Aalto University
School of Business

Data Mining in Tax Administration - Using Analytics to Enhance Tax Compliance

Information Systems Science

Master's thesis

Jani Martikainen

2012

Department of Information and Service Economy
Aalto University
School of Business

AALTO UNIVERSITY SCHOOL OF ECONOMICS
Department of Information and Service Economy
Master's thesis
Jani Martikainen

ABSTRACT
28.10.2012

ABSTRACT

Objectives of the Study

The overall objective of the study is to explore how tax administrations could make use of data mining to enhance tax compliance among the taxpayers. Tax compliance refers here to the taxpayers fulfilling their registration, filing, reporting and payment obligations correctly and at the right time. The potential role and benefits of data mining in tax administrations are clarified in view of the general operational framework, technology and organisation. The study also seeks to shed light on the potential implications of data mining on the relationship between a tax administration and a taxpayer in view of the agency theory.

Academic background and methodology

The literature review sets the scene for the study by drawing on earlier research in data mining and agency theory. Practical considerations of applying data mining, including lessons learned and plans for the future, are gathered from selected tax administrations. An important part of the methodological framework is the action research component: the author's employment at the Finnish Tax Administration, and the assignment given to him to run a feasibility study project of data mining in the organisation, establish a truly participatory setting. The author works with his colleagues to propose new methods and tools to help their community improve its work practices.

Findings and conclusions

Two particularly prominent uses for data mining are identified within a tax administration's operational framework: (1) Tax administrations can build up truly risk-based workflows for processing the registration, filing, reporting and payment transactions that the taxpayers make or should make. A comprehensive risk rating, based on data mining modeling, can be applied to each transaction so that all available relevant data are utilised. As a result, high-risk transactions can be flagged for case-specific treatment while low-risk transactions can move on to automated routine processing. (2) Tax administrations can segment the taxpayers and identify segment-specific compliance profiles in terms of diverse abilities and propensities to comply. This helps tax administrations better design and target their services and compliance actions.

The findings give grounds to believe that data mining can in many ways suggested by the agency theory advance a tax administration's goals in the agency relationship vis-à-vis a taxpayer.

Keywords

Data mining, analytics, advanced analytics, predictive analytics, public administration, agency theory, principal-agent theory, tax administration, taxation

AALTO-YLIOPISTON KAUPPAKORKEAKOULU
Tieto- ja palvelutalouden laitos
Pro Gradu -tutkielma
Jani Martikainen

TIIVISTELMÄ
28.10.2012

ABSTRAKTI

Tutkimuksen tavoitteet

Tutkimuksen tavoitteena on kartoittaa, miten veroviranomaiset voisivat hyödyntää tiedonlouhintaa veronmaksajien oikein toimimisen edistämiseksi. Oikein toimimisella tarkoitetaan tässä verotukseen liittyvien rekisteröinti-, ilmoitus- ja maksuvelvollisuuksien oikeansisältöistä ja oikea-aikaista noudattamista. Tiedonlouhinnan roolia ja hyödyntämismahdollisuuksia tarkastellaan veroviranomaisen operatiivisen toiminnan, tekniikan ja itse tiedonlouhintatyön organisoinnin näkökulmista. Lisäksi tutkitaan tiedonlouhinnan mahdollisia vaikutuksia veroviranomainen–veronmaksaja-asetelmaan agenttiteorian näkökulmasta.

Kirjallisuuskatsaus ja metodologia

Kirjallisuuskatsauksessa tuodaan esille tämän työn kannalta olennaisia havaintoja tiedonlouhintaa ja agenttiteoriaa käsittelevistä aiemmista tutkimuksista. Keskeisiä lähteitä ovat myös eri maiden veroviranomaisilta saadut tiedot tiedonlouhinnan kokemuksista ja suunnitelmista. Tärkeä osa tutkimuksen metodologista kokonaisuutta on sen toimintatutkimuskomponentti: tutkija työskentelee Verohallinnossa, jossa hänen johdettavakseen määrätyn tiedonlouhinnan esiselvitysprojektin myötä tutkimukseen on saatu todellisen kehittämistyön näkökulma olosuhteissa, joissa tutkija osallistuu toimintaan ja on mukana organisaation arkipäivässä.

Tulokset ja päätelmät

Tutkimuksessa kuvataan veroviranomaisen toiminnallinen viitekehys ja tunnistetaan siitä kaksi erityistä tiedonlouhinnan hyödyntämismahdollisuutta: (1) Veronmaksajien suorittamien rekisteröinti-, ilmoitus- ja maksutapahtumien viranomaiskäsittelyn kulku voidaan määrittää riskiperusteisesti. Tapahtuman analyyttinen kokonaistarkastelu yhdessä veronmaksajaa koskevan muun tiedon kanssa auttaa tunnistamaan, mitkä tapahtumat voidaan prosessoida koneeellisesti ja mitkä on syytä käsitellä manuaalisesti. (2) Veroviranomaiset voivat segmentoida veronmaksajat tiedonlouhinnan avulla ja tunnistaa segmenteittäin erilaisia asiakastarpeita sekä asiointikyvykkyyden ja veromyönteisyyden asteita. Nämä tiedot auttavat kohdentamaan veroviranomaisen palveluita sekä ohjaus- ja valvontatoimenpiteitä tarkoituksenmukaisemmin.

Tutkimustulosten perusteella tiedonlouhinta voi usella agenttiteoriassa esitetyllä tavalla parantaa veroviranomaisen tavoitteiden toteutumista veroviranomainen–veronmaksaja-asetelmassa.

Avainsanat

Tiedonlouhinta, analytiikka, edistyksellinen analytiikka, ennustava analytiikka, julkishallinto, agenttiteoria, päämies–agentti-suhde, veroviranomainen, verotus

ACKNOWLEDGEMENTS

I have been fortunate to combine a challenging work at the Tax Risk Management Unit of the Finnish Tax Administration with studies in the Information and Service Management Master's Programme of the Aalto University School of Economics. This thesis is a result of that symbiotic setting where inspiration has flown in both directions.

This thesis is, also, a by-product of approximately one year's brainstorming and exchanging of ideas and experiences between me and my colleagues, both domestically and internationally. I have tried my best to catch, analyse and put together here the essence of the endless fruitful conversations and debates that we have had on the use of advanced analytics in tax administration.

I want to express my deepest gratitude to all my Finnish and international colleagues who have contributed, either directly or indirectly, to this thesis. I believe I could not have got better ingredients for it. I hope the results will serve our common goal of enhancing tax compliance.

Helsinki, October 28, 2012

Jani Martikainen

TABLE OF CONTENTS

LIST OF FIGURES

1 INTRODUCTION

Taxation is an information intensive domain that involves processing of vast amounts of data concerning a large number of taxpayers. Tax administrations need these data to calculate or validate the right taxes for the taxpayers, be it natural persons, businesses and other juridical persons. In many countries both the taxpayers themselves and certain third parties are obliged to provide tax administrations with data for taxation purposes.

Just and equitable taxation constitutes one of the cornerstones of a well functioning welfare state. It is important that taxpayers know how to comply with the tax laws and, ideally, have no choice but to comply. Tax administrations can play a central role in attaining a high level of tax compliance among the taxpayers.

A central motivation behind this study is the belief that the vast amounts of data accumulated in the tax administrations' repositories are currently underutilised. To this end, an underlying assumption here is that data mining could provide powerful techniques for tax administrations to discover useful knowledge in support of their compliance enhancing agendas.

Taxation is carried out in a relationship between taxpayer and tax administration where the latter represents the state (and other public institutions). The *state* and the *governed* setting represents one form of a so-called agency relationship. The agency theory is concerned with resolving problems that can occur in agency relationships. Earlier research in agency theory will therefore be discussed and reflected against the setting and observations of this study.

1.1 Research problem

The overall research problem of this study can be stated as follows:

- How can data mining help tax administrations attain better tax compliance among the taxpayers?

In view of the research problem, it is important to understand that the functions of a tax administration go beyond tax audits and other retroactive controls. Today's tax administrations are essentially service organisations where taxpayers are truly considered as customers. It is the tax administration's task to create and maintain an enabling environment for its customers to be able to comply with applicable tax laws and regulations with minimum effort. The research problem should thus be viewed in relation to the tax administrations' overall playground that involves both *enabling* and *ensuring* compliance.

A large customer base with diverse abilities, behaviours, attitudes and motives among the taxpayers poses a permanent challenge probably for all tax administrations. Different customers need different types of attention. At the same time, many tax administrations face increasing efficiency and effectiveness requirements. Scarce resources must be allocated where they bring best return. The above considerations must be duly addressed in connection with the research problem.

Furthermore, an important consideration in this study is the role and potential implications of data mining on the relationship between a tax administration and a taxpayer in view of the agency theory. An interesting part of the research problem lies in looking into how data mining would fit in the problem-solving framework that the agency theory suggests for resolving problems in agency relationships, such as the one between a tax administration and a taxpayer.

Research issues

In spite of a seemingly promising potential, data mining has thus far been applied in tax administrations to a relatively limited extent. The overall objective of this study is to gain understanding of how data mining applications could help tax administrations accomplish their

general mission, to get the right tax at the right time, more efficiently and more effectively. This can be decomposed into the following more specific research issues:

- Research issue 1: Drafting a general operational framework of a tax administration, and identifying where in it data mining would have the biggest potential to bring added value
- Research issue 2: Identifying, in general terms, the required technology for a large-scale adoption of data mining in tax administration
- Research issue 3: Gathering available empirical evidence of data mining applications in tax administrations
- Research issue 4: Shedding light on the potential implications of applying data mining in tax administration in view of the agency theory

This thesis is not directly commissioned but there is a close linkage to the Finnish Tax Administration, owing to the author's employment therein. The author was assigned to run, from March to August 2012, a feasibility study project of the potential benefits of data mining for the Finnish Tax Administration. This thesis is, partly, a by-product of the said assignment. The thesis has, however, a more general view than addressing solely the needs of the Finnish Tax Administration.

1.2 Research methodology

The aims of the study entail a pragmatic research orientation. The research methods include the following:

- Reviewing books, articles, research papers and market information about data mining and its applications.
- Reviewing relevant research papers discussing the agency theory.
- Gathering selected tax administrations' experiences and plans concerning data mining. To this end, the author's participation in the OECD Conference on the Use of Advanced Analytics in Dublin in 29-30 November 2011, and further communication with some of the Conference delegates, provide an invaluable source of information.
- Drawing on the observations and findings of a feasibility study project of data mining in the Finnish Tax Administration in March-August 2012. Here, the research methodology resembles *action research*: the author looks into the research problem together with his colleagues, the leading domain experts and analysts in the Finnish Tax Administration, with the aim of proposing new methods and tools to help their community improve its work practices.

2 LITERATURE REVIEW

Data mining is a relatively young discipline. Its short history and inter-disciplinary nature are reflected in some vagueness in how it is defined and positioned vis-à-vis certain nearby concepts. This section first introduces some key concepts in the context of this study and then discusses certain methodological and technological aspects that merit consideration here. Thereafter we will briefly touch upon earlier academic research on data mining in tax administration. The last chapter of this section is devoted to agency theory, setting the scene for understanding how this theory can be applied in the context of tax administration.

2.1 Data mining and nearby concepts explained

There is no single comprehensive or universally accepted definition for data mining. It does not serve the aims of this study to go into detailed discussion on the nuances of the numerous definitions, acronyms and buzzwords that are associated with data mining and the related concepts. The key terms in view of this study are explained briefly below.

Data mining is defined by Hand et al. (2011, 1) as the "analysis of (often large) observational data sets to find unsuspected relationships and to summarise the data in novel ways that are both understandable and useful to the data owner". *Observational* here implies that the data under investigation have already been collected for some other purpose than the data mining analysis. Turban et al. (2011, 21) include the *process* dimension in the definition, introducing data mining as a "process of searching for unknown relationships or information in large databases or data warehouses, using intelligent tools such as neural computing, predictive analytics techniques, or advanced statistical methods".

Han and Kamber (2006, 2-4) see data mining as an outcome of the natural evolution of information technology (IT). They regard data collection and database creation, data management, and advanced data analysis, as the major phases in the evolution of the IT functionalities since the 1960s. The advanced data analysis phase, involving data warehousing and data mining, came about in the late 1980s. Data warehousing includes online analytical processing (OLAP) techniques that support multidimensional analysis and decision-making with summarisation, consolidation and aggregation features. Data mining has the potential to go

deeper in terms of insightfulness of the findings – the aim is to uncover important data patterns that help extract valuable knowledge embedded in vast amounts of data.

According to Sayad (2012), "data mining is about explaining the past and predicting the future by means of data analysis. Data mining is a multi-disciplinary field which combines statistics, machine learning, artificial intelligence and database technology." (See Figure 1.)

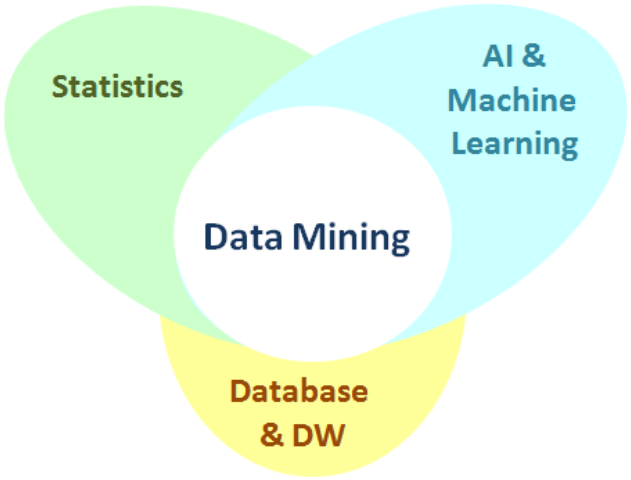

Figure 1: Disciplines combined in data mining (Sayad 2012).

Sayad's (2012) brief definitions of the disciplines help understand their interplay in data mining:

- *Statistics* – "the science of collecting, classifying, summarizing, organizing, analyzing, and interpreting data"
- *Artificial intelligence (AI)* – "the study of computer algorithms dealing with the simulation of intelligent behaviors in order to perform those activities that are normally thought to require intelligence"
- *Machine learning* – "the study of computer algorithms to learn in order to improve automatically through experience"
- *Database* – "the science and technology of collecting, storing and managing data so users can retrieve, add, update or remove such data"

- *Data warehousing* – "the science and technology of collecting, storing and managing data with advanced multidimensional reporting services in support of the decision making processes"

The EU Compliance Risk Management Guide for Tax Administrations (2010, 109) provides the following description for data mining: "An analytic process designed to explore data in search of consistent patterns and/or systematic relationships between variables, and then to validate the findings by applying the detected patterns to new subsets of data."

Data mining algorithm is "a well-defined procedure that takes data as input and produces output in the form of models or patterns" (Hand et al., 2001, 141). The procedure is made up of a finite set of rules. It must terminate after some finite number of steps and produce an output.

