digital government context, this problem can be related to several policy domains. This approach relates to the evidence-based approach, introduced by the UK government in the 1990s (Tsoukias et al., 2013). This approach is about governing based on ‘”facts” and thus using “evidence” to support decisions (De Marchi et al., 2016). However, evidence-based policy making fails to address some challenges like sufficiently solving the mistrust between citizens and policy makers or involving opinion groups and individual citizens enough (Tsoukias et al., 2013). This failure can be the result of the fact that multiple conclusions can be drawn from the same “fact”, or because the “evidence” policy makers use, is subjectively put in a hierarchy by them, which makes the term “evidence” non-neutral (De Marchi et al., 2016). Policy analytics(Gil-Garcia et al., 2018; Hamza and Mellouli, 2018), is a term describing a phenomenon where data and analytical techniques are used to support policy decisions. In this chapter, a specific policy domain is investigated (fraud detection) and will be detailed in Section 2.3.

The second step of pre-processing relate to the identification and selection of data. A lot of data can be used to support policy-making: ranging from data in simple, well-structured databases to unstructured data like pictures, audio or social-media data and other data from web 2.0 involving textual data, structural data (metadata) and temporal data(Chung and Zeng, 2018)**.** From this data, things like policy-relevant information, knowledge and ideas from citizens can be deducted (Loukis, 2018). Data is a key ingredient for any analytical exercise and experience in different fields more data is better for the analysis (Baesens et al., 2015). The pre-processing of the data (from collecting to transforming) is thus essential to avoid the Garbage In – Garbage Out[[1]](#footnote-1) principle. Data can come from numerous sources (Baesens et al., 2015): transactional data, contractual data, sociodemographic data, surveys, behavioural data, expert-based data, textual data, etc.

Five differentiating characteristics of Big Data Analytics are the following (Klievink et al., 2017): making combinations of multiple large external and internal datasets and using these, combining and using structured and unstructured data while analysing, using incoming data streams in (near) real-time, applying advanced analytics and algorithms, using existing datasets in an innovative way. This also implies that collections of data are in fact “big data” if they cannot be handled by conventional data processing. Some frequently mentioned challenges about big data in government are data integration across departmental silos, developing data standards, analysing unstructured data, establishing sufficient control towers, implementing regulations regarding compliance and security, dealing with access rights, building the infrastructure, archiving and preservation, accepting change and addressing privacy issues (Bertot et al., 2014; Joseph and Johnson, 2013; Kim et al., 2014).

One possible source of data is the government itself. Open Government Data (OGD) can help Big Data projects succeed by ensuring publicly accessible datasets through managed processes. OGD is not necessarily Big Data and Big Data contains more than only OGD. Governments can publish so-called OGD catalogues or portals, in that way providing a single point of access for governmental data (Erickson et al., 2013; Kalampokis et al., 2013). OGD is data that anyone can freely (not necessarily for free) use, reuse and redistribute (Kalampokis et al., 2013; Ubaldi, 2013). These OGD initiatives increase government transparency and accountability, but also involve many challenges (Kalampokis et al., 2013). Kalampokis et al. (2013) also claim that ideally, OGD is linked, meaning that the data should be machine-readable, with explicitly defined meaning, and linked to other external datasets.

The last step of the pre-processing stage relate to the cleansing and transformation of the data so that the analysis and use can be properly performed (Baesens et al., 2015).. In this stage, the data can be gathered (e.g. in a data mart or data warehouse) so that exploratory analysis can be performed. This is usually followed by a cleaning of the data to remove potential inconsistencies such as duplicate data, missing vales or outliers. For instance, missing value can occur because of undisclosed data or merging errors. Outliers are extreme observations that are differ greatly from the dataset. They should be detected and, if needed, treated. Finally, (e.g. missing values, outliers or duplicate data).

As the pre-processing stage consists in many activities (data collection, cleaning or transformation) (Archenaa and Anita, 2015), it can be supported by diverse management tools or data processing applications. For instance, Hadoop is an open source java-based framework that can help to manage, integrate and store these big data sets (Archenaa and Anita, 2015; Joseph and Johnson, 2013; Kim et al., 2014; Ubaldi, 2013). These management tools can also help analysing the big data, as described below in the “data analytics” section. Janssen and Kuk (2016) argue that there is a need for transparency about how the algorithm operates.

### Data Analytics

In this stage, the pre-processed data will be employed to build an analytical model. More and more, big data is widely being used by governments for identifying and analysing problems (Bertot et al., 2014). In order to do so, analytics make intensively use of the data described in the previous section. Analytics is a very broad umbrella term that encompass several techniques (Davenport and Jarvenpaa, 2008). Involved here are amongst others, statistics, data mining, business intelligence, operational research and decision analysis, machine learning and computer science, as well as disciplines like sociology, psychology and economics (Daniell et al., 2016; De Marchi et al., 2016; Tsoukias et al., 2013). The intention of analytics is to create useful information and knowledge to make good decisions. There are different types of analytics, like descriptive, predictive, explanatory and decisive analytics (Daniell et al., 2016; Davenport and Jarvenpaa, 2008)

Several analytics techniques exist to exploit the data and support policymaking. These techniques are labelled under several terms. *Data Mining* techniques are focused on finding existing patterns in datasets and making sense of the data at hand. Data mining is more of a manual process with human intervention and a focus on decision-making. When applied to business decisions, data mining is part of the global process labelled as *Business Intelligence*. Negash (2004) defines Business Intelligence as “*systems that combine data gathering, data storage, and knowledge management with analytical tools to present complex internal and competitive information to planners and decision makers*”.

*Machine Learning*, a sub-field of *AI*, is a field where the analysis aims to be automated by a machine. Machine learning systems are autonomous and can operate without human intervention and can learn and identify patterns to make decisions and to reach different conclusions based on the analysis of different situations (Sousa et al., 2019). Bishop (2006) defines Machine Learning as “*the automatic discovery of regularities in data through the use of computer algorithms to take actions such as classification*”. Tom Mitchell (1997) defines a Machine Learning problem, as a computer program that learns from experience E (e.g. watching our classification of mails), with respect to some task T (e.g. classifying mails) and some performance measure P (e.g. number of mails correctly classified), if its performance in T, as measured by P, improves with experience E. Machine learning can be supervised or unsupervised (Han et al., 2012). In *supervised* machine learning, a function maps an input to an output based on example input-output pairs (e.g. predicting the house of a price, based on the surface, based on existing price/surface mappings). The function can predict continuous values (regression) or discrete values (classification). Therefore, training examples are needed to “teach” to algorithm to perform a task. In *unsupervised* learning, the aim is to establish the existence of classes or clusters in the data given a set of measurements, observations, etc. (e.g. Google news groups news into broad categories and topics automatically).

Compared to Data Mining, machine learning goes one-step further and intends to predict future outcomes based on the pre-existing datasets. The two techniques remain complementary, as Machine Learning will use Data Mining techniques as resources. The understanding of this diverse analytics terms is essential to understand the foundations of Fraud Analytics, detailed in the following section.

### Post-Processing: Use of Big Data Analytics for Fraud Detection

According to the Oxford Dictionary, a fraud can be defined as a “*wrongful or criminal deception intended to result in financial or personal gain*”. De Roux et al. (2018) define a tax fraud as the intentional act of lying on a tax return form with intent to lower one’s tax liability and social security infringement as the intentional act of receiving social security payments someone is not entitled to.

