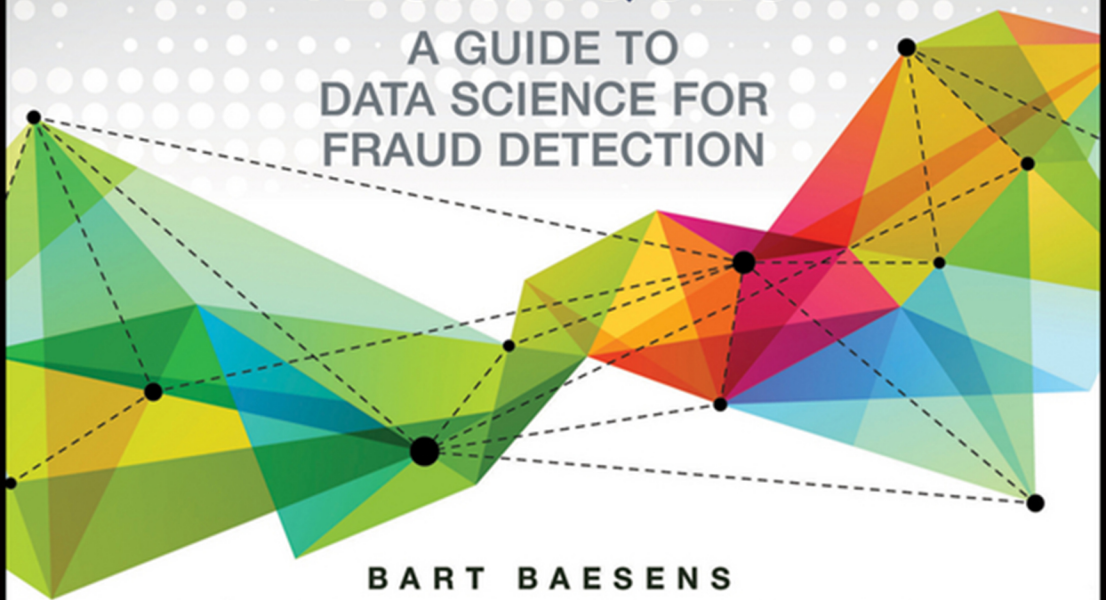
FRAUD ANALYTICS

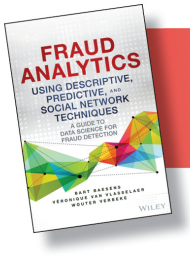
USING DESCRIPTIVE, PREDICTIVE, AND SOCIAL NETWORK TECHNIQUES

A GUIDE TO DATA SCIENCE FOR FRAUD DETECTION



BART BAESENS
VÉRONIQUE VAN VLASSELAER
WOUTER VERBEKE

WILEY



From *Fraud Analytics Using Descriptive, Predictive, and Social Network Techniques.*
Full book available for purchase here.

Contents

List of Figures xv

Foreword xxiii

Preface xxv

Acknowledgments xxix

Chapter 1 Fraud: Detection, Prevention, and Analytics! 1
- Introduction 2
- Fraud! 2
- Fraud Detection and Prevention 10
- Big Data for Fraud Detection 15
- Data-Driven Fraud Detection 17
- Fraud-Detection Techniques 19
- Fraud Cycle 22
- The Fraud Analytics Process Model 26
- Fraud Data Scientists 30
  - A Fraud Data Scientist Should Have Solid Quantitative Skills 30
  - A Fraud Data Scientist Should Be a Good Programmer 31
  - A Fraud Data Scientist Should Excel in Communication and Visualization Skills 31
  - A Fraud Data Scientist Should Have a Solid Business Understanding 32
  - A Fraud Data Scientist Should Be Creative 32
- A Scientific Perspective on Fraud 33
- References 35

Chapter 2 Data Collection, Sampling, and Preprocessing 37
- Introduction 38
- Types of Data Sources 38
- Merging Data Sources 43
- Sampling 45
- Types of Data Elements 46