Model is "a high-level, global description of a data set. It takes a large sample perspective. It may be descriptive – summarising the data in a convenient and concise way – or it may be inferential, allowing one to make some statement about the population from which the data were drawn or about likely future data values" (Hand et al., 2001, 165). Another definition says that a model, in a data mining context, "refers to an algorithm as applied to a data set, complete with its settings (many of the algorithms have parameters that the user can adjust)". (Shmueli et al., 2010, 7)

Pattern is "a set of measurements on an observation", for example the height, weight and age of a person. Here, *observation* is the "unit of analysis on which the measurements are taken", such as a customer or a transaction. (Shmueli et al., 2010, 8)

Knowledge discovery in databases (KDD) is generally regarded as a synonym for data mining but there is also an interpretation that data mining constitutes only one step in a wider KDD process. According to Han and Kamber (2006, 5-7), the KDD process steps are: (1) data cleaning – removing noise and inconsistent data, (2) data integration – combining multiple data sources, (3) data selection – retrieving relevant data from the database, (4) data transformation – transforming or consolidating data into appropriate form for the knowledge discovery task at hand, *(5) data mining* – applying intelligent methods to extract data patterns, (6) pattern

evaluation – identifying truly interesting patterns that represent knowledge, and (7) knowledge presentation – presenting the discovered knowledge to the user.

It is obvious from the research problem formulation that this study discusses data mining in a wide, yet practical, sense. It is important to incorporate the process dimension to be able to meet the aims of the study. There is thus no need to distinguish between data mining and KDD here. Neither is there a need to formulate a precise definition or coverage for data mining. It suffices to say that this study looks broadly into the possibilities of the interplay of statistics, mathematics and information technology, in enabling the processing of vast amounts of data and subsequent insightful analysis. If one had to select one of the above referenced definitions, it would be safest to stick to the one provided by the EU Compliance Risk Management Guide.

Having brought some clarity to the central theme of the study above, let us next cover some nearby and partly overlapping concepts.

Business intelligence (BI) is an umbrella concept that covers architectures, applications, databases, analytical tools and methodologies used for decision support. Analyses of historical and current data, situations and performances give decision makers valuable insight and enable them to make more informed and better decisions. From a process point of view BI can be seen as a transformation where data are first refined to information, then to decisions, and finally to actions. The architecture of BI has four main components: (1) data warehouse, with its source data; (2) business analytics – a collection of tools for manipulating, mining and analysing the data in the data warehouse; (3) business performance management – methodology and applications for monitoring and analysing organizational performance; and (4) user interface, such as a dashboard. (Turban et al., 2011, 19-22).

Automated decision systems (ADS) are essentially ***rule-based systems*** that provide a solution, usually in one functional area, to a specific repetitive problem. ADS generally support the decision-making of frontline employees who must make quick decisions in structured problems on the basis of available customer information (Turban et al., 2011, 13-15).

Online transaction processing (OLTP) systems are operational database systems that typically cover a bulk of the day-to-day operations of an organisation, such as purchasing, inventory,

banking, payroll, registration and accounting. OLTP systems are transaction oriented, that is, they allow users to enter, view, modify or delete data regarding one transaction (event) at a time. OLTP systems are typically used by clerks or other frontline employees. The organisation's clients may also have limited access to see and edit some of the data concerning themselves. If query processing tasks need to be performed in the OLTP systems, these are typically done by IT professionals. (Han and Kamber 2006, 108-109)

Let us make a brief synthesis of the above three concepts (BI, ADS, OLTP) in the context of this study: BI and ADS have to do with decision-making, but on different levels. BI is an umbrella concept that has a managerial perspective. Data mining may or may not be incorporated in its business analytics suite. ADS, in turn, are in tax administration typically integrated in the operational systems, as will be seen in Chapter 3.2. In many tax matters, the taxpayers' transaction data go through an ADS whose rules determine how the case will be handled in the tax administration. The tax administration's operational systems that the clerks use in their daily work are generally OLTP type of systems, with ADS subsystems embedded where needed, as explained above.

2.2 Data mining as a process

In this study data mining is understood as a process, ideally integrated in a tax administration's day-to-day business. In this context, the challenge is to find insightful knowledge from large data sets and to make this knowledge actionable. There are some established ways to set out the data mining process. One such way was briefly mentioned in Chapter 2.1 in connection with the KDD explanation. Another well established process model is described below.

Cross-Industry Standard Process for Data Mining (CRISP-DM), illustrated in Figure 2, is the IBM Corporation's way to guide data mining efforts. It is a process model that provides an overview of the data mining life cycle. (IBM 2011, 1)

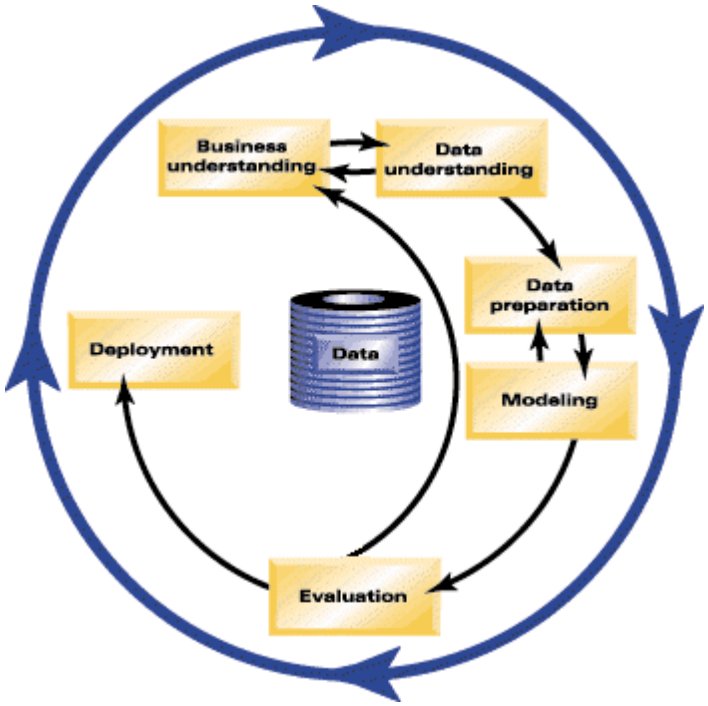

Figure 2: CRISP-DM process model (IBM 2011, 1).

As shown in Figure 2, the CRISP-DM consists of six phases setting out the process from start ("Business understanding") to finish ("Deployment"). As evident from Figure 2, the sequence of the phases is not strict but the projects can move back and forth between phases. Arrows indicate the most important and frequent dependencies between phases. (IBM 2011, 1)

1. ***Business understanding*** phase seeks to clarify the problems, goals and resources. This phase is about gathering relevant background information, stating and documenting in concrete terms the expected gains with data mining in the form of business success criteria, and producing a plan for the data mining project. (IBM 2011, 4-12)
2. ***Data understanding*** phase includes a closer look at the data available for mining and a consideration whether any supplemental data should be acquired. Issues such as the adequacy of data for the intended purpose, relevance of attributes, consistency of data formats across different sources, and the way of handling missing values, are considered. (IBM 2011, 13-18)
3. ***Data preparation*** is often the most time-consuming phase and usually takes 50-70 % of a project's time and effort. This is the phase where actual data sets for mining are formed, including the merging of all interconnected data and/or records, selecting sample subsets of data, deriving new attributes, sorting the data for modeling, cleaning data from errors and missing or inconsistent values, and splitting the data into training and test sets. (IBM 2011, 19-24)
4. ***Modeling*** is where the analysis tools are applied to the prepared data, and the results begin to shed light on the business problem posed in the business understanding phase. To start with, data miners typically apply several modeling techniques in parallel, using the default parameters, and then fine-tune the parameters or revert to the data preparation phase for manipulations required by the model(s) of choice. Modeling is thus usually conducted in multiple iterations. (IBM 2011, 25-31)
5. ***Evaluation*** is about reflecting the results obtained in the modeling phase against the business success criteria established at the beginning of the project. The idea is to ensure that the organisation can make use of the results. A key consideration here is the applicability of the conclusions or inferences drawn from the models and the data mining process to the business goals. (IBM 2011, 32-34)
6. ***Deployment*** conveys the new insights into improvements in the organisation's business processes. This can mean a formal integration such as the implementation of a data mining model producing scores that are then read into certain information systems. An important further consideration here is how to measure and monitor the validity and

accuracy of each model. Alternatively, deployment can mean using the insights gained from data mining for business planning and decision-making. (IBM 2011, 35-39)

2.3 Data mining tasks and functionalities

Despite being a relatively young discipline data mining applications abound in business and science. Data Mining applications are commonly used to accomplish enhanced operational efficiency, improved marketing campaigns, management of risk, detection of problems, identification of fraud, as well as support for research and development (Myatt & Johnson 2009, 165-166).

Data mining generally involves two kinds of tasks: descriptive and predictive. *Descriptive* tasks characterise the general properties of the data and often summarise the data in a convenient and concise form. *Predictive* tasks perform inference on the current data in order to make predictions about the likely future data values. (Han and Kamber 2006, 21)

Han and Kamber (2006, 21-27) describe six types of data mining functionalities and the corresponding typical patterns that can be mined:

1. ***Class description: characterisation and discrimination***. Class description addresses data classes that have been established by dividing the objects of the data set under study into classes on the basis of certain predetermined criteria. Data characterisation involves summarisation of the general features of a selected class (target class). Data discrimination involves comparison of the general features of target class objects with the general features of objects from one or more other classes.
2. ***Mining frequent patterns, associations and correlations***. Frequent patterns can be item sets or subsequences that occur frequently in data. A frequent item set is a set of items that appear frequently together, typically in a transactional data set. A frequent subsequence is a sequence of events that occur frequently in a certain order. Data mining can be used to find association rules that indicate frequent patterns. The findings are usually characterised with indicators such as *confidence*, that is, the likelihood of item/event Y appearing if item/event X appears, and *support*, that is, the share of cases where both X and Y appear of all cases under analysis. Typically, certain minimum confidence and support thresholds must be met to qualify the association rules as interesting. Additional analysis can reveal interesting statistical correlations between associated attribute-value pairs.

3. ***Classification and prediction***. Classification aims at distinguishing between data classes and thereby determining the likely classes of objects whose class labels are unknown. Data mining "learns" to assign a class label for objects whose class labels are unknown on the basis of an analysis of objects with known class labels. While *classification* deals with categorical variables, *prediction* is an analogous functionality dealing with continuous-valued variables whereby unknown numerical data values are predicted on the basis of an analysis of objects with known numerical data values.
4. ***Cluster analysis***. Cluster analysis seeks similarities and dissimilarities between the data objects under study and groups them into clusters based on an aggregate similarity-dissimilarity evaluation. Clusters are formed so that objects within a cluster have maximal similarity with one another but are maximally different from objects in other clusters. Clustering does not use any predetermined class labels (as these are generally not known), but classes and labels for them can be derived on the basis of the results of cluster analysis.
5. ***Outlier analysis***. Outliers are data objects that stand out from the mainstream data. They represent rare events. Outliers can be detected by using statistical tests that assume a distribution or probability model for the data, or by using distance measures from the main clusters of the data set under study.
6. ***Evolution analysis***. Data evolution analysis models regularities or trends of the behaviour of the objects under study over time. Such regularities may help predict future behaviour. Evolution analysis may include elements of the above described data mining functionalities, applied for time-related data, but there are also distinct features such as time-series data analysis, sequence or periodicity pattern matching, and similarity-based data analysis.

Sayad (2012) presents a somewhat different approach for classifying the data mining tasks. He first distinguishes between *exploration*, that is, explaining the past, and *modeling*, that is, predicting the future. Exploration is about applying statistical and visualisation techniques to describe the data and thereby bring important aspects into focus for further analysis. Modeling is the process by which a model is created to predict an outcome. Modeling can be predictive or descriptive, or it can discover association rules. If the outcome of *predictive modeling* is categorical, it is called *classification*, and if the outcome is numerical, it is called *regression*.

Descriptive modeling, or *clustering*, is the assignment of observations into clusters so that observations in the same cluster are highly similar to one another. Finally, *association rules*, which Sayad also classifies under modeling, can find interesting associations amongst observations. Classification, regression, clustering and association rules are also commonly referred to as the basic tasks of data mining.

Shmueli et al. (2010, 13-14) propose another slightly differing categorisation of data mining tasks, the "core ideas in data mining": *classification, prediction, association rules* (also known as *affinity analysis*), *data reduction, data exploration* and *data visualisation*. Classification, prediction and to some extent affinity analysis constitute the analytical methods employed in *predictive analytics*.

The above referenced sources do not make a clear distinction between data mining *tasks* and *functionalities*. On the next level down in the concept hierarchy come the *methods*. A given data mining task (functionality) can be performed by applying one or more data mining methods (also called data mining techniques). Examples of methods/techniques are decision tree, logistic regression, k-nearest neighbours, k-means, support vector machine and self-organising map. It goes beyond the scope of this study to dig into individual methods. Any given method, in turn, can be implemented with one or more alternative *algorithms*.

There is at least one more fundamental divider of data mining techniques. It is the distinction between *supervised* and *unsupervised* methods. In the supervised methods, the data mining algorithm first examines a sample data set where the values of the outcome variable of interest are known. The algorithm learns from this *training data* the relationship between predictor variables and the outcome variable. Once the learning has taken place, the algorithm is applied to another sample data set where the outcome is also known, but not revealed to the algorithm (validation data), to see its performance in predicting the outcome. Having validated the performance, the model can then be used to classify or predict the outcome of interest in new cases where the outcome is unknown. In case of unsupervised methods, there is no outcome variable to predict or classify. Supervised learning is used in classification and prediction whereas association rules and clustering techniques are unsupervised methods. (Shmueli et al. 2010, 15)

2.4 Data mining tools and related technology

In the same way as there is no consensus about the definition of the data mining concept, there is no universally acknowledged taxonomy regarding the product/market landscape. Different researchers and vendors have different approaches.

Product/market landscape

Mikut and Reischl (2011) discuss the development and categorisation of data mining tools. They describe a typical life cycle of how an individual new data mining method advances from a theoretical paper through a research prototype phase to an open source or commercial software package. The largest commercial success stories have resulted from step-wise integration of data mining methods into established commercial statistics tools, such as the products of SPSS and SAS companies, both founded in the 1970s. Acquisitions and subsequent renaming and integration of product lines have kept shaping the product/market landscape. SAS and SPSS, the latter currently belonging to IBM Corporation, are reportedly still the leading commercial advanced analytics vendors, with over 50 % combined market share in 2011 (IDC 2012, 14).

To facilitate the tool selection, Mikut and Reischl (2011) present a categorisation of data mining software into nine types, as listed below. The categorisation criteria are based on different user groups, data structures, data mining tasks and methods, visualisation and interaction styles, import and export options for data and models, platforms, and license policies.

1. ***Data mining suites*** focus broadly on data mining and include numerous methods. The application focus is wide and not restricted to any specific field. Coupling to business solutions, import and export of models, reporting, and a variety of platforms are supported.
2. ***Business intelligence packages*** have no special focus on data mining, but include basic mining functionality. They have a highly developed reporting functionality and database coupling. Implementation is via a client/server architecture.
3. ***Mathematical packages*** provide a large and extendable set of algorithms and visualisation routines.
4. ***Integration packages*** are extendable bundles of many different open-source algorithms.

5. ***Add-ons for other tools***, such as Excel, have limited but potentially useful functionality.
6. ***Data mining libraries*** implement data mining methods as a bundle of functions.
7. ***Specialties*** are otherwise similar to data mining suites but implement only one specific family of methods.
8. ***Research oriented implementations*** of new and innovative algorithms.
9. ***Solutions*** represent a group of tools that are customised to certain narrow application fields, such as text mining.