Traditionally, tax authorities tackled tax fraud with two approaches (Castellón González and Velásquez, 2013). The *auditor experience* approach randomly selects tax declaration and audit them based on experience and domain knowledge. The *rule-based system* approach applies “if-then” rules to detect fraud cases. These rules are burdensome to develop as experts have to review and generalize fraud characteristics after they identify them. The two main issues with these techniques are that they exclusively rely on past experiences and dismiss new fraud mechanisms and are based on subjective judgment of experts.

Policy Analytics have been applied in previous research to detect tax fraud (Van Vlasselaer et al., 2017; Yu et al., 2003) and has been labelled as “Fraud Analytics” (Baesens et al., 2015). The detection of fraud can be considered as the use of analytics in the specific step of “Implementation” in the policy lifecycle (Pencheva et al., 2018). The use of analytics techniques can also be inserted in a more global approach consisting of (Baesens et al., 2015):

* Fraud detection: Applying detection models on new, unseen observations and assigning a fraud risk to every observation;
* Fraud investigation: A human expert is often required to investigate suspicious, flagged cases given the involved subtlety and complexity;
* Fraud confirmation: Determining true fraud label, possibly involving field research;
* Fraud prevention: Preventing fraud to be committed in the future. This might even result in detecting fraud even before the fraudster knows s/he will commit fraud;

As Data mining allows finding patterns and anomalies in large amounts of data, it is helpful for fraud detection. However, traditional data mining techniques are not adequate to detect advanced types of fraud due to their inherent characteristics (Van Vlasselaer et al., 2017):

* Uncommon: the datasets are skewed as a limited number of fraud cases are identified;
* Well-considered: frauds are planned, well though off and probably present in more than one attribute in the data;
* Time-evolving: fraudsters are adaptive and learn from past mistakes;
* Carefully organized: fraudsters have allies and transfer knowledge with each other to commit fraud without being detected;
* Imperceptibly concealed: data can be overlapping as fraudsters may have same characteristics as legitimate companies.

Therefore, advanced techniques based on machine learning are needed for the detection of more complex frauds as reported in (De Roux et al., 2018; Van Vlasselaer et al., 2017). Recent works in tax fraud detection are based on supervised machine learning techniques with labelled, audit-assisted data. A typical example in a fraud detection setting, using regression is predicting the amount of fraud. In classification, the target is categorical which means that it can only take on a limited set of predefined values. In binary classification, only two classes are considered (e.g., fraud versus no-fraud) whereas in multiclass classification, the target can belong to more than two classes (e.g., severe fraud, medium fraud, no fraud). Predictive analytics (supervised learning) can however also be useful to help explain the anomalies found by descriptive analytics (unsupervised learning). However, the “follow and auditing” approach can be slow and costly.

Indeed, several authors argue for the application of unsupervised techniques (De Roux et al., 2018). For instance, unsupervised techniques could find behaviour that deviates from normal behaviour and finding outliers/anomalies (Baesens et al., 2015). They are unsupervised, as they do not need observations to be labelled as fraudulent or non-fraudulent. Anomalies do not necessarily represent fraudulent observations. Hence, the usage of unsupervised learning for fraud detection requires extensive follow-up and validation of the identified, suspicious observations.

A third type of technique that can be used are social network analysis where fraudulent activities can be identified within a network of linked entities (Baesens et al., 2015). Indeed, fraud is social in nature fraud. The probability of someone committing fraud depends on the people he or she is connected to. These are the so-called guilt-by-associations (Koutra et al., 2011). An optimal fraud detection will make use, in a complementary manner, of these different techniques. A good example of big data analysis to detect tax fraud is the British Connect System[[2]](#footnote-2) (Maciejewski, 2017).

### Research Gaps

Following this background overview of the literature, we can identify two main research gaps this chapter intends to tackle.

First, there is a lack of a formalized process for Fraud Analytics by governments that integrate big data and AI solutions (Research Gap 1). Indeed, the big data analytics process has been well explored in the literature but its application in the fraud detection domain remains to be formalized. The existing knowledge base focuses more about the diverse techniques used in fraud analytics but less about describing the global processes that formalize the integration of digital technologies in business processes. This gap will be tackled in section 4.1 by presenting the main available insights from two practical cases.

Second, this section showed that the big data analytics process is not only technical but also focuses on interactions between data scientists and domain experts. To make the fraud detection process as efficient as possible, a close collaboration should be ensured between IT experts in charge of the data analytics and the business domain experts, knowledgeable in fraud detection, investigation or legal constraints. Indeed, as (Viaene, 2013) underlines it, IT experts (data scientists, computer scientists, data miners, etc.) are not always knowledgeable in the *business domain* as they experiment the data in the *model domain*. Domain experts should assess and validate the output of the data experiments and algorithms as much as possible. The continuous conversation between those two stakeholders’ groups also allows hidden assumptions or possible bias to be tackled as soon as possible. The alignment between two key stakeholders’ groups (Business and IT) is thus a key success factor in analytics projects. However, as there is little research about how this challenge takes place in practice (Research Gap 2), we investigate it in-depth in section 4.2.

## Methodology

To formalize the Fraud Analytics Process (Research Gap 1), we first identify in the literature how the big data analytics domains have been applied to fraud detection. This step is summarized in Section 2. This allows developing a theoretical lens to examine two case studies within the Belgian Federal government: the detection of tax frauds by the Federal Public Service (FPS) Finances and the detection of social security infringements by the Social Security Institutions (SSIs). By analysing these case studies, the organizational challenge of Business-IT alignment can be clearly illustrated as well (Research gap 2). Yin (2014) mentions that case study research is relevant to examine a current phenomenon when “how” questions are raised and where the researcher has no intervention in the case. Furthermore, two case studies were selected to improve the external validity of the research and draw general conclusions about the fraud analytic process in Belgium.

Data from the cases were extracted through semi-structured interviews. The main insights from the literature were used to devise an interview guide. Indeed, this qualitative method is effective when covering a complex topic in detail (Baarda et al., 1996; Boyce and Neale, 2006). Moreover, this technique is relevant for our research questions, as it centres around the expertise of the practitioners, and not on the validation of the knowledge of the researchers. The complete interview guide can be found in the Appendix section. The goal of this interview guide is to better understand the Fraud Analytics process, and its related challenges, implemented in the Belgian Federal governments to detect fraud. It is structured around the three big data analytics phase described in (Klievink et al., 2017). The data collected through this guide was used to understand the impact of legal requirements on the fraud analytics process and to understand the organizational alignment challenge in Fraud Analytics.

In total, 21 interviews were performed with stakeholders from the two backgrounds of interest: IT and Business profiles. This enabled us to create a complete understanding of the fraud analytics process and of the alignment challenge. The full list of interviewees can be found in Table 1.