Visual Data Exploration and Exploratory Statistical Analysis 47
Benford's Law 48
Descriptive Statistics 51
Missing Values 52
Outlier Detection and Treatment 53
Red Flags 57
Standardizing Data 59
Categorization 60
Weights of Evidence Coding 63
Variable Selection 65
Principal Components Analysis 68
RIDITs 72
PRIDIT Analysis 73
Segmentation 74
References 75

Chapter 3 Descriptive Analytics for Fraud Detection 77
Introduction 78
Graphical Outlier Detection Procedures 79
Statistical Outlier Detection Procedures 83
Break-Point Analysis 84
Peer-Group Analysis 85
Association Rule Analysis 87
Clustering 89
Introduction 89
Distance Metrics 90
Hierarchical Clustering 94
Example of Hierarchical Clustering Procedures 97
k-Means Clustering 104
Self-Organizing Maps 109
Clustering with Constraints 111
Evaluating and Interpreting Clustering Solutions 114
One-Class SVMs 117
References 118

Chapter 4 Predictive Analytics for Fraud Detection 121
Introduction 122
Target Definition 123
Linear Regression 125
Logistic Regression 127
Basic Concepts 127
Logistic Regression Properties 129
Building a Logistic Regression Scorecard 131

Variable Selection for Linear and Logistic Regression 133
Decision Trees 136
Basic Concepts 136
Splitting Decision 137
Stopping Decision 140
Decision Tree Properties 141
Regression Trees 142
Using Decision Trees in Fraud Analytics 143
Neural Networks 144
Basic Concepts 144
Weight Learning 147
Opening the Neural Network Black Box 150
Support Vector Machines 155
Linear Programming 155
The Linear Separable Case 156
The Linear Nonseparable Case 159
The Nonlinear SVM Classifier 160
SVMs for Regression 161
Opening the SVM Black Box 163
Ensemble Methods 164
Bagging 164
Boosting 165
Random Forests 166
Evaluating Ensemble Methods 167
Multiclass Classification Techniques 168
Multiclass Logistic Regression 168
Multiclass Decision Trees 170
Multiclass Neural Networks 170
Multiclass Support Vector Machines 171
Evaluating Predictive Models 172
Splitting Up the Data Set 172
Performance Measures for Classification Models 176
Performance Measures for Regression Models 185
Other Performance Measures for Predictive Analytical Models 188
Developing Predictive Models for Skewed Data Sets 189
Varying the Sample Window 190
Undersampling and Oversampling 190
Synthetic Minority Oversampling Technique (SMOTE) 192
Likelihood Approach 194
Adjusting Posterior Probabilities 197
Cost-sensitive Learning 198
Fraud Performance Benchmarks 200
References 201

Chapter 5 Social Network Analysis for Fraud Detection 207
Networks: Form, Components, Characteristics, and Their Applications 209
Social Networks 211
Network Components 214
Network Representation 219
Is Fraud a Social Phenomenon? An Introduction to Homophily 222
Impact of the Neighborhood: Metrics 227
Neighborhood Metrics 228
Centrality Metrics 238
Collective Inference Algorithms 246
Featurization: Summary Overview 254
Community Mining: Finding Groups of Fraudsters 254
Extending the Graph: Toward a Bipartite Representation 266
Multipartite Graphs 269
Case Study: Gotcha! 270
References 277

Chapter 6 Fraud Analytics: Post-Processing 279
Introduction 280
The Analytical Fraud Model Life Cycle 280
Model Representation 281
Traffic Light Indicator Approach 282
Decision Tables 283
Selecting the Sample to Investigate 286
Fraud Alert and Case Management 290
Visual Analytics 296
Backtesting Analytical Fraud Models 302
Introduction 302
Backtesting Data Stability 302
Backtesting Model Stability 305
Backtesting Model Calibration 308
Model Design and Documentation 311
References 312

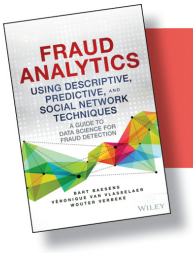
Chapter 7 Fraud Analytics: A Broader Perspective 313
Introduction 314
Data Quality 314
Data-Quality Issues 314
Data-Quality Programs and Management 315
Privacy 317
The RACI Matrix 318
Accessing Internal Data 319