IDC (2012, 2-4) defines business analytics market as an aggregation of several software tools and application markets, consisting of three primary segments: (1) performance management and analytic applications; (2) business intelligence and analytic tools; and (3) data warehouse platform software. The second segment is obviously of highest relevance in view of this study. It has four subsegments: (a) query, reporting, and analysis tools; (b) advanced analytics tools; (c) spatial information analytics tools; and (d) content analysis tools. In IDC's business analytics taxonomy data mining is placed in the (b) subsegment together with statistics.

Popular brands

KDnuggets.com, data mining community's online resource site operating since 1997, arranges annual software polls about the usage of different tools in real projects. The 13th such poll was held in 2012, attracting 798 participants. For the first time, the number of users of open source software exceeded the number of users of commercial software. 28 % of voters used commercial software but not open source software, 30 % used open source software but not commercial, and 41 % used both. The most popular tools according to the poll were R*, Excel, and RapidMiner*, followed by Knime*, Weka*, Statistica, SAS, Matlab and IBM SPSS (open source tools marked with asterisk). (KDnuggets.com, 2012)

The results of another online survey from year 2011, conducted by Rexer Analytics, a US based analytics and CRM consulting firm, among 1319 data miners from over 60 countries, are generally in line with those of the above referenced KDnuggets poll: R* is the most popular tool, followed by SAS, IBM SPSS, Weka*, Statistica, RapidMiner*, Matlab and Knime* (open source tools marked with asterisk). (Rexer Analytics, 2012)

Data mining tools in relation to an organisation's overall information system architecture

Systematic business-driven data mining must be supported with proper tools and technical infrastructure. Technology must support the aggregation and processing of large data sets as well as the appropriate delivery and deployment of the mining results.

The day-to-day operations of an organisation typically involve the use of application-specific online transaction processing (OLTP) systems. They perform concurrent operational processing and store the corresponding data in the respective repositories. These repositories could in principle constitute a source of input data for mining, but direct data retrieval from them involves certain problems and risks: firstly, the coverage, quality and structure of the data in operational databases may not meet the requirements of mining tasks; secondly, the complex queries, often needed for mining, may substantially degrade the performance of operational tasks. (Han and Kamber 2006, 108-110).

A typical solution to the above problems is data warehousing. A data warehouse is a central repository where data from selected sources are regularly uploaded through an ETL procedure (ETL = extract, transform, load). The sources can be internal operational databases or external sources alike. The purpose of the ETL procedure is to ensure integrity, consistency and suitability of the uploaded data for its intended further use, such as analytics.

The need for separate data warehouses may however be diminishing. Vendors of operational databases are beginning to optimise their systems to analytics queries. As this trend continues, the separation between OLTP and analytical processing systems may decrease. (Han and Kamber 2006, 110).

2.5 Data mining in audit target selection

The Tax Auditing Unit of the Finnish Tax Administration participated in 2008-2011 in a research project called *Titan*, led by Professor Barbro Back from Åbo Akademi University. The aim of the project was to develop new information systems based on intelligent computational models to serve as decision aid for managers and stakeholders. The tax audit target selection process of the Finnish Tax Administration was one of the subjects of study under the auspices of the Titan project. One concrete project output here was Minna Tähtinen's licentiate thesis *Data Mining in Tax Auditing*. (Åbo Akademi University, 2012)

Tähtinen (2011) studies the potential of data mining to support the selection of companies for tax audit in today's ever growing complexity of business relationships. The way that tax auditors have identified the audit targets to date, relying on their past experience to pose queries to the database of financial reports, may not pick the best candidates for tax audit, due to the multitude of indicators for possible tax evasion. Another drawback by the current selection approach is that it focuses on only one company at a time while it would be worthwhile to view several companies simultaneously. (Tähtinen 2011, 2, 10, 14)

One objective in Tähtinen's thesis is to develop a generalisable model for identifying companies that merit a tax audit. Such companies form the *target group*. Data mining is first used to define the profiles of companies that have been chosen for audit with a reason, that is, where audits have yielded additional taxes. The features that differentiate these profiles from those of the companies with no need for audit are interesting. (Tähtinen 2011, 12)

Self-organising map (SOM) based clustering is applied as a data mining tool to find similarities in the audited companies of the target group. The SOM is a form of neural network frequently used in data mining tasks. The SOM algorithm projects multidimensional data onto a two-dimensional map and divides the observations into clusters. The SOM thus combines two data mining tasks: clustering and visualisation. The goal is to create a self-organising map where one cluster, called the *key cluster*, should ideally include the majority of the target group companies. (Tähtinen 2011, 12-13, 33)

Tähtinen uses taxation data of more than 5,000 partnership companies from year 2004 to build the model. In the data cleaning phase the data set was reduced to some 4,000 companies of which approximately 100 belonged to the target group. Three variants of the model were built and compared. At best, 93 % of the auditing result could have been collected from the companies placed in the key cluster, but on the other hand, all model variants were quite generous in placing also companies not in need of audit in that cluster. The model variants generally appeared to perform better in catching big evaders than distinguishing between the companies where audit is not needed and those where auditing result was low. All in all, the study concluded that the self-organising map could function as a tool to support the audit target selection, but the application area is very complex and further research is needed. (Tähtinen 2011, 40-53, 65-67)

2.6 Agency theory

"The relationship of agency is one of the oldest and commonest codified modes of social interaction. We will say that an agency relationship has arisen between two (or more) parties when one, designated as the agent, acts for, on behalf of, or as representative for the other, designated as the principal, in a particular domain of decision problems. Examples of agency are universal. Essentially all contractual arrangements, as between employer and employee or the state and the governed, for example, contain important elements of agency." (Ross, 1973, 134)

Agency theory is concerned with resolving problems that can occur in agency relationships. These problems are generally divided into agency problems and risk-sharing problems. Agency problems arise when the agent's and the principal's goals or interests conflict, and it is difficult or expensive for the principal to verify the agent's actual performance or behaviour. Risk-sharing problems stem from the fact that the principal and the agent may have different attitudes toward risk, resulting in different preferences for actions. (Eisenhardt 1989, 58)

One central assumption of agency theory is that there is an information asymmetry between the principal and the agent, to the advantage of the latter. Agents tend to exploit these information imbalances to maximise their own interests at the expense of their principals' interests. In response, principals try to bridge the information asymmetries by monitoring the agents. Principals may also offer incentives or impose sanctions in an effort to align the agents' interests with those of their own. All these efforts, coupled with the fact that the agent's performance may still not fully meet the principal's interests, constitute agency costs. In some occasions, the agent's concerns on his or her own reputation may create self-regulation on his or her part to comply with the principal's interests. (Shapiro, 2005, 264-265)

Eisenhardt (1989, 57-74) links the roots of the agency theory to information economics and discusses the central role of information systems in curbing agent opportunism. Here, information systems refer to budgeting systems, reporting procedures, boards of directors and other supervision arrangements. The agent is likely to behave in the interests of the principal after realising that he or she cannot deceive the principal. Agency theory generally regards information as a commodity, implying that principals can invest in information systems in order to control agent opportunism. Eisenhardt also pays attention to the duration of the agency

relationship, noting that in a long-term agency relationship the principal will learn about the agent and will thus be able to assess behaviour more readily.

The above discussion of agency theory is mainly related to its economics paradigm. Typical settings here include the relationships between owners and managers of a firm and between employers and employees. These relationships are in essence contracts where the incentives, monitoring mechanisms, and other forms of control undertaken to minimise the agency costs form the elements of the contract (Shapiro, 2005, 265-266). It is the problematic nature of the contract between principal and agent that constitutes the object of analysis in the agency theory in the economics paradigm (Eisenhardt 1989, 59). Typical problem domains include compensation, regulation, leadership and vertical integration (Eisenhardt 1989, 59).

Shapiro (2005, 266-267) broadens the discussion of agency theory from its classic economics paradigm to a wider context of social sciences, in particular to sociology. She notes that certain assumptions made in the economics paradigm are generally relaxed in the scholarship on agency outside of economics. First and foremost, the assumption that organisational structures and networks could be reduced to dyads of rational, self-interested and utility-maximising individuals, should be rejected. Shapiro emphasises the role of organisational aspects in and around agency relationships. Actors are not just principals or agents, but often both at the same time, even in the same transaction or hierarchical structure.

In her discussion of agency theory beyond the classic economics paradigm Shapiro (2005, 278-280) further notes that, instead of the solitary principal and agent setting, typically held in economics, there are often multiple principals and/or agents involved in an agency relationship. This may entail conflicting interests and/or different attitudes toward risk even in the same side of the relationship. This, in turn, may cause mixed messages, conflicting instructions or vague contract designs. Even when putting his or her own interests aside, the agent may face the problem of maneuvering between the principals' diverse and competing, possibly irreconcilable, interests. Another potential consequence of the multiple principals and agents setting is an altered information balance. Information leakages from among competing agents, for instance, may reduce the information asymmetry between the two sides of the agency relationship.

Shapiro (2005, 279) refers to misrepresentation and deception, as forms of lying, and to misappropriation, self-dealing and corruption, as forms of stealing, by those in position of trust (that is, agents) as core elements of white-collar crime. Understanding how the structural properties of agency relationships facilitate the misconduct and confound systems of social control is central to agency theory models regarding policing and sanctioning of agent opportunism. Shapiro (2005, 280) also notes that many monitoring arrangements constitute themselves agency relationships. The monitors act on behalf of some set of principals. They shirk, engage in corruption, or may simply monitor wrong things.

Shapiro (2005, 270-272) also discusses the relevance of agency theory in the context of the political system. Here, a complex network of relationships can be observed between actors such as the citizens, nation states, elected officials, lawmakers, members of the executive branch, administrative agencies, civil servants and courts. These actors concurrently play principal and agent roles within and across diverse organisations. Eisenhardt (1989, 58) notes that the agency structure is applicable in a variety of settings, ranging from macro level issues such as regulatory policy to micro level dyad phenomena such as lying and other expressions of self-interest.

Waterman and Meier (1998) have studied the relationship between the bureaucracy and its political environment. They base their analysis on agency theory but draw attention to certain distinctions between the classic economic and the *institutional* or *regulatory* principal-agent models. In the institutional setting three main categories of actors are distinguished: the general public constitutes the principal of the elected politicians who in turn control the bureaucrats. One distinction between the economic and institutional models is that, in the latter, politicians as principals may not monitor their agents as intensively as principals in the economic model do. This is explained by the assumption that politicians are unlikely to directly bear the cost incurred by the bureaucrats' misbehaviour as the bulk of that cost is passed along to the general public. Consequences may possibly follow, though, through elections, but only if the general public becomes aware of the politicians' lax oversight. (Waterman and Meier 1998, 175)

Waterman and Meier (1998) seek to apply the principal-agent model in the political control of bureaucracy. They question the assumptions of goal conflict and information asymmetry. Instead of the classic way of treating the goal conflict and information imbalance as constants, they

suggest a composition of three variables: (1) goal conflict or consensus; (2) the principal's information level, low or high; and (3) the agent's information level, low or high. The possible combinations across the above variables add up to eight (2 x 2 x 2 = 8) cases of which the classic agency theory represents the one with goal conflict, the principal having low information level, and the agent having high information level. Waterman and Meier (1998, 185) argue that in the bureaucratic setting, where the focus is typically on policy instead of profit, goal conflict may not always exist. In such circumstances the need for policing and monitoring should be reduced.

Castañer (2011) discusses the appropriateness of applying agency theory to public administration. According to him, the (voting) population is in democracies sovereign and thus the principal which elects individuals to represent it as well as to directly or indirectly lead each public administration. The population is the sovereign owner of the public assets, and the elected politicians are its agents. Politicians are elected on the basis of electoral programme which then in a way becomes a contract between the population and the successful candidates. Approval of the use of the administration's assets (budget) is subject to validation by the elected representatives of the population through parliamentary vote. Thus, people governing public administrations are also agents of the population. Castañer's discussion ends with open questions concerning the population's possibilities to deal with the agency problem, with the notion that politicians may also be opportunistic.

3 TAX ADMINISTRATION

Tax administration has a dual meaning in this study: First, as a general concept, it refers to the entire range of operations that a mandated government entity runs in order to implement and enforce the tax laws and regulations. Second, tax administration is also the government entity referred to above. For instance in Finland the official English translation of this entity is the *Finnish Tax Administration*. Tax administrations have varying mandates, structures and naming conventions across different countries.

A tax administration's core business is, generally speaking, to get the right tax at the right time from the right taxpayers, and to make the funds timely available for the right tax recipients (the state, municipalities, congregations, and others). Tax laws and regulations determine from whom, how much, and when, tax is due. The laws and regulations set forth certain registration, filing, reporting and payment obligations that the taxpayers must observe.

Tax laws and regulations are, however, not always simple and easy to comply with. On the other hand there are always citizens and organisations deliberately seeking ways to avoid or evade taxes. A tax administration is there to implement the tax laws and regulations. It should ideally make dealing with tax issues as easy as possible for everybody, help those who have difficulties, and ensure that all taxpayers pay their lawful share.

In developed economies, taxation is based on tax returns and other data submissions that the taxpayers must give regularly or in the occurrence of certain events to the tax administration. The tax administration's systems process the data and calculate or validate the taxes due. Tax administrations generally receive data also from certain third parties for comparison purposes and for prepopulating the return forms for certain taxpayer groups. For instance in Finland the natural persons' income tax returns come pre-filled with data from third parties such as employers, banks and labour unions. In corporate taxation the tendency is towards self-assessment where the taxpayer him-/her-/itself calculates the tax, and it is the tax administration's task to validate it, either as such or corrected, pursuant to the verification and control measures.

3.1 Business processes

Today's tax administrations tend to have their operations arranged in processes. While *taxation* is obviously the almost all-encompassing core business process, there are several other important processes around it to make taxation possible and, furthermore, to make it efficient and effective. The main business processes of the Finnish Tax Administration are shown in Figure 3, followed by brief descriptions (English translations of the process titles are unofficial). (Huhtanen 10.8.2012, interview)

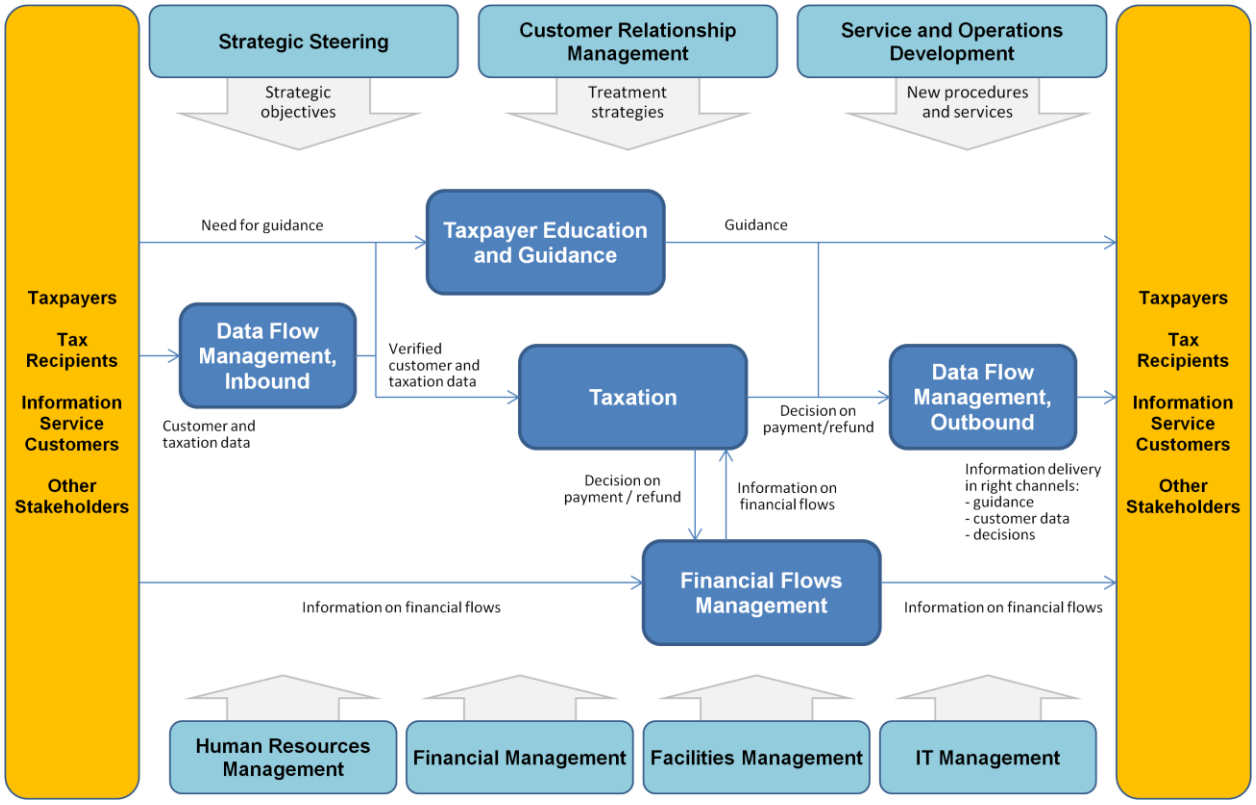

Figure 3: Business processes of the Finnish Tax Administration (Finnish Tax Administration 2012).