*Table 1. List of Interviewees*

|  |  |  |
| --- | --- | --- |
| Case | Function | Organisation |
| Tax Fraud | Business | FPS Finances |
| Tax Fraud | IT | FPS Finances |
| Tax Fraud | Business | FPS Finances |
| Tax Fraud | IT | FPS Finances |
| Tax Fraud | IT | FPS Finances |
| Tax Fraud | IT | FPS Finances |
| Tax Fraud | Business | FPS Finances |
| Tax Fraud | Business | FPS Finances |
| Social Security | IT | Smals (Private company) |
| Social Security | IT | Smals (Private company) |
| Social Security | IT | Smals (Private company) |
| Social Security | Business | CBSS (Social Security Database) |
| Social Security | Business | CBSS (Social Security Data) |
| Social Security | Business | ONEM (Job allocations) |
| Social Security | Business | INAMI (Health allocations) |
| Social Security | IT | INAMI (Health allocations) |
| Social Security | Business | INAMI (Health allocations) |
| Social Security | Business | INAMI (Health allocations) |
| Social Security | Business | INAMI (Health allocations) |
| Social Security | IT | ONSS (Social Security coordination) |
| Social Security | IT | ONSS (Social Security coordination) |

Following the recommendations of (Yin, 2014), we analysed the data related to the two cases having four quality criteria in mind: construct validity, internal validity, external validity and reliability. We ensured construct validity by validating, with the interviewees, the main findings from the study. Regarding internal validity, the causal relationship between the business-IT alignment challenge and the fraud analytics process was once again ensured by the confirmation interviews. Although interviews could have been performed with complementary organisations (e.g. other SSIs or administrations at different government level) to increase this validity, this was not done here as we had reached code saturation. Regarding external validity, we ensured a replication logic by taking two case studies to study the fraud analytics process. Finally, to ensure reliability, we have stored all the interview reports in a case study database in the research data repository of our university.

The method used to analyse the findings extracted from the research data was *Thematic Content Analysis* (Anderson, 2007). The analysis started with summarising the interviews in a data memo. In order to code the data, we skim the transcripts and highlight relevant sentences based on the research questions. Then, we insert the codes into a table divided by the main stages of big data analytics. This method enables us to link similar themes from every interview to each other, making it easier to analyse what was expressed and how it compares with other findings. We categorise the results from our interviews (textual data) to concrete steps of the fraud analytics process as well as the related challenges. We performed interviews until we reached code saturation, meaning that no new codes were identified after five consecutive interviews (Guest et al., 2006).

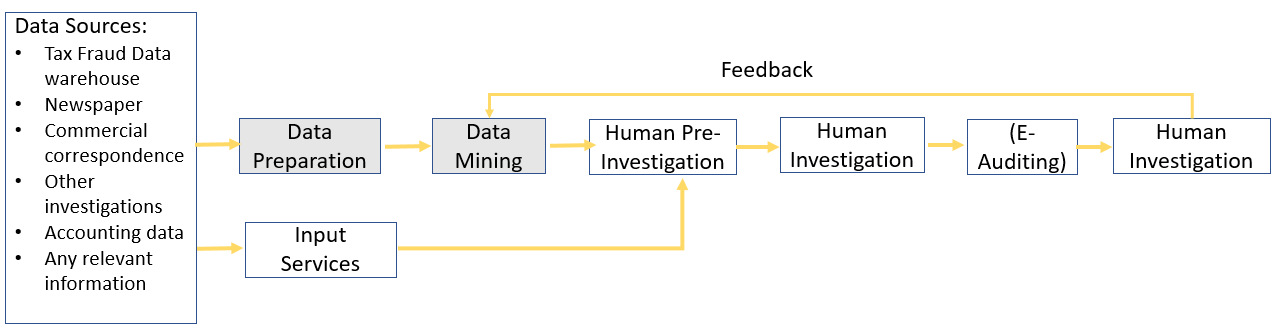
## Results

### Formalizing the Fraud Analytics Process: Insights from the case studies

From the 21 interviews we performed and analysed, we can formalize the fraud analytics processes of the two studied case studies (i.e. the detection of tax frauds and the detection of social security infringements).

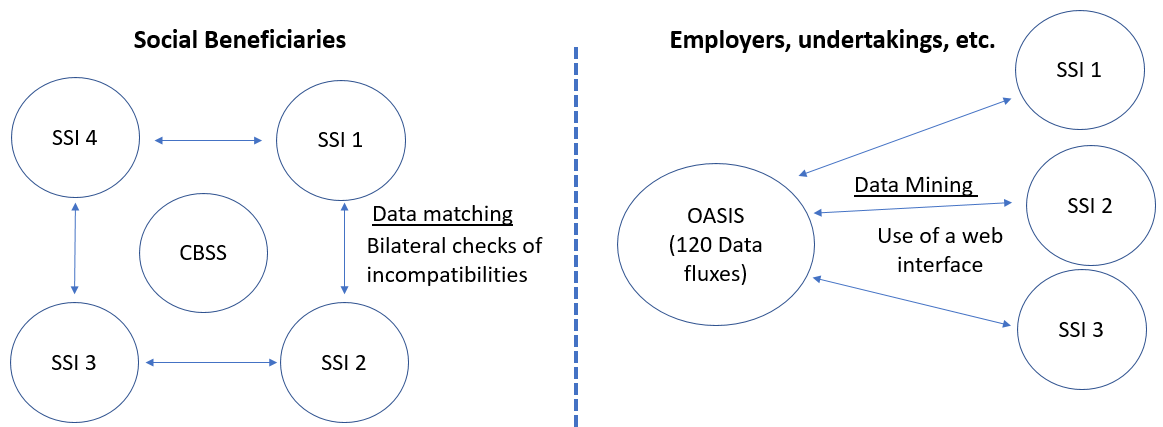
Regarding the tax fraud detection process, data is first extracted from several sources and prepared for analysis. Then, data mining is used to signal potentially fraudulent cases that need to be further examined. These two tasks (in grey) are performed by data miners. Then, at the pre-investigation stage, the signals derived from the data mining tasks are enriched with data from other sources, and it is decided whether a proper investigation should be started. Finally, in the investigation stage, some of the potentially fraudulent cases are examined in-depth, with the support of analytics (e.g. text mining) to explore a large quantity of unstructured data. This stage is also referred to as e-auditing. Inspectors perform these inspection tasks and then give feedback to data miners about the relevance of the signals. It must be noted that cases to be investigated are sometimes also suggested by “Input services” that manually detect cases to be further investigated. The fraud analytics process, applied to tax fraud detection, is presented in a simplified version in Figure 2.

*Figure 2. Fraud Analytics Process - Tax Fraud Detection*



For the social security infringement detection process, it is important to understand that a “Social Security Network” was created by the law of 15 January 1990[[3]](#footnote-3), in which all the Belgian Federal public SSIs are structured around the “Crossroad Bank for Social Security” (CBSS) (Degrave, 2020). The CBSS acts as the core of the network, and the SSIs are the nodes[[4]](#footnote-4). While these SSIs remain in control of their authoritative sources of personal social data, the CBSS acts as the central actor for the data sharing between them.[[5]](#footnote-5) The CBSS thus does not itself store any data, but rather acts as a “gatekeeper” that checks that an SSI has the right to access data stored on one of the nodes of the network (another SSI). Regarding social security fraud, there is a difference between the types of techniques used to detect fraud committed by beneficiaries of social allocations, on the one hand, and employers, health institutions, independent workers, etc., on the other hand. For the former, SSIs mainly rely on data matching techniques via bilateral crosschecks from other SSIs’ databases, aimed at identifying incompatibilities in terms of allocations. These are performed either before or after the payment of the allocation. For the latter, social security institutions mainly rely on data mining techniques, using the OASIS data warehouse, where larger quantities of pseudonymised data are compiled. Moreover, one SSI is currently developing a Big Data Analytics Platform to improve the data governance mechanisms between SSIs, notably to tackle social fraud. The social security infringement detection process is presented in a simplified version in Figure 3.