Label-Based Access Control (LBAC) 324
Accessing External Data 325
Capital Calculation for Fraud Loss 326
Expected and Unexpected Losses 327
Aggregate Loss Distribution 329
Capital Calculation for Fraud Loss Using Monte Carlo Simulation 331
An Economic Perspective on Fraud Analytics 334
Total Cost of Ownership 334
Return on Investment 335
In Versus Outsourcing 337
Modeling Extensions 338
Forecasting 338
Text Analytics 340
The Internet of Things 342
Corporate Fraud Governance 344
References 346

About the Authors 347

Index 349

From *Fraud Analytics Using Descriptive, Predictive, and Social Network Techniques: A Guide to Data Science for Fraud Detection*, by Bart Baesens, Véronique Van Vlasselaer, and Wouter Verbeke. Copyright © 2015, SAS Institute Inc., Cary, North Carolina, USA. ALL RIGHTS RESERVED.



From *Fraud Analytics Using Descriptive, Predictive, and Social Network Techniques*.
Full book available for purchase here.

CHAPTER 1

Fraud: Detection, Prevention, and Analytics!

INTRODUCTION

In this first chapter, we set the scene for what's ahead by introducing fraud analytics using descriptive, predictive, and social network techniques. We start off by defining and characterizing fraud and discuss different types of fraud. Next, fraud detection and prevention is discussed as a means to address and limit the amount and overall impact of fraud. Big data and analytics provide powerful tools that may improve an organization's fraud detection system. We discuss in detail how and why these tools complement traditional expert-based fraud-detection approaches. Subsequently, the fraud analytics process model is introduced, providing a high-level overview of the steps that are followed in developing and implementing a data-driven fraud-detection system. The chapter concludes by discussing the characteristics and skills of a good fraud data scientist, followed by a scientific perspective on the topic.

FRAUD!

Since a thorough discussion or investigation requires clear and precise definitions of the subject of interest, this first section starts by defining fraud and by highlighting a number of essential characteristics. Subsequently, an explanatory conceptual model will be introduced that provides deeper insight in the underlying drivers of fraudsters, the individuals committing fraud. Insight in the field of application—or in other words, expert knowledge—is crucial for analytics to be successfully applied in any setting, and matters eventually as much as technical skill. Expert knowledge or insight in the problem at hand helps an analyst in gathering and processing the right information in the right manner, and to customize data allowing analytical techniques to perform as well as possible in detecting fraud.

The *Oxford Dictionary* defines fraud as follows:

> **Wrongful or criminal deception intended to result in financial or personal gain.**

On the one hand, this definition captures the essence of fraud and covers the many different forms and types of fraud that will be

discussed in this book. On the other hand, it does not very precisely describe the nature and characteristics of fraud, and as such, does not provide much direction for discussing the requirements of a fraud detection system. A more useful definition will be provided below.

Fraud is definitely not a recent phenomenon unique to modern society, nor is it even unique to mankind. Animal species also engage in what could be called *fraudulent activities*, although maybe we should classify the behavior as displayed by, for instance, chameleons, stick insects, apes, and others rather as *manipulative behavior* instead of fraudulent activities, since *wrongful* or *criminal* are human categories or concepts that do not straightforwardly apply to animals. Indeed, whether activities are *wrongful* or *criminal* depends on the applicable rules or legislation, which defines explicitly and formally these categories that are required in order to be able to classify behavior as being fraudulent.

A more thorough and detailed characterization of the multifaceted phenomenon of fraud is provided by Van Vlasselaer et al. (2015):

> **Fraud is an uncommon, well-considered, imperceptibly concealed, time-evolving and often carefully organized crime which appears in many types of forms.**

This definition highlights five characteristics that are associated with particular challenges related to developing a fraud-detection system, which is the main topic of this book. The first emphasized characteristic and associated challenge concerns the fact that fraud is ***uncommon***. Independent of the exact setting or application, only a minority of the involved population of cases typically concerns fraud, of which furthermore only a limited number will be known to concern fraud. This makes it difficult to both detect fraud, since the fraudulent cases are covered by the nonfraudulent ones, as well as to learn from historical cases to build a powerful fraud-detection system since only few examples are available.