Brief descriptions of the main business processes: (Huhtanen 10.8.2012, interview)

- The ***Taxpayer Education and Guidance*** process determines how the tax administration seeks to increase the taxpayers' ability to deal with their taxation matters correctly. Typical tax administration's measures here include information dissemination in various

forms and channels, targeted educational campaigns, and personalised guidance in complicated situations. The principal aim is to prevent the taxpayers' mistakes before they occur.

- The ***Data Flow Management*** process is the tax administration's "logistics" process. It defines how inbound data, the tax administration's "raw material", are received from various sources and interfaces, and how these data are transmitted to operational systems. In the outbound end, the process defines how the outputs of the Taxation and the Taxpayer Education and Guidance processes are prepared for delivery in the appropriate channels to the recipients.
- The ***Taxation*** process covers the tax administration's procedures to attend a taxpayer from the moment a ground for taxation emerges until it discontinues. The process lays down the workflow in customer registration issues, in validating the taxable income and taxes due, in conducting tax audits, as well as in overseeing that taxes are duly paid and in pursuing collection measures for the indebted taxpayers.
- The ***Financial Flows Management*** process defines the procedures in distributing the tax revenue among the tax recipients (the state, municipalities, congregations, and others) and transmitting the respective funds to them.
- The ***Strategic Steering Process*** covers the environmental scanning, strategic planning, medium-term business planning, short-term operational planning, resources planning, operational target setting, as well as the follow-up of the strategy implementation and effectiveness.
- The ***Customer Relationship Management*** process determines how taxpayer behaviour is observed and analysed, how these analyses are used for taxpayer segmentation, and how segment-specific treatment strategies are designed. The treatment strategies are deployed in the Taxation and the Taxpayer Education and Guidance processes.
- The ***Services and Operations Development*** process outlines how the tax administration's services and operations are developed. The starting point here is typically a development initiative together with a corresponding needs assessment. The outputs can be, for instance, new internal procedures, new services for taxpayers, or new treatment measures to address certain tax risks.

3.2 Information systems

Tax administration is an information intensive domain that involves processing of vast amounts of data concerning a large number of taxpayers. Tax administrations have been in many countries among the forerunners of computerisation in the public sector. Owing to the early start and the tradition of building systems with relatively narrow scopes, typically one for each tax type or functional area, many tax administrations have today in place complex bundles of legacy systems. For instance, the Finnish Tax Administration currently runs dozens of operational systems, the oldest of which date from the 1980s. (Lehtinen 20.8.2012, interview)

Information systems in tax administration have traditionally been tailor-made solutions, designed according to country-specific requirements. The systems typically serve certain narrowly defined purposes, and are rigid to alterations. As a rule, frontline employees use the systems to view or modify data regarding one customer or one transaction at a time. The use cases are designed for transaction-specific processing, not for viewing or handling several customers' data simultaneously. The systems perform concurrent operational processing and store the data in application-specific repositories. (Lehtinen 20.8.2012, interview)

Examples of the systems currently in use in the Finnish Tax Administration are listed below. The list is not exhaustive as it is meant for illustration purposes. (Lehtinen 20.8.2012, interview)

- There is a dedicated system for handling the basic customer data such as the taxpayers' names, addresses, identity numbers and status with respect to different tax liabilities. This system features connections to the Finnish Population Information System and to the Finnish Trade Register, enabling automatic creation of most customers in the system when they are born, as well as certain automatic updates. Certain attributes, such as the tax liability status information, are determined on the basis of the customer's transactions with the tax administration.
- Corporate taxation uses a dedicated system to process the business income tax data of corporate entities such as limited liability companies. The customers typically submit their tax returns annually to the Tax Administration, which performs the income tax assessment using the system. Certain logical and technical controls, run in the system,

constitute an integral part of the assessment. There is a rule-based subsystem which picks up cases for manual scrutiny if the rules fire. In case no rules fire the assessment is finished automatically without human intervention.

- There is a dedicated system for processing the data related to the value added tax, employer contributions and some other charges reported on periodic tax returns. Business customers typically submit these returns monthly. In the same way as in business income taxation, there are rule-based controls that pick up cases for manual review if certain rules fire.
- There are separate systems for managing individuals' taxation, real estate tax, transfer tax, and inheritance and gift tax.
- There is a dedicated system for implementing and following up the payment of taxes.
- There is a dedicated system for managing the collection measures of the indebted customers who have failed to settle their taxes in time.

The fragmented system architecture poses challenges for generating an all-encompassing view of the tax matters of an individual customer, as well as for gathering aggregate data for business and risk analyses. The possibility to put together data across different tax types and functional areas would be particularly important for tax risk management and auditing purposes. Auditors would need a wide array of taxation related data, firstly, to be able to select the most relevant targets for audit, and secondly, while already conducting an audit, to be able to dig into details in the most alarming issues. (Lehtinen 20.8.2012, interview)

To tackle the challenge of data fragmentation across different systems, tax administrations in some countries, including Finland, have embarked on building data warehouses. Data from different internal source systems, possibly complemented with external sources, are extracted, transformed and loaded in a unified format into a central repository. In the Finnish Tax Administration, the Tax Auditing Unit initiated a data warehouse project in 2006. Today the auditors utilise the data warehouse in audit target selection by applying experience based rules across the wide array of data. (Lehtinen 20.8.2012, interview)

A relatively recent development is the appearance of commercial off-the-shelf (COTS) tax administration products in the market. The idea is to provide one integrated platform for

administering all (or most) tax types and functions of the tax administration, basically in the same way as Enterprise Resource Planning (ERP) solutions do in private sector corporations. Given the ever growing complexity and maintenance costs of the legacy systems, there is a temptation for tax administrations to look into the possibilities of the new generation COTS products. They feature built-in flexibility to accommodate country-specific requirements and changes when new laws and regulations are enacted. (Lehtinen 20.8.2012, interview)

3.3 Tax compliance management and tax risk management

Tax compliance management and tax risk management are relatively recently introduced intertwined approaches in operationalising a tax administration's strategy. They started gaining foothold since the late 1990s in the Organisation for Economic Co-operation and Development (OECD) and the European Union (EU) fora. The OECD published its Guidance Note on Compliance Risk Management in 2004 (hereinafter referred to as "OECD Guidance Note"), and the EU followed suit in 2006, publishing the Risk Management Guide for Tax Administrations (renamed as Compliance Risk Management Guide for Tax Administrations in the 2010 update, hereinafter referred to as "EU Guide").

Tax compliance management is essentially about optimising the use of resources allocated to a tax administration in order to maximise the overall level of compliance with the tax laws (OECD 2004, 6). *Compliance* here relates to the extent to which taxpayers meet the obligations placed on them by law. The OECD Guidance Note (2004, 7) identifies four broad categories of taxpayer obligations:

- ***Registration*** in the system
- Timely ***filing*** or lodgment of requisite taxation information
- ***Reporting*** of complete and accurate information (incorporating good record keeping)
- ***Payment*** of taxation obligations on time

A taxpayer's failure to meet any of the above obligations may be considered non-compliance. It may be due to unintentional error or intentional fraud. The compliance approach recognises that the taxpayers have diverse capabilities, behaviours, attitudes and motives, and that there is a need to adjust and target the tax administration's services and interventions accordingly. The OECD Guidance Note (2004, 10) encourages tax administrations "to give greater attention to understanding the factors that shape taxpayers' compliance behaviour so that a potentially more effective set of responses – ones that deal with the underlying non-compliant behaviour rather than focussing on treating the symptoms – can be crafted and implemented".

The OECD Guidance Note (2004, 41) presents, with reference to Dr Valerie Braithwaite's research, four motivational postures that characterise the way individuals relate to tax

administration and the tax system. These postures are based on sets of values, beliefs and attitudes adopted by the person. A so-called compliance pyramid, shown in Figure 4 below, is a widely used way in tax administrations to illustrate the four motivational postures, together with corresponding compliance strategies.

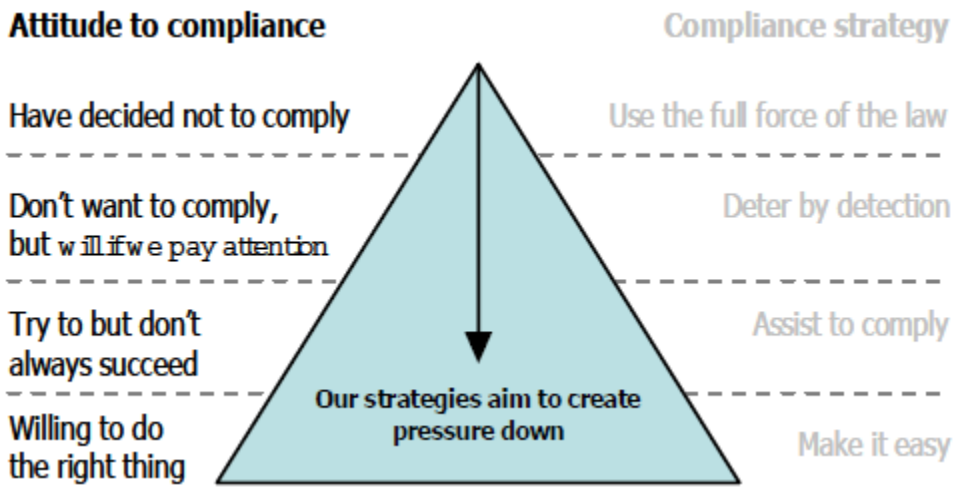

Figure 4: Compliance pyramid, spectrum of taxpayer attitudes to compliance (OECD 2004, 41)

With targeted activities, suited according to motivational postures, tax administrations can stimulate compliance and constrain the motivation to resist or evade compliance. It is important to note, however, that an individual taxpayer may adopt any of the attitudes in Figure 4 at different times, or adopt all of them simultaneously in relation to different issues. A tax administration's strategy should be designed so as to create an overall pressure for taxpayers to move down in the pyramid. (OECD 2004, 41-42)

Tax risk management, also known as ***compliance risk management***, looks at risks that affect compliance with registration, filing, reporting and payment requirements. A tax compliance risk involves a taxpayer's failure to properly register in the tax system, to properly file taxation information, to properly report tax liabilities, or to properly pay taxation obligations (OECD 2004, 8). The EU Guide (2010, 109) has a more general approach, defining a risk as "the threat or probability that an action or event will adversely affect an organisation's ability to achieve its

objectives". The Finnish Tax Administration leans on both the OECD and EU approaches and defines a tax risk as a "threat, attributable to taxpayer behaviour, that the Finnish Tax Administration's objectives are not met". Another localisation aspect in Finland's case lies in the categorisation of taxpayer obligations: the Finnish Tax Administration does not distinguish between *filing* and *reporting* obligations but combines these two into an *obligation to declare*. (Mäkelä 28.9.2012, interview)

The tax compliance approach sets the scene for tax risk management but the latter goes deeper in setting out a process for the identification, analysis and mitigation of tax risks. The EU Guide is in this respect more specific than the OECD Guidance Note. The EU Guide (2010, 5) defines the compliance risk management as "a systematic process in which a tax administration makes deliberate choices on which treatment instruments could be used to effectively stimulate compliance and prevent non-compliance, based on the knowledge of all taxpayers (behaviour) and related to the available capacity". Furthermore, the objective of applying compliance risk management is "to enable a tax administration to accomplish its mission(s) by facilitating management to make better decisions" (EU 2010, 7).

Despite slightly differing formulations of the process steps, the OECD Guidance Note and the EU Guide define the tax risk management process, by and large, in a uniform way. The steps, as formulated in the EU Guide, are shown in Figure 5 and will be explained thereafter. The EU Guide (2010, 10-11) presupposes that the compliance risk management way of thinking will in the long term change tax administrations in a profound way: the traditional repressive approach associated with fraud detection and sanctions is being replaced by a holistic and co-operative approach to ensure compliance with new forms of treatment. This paradigm shift involves, among other issues, influencing taxpayer behaviour, targeted proactive and preventive measures, and internal risk intelligence work in tax administrations to facilitate the integration of their planning and decision processes.