*Figure 3. Fraud Analytics Process - Social Security Infringements Detection*



### Business-IT Alignment Challenges in the Fraud Analytics Process

In Figure 4, a simplified Fraud Analytics process, abstracting the two above studied processes, is presented. This simplified process highlights the moments where alignment between the business and IT sides is necessary to ensure an efficient fraud detection. As identified in the interviews, alignment is particularly necessary for three key moments:

1. Requirements identification: How can the business team translated their needs in technical requirements for the IT team?
2. Presentation: How can the IT team present the output of the data analytics phase to the business team?
3. Feedback: How can the business team give feedback about the output of analytics to the IT team?

*Figure 4. Simplified Fraud Analytics process highlighting the need for alignment*



In the two cases studied above, several interviewees highlighted the lack of alignment for these three key interaction moments:

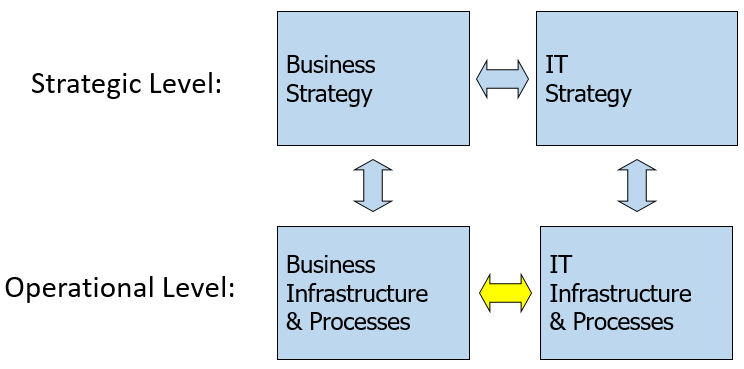
1. Business teams sometimes lack understanding about what the IT team needs. Furthermore, the data miners are often not fraud specialists and the needs from the business sometimes lack understandability and feel fragmented. As one interviewee nicely puts it: “*Business talks fraud and IT talks statistics*”. The IT teams can interpret this lack of communication in several ways: business experts may feel that the IT teams threaten their job, business experts do not know the relevance of the miners, business expert do not want to change their traditional way of working. Furthermore, one investigator mentioned that it is essential that “*data scientists receive insights from the field due to the changing nature of fraud, as it impacts what the data mining needs to look for*”.
2. There are no standard ways to present the output of the analysis to business domain experts. It strongly varies depending on the IT team. Some IT teams try to make the output by “*spending a lot of time to make the analysis as readable as possible*”. For instance, they provide an excel sheet with clear headings or point separators. The goal is that the business experts immediately understand the output of the analytics phase.
3. Once this output is sent, the IT teams do not receive follow-up from the business experts. As one data miner puts it, she can “*only can assume that what she sent out is good*”. They are not aware about what happens with the data, which is frustrating, as they do not know if the business experts appreciate the output of the analytics phase. They welcome feedback to improve their output. In the current situation, IT teams can only assume to what extent the output is satisfying. However, it is possible that the business experts add data manually afterwards and lose time even though the IT teams could easily do it.

## Recommendations for Alignment in Fraud Analytics

Based on the identified Fraud Analytics processes of the previous section, we better understand the organizational challenge related to collaboration between domain experts and data scientists. Therefore, in this section, we investigate the leads for solution from three research areas. Based on these leads, we suggest a tentative methodology to align data analytics and business domain in fraud analytics.

Even though the key questions raised in Section 4.2. might have specific answers, the literature about Business-IT alignment is promising to tackle the alignment challenges in an overarching manner. As shown in Figure 5, (Avison et al., 2004) underline that this alignment can take place at two levels: the alignment between the business strategy and the IT strategy, and the alignment between the operational business processes and IT processes. In the case of interest in this chapter, we focus on fostering the operational alignment between the business (e.g. controllers, investigators) and IT processes (e.g. data scientists, data miners), in yellow in Figure 5. In the following sub-sections, we present three ways forward to foster this alignment for Fraud Analytics.

*Figure 5. Simplified Strategic alignment model by (Avison et al., 2004)*



### Agile Methods

Traditional systems development approaches, such as the Waterfall model, seemed to prevail for a long time in governments. No complete study has been found on current software development practices in governments but authors have underlined the predominance of the waterfall model (Følstad et al., 2004; Pardo and Scholl, 2002). Such methods highly rely on thorough planning and process standardization and assume that the requirements remain static throughout the development process. They prevent public organizations from quickly adapting their processes to foster collaboration and participation of all stakeholders in the process. Nonetheless, over the last decade, some governmental organizations are becoming interested in a number of new techniques and approaches, such as agile development, to stimulate a more collaborative work environment in governments. Agile software development refers to a group of flexible and lightweight methodologies that rely on a set of principles and practices for the development of software (e.g., time-boxed iterations, customer involvement, daily meetings, continuous process improvement…) (Beck et al., 2001). Agile methods share a number of principles that drive the development process of practitioners. These 12 Agile Principles (AP) are described in the Agile manifesto (Beck et al., 2001) and are listed below for the sake of clarity:

* AP1*: Customer satisfaction by early and continuous delivery of valuable software*
* AP2*: Welcome changing requirements, even in late development*
* AP3*: Working software is delivered frequently, weeks rather than months*
* AP4*: Close, daily cooperation between business people and developers*
* AP5*: Projects are built around motivated individuals, who should be trusted*
* AP6*: Face-to-face conversation is the best form of communication*
* AP7*: Working software is the primary measure of progress*
* AP8*: Sustainable development, able to maintain a constant pace*
* AP9*: Continuous attention to technical excellence and good design*
* AP10*: Simplicity—the art of maximizing the amount of work not done—is essential*
* AP11*: Best architectures, requirements, and designs emerge from self-organizing teams*
* AP12*: Regularly, the team reflects on how to become more effective, and adjusts accordingly*

These general principles have been implemented through more formalized methods such as Extreme Programming (XP), SCRUM, Feature Driven Development, Dynamic Systems Development Method (DSDM), Lean Development/Management (Cohen et al., 2003). These methods can deliver high value for stimulating alignment between business and IT in Fraud Analytics instead of the sequential process shown in Figure 4. This process, that can be referred to as “Agile Analytics”, is an emerging research area (Collier, 2015)

### Participation Methods

Another lead for solution to foster alignment between business and IT in Fraud Analytics resides in the user participation in information systems literature. The information system research field has long proven that an increased user satisfaction and early involvement in the development process (e.g. requirements engineering activities) improves system quality (Hartwick and Barki, 1994). The importance of user participation and customer involvement has been underlined by the evolution of the traditional Waterfall software development methods to agile methods. These methods advocate more customer involvement in software development as shown in the agile principles AP1, AP2 and AP4 above. Indeed, the data analytics performed by the IT team can be considered as a software development project with a key phase of requirements’ identification. This identification can greatly be improved thanks to participation methods with the business team members as users.