In fact, fraudsters exactly try to blend in and not to behave different from others in order not to get noticed and to remain covered by nonfraudsters. This effectively makes fraud ***imperceptibly concealed***, since fraudsters do succeed in hiding by ***well considering*** and planning how

to precisely commit fraud. Their behavior is definitely not impulsive and unplanned, since if it were, detection would be far easier.

They also adapt and refine their methods, which they need to do in order to remain undetected. Fraud-detection systems improve and learn by example. Therefore, the techniques and tricks fraudsters adopt ***evolve in time*** along with, or better ahead of fraud-detection mechanisms. This cat-and-mouse play between fraudsters and fraud fighters may seem to be an endless game, yet there is no alternative solution so far. By adopting and developing advanced fraud-detection and prevention mechanisms, organizations do manage to reduce losses due to fraud because fraudsters, like other criminals, tend to look for the easy way and will look for other, easier opportunities. Therefore, fighting fraud by building advanced and powerful detection systems is definitely not a pointless effort, but admittedly, it is very likely an effort without end.

Fraud is often as well a ***carefully organized*** crime, meaning that fraudsters often do not operate independently, have allies, and may induce copycats. Moreover, several fraud types such as money laundering and carousel fraud involve complex structures that are set up in order to commit fraud in an organized manner. This makes fraud not to be an isolated event, and as such in order to detect fraud the context (e.g., the social network of fraudsters) should be taken into account. Research shows that fraudulent companies indeed are more connected to other fraudulent companies than to nonfraudulent companies, as shown in a company tax-evasion case study by Van Vlasselaer et al. (2015). Social network analytics for fraud detection, as discussed in Chapter 5, appears to be a powerful tool for unmasking fraud by making clever use of contextual information describing the network or environment of an entity.

A final element in the description of fraud provided by Van Vlasselaer et al. indicates the ***many different types of forms*** in which fraud occurs. This both refers to the wide set of techniques and approaches used by fraudsters as well as to the many different settings in which fraud occurs or economic activities that are susceptible to fraud. Table 1.1 provides a nonexhaustive overview and description of a number of *important* fraud types—*important* being defined in terms of frequency of occurrence as well as the total monetary value involved.

Table 1.1 Nonexhaustive List of Fraud Categories and Types

Credit card fraud	In credit card fraud there is an unauthorized taking of another's credit. Some common credit card fraud subtypes are counterfeiting credit cards (for the definition of counterfeit, see below), using lost or stolen cards, or fraudulently acquiring credit through mail (definition adopted from definitions.uslegal.com). Two subtypes can been identified, as described by Bolton and Hand (2002): (1) Application fraud, involving individuals obtaining new credit cards from issuing companies by using false personal information, and then spending as much as possible in a short space of time; (2) Behavioral fraud, where details of legitimate cards are obtained fraudulently and sales are made on a "Cardholder Not Present" basis. This does not necessarily require stealing the physical card, only stealing the card credentials. Behavioral fraud concerns most of the credit card fraud. Also, debit card fraud occurs, although less frequent. Credit card fraud is a form of identity theft, as will be defined below.
Insurance fraud	Broad category-spanning fraud related to any type of insurance, both from the side of the buyer or seller of an insurance contract. Insurance fraud from the issuer (seller) includes selling policies from nonexistent companies, failing to submit premiums and churning policies to create more commissions. Buyer fraud includes exaggerated claims (property insurance: obtaining payment that is worth more than the value of the property destroyed), falsified medical history (healthcare insurance: fake injuries), postdated policies, faked death, kidnapping or murder (life insurance fraud), and faked damage (automobile insurance: staged collision) (definition adopted from www.investopedia.com).
Corruption	Corruption is the misuse of entrusted power (by heritage, education, marriage, election, appointment, or whatever else) for private gain. This definition is similar to the definition of fraud provided by the Oxford Dictionary discussed before, in that the objective is personal gain. It is different in that it focuses on misuse of entrusted *power*. The definition covers as such a broad range of different subtypes of corruption, so does not only cover corruption by a politician or a public servant, but also, for example, by the CEO or CFO of a company, the notary public, the team leader at a workplace, the administrator or admissions-officer to a private school or hospital, the coach of a soccer team, and so on (definition adopted from www.corruptie.org).
Counterfeit	An imitation intended to be passed off fraudulently or deceptively as genuine. Counterfeit typically concerns valuable objects, credit cards, identity cards, popular products, money, etc. (definition adopted from www.dictionary.com).