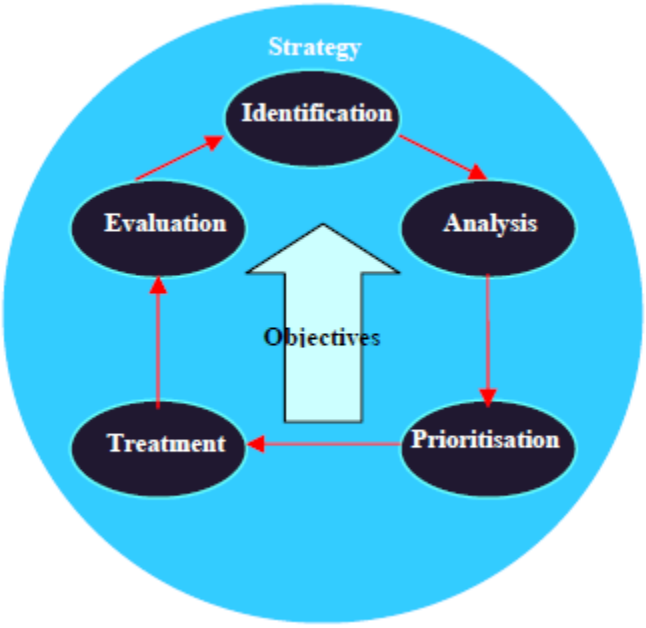

Figure 5: Tax risk management process (EU 2010, 9)

Identification. Identification of tax risks is the first step of the tax risk management process: risks that prevent a tax administration from achieving its objectives are identified at different levels. The EU Guide suggests starting on a general level and then drilling down to details. A high level composition of risky areas and groups/segments of taxpayers with different compliance levels provides direction for identification of risky activities. The output of the identification process step is a list of potential risks, pointing out areas that merit analysis. (EU 2010, 25-31)

Analysis. Selected identified tax risks are analysed to attain knowledge (intelligence) about the risks and the related taxpayer behaviour. As a result, the characteristics of the taxpayers involved in the respective tax risk will be discovered, the reasons behind taxpayer behaviour will be understood, the likelihood and frequency of the risk will be estimated, the consequences of the risk materialising will be projected, possible treatment options and their resource requirements will be established, and the potential impact on the tax administration's objectives will be assessed. (EU 2010, 32-35)

Prioritisation. The third process step is the prioritisation of risks to be treated, including the corresponding selection of taxpayers. The most important treatment strategies are risk reduction, that is, minimising the frequency and/or extent of the risk in a coming period, and risk covering, that is, neutralising the impact of an occurred risk. Risk reduction activities are generally focused on taxpayer groupings while risk covering activities are usually related to an individual taxpayer. Prioritisation is about allocating the tax administration's available resources optimally between the available treatment options. The output of prioritisation is a treatment plan setting out risks to be treated as well as the treatment choices. (EU 2010, 35-39)

Treatment. The EU Guide (2010, 40) defines risk treatment as "the process in which the negative impact of the risk on the administration's objectives is neutralised". The treatment plan is translated into activities that can include educating, helping, guiding, supporting, encouraging and influencing taxpayers to be compliant, and dealing with those that remain non-compliant. The treatment can result in full elimination of the risk, in a lower risk level following the undertaken risk reduction measures, or in corrections made to redress the taxpayers' non-compliant behaviours. (EU 2010, 40-49)

Evaluation. The fifth and final process step is evaluation. It can take place on different levels: on global level the focus is on establishing how the tax administration meets its long-term objectives. On activity level the effectiveness of different types of activities can be established so that future treatment plans can be designed with this knowledge in mind. The evaluation step closes the process cycle and contributes to learning and thereby working smarter in the future. (EU 2010, 50)

The OECD Guidance Note (2004, 34) mentions data mining as an example of how technology can be used to supplement human experience in detecting non-compliance. More specifically, the potential of data mining is recognised in distinguishing the characteristics of taxpayers that have been non-compliant, usually on the basis of past audit results, from those of the compliant ones. By analysing thousands of characteristics simultaneously, such patterns may be found in the data that they give rise to completely or partly new criteria for identifying non-compliance. Seeing from today's perspective, this is still a highly relevant application area of data mining in tax

administrations but represents only one area within a wide range of opportunities (Mäkelä 28.9.2012, interview).

The OECD Guidance Note (2004, 13-15) touches upon certain key internal capabilities of a tax administration that are important for compliance risk management to be effective: risk management methodology should be central to organisational reporting, governance and decision-making processes; risk management should be embedded in the conduct of the core business processes; cross-organisational mechanisms and forums should be utilised to ensure the integration of organisational strategies and the maximum use of intelligence; there should be an ability to receive multiple pieces of disparate information and to combine and interpret them to form intelligence; and a sufficient investment should be made in analytical and research competencies.

3.4 Tax gap

The identification and analysis steps of the tax risk management process may include an assessment of the tax gap, seeking to answer the question of 'what is missing', and how this gap is broken down between risk areas and/or taxpayer groupings.

The EU Compliance Risk Management Guide for Tax Administrations (2010, 109) defines the tax gap as "the potential tax yield minus the actual tax revenues". In other words it is the difference between the amount of taxes that should be collected, pursuant to tax laws and regulations, and the amount of taxes actually collected. In practical terms, tax gap represents the missing tax revenue due to different types of fraud and errors made by taxpayers. As one of the main objectives of a tax administration is to ensure that right amounts of taxes are paid, the tax gap is a fundamental measure of the financial extent of tax risks.

In 2007, the Swedish Tax Agency estimated that Sweden's annual tax gap would amount to approximately 133 billion Swedish crowns (equivalent to approximately 14 billion euro). This corresponds to roughly 5 % of the gross domestic product and 10 % of total tax revenues. The largest part of the tax gap, 66 billion Swedish crowns, is attributable to black work of which two thirds can be assigned to micro businesses. (EU 2010, 66)

A breakdown of the total tax gap into different types of non-compliance, economic sectors, and categories of taxpayers could help tax administrations see areas that should be addressed. Estimation of the tax gap in certain strategically interesting tax risk areas is the approach adopted by the Finnish Tax Administration but no aggregate tax gap estimate has been established there. It is however interesting to note that the tax control in connection with taxation brings some 1.7-2 billion euro in taxes collected every year, and taxes collected as a result of tax audits bring another 300 million euro (Finnish Tax Administration, 2012, 7). These figures can be compared with the net cost level of some 400 million of the Finnish Tax Administration (Finnish Tax Administration, 2012, 14). Even if the above figures do not necessarily speak anything about the total tax gap, it is reasonable to assume that, without the impact of tax control and audits, the tax gap could be some 2–2.3 billion euro bigger than whatever it is. (Mäkelä 28.9.2012, interview)

4 THEORETICAL FRAMEWORK

The theoretical framework seeks to adopt the agency theory to the setting of this study. A number of agency relationships can be found around tax administration, as illustrated in Figure 6.

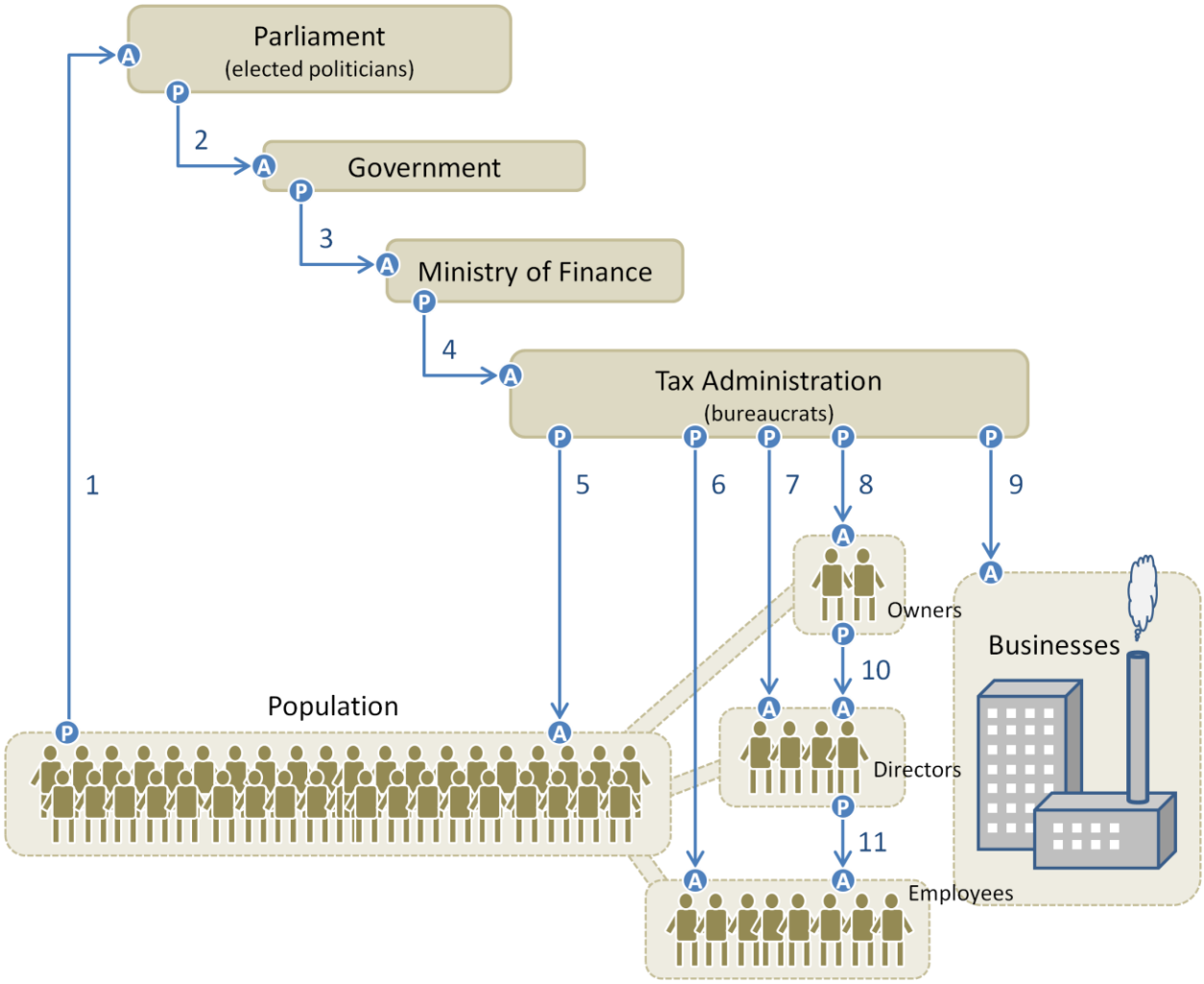

Figure 6: Agency relationships between tax administration and some of its stakeholders. "P" denotes to principal, "A" to agent.

In principle the highest, though rather abstract, level of agency relationship can be observed between the citizen and the state as a whole. In that relationship the parties' roles can be looked in different ways: On the one hand, in a welfare state a citizen can be seen as a principal, entitled

to certain services that the state must provide. On the other hand, the classic *state* and the *governed* setting implies reversed roles, acknowledging the right of the state as a principal to demand from its citizens certain services or contributions, such as paying taxes, to make the welfare happen.

A number of agency relationships can be identified when decomposing the relationship between citizen and state to reflect typical decision-making and governance structures in a democratic market economy. In the context of this study the primary interest lies on the relationships marked with numbers 5 to 9 in Figure 6, that is, those where the tax administration is the principal and requires that taxpayers, be it natural or juridical persons, fulfil certain registration, filing, reporting and payment obligations. Tax administrations are authorised to exercise certain power to ensure the taxpayers' compliance with these requirements, but on the other hand, they are generally also responsible for providing facilities, services and guidance to enable the taxpayers to meet their obligations.

When viewing the above established tax administration (principal) – taxpayer (agent) setting against the literature review of agency theory (Chapter 2.6), one cannot disregard the question of the nature of the contract between the parties. The contract here is obviously not based on the parties' free will or voluntary commitment in the same sense as the classic economics paradigm suggests. Nevertheless, both parties have economic interests at stake: the tax administration seeks to minimise the tax gap and the taxpayer usually seeks to minimise his/her/its tax burden. It must also be noted that literature has validated the applicability of agency theory even without the presence of economically motivated incentives.

After ascertaining above that the agency theory can indeed be applied to the tax administration–taxpayer setting, let us move on to leveraging the theory and establishing potential functions that data mining could have in helping tax administrations meet their goals.

Taxation in developed economies is based on the taxpayer's declaration or self-assessment of taxes which the tax administration processes, and if need be adjusts, to arrive at a taxable income and taxes due. As long as there is an information gap between the tax administration and the taxpayer, there is room for the taxpayers to misrepresent their income or deductions. As noted in the literature review, information systems can help reduce the information gap. The assumption

here is that data mining in particular provides novel and efficient ways to reduce that gap. Tax administrations can thereby improve their control measures and the hit rate of detecting fraud and errors. In addition to the direct benefit of more effective controls as such, there is also the potential benefit of less attempted fraud, as a result of the raising awareness of the tax administrations' better controls. As noted in the literature review, the agent (here, the taxpayer) is likely to behave in the interests of the principal (here, the tax administration) after realising that the principal cannot be deceived.

Another point raised in the literature review is the principal's chance of learning from the agent's past behaviour. Here, it seems obvious that data mining could support tax administrations in using historical data to establish behavioural patterns that predict the taxpayers' current or future behaviour. This could bring new insight into establishing smarter controls and/or building preventive measures that would nullify the taxpayer's potential misbehaviour before it even emerges. While the tax administrations' current ways of utilising historical data are generally limited to relatively simple comparison analyses in individual cases where error or fraud has already been otherwise detected, data mining could provide novel ways to boost the effectiveness of the control and taxation apparatus in a wider scale.

As also noted in the literature review, reducing the goal conflict between a principal and an agent generally implies better performance by the agents and thus a reduced need to monitor them. An obvious way to reduce the goal conflict would be to influence the taxpayers' attitudes, or broadly speaking, the tax moral. The more committed the people are to paying their taxes, the better the overall compliance would be, obviously. Attempts to influence the tax moral would presumably be more effective if the target groups' existing behaviours, needs and other relevant circumstances are taken into account. In the same way as marketing campaigns are tailored to different customer segments in the commercial arena, the assumption here is that appropriate taxpayer segmentation could help tax administrations conduct their influencing activities more effectively.

One interesting characteristic of the tax administration – taxpayer setting is the multitude of agents, and other networks of agency relationships among them (see Figure 6 above). While the tax administration is a common principal to a number of taxpayers, the same taxpayers form

networks of agency relationships among themselves in terms of ownership structures, business relations, employment and other arrangements. This creates a potential opportunity for the tax administration to reduce the information gap by putting together data from different interlinked taxpayers and verifying the consistency. The challenge for the tax administration is to harness the multiple agent settings and find the most knowledgeable and trustworthy agents.

5 EMPIRICAL PART

The empirical part draws on lessons learned in the international co-operation platforms of tax administrations as well as on the feasibility study of data mining that the author was in charge of in the Finnish Tax Administration in March-August 2012. Chapter 5.1 summarises the international experience, and Chapter 5.2 outlines the main phases and key outcomes of the feasibility study project.

5.1 Experiences from data mining in certain tax administrations

OECD Conference on the Use of Advanced Analytics in Tax Administrations

The Organisation for Economic Co-operation and Development (OECD) organised a conference on the use of advanced analytics in tax administrations in Dublin in November 2011, bringing together decision makers and analytics practitioners from 22 tax administrations. The event was intended for sharing insights on advanced analytics best practices and experience amongst tax administrations. The following bullet points summarise the author's key observations, relevant to this thesis, from the conference presentations and subsequent discussions with some of the delegates.

- ***Ms. Josephine Feehily***, Chairman of ***Revenue Irish Tax and Customs***, noted in her opening remarks of the conference that the event came about from realising in early 2001 the meaningfulness of sharing lessons learned about using advanced analytics on the OECD level. In further discussion it appeared that some administrations were using advanced analytics or investigating the possibilities presented by analytics. Many showed interest in exploring the use of analytics further. In Ireland's case a decisive trigger for embarking on the use of analytics was the reform targets set by the Government in 2008, including administrative budget cuts of 20 % and staff reduction of 10 %. This pushed Revenue to work smarter by refining its systems, processes and information sources. Revenue commenced using analytics, initially to improve audit and compliance interventions, but ultimately to be more effective at its business of collecting tax and operating Customs controls. The use of analytics has moved to an industrial scale in 2011 with the collaboration of Accenture.