This increased user participation has historically been implemented in the form of three main practices, giving different power of decision to the users: participatory design, user-centered design and user-innovation. Based on (Holgersson et al., 2012), we provide the following definitions for these three concepts. *Participatory Design* (PD) advocates an approach where good ideas are as likely to come from the user groups than from the decision-makers (Schuler and Namioka, 1993). In that regard, users and system developers are considered partners in the development process. In the context of this approach, the users can contribute as advisors (by assessing prototypes), as representatives to represent a particular user group or as all-inclusive participants where all users contribute to the development work. *User-Centered Design* (UCD) emerged in the human-computer interaction field and underlines the important impact of user needs on the design of the interface (Abras et al., 2004). Contrary to the previous approach, users and developers are not seen as equal because users only provide knowledge to the developers who, consequently, takes into account this business domain knowledge. For instance, the developers could organize focus groups to gather this knowledge but still have the power to take all decision unilaterally. *User Innovation (UI)* is the extreme counterpart of non-participation where the problem identification and design solutions emerge directly from the user, or more specifically from the “lead users” group. This sub-group refers to users that have strong needs that will become more general in the marketplaces in the future (von Hippel, 1986).

These three main practices can be implemented through several participation methods, directly relevant in an analytics project:

* *Interviews*: Interviews constitute a direct interaction method often used in the context of requirements engineering. Through this method, the IT team can ask specific questions to understand the needs of the business team for the analytics phase.
* *Representation in the project team*: In order to give more influence to users, Chan and Pan (2008) advocate the identification of salient intermediaries representing the users in all development stages. With this method, a representative from the business team (e.g. the lead fraud investigator) could be involved in every step of the analytics project for feedback.
* *User workshops*: The organization of workshops to interact with a selected group of representative users is a method often used in the requirements identification stage to elicit innovation solutions (Følstad et al., 2004; Oostveen et al., 2004). In more recent research, these workshop are organized thanks to creativity techniques such as visualization tools or improvisation principles (Mahaux and Maiden, 2008). Design Thinking also represents a creative technique that can enhance user workshops.
* *Answer to surveys*: Surveys are used for a number of purposes (market evaluation, research…) but also for the large-scale participation of users, mainly in the evaluation phase of software development. De Róiste (2013) provides insights about this evaluation by users through online surveys, phone or in person surveys. Large-scale surveys can be useful to help the IT team received feedback about the output of the analytics phase.
* *Dedicated Software***:** In order to facilitate the large-scale participation of users, practitioners can develop dedicated software (that can take the form of platforms, applications,…) to to gather users’ ideas and needs (Berntzen and Johannessen, 2016). Crowd-centric Requirements Engineering (CCRE) platforms apply the crowdsourcing paradigm in all phases of requirements engineering such as elicitation, negotiation and prioritization (Snijders et al., 2015).
* *Prototyping*: Prototyping is a method often used to present a non-finished product to its potential users. van Velsen et al. (2009) suggested a user-centric requirement engineering method for the design of online services with a rapid prototyping tested through focus groups, interviews or citizen walkthrough. These iterative developments of the output of analytics can deliver value to foster alignment.

### Design Thinking

In this section, we expand on the use of design thinking in analytics to stimulate alignment in Fraud Analytics. Design Thinking can be considered as a specific combination of the user innovation practice as well as the interview, prototyping and workshop participation methods. Indeed, a promising lead to stimulate the collaboration between IT (the developers) and business domain experts (the users) resides in the use of the design thinking methods as reported in (Chongwatpol, 2020). Design Thinking is a creative process to develop solutions aligned with users’ needs, combining several participation methods. It is applied in many contexts and recently investigated for analytics. These steps are generic but have recently shown potential for improving analytics projects (Chongwatpol, 2020):

* *Empathize*: understanding the business perspective. In this step, interviews with a representative set of stakeholders are performed to understand their needs, experience, and motivations.
* *Define*: unpacking and structuring the key findings from the interviews. In this step, the main challenges of the stakeholders are identified and features (or specific data analysis techniques) are mapped to solve them. Furthermore, the challenges are formulated as point-of-view statements to make them meaningful. If necessary, personas (fictional users with clear needs to be addressed by the solution) can be used.
* *Ideate*: brainstorming about which techniques or combination of techniques are relevant. In this step, workshops are used to design alternatives analytics solutions by thinking creatively through brainstorming, storyboards, mind maps, and idea clustering and selection. The ideas must be of large quantity and diversity of ideas so that unexpected techniques are uncovered.
* *Prototype* : finalizing the ideas for the solution space to evaluate practicability. In this step, low-fidelity analytics prototype (using mock-ups of visualizations for instance) allow to demonstrate potential solutions and leave the details for later.
* *Test*: receiving feedback on the intermediary solution. In this step, feedback is collected through interviews about the prototype to better understand users’ opinion, experience and feedback.

### Towards a Methodology for Business-IT Alignment in Fraud Analytics

Drawing from the insights gathered from the agile methods, participation methods and design thinking literatures, we suggest a tentative methodology for Business-IT alignment in Fraud Analytics. Figure 6 represents this methodology visually.

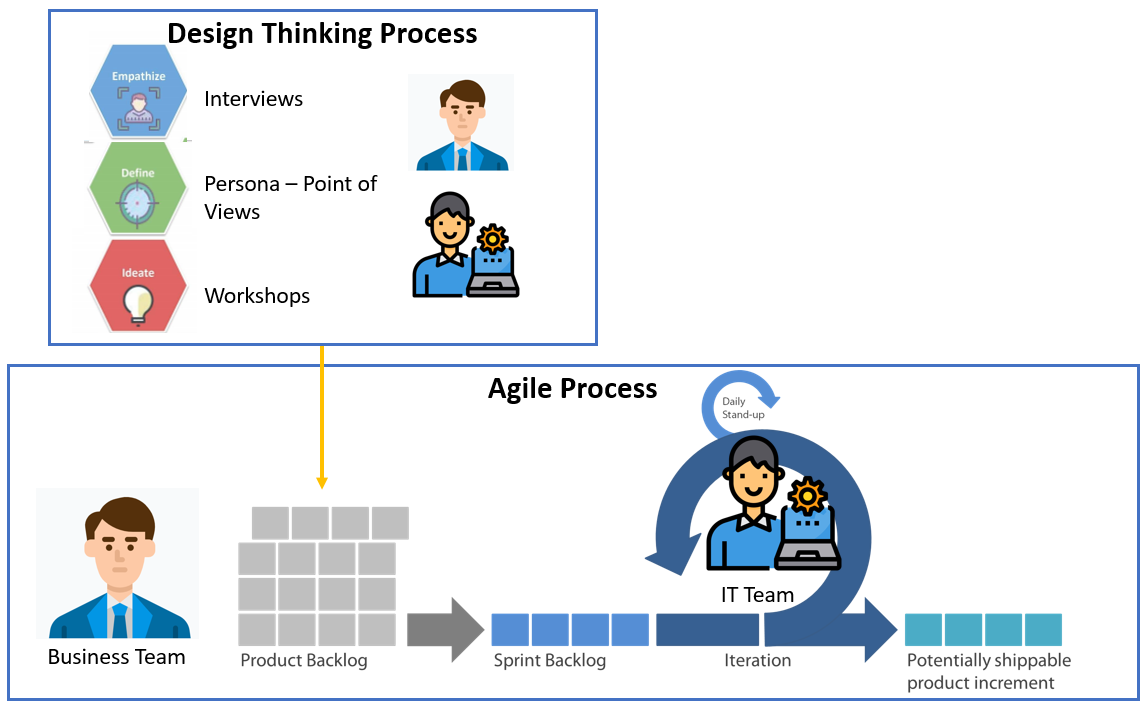


Figure 6. Methodology for Business-IT Alignment in Fraud Analytics

Drawing from the SCRUM agile methodology (Schwaber and Sutherland, 2017), the business team (fraud investigators) can be considered as the product owners of the analytics process and the IT team (the data scientists) can be considered as the development team in charge of iterative provision of analytics output. The methodology is composed of the following steps:

1. The methodology starts with the Design Thinking process where Business and IT work together. The business-side is interviewed by the IT-side *(“Empathize” stage, “Interview” participation method)* so that clear requirements are defined (“Define” stage). Furthermore, the organization of ideation workshops allows thinking creatively about analytics solutions to address the requirements *(“Ideate” stage, “Workshop” participation method).* The refined leads for solutions are bundled into a product backlog that the business team is in charge of. This backlog summarizes the requirements to be addressed by the IT team in a hierarchical manner.
2. The IT team self-organizes and selects which elements from the backlog they will tackle in a sprint (time-boxed period of work, often 2-4 weeks). The team then works iteratively *(“Prototype” stage, “Prototyping” participation method)*, with daily stand-ups to reflect on the work done.
3. After the sprint, the IT team produce potentially shippable product increments submitted for feedback to the business team *(“Test” stage)*. Based on the feedback received through several iterations, the analytics will be more aligned to the needs of the fraud investigators and this will foster collaboration.

Step 1 allows business and IT to formalize clear requirements through relevant participation methods as well as providing a usable method to manage them (the product backlog). In that sense, it helps to address the “Requirements identification” alignment moment. Step 2 and the prototyping approach allow iteratively improving the way the IT team presents the output of analytics to the business team. In that sense, it helps to address the “Presentation” alignment moment. Step 3 formalizes feedback within the analytics process so that there is communication between the two teams and self-reflection about the overall process. This helps to address the “Feedback” alignment moment.

Despite the promises of this generic methodology, a particular attention should be paid on its customization depending on the context. Indeed, the implementation of the methodology in a government context may raise challenges because of the intrinsic characteristics of government: need for regulatory compliance, lack of strategic or operational support, reluctance to change, etc. Indeed, the same challenge was met for the implementation of agile methods. As a consequence, a growing line of research has identified that practitioners also use tailored methods that fit the specificities of their organizations (Campanelli and Parreiras, 2015). The same customization approach should be followed to iteratively improve and adapt the alignment methodology to each government organization.

## Conclusion

Business-IT alignment has for long been considered as a key success factor for the digital governance of public administrations as it allows the IT operations and strategies to support the business operations and strategies (Avison et al., 2004). However, the introduction of more advanced technologies (such as Big Data, Artificial Intelligence or Machine Learning) makes the need for alignment increasingly important. Indeed, as we have seen in this chapter, the introduction of analytics in the fraud detection domain does not simply affect the teams on the IT-side of the organization but consists in a more global process where collaboration between business and IT teams is essential. Ensuring a dialog between data scientists, data miners, legal specialists and tax investigators in public administrations becomes increasingly challenging as fraud become more and more difficult to detect while the analytics used also become increasingly complex. The organizational alignment between diverse profiles is a pre-condition to ensure a proper data-driven governance in public administrations.

In order to contribute to this need for alignment, this chapter formalizes the Fraud Analytics process, illustrates one key organizational challenge (the alignment between business and IT teams) and provides leads for solutions to foster this alignment in Fraud Analytics. For this, we examine two case studies from the Belgian Federal government through 21 interviews: the detection of tax frauds and of social security infringements. This chapter is relevant for researchers as it formalizes the Fraud Analytics process for both cases and then suggests three leads for solution, under-investigated for the fraud analytics domain, to foster business-IT alignment: agile methods, participation methods and design thinking. Furthermore, it is relevant for practitioners as the leads for solutions are bundled into a usable methodology, directly actionable to improve the effectiveness of fraud analytics. This managerial solution can be considered as an effective governance tool in public administrations to increase organizational alignment in data-driven public administrations.

The research performed in this chapter also has two main inherent limitations that introduce leads for further research. First, only two cases were examined to formalize the Fraud Analytics process in government. These cases come from Belgium and an international comparison with other countries would be necessary to increase external validity. Second, three specific leads for solutions were selected for discussion in this chapter. Other leads are promising and deserve the attention of the interested future researchers: the use of data logistics platforms to promote collaboration between business and IT, the coaching of stakeholders for cultural change to increase collaboration or the introduction of business analyst profiles to make the link between business and IT.

## References

Abras, C., Maloney-Krichmar, D., Preece, J., 2004. User-centered design. Bainbridge, W. Encycl.

Anderson, R., 2007. Thematic Content Analysis (TCA): Descriptive Presentation of Qualitative Data Using Microsoft Word. Descr. Present. Qual. data 1–4.

Archenaa, J., Anita, E.A.M., 2015. A survey of big data analytics in healthcare and government, in: Procedia Computer Science. pp. 408–413. https://doi.org/10.1016/j.procs.2015.04.021

Avison, D., Jones, J., Powell, P., Wilson, D., 2004. Using and validating the strategic alignment model. J. Strateg. Inf. Syst. 13, 223–246. https://doi.org/10.1016/j.jsis.2004.08.002

Baarda, B. (Dirk B., Goede, M.P.M. de (Matthëus P.M., Meer-Middelburg, A.G.E. van der, 1996. Basisboek open interviewen : praktische handleiding voor het voorbereiden en afnemen van open interviews [Book about basics open interviewing: a practical guidance for preparing and conducting open interviews]. Stenfert Kroese.

Baesens, B., Vlasselaer, V. Van, Verbeke, W., 2015. Fraud Analytics Using Descriptive, Predictive, and Social Network Techniques, Fraud Analytics Using Descriptive, Predictive, and Social Network Techniques. https://doi.org/10.1002/9781119146841

Beck, K., Beedle, M., Van Bennekum, A., Cockburn, A., Cunningham, W., Fowler, M., Grenning, J., Highsmith, J., Hunt, A., Jeffries, R., Kern, J., Marick, B., Martin, R.C., Mellor, S., Schwaber, K., Sutherland, J., Thomas, D., 2001. Agile Manifesto. Softw. Dev. https://doi.org/10.1177/004057368303900411

Berntzen, L., Johannessen, M.R., 2016. The Role of Citizen Participation in Municipal Smart City Projects: Lessons Learned from Norway, in: Smarter as the New Urban Agenda. Springer International Publishing, Switzerland, pp. 299–314. https://doi.org/10.1007/978-3-319-17620-8

Bertot, J.C., Gorham, U., Jaeger, P.T., Sarin, L.C., Choi, H., 2014. Big data, open government and e-government: Issues, policies and recommendations. Inf. Polity 19, 5–16. https://doi.org/10.3233/IP-140328

Bishop, C.M., 2006. Pattern Recoginiton and Machine Learning, Information Science and Statistics.