- ***Dr Rohan Baxter***, Senior Data Miner from the ***Australian Tax Office (ATO)*** showcased the ATO analytics capability, which consists of predictive risk models, large-scale visualisations, use of third-party data and social network analysis. One application area was presented in more detail: using analytics to decide collection actions for ATO's indebted taxpayers more effectively. Scores reflecting a taxpayer's propensity and capacity to repay a debt are computed. The scores place the taxpayer in one of the four quadrants of a 2x2 matrix. Each quadrant comes with a suggested treatment strategy: *remind* those who have both high propensity and high capacity to pay; *watch* those who have high propensity but low capacity to pay; *write-off* the debts of those who have both low propensity and low capacity to pay; and *recover* the debts of those who have high capacity but low propensity to pay. Dr Baxter also outlined key future trends in the area of real-time analytics and the use of analytics to support decision-making throughout a complex process ("end-to-end analytics"). While the current way of using analytics generally *informs* the business process, the future vision is to have analytics *embedded* in its every step. This would enable tax administrations to tailor treatment to the individual and, on the basis of feedback loops, keep up continuous optimisation of the steps from case creation to case completion.
- ***Mr. Mads Krogh Nielsen***, Special Adviser and Project Leader from the ***Danish Ministry of Taxation (SKAT)***, showed in his presentation "Using data mining to collect taxes" how to create a scoring model for segmenting debtors in an automated collection system. This new system, planned to be finished in 2012, covers all types of arrears to the public sector. A data mining model, which will be an integral part of the system, shows the expected future payment behaviour of the debtor, and accordingly proposes an appropriate treatment track for him/her/it. SKAT expects to achieve striking reductions in the collection costs and higher service level due to the automatisation, and thereby also standardisation, of the process cycle. In post-conference discussions with Mr. Nielsen, it appeared that SKAT expects over 100 man years resource savings.
- ***Mr. Dean Silverman***, Senior Advisor to the Commissioner of the ***US Internal Revenue Service (IRS)*** told in his presentation "Making analytics pay, making analytics mainstream" that IRS pursues a more data-driven and analytical culture within its core compliance activities. Capacity is being built in the agency to continually analyse and

apply data to evolve compliance programmes. One recent initiative is the deployment of analytics to process the card payment data that a new third party reporting requirement bring. Data submitted by the merchants can now be analysed against the data reported by the merchant acquirers. This new requirement is particularly intended to reduce the lack of transparency on the part of small businesses which comprise the biggest part of the US tax gap. According to Mr. Silverman the keys to success with data analytics in tax administration are integration with the business, big-picture problem-solving orientation, ownership at the top, and long-term human resources leadership.

- ***Mr. Declan Rigney***, Senior Manager from ***Revenue Irish Tax and Customs***, covered in his presentation the key features of Revenue's recent deployment of a real-time risk framework into the processing of employee tax credit claims. Here, predictive analytics is embedded into the functionality of Revenue's Pay As You Earn (PAYE) system which processes income taxes of over two million people's salaries. In PAYE's online platform employers calculate and deduct taxes from their employees' salaries, as advised by Revenue. The employees can manage certain aspects of their tax affairs in PAYE, among other issues they can claim additional tax credits besides the initial basic credits granted at the start of the year. Upon Revenue's validation the employer gets instruction via PAYE to make a corresponding refund to the employee. Revenue's aspiration for higher customer service standards, in terms of accurate and timely processing of the claims, and prevention of fraudulent and erroneous refunds were among the triggers for predictive analytics here. During seven months, from January to July 2011, an analytical model predictive of fraudulent behaviour in tax credit claims was developed, and the corresponding complex rules were integrated into Revenue's core systems. As a result, 50 % increase in accuracy of stopped cases was achieved. The project paid for itself within eight months.
- ***Dr. Duncan Cleary***, Statistician from ***Revenue Irish Tax and Customs***, gave an overview of the added value of advanced analytics to the work of Revenue by making better use of its data and gaining improved understanding of the taxpayer population. Results from evidence based projects are used to better target services to customers and to improve compliance. At the same time costs can be reduced due to better resource allocation. Dr Cleary elaborated on a case study where data mining is used to assist better

target selection for tax audit. Building on banking analogy of predicting the likelihood of a case defaulting on a loan, credit scoring techniques are used here to predict the likelihood of a case yielding in the event of an audit. The prediction is based on the profiles of the cases and the profiles of audit cases that have yielded in the past. Evaluation cases have shown a hit rate of approximately 75 % and a significant correlation between the predicted and actual outcome of the audits.

The use of analytics in the Australian Taxation Office

Commissioner Michael D'Ascenzo from the Australian Taxation Office (ATO) presented a broad range of applications of analytics used in the ATO and the major results achieved by them in a speech for the Australian Institute of Company Directors in June 2012 (Australian Taxation Office, 2012). The ATO has pioneered a dedicated analytics capability to tackle the information challenge posed by 40 million taxpayer transactions and half a billion data items from third parties each year. Analytics improves ATO's ability to differentiate if and how they respond to a taxpayer transaction, based on what is known about the taxpayer. This information helps the ATO personalise their support and compliance strategies.

- ***Lodgment models*** predict, firstly, the propensity of a taxpayer to lodge tax return late within a certain time period, and secondly, the risk to revenue posed by non-lodgment of a tax return or an activity statement. These models help the ATO to focus their follow-up action on cases that are not likely to lodge, not even with a delay, and whose non-lodgment poses a big enough risk to tax revenue to justify the use of resources for follow-up action.
- ***Debt models*** enable the ATO to manage their collectible debt with fit for purpose responses. The risk scores for a taxpayer's propensity to repay a debt, and their capacity to pay, enable the ATO to determine an appropriate support or compliance strategy. Income tax return data are used to compute the risk scores. The models help the ATO identify a self-finaliser segment, where no action is needed, as well as the cases where the pursuit of the debt is uneconomical, and finally, the cases where tough measures are needed quickly.

- ***Models to detect high-risk refunds*** have helped the ATO to prevent refunds worth $665 million from being issued incorrectly in 2010-11. The ATO's refund models include a scalable social network discovery algorithm which detects networks among individuals, companies, partnerships or tax return forms. The models are updated and refined to improve detection and extend recognition of new and emerging frauds. The models also speed up the processing of legitimate cases as the ATO can reduce consideration of cases that do not require further review by up to 30 %.
- ***Clustering techniques*** contribute to effective decision-making by helping the ATO better understand large populations of taxpayers. Techniques such as self-organising maps are used to identify and visualise different clusters within the populations, representing taxpayers who behave in a sufficiently similar way. Data-driven methods have yielded more "natural" groupings of taxpayers, allowing the ATO to group a large population into a series of relevant clusters and better understand the tax issues and service needs of each cluster.

5.2 Feasibility study of data mining in the Finnish Tax Administration

The General Director of the Finnish Tax Administration set in March 2012 a feasibility study project to assess the potential benefits of data mining. The main objective of the project was to make a justified recommendation regarding a larger-scale adoption of advanced analytics. The author of this thesis, working as a Senior Analyst in the organisation, was assigned as a project manager in the said study.

The feasibility study project was instrumental in gathering invaluable inputs from leading experts and managers of the Finnish Tax Administration. The project enabled the author to look into the research problem of this thesis together with his colleagues who have extensive domain knowledge and experience. Some data mining tool vendors were also consulted.

This chapter outlines the main activities as well as the key observations and lessons learned during the project in view of the research problem of this thesis. Figure 7 below shows the main activities along the six-month time frame of the project, from March till August 2012.

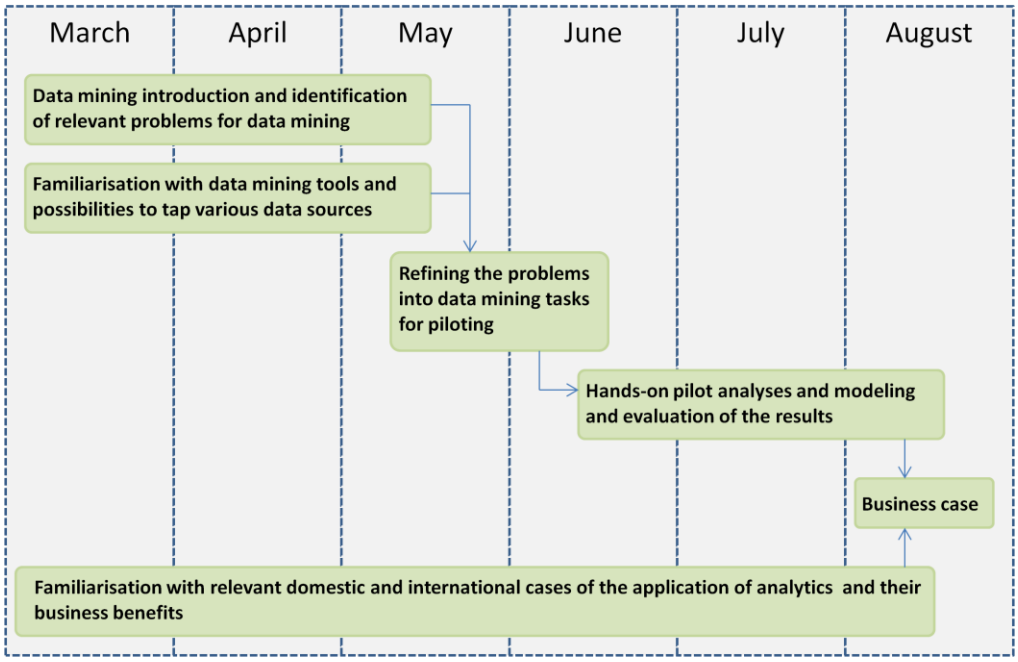

Figure 7: Timeline of the feasibility study of data mining in the Finnish Tax Administration in 2012.

Forming a business-minded project team

One important consideration at the start of the project was the composition of the project team. To ensure as comprehensive as possible a picture of the potential application areas and advantages of data mining, representatives from all main business units dealing with taxation were requested to join the team. Brief descriptions of these units, as formulated in the Finnish Tax Administration's official website, are given below. (Finnish Tax Administration, 2012)

- The ***Individual Taxation Unit*** guides and serves private customers as well as business owners and self-employed persons. It also manages customer information, income taxation and withholding, tax control in connection with taxation as well as inheritance, gift, asset transfer and real estate taxation.
- The ***Corporate Taxation Unit*** is responsible for providing guidance and services for limited companies and other corporate customers, customer information and tax control in connection with taxation.
- The ***Tax Collection Unit*** carries out tasks related to the payment, collection, recovery and remittance of taxes and the tax account procedure.
- The ***Tax Auditing Unit*** directs tax auditing activities as a part of tax control and performs tax audits, supervises internal EU trade and performs other tax control duties.

Apart from the domain expertise brought in by the business unit representatives, there were also analysts from the Tax Risk Management Unit in the project team. This is a horisontal support unit in charge of the development of tax risk management methods and tools. During the project it turned out that the cross-functional composition of the team contributed greatly to the project's achievements. Presumably, this early involvement of the business units will also help pave the way for proceeding with the possible wider adoption of analytics in the business processes.

Data mining introduction and identification of relevant business problems for data mining

The first half of the project period included two parallel activities: firstly, gaining sufficient understanding of data mining to be able to identify relevant business problems to look into with it, and secondly, familiarisation with data mining tools and the possibilities to connect various data sources with them.

The general introduction of data mining and the identification of business problems took place through a series of workshops. In the beginning two project team members (the author of this thesis included) shared their prior knowledge and earlier exposure to data mining with the other project team members. Here, an important source of inspiration was the November 2011 OECD Conference on the Use of Advanced Analytics in Tax Administrations. After introductory sessions the project team, divided into sub-groups, started to identify business problems to look into. Interesting problem areas were relatively easy to come by, as there were already a number of topics awaiting investigation in the business units' agendas. The following topics for investigation themes were formulated:

- Profiling of taxpayers who report exceptionally low personal income
- Profiling of taxpayers whose income tax declarations have been corrected in desk audits
- Analysing the audited firms' post-audit behaviour after discovering black labour in the audit
- Analysing the distribution of relative business income tax burden across the main segments of corporate customers
- Analysing the tax revenue impact of a proposed new law on limiting large corporations' entitlement to deduct interest expenses from taxable business income
- Developing the audit target selection criteria in a selected industry sector
- Analysing the reasons behind the growth of tax arrears from 2010 to 2011

Familiarisation with data mining tools and possibilities to tap various data sources

Familiarisation with data mining tools in the Finnish Tax Administration had started already in the second half of 2011 with a series of tool presentations. When designing the feasibility study project in early 2012, a decision was made that the project should aim a step further: the project team should get hands-on experience and be able to show in concrete terms some real results based on actual data. Moreover, one area of interest was to test the connectivity of the internal data repositories with the data mining tools.

Data mining tools for pilot use were selected on the basis of their popularity in European tax administrations. The vendors whose tools were selected for pilot use provided the project team with basic training in using the respective tools. Follow-on workshops were arranged to provide

further guidance and support after the project team members had already done some first trials with the tools. In the follow-on workshops the project team members could pose specific questions on the tasks they were working with. This kind of learning by doing approach worked relatively well, given the tight time frame of the project.

Experiments to connect certain internal data repositories to the data mining tools were conducted in co-operation with the database specialists of the Finnish Tax Administration. These experiments were successful in providing access to the intended sources of data for the purposes of the feasibility study project.

Refining the business problems into data mining tasks for piloting

Having established the business problems to look into, they were refined into concrete data mining tasks. This involved considering how data mining techniques could help address the problems, and specifying the source data that could best serve the respective analyses and modeling. An immediate further step was the preparation of the source data for mining.

Hands-on pilot analyses and modeling

Following the data mining task formulations, hands-on pilot analyses and modeling started in sub-groups of the project team. Progress was made in all intended tasks in summer 2012. The main purpose of the hands-on experiments was to understand in concrete terms the usage logic of data mining tools and to see in practise what kind of results such tools could generate.

Given the tight time frame of the project and the limited availability of the data mining tools during it, it was not expected that the results of the pilot data mining tasks would bring much new subject matter insight. It was acknowledged early on that the experiments within the scope of the project would probably at best prove or verify some previously held assumptions or beliefs. In general that turned out to be the case, though some promising-looking predictive models were also generated.

Building business case

One of the objectives of the feasibility study project was to build a business case weighing the potential benefits and costs of large-scale adoption of advanced analytics in the Finnish Tax

Administration. As the results of the pilot data mining tasks were not intended for operational deployment, these exercises could yield only hypothetical input for the business case.

A big part of the potential benefits for the business case were drawn on international experience: better focused preventive work and services through customer segmentation, improved hit rate of compliance actions, and resource savings through business process automation have been demonstrated by the experiences of Australia, Denmark and Ireland, for instance.

Familiarisation with relevant domestic and international cases of the application of analytics

During the whole project period the project team gathered information on relevant cases that could contribute to building the business case. In Finland, visits were paid to organisations such as a financing company that provides its clients with loans and venture capital investments, and a nationwide mail services provider. In the latter, analytics is used, among other issues, to predict the mail volumes in different branch offices of the organisation to help adjust the allocation of human resources accordingly (Viherä 14.6.2012, interview). Internationally, visits were paid to the Danish Ministry of Taxation (SKAT) and to the Norwegian Tax Administration (Skatteetaten). And as noted earlier, an important source was the November 2011 OECD Conference on the Use of Advanced Analytics in Tax Administrations where the business benefits of analytics were manifested in a number of presentations.

Conclusions of the feasibility study

Towards the end of August 2012, the feasibility study project team compiled their findings and drew conclusions. By and large the project team considered the experiences gained during the project encouraging for a possible larger scale adoption of advanced analytics in the Finnish Tax Administration.

6 FINDINGS

This section seeks to respond to the research problem and to meet the aims of the study by making a synthesis of the observations.

Research problem: How can data mining help tax administrations attain better tax compliance among the taxpayers?