Boyce, C., Neale, P., 2006. Conducting in-depth interviews: A Guide for designing and conducting in-depth interviews. Evaluation 2, 1–16. https://doi.org/10.1080/14616730210154225

Campanelli, A.S., Parreiras, F.S., 2015. Agile methods tailoring - A systematic literature review. J. Syst. Softw. 110, 85–100. https://doi.org/10.1016/j.jss.2015.08.035

Castellón González, P., Velásquez, J.D., 2013. Characterization and detection of taxpayers with false invoices using data mining techniques. Expert Syst. Appl. https://doi.org/10.1016/j.eswa.2012.08.051

Chan, C.M.L., Pan, S.L., 2008. User engagement in e-government systems implementation: A comparative case study of two Singaporean e-government initiatives. J. Strateg. Inf. Syst. 17, 124–139. https://doi.org/10.1016/j.jsis.2007.12.003

Chantillon, M., Simonosfki, A., Tombal, T., Kruk, R., Crompvoets, J., Snoeck, M., 2020. Analysing e-government through the multi-level governance lens - An exploratory study in Belgium, in: CEUR Workshop Proceedings.

Chongwatpol, J., 2020. Operationalizing Design Thinking in Business Intelligence and Analytics Projects. Decis. Sci. J. Innov. Educ. 18, 409–433. https://doi.org/10.1111/dsji.12217

Chung, W., Zeng, D., 2018. Social-media-based policy informatics: Cyber-surveillance for homeland security and public health informatics, in: Public Administration and Information Technology. Springer, pp. 363–385. https://doi.org/10.1007/978-3-319-61762-6\_16

Cohen, D., Lindval, M., Costa, P., 2003. Agile Software Development, DACS SOAR Report. Maryland. https://doi.org/10.1145/1978802.1978803

Collier, K., 2015. Agile Analytics: A Value-Driven Approach to Business Intelligence and Data Warehousing., Agile Software Development Series.

Daniell, K.A., Morton, A., Ríos Insua, D., 2016. Policy analysis and policy analytics. Ann. Oper. Res. 236, 1–13. https://doi.org/10.1007/s10479-015-1902-9

Davenport, B.T.H., Jarvenpaa, S.L., 2008. The Strategic Use of Analytics in Government, IBM Center for The Business of Government. IBM Center for The Business of Government.

De Marchi, G., Lucertini, G., Tsoukiàs, A., 2016. From evidence-based policy making to policy analytics. Ann. Oper. Res. 236, 15–38. https://doi.org/10.1007/s10479-014-1578-6

De Róiste, M., 2013. Bringing in the users: The role for usability evaluation in eGovernment. Gov. Inf. Q. 30, 441–449. https://doi.org/10.1016/j.giq.2013.05.007

De Roux, D., Pérez, B., Moreno, A., Del Pilar Villamil, M., Figueroa, C., 2018. Tax fraud detection for under-reporting declarations using an unsupervised machine learning approach, in: Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. https://doi.org/10.1145/3219819.3219878

Degrave, E., 2020. The Use of Secret Algorithms to Combat Social Fraud in Belgium. Eur. Rev. Digit. Adm. Law 1, 167–177.

Desouza, K.C., Jacob, B., 2017. Big Data in the Public Sector: Lessons for Practitioners and Scholars. Adm. Soc. 49, 1043–1064. https://doi.org/10.1177/0095399714555751

Erickson, J.S., Viswanathan, A., Shinavier, J., Shi, Y., Hendler, J.A., 2013. Open government data: A data analytics approach. IEEE Intell. Syst. 28, 19–23. https://doi.org/10.1109/MIS.2013.134

Følstad, A., Jørgensen, H.D., Krogstie, J., 2004. User involvement in e-government development projects. third Nord. Conf. Human-computer Interact. 82, 217–224. https://doi.org/10.4018/978-1-59904-027-1.ch016

Gil-Garcia, J.R., Pardo, T.A., Luna-Reyes, L.F., 2018. Policy analytics: Definitions, components, methods, and illustrative examples, in: Policy Analytics, Modelling, and Informatics. Springer, pp. 1–16. https://doi.org/10.1007/978-3-319-61762-6\_1

Guest, G., Bunce, A., Johnson, L., 2006. How Many Interviews Are Enough?: An Experiment with Data Saturation and Variability. Field methods 18, 59–82. https://doi.org/10.1177/1525822x05279903

Hamza, K., Mellouli, S., 2018. Background on frameworks for policy analytics, in: Public Administration and Information Technology. Springer, pp. 19–37. https://doi.org/10.1007/978-3-319-61762-6\_2

Han, J., Kamber, M., Pei, J., 2012. Data Mining: Concepts and Techniques, Data Mining: Concepts and Techniques. https://doi.org/10.1016/C2009-0-61819-5

Hartwick, J., Barki, H., 1994. Explaining the role of user participation in information system use. Manage. Sci. 40, 440–465. https://doi.org/10.1287/mnsc.40.4.440

Henderson, J.C., Venkatraman, N., 1999. Strategic alignment: leveraging information technology for transforming organizations. IBM Syst. J. https://doi.org/10.1147/SJ.1999.5387096

Holgersson, J., Karlsson, F., Holgersson, J., Söderström, E., Hedström, K., 2012. Exploring user participation approaches in public e-service development. Gov. Inf. Q. 29, 158–168. https://doi.org/10.1016/j.giq.2011.07.009

Janssen, M., Kuk, G., 2016. The challenges and limits of big data algorithms in technocratic governance. Gov. Inf. Q. 33, 371–377. https://doi.org/http://dx.doi.org/10.1016/j.giq.2016.08.011

Joseph, R.C., Johnson, N.A., 2013. Big data and transformational government. IT Prof. 15, 43–48. https://doi.org/10.1109/MITP.2013.61

Kalampokis, E., Tambouris, E., Tarabanis, K., 2013. Linked Open Government Data Analytics, in: Wimmer, M.A., Janssen, M., Scholl, H.J. (Eds.), Electronic Government: Proceedings of the 12th IFIP WG 8.5 International Conference, EGOV 2013, Lecture Notes in Computer Science. Koblenz, Germany, pp. 99–110.

Kim, G.H., Trimi, S., Chung, J.H., 2014. Big-data Applications in the Government Sector. Commun. ACM 57, 78–85. https://doi.org/10.1145/2500873

Klievink, B., Romijn, B.J., Cunningham, S., de Bruijn, H., 2017. Big data in the public sector: Uncertainties and readiness. Inf. Syst. Front. 19, 267–283. https://doi.org/10.1007/s10796-016-9686-2

Koutra, D., Ke, T.Y., Kang, U., Chau, D.H., Pao, H.K.K., Faloutsos, C., 2011. Unifying guilt-by-association approaches: Theorems and fast algorithms, in: Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). https://doi.org/10.1007/978-3-642-23783-6\_16

Loukis, E.N., 2018. Citizen-sourcing for public policy making: Theoretical foundations, methods and evaluation, in: Public Administration and Information Technology. Springer, pp. 179–203. https://doi.org/10.1007/978-3-319-61762-6\_8

Maciejewski, M., 2017. To do more, better, faster and more cheaply: using big data in public administration. Int. Rev. Adm. Sci. https://doi.org/10.1177/0020852316640058

Mahaux, M., Maiden, N., 2008. Theater improvisers know the requirements game. IEEE Softw. 25, 68–69. https://doi.org/10.1109/MS.2008.128

Negash, S., 2004. Communications of the Association for Information Systems Business Intelligence BUSINESS INTELLIGENCE. Commun. Assoc. Inf. Syst. https://doi.org/10.1007/978-3-540-48716-6\_9

Oostveen, A.-M., Van Den Besselaar, P., Besselaar, P. van den, 2004. From small scale to large scale user participation: a case study of participatory design in e-government systems, in: Eighth Conference on Participatory Design: Artful Integration: Interweaving Media, Materials and Practices. ACM Press, Toronto, Ontario, Canada, pp. 173–182. https://doi.org/10.1145/1011870.1011891

Pardo, T., Scholl, H.J., 2002. Walking atop the Cliffs--Avoiding Failure and Reducing Risk in Large-Scale E-government Projects, in: Proceedings of the 35th Hawaii International Conference on System Sciences (HICSS-35). Computer Societry Press, Island of Hawaii (Big Island), p. 124b (1-10).