There is a pile of evidence indicating that advanced analytics, in particular data mining, can play a significant role in helping tax administrations enhance tax compliance. As Michael D'Ascenzo, Commissioner of Taxation from the Australian Taxation Office, has put it:

> "The power of analytics is fast becoming central to supporting and protecting taxpayers, tailoring service delivery and operating an efficient tax and superannuation administration." (Australian Taxation Office, 2012)

The above quote essentially says it all. Firstly, analytics *supports* taxpayers. This support can be realised in the form of tailored services and guidance which become possible as the use of analytics enables tax administrations to group large populations of taxpayers into relevant segments and better understand each segment, their service needs and tax issues. Services and guidance designed to address the segment-specific circumstances enhance tax compliance as the taxpayers are better able, and less burdened, to comply with tax laws and regulations.

Secondly, analytics *protects* taxpayers. This protection is realised mainly through safeguarding the equal treatment of taxpayers. Analytics is used to identify, analyse and predict areas of non-compliance, in other words events where taxpayers perform registration, filing, reporting or payment incorrectly or neglect their obligations. Better prevention and detection of such events ensures that as many as possible pay their lawful share, and safeguards the aggregate public revenue. Tax administrations thus champion the majority of taxpayers who exhibit good citizenship. Another aspect of protecting taxpayers is saving the public funds, in effect the taxpayers' money, by using analytics to increase cost efficiency in tax administration through better informed decisions in allocating resources and steering the workflows in the core business processes.

Research issue 1: Drafting a general operational framework of a tax administration, and identifying where in it data mining would have the biggest potential to bring added value

The core business processes of a tax administration serve two high-level aims: firstly, to *enable* as high as possible a level of tax compliance on the taxpayers' own initiative, and secondly, to *ensure* as high as possible, ideally 100 % level of compliance. This big picture is by and large similar across all modern tax administrations, but the scopes and limits of their mandates vary country by country. Figure 8 below shows one way, developed by the author in the course of the research work, to outline the general operational framework of a modern tax administration.

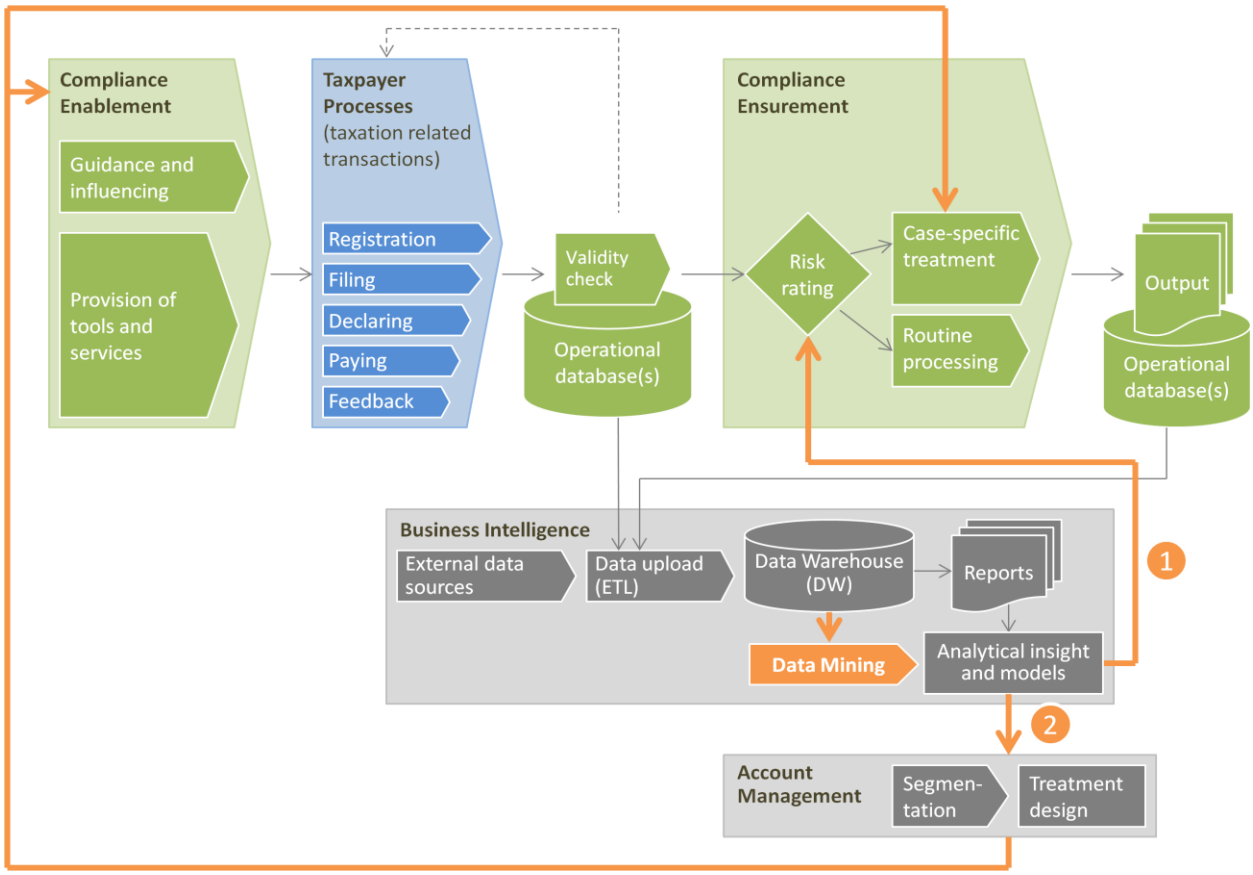

Figure 8: General operational framework of a tax administration. *Green shapes represent core business processes. The blue area covers the taxpayers' transactions. Grey shapes cover support processes that provide steering to the workflows of the core business processes.*

The framework shown in Figure 8 has proven generic enough to serve as a reference and basis for exchanging views and experiences among tax officials across different countries.

Compliance enablement refers to activities that a tax administration performs in order to create and maintain an enabling environment for taxpayers to comply with applicable tax laws and regulations. The measures here include, for example, the provision of information, instructions and guidance for taxpayers in diverse forms and channels; the provision of clear, potentially prepopulated, tax forms; and easy-to-use online services. The focus is on preventive measures that eliminate as many as possible events of non-compliance before they occur. One aspect here is also influencing the taxpayers and thereby encouraging them to comply on their own initiative. Tax administrations are increasingly interested in behavioural sciences in order to foster favourable attitudes to paying taxes.

Taxpayer processes include the taxation related transactions that a taxpayer must do to fulfil the applicable registration, filing, reporting and payment requirements. A taxpayer can naturally also give feedback to the tax administration. The transactions involve, as a rule, submission of certain data to the tax administration. Once the data reach the tax administration's operational system, certain validity checks are usually applied to identify submissions with obvious inconsistencies, and in these cases a new submission may be prompted. Transactions with consistent, technically and formally valid data are stored in the operational database and proceed in the tax administration's workflow.

Compliance ensurement refers to the measures that a tax administration takes to ensure that taxpayers indeed duly fulfil their registration, filing, reporting and payment obligations. The big volume of the transactions does not allow the tax administration to review all of them manually. Therefore, selection must be made to determine which cases merit a case-specific treatment, and which cases can move on to automated routine processing. Case-specific treatment refers here to any manual investigation and/or compliance intervention by the tax administration, such as a phone inquiry, a written request for additional information, or a tax audit. Regardless of the route of a transaction within the Compliance ensurement stage, sooner or later it results in certain output, such as a register entry, validated tax assessment, or an adjustment of taxable income.

Business intelligence (BI) represents the "brains" of the framework. Certain data from the operational transaction flow are regularly uploaded, possibly to a data warehouse, and made available for reports and analyses, together with data from certain external sources. The analyses generate business insight that helps the tax administration work smarter, as outlined below in the descriptions of Arrows 1 and 2.

Account management is an immediate beneficiary of a bulk of the business insight provided by the BI. Cluster analysis can yield meaningful taxpayer segments whose needs or behaviour can be effectively addressed by tailored treatment strategies.

This study argues that the two numbered orange arrows in Figure 8 indicate where data mining's most prominent potential to bring added value to a tax administration's work lies.

Arrow 1: Using data mining to determine optimal, risk-based transaction processing

The risk rating diamond in Figure 8 incorporates a selection mechanism that divides the flow of incoming taxpayer transactions, as well as omissions of mandatory transactions, into two broad types of processing tracks: automated routine processing and case-specific treatment. The latter can also be partly automated but it involves at least some kind of manual review or human intervention. On what basis the selection mechanism works is an important question in view of the whole tax administration's performance. This question can be framed as the *selection problem* for case-specific treatment.

The way of addressing the selection problem in tax administrations has evolved over time. The dominating solution today are rule-based systems where certain business rules, programmed into the system, trigger case-specific treatment if the conditions set by the rules are met. This would be a workable method provided that the validity of the rules can be proven. In reality, however, the rules tend to be formed on heuristic basis, and many of them tend to detect superficial logical or technical inconsistencies rather than behavioural patterns or other more profound determinants of non-compliance. Furthermore, any single rule typically covers only a narrow area, checking if some specific issue, possibly only a formality, is in order. The multitude of such narrowly defined rules may at worst undermine the effectiveness of tax control with large quantities of

insignificant or meaningless cases being picked up for investigation. This would tie up the tax administration's resources and prevent the officials from seeing the forest for the trees.

Data mining is arguably a solution to many of the shortcomings of the traditional way of setting the rules in a rule-based system:

- With data mining, and a proper technical infrastructure around it, cross-cutting analyses combining data across diverse internal and external sources can be made. This augments the knowledge base in comparison with the traditional narrowly focused rules and thus contributes to making fewer but better justified and more effective selections.
- Data mining can improve the hit rate of the selection mechanism with models trained on the basis of past successful selections. Data mining techniques have the ability to learn the typical characteristics and/or patterns that are common to cases where significant errors or fraud have been discovered in the past. This insight augments the knowledge base used in the selection and is thus likely to improve the hit rate.
- Data mining provides an unbiased approach to identifying the most significant determinants of error, fraud or other incident of interest. A wide range of independent variables can be used as a starting point in modeling. Data mining tools provide functionalities to compare the relative significance levels of numerous variables in a given task.
- When solving a business problem with data mining, business rules can be generated as a by-product of the modeling process. There is no need for separate encoding of the rules. Many data mining tools can write the rules in a standard computing language such as the PMML (Predictive Model Markup Language). If the operational system reads the same language, the rules can be flexibly embedded and kept up-to-date in it.

The literature review and the observations of the empirical part endorse the above reasoning and give grounds to believe that data mining can improve the hit rate of the selection for case-specific treatment. This means that a higher percentage of the transactions flagged for case-specific treatment turn out justifiably flagged in reality. Besides the ability to flag the cases with highest likelihood of being successful targets for compliance actions, data mining models can also prioritise among them, and predict the monetary yields of the proposed compliance actions.

By putting together the likelihood scores of being a successful target, the predicted monetary yields, and the tax administration's resource constraints, the tax administration can optimise its selection strategy to focus on the cases which are likely to result in the best overall return.

The above discussion addressed selection for case-specific treatment in tax control where non-compliance generally occurs in filing or reporting. The potential of data mining in improving the processing track selection however goes beyond tax control. The Danish and the Australian cases, introduced in the empirical part, illustrate the power of data mining in determining treatment tracks for cases of payment non-compliance, in other words for the taxpayers who have payment arrears.

The fourth possible area of non-compliance is registration. No reference of using data mining in that area could be identified when compiling this study. The logic says however that registration would be an equally potential area where non-compliance could be tackled with the help of data mining. In on-the-job discussions with the domain experts in the Finnish Tax Administration it has turned out that registration would indeed be a worthwhile area to look into, not least because its current workflows are heavily resource consuming.

Arrow 2: Using data mining for taxpayer segmentation and better targeted attention

A large customer base with diverse abilities, behaviours and attitudes among the taxpayers poses a challenge for tax administrations. Different customers need different types of services and attention. This challenge can be framed as the *diversity problem* of the taxpayer population.

To address the diversity problem some tax administrations have embarked on taxpayer segmentation. The idea is to identify taxpayer groups whose members possess sufficiently similar characteristics, behaviours and/or needs that imply sufficiently significant common tax risks or service needs so that they merit a tailored response by the tax administration. The applications of analytics in the Australian Taxation Office (ATO), introduced in the empirical part, include clustering techniques that help the ATO to better understand large populations of taxpayers. Data-driven methods yield "natural" groupings of taxpayers, allowing the ATO to group a large population into a series of relevant clusters and better understand the tax issues and service needs of each cluster.

As Arrow 2 in Figure 8 above indicates, a typical starting point of using analytical insight in Account management is to determine the relevant segments and to design appropriate treatment strategies for each segment. Both the ATO case and the Finnish study on audit target selection (Chapter 2.5) lean on the power of self-organising maps in identifying and visualising different clusters within the populations, representing taxpayers who behave in a similar way.

The orange follow-on part of Arrow 2 in Figure 8 goes on from Account management to Compliance enablement and Compliance ensurement. The message here is that the segments are not formed just for the sake of segmentation but for drawing on the shared needs and behaviour within the segment to design relevant response in terms of services and compliance actions.

The compliance actions can aim at preventing non-compliance from occuring in the first place, through measures such as targeted educational campaigns or outbound calls giving targeted guidance. On the other hand, the traditional compliance ensurement measures such as tax audits are not likely to extinct in the foreseeable future. In some cases tax audits may be triggered by a single taxation transaction. Such cases belong to the transaction-driven processing track selection described under Arrow 1. As a rule, however, taxpayers for tax audits are selected on the basis of more comprehensive analyses of their compliance behaviour where transaction level data are aggregated onto the customer level and then possibly reflected against the norms of the segment(s) that the taxpayer belongs to. Arrow 2 is thus the dominating route for tax audit target selection.

Recent discussions in the Finnish Tax Administration tend to speak in favour of strengthening the segmentation approach and segment-driven compliance actions as the way of tackling tax risks. It seems however inevitable that there will always be tax risks whose occurrence cannot be connected with any specific predetermined segment(s). The tax administrations will therefore need to maintain and run in parallel the transaction-based treatment track selection and the segmentation-based targeting of compliance actions.

Research issue 2: Identifying, in general terms, the required technology for a large-scale adoption of data mining in tax administration

The literature review suggests that the technology solution of large-scale adoption of data mining would consist of three components: (1) data sources, most notably the operational systems with their databases; (2) data warehouse, a central data repository to which certain data from the operational systems, and possibly from certain external sources, are regularly uploaded; and (3) the data mining tools for conducting the actual analytical and modeling tasks.

Expert views gathered in connection with the feasibility study of data mining in the Finnish Tax Administration mainly endorse the above described high-level system architecture. There are, however, some arguments in favour of higher reliance on operational systems, or more specifically, the centralised copy database of the main operational systems. The rationale for these arguments is the fact that the copy database has always fresher and more comprehensive data than the data warehouse. The main hindrance for using the copy database is its complex structure as it has not been designed for analytics query purposes.

The feasibility study project ended up with a compromise recommendation: the data warehouse could serve as a main internal source for analytics, with its content being gradually extended to cover the most frequent query needs, while the copy database could be available as a "reserve option" for cases where needs emerge for so fresh or specific data that they are not covered by the data warehouse. In connection with the possible future introduction of a new generation integrated tax administration system the setting may change. Such a new integrated system may feature a (copy) database structure that serves the analytics queries in a more straightforward way.

As far as the data mining tools are concerned, it was a straightforward conclusion in the feasibility study that a client/server solution should be acquired for a large-scale adoption of data mining in an organisation like the Finnish Tax Administration. Such a solution is the only feasible way to do the "heavy-duty modeling" among a dozen analysts and keep the work organised.

Research issue 3: Gathering available empirical evidence of data mining applications in tax administrations

The OECD Conference on the use of advanced analytics in tax administrations in November 2011 was a compelling manifestation for the power of data mining in addressing the tax compliance challenge. Even if representatives of only less than half of the participating 22 tax administrations in the Conference made a formal presentation, representatives of many of the remaining administrations appeared to be either running their early experiments or planning to do so in the near future. All 22 countries showed keen interest and active participation in group discussions.

The group discussions held in the Conference also touched upon the organisational aspect of arranging the day-to-day conduct of analytics in tax administrations. It was acknowledged that at least as importantly as analytics is about technology, it is about people and co-operation among them. It was evident from the group discussions that there are probably as many ways to organise the analytics function as there are tax administrations. One size does not fit all, nor does any one structure. What tended to be a dominating high-level arrangement, though, was the concentration of analytical expertise into some kind of centralised unit.

Evidence reported in Chapters 5.1 and 2.6 discuss the use of data mining in addressing the following business problems:

- Determining how to respond to omissions of tax returns (Australia)
- Detecting high-risk tax refund claims (Australia)
- Determining appropriate collection actions for payment arrears (Australia and Denmark)
- Identifying meaningful taxpayer segments for tailored services and/or compliance action (Australia)
- Using card payment data, reported by third parties, to improve transparency on the part of small businesses (USA)
- Developing audit target selection to improve the hit rate of the audits and to identify high-yield cases (Ireland and Finland)
- Processing the salary earners' tax credit claims to speed up the service process and to detect erroneous claims (Ireland's PAYE system)

While there is already evidence demonstrating the power of analytics in certain specific applications within tax administrations, there is an increasing acknowledgement of the need to embed analytics in a comprehensive way in the day-to-day operations (rather than developing the deployment capacity of individual models). In particular the US and Australian delegates of the OECD Conference urged tax administrations to re-engineer their business processes so that analytics would genuinely become "business as usual".

Research issue 4: Shedding light on the potential implications of applying data mining in tax administration in view of the agency theory

The literature review introduced the agency theory, and the theoretical framework put it together with data mining in the context of tax administration. Furthermore, the theoretical framework gave rise to a series of logically derived assumptions regarding the effects of data mining on the agency relationship between a tax administration and a taxpayer. The following bullet points provide a synthesis of the most noteworthy assumptions and the corresponding justifications, with references to concrete supporting evidence, where available:

- Data mining can provide novel and effective ways for tax administrations to develop their information systems and thereby reduce the information asymmetry vis-à-vis taxpayers. This is mainly due to being able to combine data from various sources and consolidate plentiful variables into easy-to-interpret, actionable classifications and predictions. Being better informed, tax administrations can improve their control measures and the hit rate of detecting fraud and errors. Ireland's initiative of developing audit target selection, for instance, provides support to this assumption (see Chapter 5.1).
- Data mining can support tax administrations in using historical data to establish behavioural patterns that predict taxpayers' current or future behaviour. This can bring new insight into developing smarter controls and/or building preventive measures to nullify a taxpayer's potential misbehaviour before it emerges or leads to harmful consequences. Ireland's both initiatives discussed in Chapter 5.1 – developing audit target selection and reviewing the salary earners' tax credit claims in the PAYE system – provide support to this assumption.

- In addition to the direct benefit of the smarter controls in actually *catching* the fraudsters, there is also the potential indirect benefit of less *attempted* fraud, owing to raising public awareness of the tax administration's better controls. Taxpayers are likely to align their behaviour in line with the applicable tax laws and regulations after realising that the tax administration cannot be deceived. The USA's initiative to use card payment data to improve small businesses' transparency is the case in point here (see Chapter 5.1).
- Data mining, with its clustering techniques, can help tax administrations segment their heterogeneous taxpayer bases into sufficiently uniform subpopulations whose needs, behaviours and attitudes can be discovered, understood and responded to. This enables tax administrations to pursue improved customer intimacy in attending specific segments with tailored communication and other measures. One important aspect here is to mitigate the goal conflict vis-à-vis taxpayers by promoting good citizenship and the high tax morale associated with it. The better the tax administration knows its audience the better chances it has to influence it. Australia's and Denmark's initiatives to determine the collection actions with the use of analytics, as well as Australia's notion of developing the overall service orientation give support to this assumption (see Chapter 5.1).
- Data mining, most notably with its social network analysis techniques, can help tax administrations get hold of the ever growing complexity of business relationships, ownership structures and other types of linkages and dependencies among and across taxpayers. This enables tax administrations to identify multiple agent settings with the possibility to cross-check or verify certain critical information across several sources. At the time of writing this thesis there was no concrete initiative of this application area at hand, but talks in the international fora, such as the OECD Conference (see Chapter 5.1), indicated that developments are progressing towards this direction in at least the Australian, Irish and Swedish tax administrations.

To sum up, the evidence of data mining applications in tax administrations that was available for this thesis, gives grounds to believe that the above assumptions would hold. Data mining thus appears to fit well in the overall framework that the earlier agency theory findings suggest for resolving problems in agency relationships.

7 CONCLUSIONS

This study has looked into the possibilities of using data mining, or more broadly analytics, in tax administrations to enhance tax compliance. Enhancing tax compliance is about making tax issues easy to deal with for the taxpayers, helping those who have difficulties, and ensuring that the taxpayers fulfil the registration, filing and reporting requirements, and pay their lawful share. Given that taxpayer populations are generally large and heterogeneous, it is important that tax administrations arrange their business processes and workflows smartly and target their limited resources wisely to address the compliance challenge.

As evident from the title of the thesis, data mining has been chosen as the flagship form of analytics in this study. Definitions for data mining abound. Here it suffices to say that data mining is a process of applying appropriate statistical and mathematical methods to large data sets, with the help of information technology, to derive business insight and actionable predictive models. Data mining is typically applied to data sets that have piled up or been captured for some other purpose than analysis or modeling. Tax administrations' data repositories are in this respect highly suitable for mining as they have been collected primarily for administrative purposes, generally not with analysis or modeling in mind.

Tax administrations are generally well computerised. Information technology however tends to serve mainly the automation of routine tasks rather than the generation of business insight. Rule-based systems are widely used to tackle certain repetitive problems, perhaps most notably to select cases for audit or control in connection with taxation. The rules in these systems tend to be formed by heuristic means, possibly relying on certain incidental experiences or established beliefs. The meaningfulness of such selection criteria can be questionable from an analytical point of view, and certainly the human brain is bound to meet its limitations when searching for interrelationships across tens or hundreds of variables among thousands of cases.

Not only the sheer volume of data but also the generally growing complexity of business relationships, ownership structures and other types of linkages and dependencies among and across taxpayers suggest that the capacity of human brain is becoming inadequate in seeing the forest for the trees.

Data mining is arguably a solution to many of the shortcomings of the traditional rule-based approach in addressing the *selection problem* of cases for tax audit and control. Data mining algorithms can "learn" from the existing known outcomes of previously processed cases and the circumstances associated with them, and apply this knowledge to predict taxpayer behaviour in future cases under sufficiently similar circumstances. Data mining helps thus tax administrations select the optimum targets, typically those with the highest likelihood of yielding additional taxes, for intervention. Besides tax audit and control, the same "learning" logic can be applied to detecting errors and fraud in taxpayer registration and payment related transactions. The findings of this thesis give grounds to believe that data mining could significantly contribute to making the tax administration's selection mechanism for compliance interventions more effective. Evidence suggests that data analysis is becoming one of the tax administrations' main tools in combating error and fraud.

Diversity problem of the taxpayer population is the second high-level business problem, or challenge, that this study proposes to tackle with data mining. Data mining's value added here lies, first and foremost, in helping create and maintain an enabling environment for compliance where taxpayers can handle their tax issues with minimum effort and where their specific support needs can be effectively addressed. Data mining techniques can help segment the large and heterogeneous taxpayer base into manageable subpopulations whose service needs, behaviours and attitudes are easier to discover, understand and respond to. There is already some evidence demonstrating the power of analytics in identifying taxpayer groups among whom compliance enhancement has taken place with tailored attention by the tax administration.

The study has also looked into the potential implications of data mining in tax administration in view of the *agency theory*. The relationship between tax administration and taxpayer constitutes an agency relationship where the former governs the latter. Non-compliance can be considered an agency problem that occurs in this relationship. Data mining fits well in the overall problem-solving framework that the agency theory suggests for resolving agency problems. Firstly, by providing the tax administration with better means to draw on multiple interconnected, yet often scattered, pieces of data, data mining can reduce the information asymmetry vis-à-vis the taxpayer. Secondly, data mining's predictive models help the tax administration anticipate where fraud or error is likely to occur. Better hit rates of compliance interventions are likely to result in

less attempted fraud in the long term. Thirdly, being better able to follow the taxpayer behaviour over time, using insight generated with data mining, the tax administration has better chances to influence the taxpayer and foster a high tax morale. Fourthly, data mining, with its social network analysis, may help the tax administration identify and understand multiple agent settings so that they can get and cross-check data from and across all relevant parties.

As evident from above, data mining has an obvious potential to become one of the tax administrations' main tools in defending public revenue. The main application areas for data mining in tax administration, as identified in this study, are (1) making a truly risk-based selection mechanism for determining cases for compliance interventions; and (2) segmenting the heterogeneous taxpayer base into manageable, sufficiently uniform subpopulations for building effective segment-specific compliance enhancing agendas. Efficiency improvements in tax administrations, with the associated cost savings, are likely to result as a by-product since both above application areas involve increased automatisation and better optimised use of expensive human resources.

The author's employment in the Finnish Tax Administration has facilitated the preparation of this thesis by, for instance, enabling spontaneous communication with leading domain experts in Finland and certain other countries. Nevertheless it must be acknowledged that the evidence gathered in this study is fragmented and by no means provides an exhaustive record of data mining applications in tax administration. A great deal of other application areas are likely to exist, including such that the respective tax administrations are not willing to make public.

While relying on support from the agency theory, this thesis has a pragmatic approach as it seeks to give concrete guidance for tax administrations on what to expect from data mining and how to embark on its adoption. The focus here is on high-level fundamental issues while the individual application areas are intentionally left on rather superficial level. The applications areas, however, constitute interesting topics for further research. After grasping the big picture from an umbrella study like this, an obvious next step would be to see in more quantitative terms the potential benefits of data mining across a range of application areas.

REFERENCES

Books and reports

Han, J. & Kamber, M. (2006), Data Mining: Concepts and Techniques. Second Edition. Morgan Kaufmann Publishers, San Francisco.

Hand D. J., Mannila H. & Smyth P. (2001), Principles of Data Mining, Massachusetts Institute of Technology, Cambridge, Massachusetts, 578 p.

Myatt G. J. & Johnson W. P. (2009), "Making Sense of Data II – a Practical Guide to Data Visualization, Advanced Data Mining Methods, and Applications", John Wiley & Sons, Inc., Hoboken, New Jersey, 291 p.

Shmueli G., Patel N. R. & Bruce P. C. (2010), Data Mining for Business Intelligence, John Wiley & Sons, Inc., Hoboken, New Jersey, 404 p.

Turban E., Sharda R. & Delen D. (2011), Decision Support and Business Intelligence Systems, Ninth Edition, Pearson, Boston, 696 p.

Tähtinen M. (2011), Data Mining in Tax Auditing, Licentiate Thesis in Information Systems, Department of Information Technologies, Åbo Akademi University, Turku, 72 p.

Articles

Eisenhardt, K. M. (1989) "Agency Theory: An Assessment and Review", *Academy of Management Review*, Vol. 14, No. 1, pp. 57-74

Mikut, R. & Reischl M. (2011) "Data mining tools", *WIREs Data Mining and Knowledge Discovery*, Vol. 1, No. 5, pp. 431-443

Ross, S. A. (1973) "The Economic Theory of Agency: The Principal's Problem", *Decision Making Under Uncertainty*, Vol. 63, No. 2, pp. 134-139

Shapiro, S. P. (2005) "Agency Theory", *Annual Review of Sociology*, Vol. 31, pp. 263-284

Waterman, R. W. & Meier, K. J. (1998) "Principal-Agent Models: An Expansion?", *Journal of Public Administration Research and Theory*, Vol. 8, No. 2, pp. 173-202

Interviews

Huhtanen Tiina-Liisa, Quality Manager, Finnish Tax Administration, Helsinki

Lehtinen Heli, Director, Finnish Tax Administration, Helsinki

Mäkelä Ari, Risk Management Director, Finnish Tax Administration, Helsinki, 28.9.2012

Nielsen Mads Krogh, Special Adviser and Project Leader, Danish Ministry of Taxation (SKAT), Copenhagen 1.2.2012

Viherä Jussi, Development Manager, Itella Oyj, Helsinki, 14.6.2012.

Internet-references

Australian Taxation Office (2012), "The effective use of analytics in public administration: The Australian Taxation Office Experience. Speech by the Commissioner of Taxation Michael D'Ascenzo for the Australian Institute of Company Directors, Luncheon - Hobart, 22 June 2012". Online. Available at: http://www.ato.gov.au/corporate/content.aspx?doc=/content/00325633.htm&pc=001/001/001/002/001&mnu=0&mfp=&st=&cy=

Castañer, X. (2011), Applying agency theory to public administration (government). Online. Available: http://people.hec.unil.ch/xcastaner/2011/06/08/applying-agency-theory-to-public-administration-government/, [3.3.2012]

European Commission (2010), Compliance Risk Management Guide for Tax Administrations http://ec.europa.eu/taxation_customs/resources/documents/common/publications/info_docs/taxation/risk_managt_guide_en.pdf

Finnish Tax Administration (2012), Presentation of the Finnish Tax Administration. Online. Available at: http://www.vero.fi/en-US/Tax_Administration/Presentation_of_the_Finnish_Tax_Administ(15847)

IBM Corporation (2011), IBM SPSS Modeler CRISP-DM Guide. Online. Available at: ftp://ftp.software.ibm.com/software/analytics/spss/documentation/modeler/14.2/en/CRISP_DM.pdf, [10.6.2012]

IDC (2012), IDC Market Analysis: Worldwide Business Analytics Software 2012–2016, Forecast and 2011 Vendor Shares. Excerpt available online at: http://idcdocserv.com/235494e_sas, [29.9.2012]

KDnuggets.com (2012), KDnuggets Software Poll, Online. Available at: http://www.kdnuggets.com/polls/2012/analytics-data-mining-big-data-software.html, [29.9.2012]

Organisation for Economic Co-operation and Development (2011), Conference on the Use of Advanced Analytics in Tax Administrations. Dublin, Ireland, 29-30 November 2011. Conference materials available at: http://www.oecd.org/ctp/taxadministration/ftaanalyticsconference.htm, [31.7.2012]

Organisation for Economic Co-operation and Development (2004), Guidence Note, Compliance Risk Management: Managing and Improving Tax Compliance http://www.oecd.org/dataoecd/44/19/33818656.pdf, [31.7.2012]

Rexer Analytics (2012), 2011 Data Miner Survey. Online. Available at: http://www.rexeranalytics.com/Data-Miner-Survey-Results-2011.html, [29.9.2012]

Sayad, S. (2012), Data Mining. Online. Available at: http://www.saedsayad.com/data_mining.htm, [8.8.2012]

Åbo Akademi University (2012), Titan. Online. Available at: http://research.it.abo.fi/research/data-mining-and-knowledge-management-laboratory/projects/titan, [18.8.2012]