Pencheva, I., Esteve, M., Mikhaylov, S.J., 2018. Big Data and AI – A transformational shift for government: So, what next for research? Public Policy Adm. https://doi.org/10.1177/0952076718780537

Schuler, D., Namioka, A., 1993. Participatory design: Principles and practices. CRC Press.

Schwaber, K., Sutherland, J., 2017. The Scrum Guide: The Definitive The Rules of the Game, Scrum.Org and ScrumInc.

Snijders, R., Ozum, A., Brinkkemper, S., Dalpiaz, F., 2015. Crowd-Centric Requirements Engineering : A method based on crowdsourcing and gamification. Dep. Inf. Comput. Sci. Utr. Univ. Tech. Rep. UU-CS-2015-004.

Sørum, H., 2011. An empirical investigation of user involvement, website quality and perceived user satisfaction in eGovernment environments. Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics) 6866 LNCS, 122–134. https://doi.org/10.1007/978-3-642-22961-9\_10

Sousa, W.G. de, Melo, E.R.P. de, Bermejo, P.H.D.S., Farias, R.A.S., Gomes, A.O., 2019. How and where is artificial intelligence in the public sector going? A literature review and research agenda. Gov. Inf. Q. 36. https://doi.org/10.1016/j.giq.2019.07.004

Tom Mitchell, 1997. Machine Learning, McGraw Hill.

Tsoukias, A., Montibeller, G., Lucertini, G., Belton, V., 2013. Policy analytics: An agenda for research and practice. EURO J. Decis. Process. 1, 115–134. https://doi.org/10.1007/s40070-013-0008-3

Ubaldi, B., 2013. Open Government Data: Towards Empirical Analysis of Open Government Data Initiatives. Paris.

van Velsen, L., van der Geest, T., ter Hedde, M., Derks, W., 2009. Requirements engineering for e-Government services: A citizen-centric approach and case study. Gov. Inf. Q. 26, 477–486. https://doi.org/10.1016/j.giq.2009.02.007

Van Vlasselaer, V., Eliassi-Rad, T., Akoglu, L., Snoeck, M., Baesens, B., 2017. GOTCHA! Network-based fraud detection for social security fraud. Manage. Sci. https://doi.org/10.1287/mnsc.2016.2489

Viaene, S., 2013. Data scientists aren’t domain experts. IT Prof. 15, 12–17. https://doi.org/10.1109/MITP.2013.93

von Hippel, E., 1986. Lead Users: A Source of Novel Product Concepts. Manage. Sci. 32, 791–805. https://doi.org/10.1287/mnsc.32.7.791

Yin, R.K., 2014. Case Study Research: Design and Methods, Sage Publi. ed, Sage Publications. https://doi.org/10.1097/FCH.0b013e31822dda9e

Yu, F., Qin, Z., Jia, X.L., 2003. Data mining application issues in fraudulent tax declaration detection, in: International Conference on Machine Learning and Cybernetics. https://doi.org/10.1109/icmlc.2003.1259872

## Appendix: Interview Guide

The goal of the following questions is to understand the fraud analytics process from the perspective of the interviewee. Furthermore, it allows eliciting the main organizational and legal challenges related to the process.

**General Questions**:

* + Can you describe your function within your organisation?
  + What is your background ? What is your main expertise in the organization ?
  + What is your role related to Data Analytics in your organisation?
  + What is your role related to Fraud Detection in your organisation ?

**Domain of Fraud**:

* + What does your organisation consider as a Tax Fraud/Social Security Infringement?
  + How do you detect these frauds? What are the different steps ?
  + What are the actors involved in the fraud investigation process ?
  + How is this policy implemented? Who are the main actors and what are they doing?

**Pre-Processing stage**:

* + Which data sources do you use for decision making? From which actors?
  + Do you combine data from several sources? What are the challenges?
  + How are the data stored?
  + What is the ideal way you would like to see the data presented to you ?
  + Do you have procedures in place to answer data subjects’ right request under the GDPR (access, rectification, erasure,…)?
  + How do you ensure the representativity of the datasets used to train the algorithms? How do you avoid biases?
  + Could you tell us more about the data warehouse to fight tax fraud (Art. 5.1 of the Law of 3 August 2012)?
  + Could you tell us more about the OASIS data warehouse (Art. 5bis of the Law of 15 January 1990)?
  + Do you define the purposes before collecting and combining the data, or do you collect and combine the data first then define the purposes for which you’ll use them?
  + How do you determine these purposes? Is it always linked to a legal obligation ?
  + How do you determine the categories of data you will collect?
  + How do you ensure that the data is kept up-to-date?
  + Are the citizens informed about these data collections/combinations? If yes how?

**Data Analytics stage**:

* + How do you investigate a specific fraud ?
  + Which techniques do you apply to analyse the data related to fraud?
  + How did you start to use these techniques ?
  + Are these technical tools developed internally or by private entities? Who holds the rights on the results?

**Post-Processing stage (Use for Fraud Detection):**

* + Do you use different techniques depending on the type of fraud ?
  + How is the output of the analytics phase presented? How does it impact your fraud investigation process?
  + Are some decisions solely based on automated processing, or is there always a human intervention in the fraud detection process (confirmation, investigation, etc…)? How significant is it?
  + Is the citizen made aware of the logic involved behind the decision (relevant factors)?
  + How can the citizen contest the decision?
  + Are you able to explain the predictions/decisions made by the algorithm?
  + What are the other main legal/organizational/technical challenges when using those techniques ?
  + To what degree do you take into account public opinion and the opinions of external stakeholders (such as businesses) on new fraud detection techniques and data-gathering methods?

**Future directions and Closing Questions**:

* + What would be for you, personally, the future of fraud detection policy in 10 years ? What would be the main factors to make it happen? What would be the main obstacles ?
  + Would you like to address another topic related to the research that we didn’t highlight in the interview?
  + Are there other stakeholders in your organisation that would be relevant for us to interview?

1. This expression refers to the idea that, in computer science, inputs of poor quality (e.g. missing data) will produce outputs of poor quality. [↑](#footnote-ref-1)
2. <https://www.telegraph.co.uk/finance/personalfinance/tax/11697816/What-does-the-taxman-know-about-you-your-finances-and-your-lifestyle.html> [↑](#footnote-ref-2)
3. Loi du 15 janvier 1990 relative à l'institution et à l'organisation d'une Banque-carrefour de la sécurité sociale, *M.B*., 22 février 1990. [↑](#footnote-ref-3)
4. <https://www.ksz-bcss.fgov.be> [↑](#footnote-ref-4)
5. Art. 3 of the Law of 15 January 1990. [↑](#footnote-ref-5)