(*continued*)

Table 1.1 (*Continued*)

Product warranty fraud	A product warranty is a type of guarantee that a manufacturer or similar party makes regarding the condition of its product, and also refers to the terms and situations in which repairs or exchanges will be made in the event that the product does not function as originally described or intended (definition adopted from www.investopedia.com). When a product fails to offer the described functionalities or displays deviating characteristics or behavior that are a consequence of the production process and not a consequence of misuse by the customer, compensation or remuneration by the manufacturer or provider can be claimed. When the conditions of the product have been altered due to the customer's use of the product, then the warranty does not apply. Intentionally wrongly claiming compensation or remuneration based on a product warranty is called product warranty fraud.
Healthcare fraud	Healthcare fraud involves the filing of dishonest healthcare claims in order to make profit. Practitioner schemes include: individuals obtaining subsidized or fully covered prescription pills that are actually unneeded and then selling them on the black market for a profit; billing by practitioners for care that they never rendered; filing duplicate claims for the same service rendered; billing for a noncovered service as a covered service; modifying medical records, and so on. Members can commit healthcare fraud by providing false information when applying for programs or services, forging or selling prescription drugs, loaning or using another's insurance card, and so on (definition adopted from www.law.cornell.edu).
Telecommunications fraud	Telecommunication fraud is the theft of telecommunication services (telephones, cell phones, computers, etc.) or the use of telecommunication services to commit other forms of fraud (definition adopted from itlaw.wikia.com). An important example concerns cloning fraud (i.e. the cloning of a phone number and the related call credit by a fraudster), which is an instance of superimposition fraud in which fraudulent usage is superimposed on (added to) the legitimate usage of an account (Fawcett and Provost 1997).
Money laundering	The process of taking the proceeds of criminal activity and making them appear legal. Laundering allows criminals to transform illegally obtained gain into seemingly legitimate funds. It is a worldwide problem, with an estimated $300 billion going through the process annually in the United States (definition adopted from legal-dictionary.thefreedictionary.com).
Click fraud	Click fraud is an illegal practice that occurs when individuals click on a website's click-through advertisements (either banner ads or paid text links) to increase the payable number of clicks to the advertiser. The illegal clicks could either be performed by having a person manually click the advertising hyperlinks or by using automated software or online bots that are programmed to click these banner ads and pay-per-click text ad links (definition adopted from www.webopedia.com).

Table 1.1 (*Continued*)

Identity theft	The crime of obtaining the personal or financial information of another person for the purpose of assuming that person's name or identity in order to make transactions or purchases. Some identity thieves sift through trash bins looking for bank account and credit card statements; other more high-tech methods involve accessing corporate databases to steal lists of customer information (definition adopted from www.investopedia.com).
Tax evasion	Tax evasion is the illegal act or practice of failing to pay taxes that are owed. In businesses, tax evasion can occur in connection with income taxes, employment taxes, sales and excise taxes, and other federal, state, and local taxes. Examples of practices that are considered tax evasion include knowingly not reporting income or underreporting income (i.e., claiming less income than you actually received from a specific source) (definition adopted from biztaxlaw.about.com).
Plagiarism	Plagiarizing is defined by *Merriam Webster*'s online dictionary as to steal and pass off (the ideas or words of another) as one's own, to use (another's production) without crediting the source, to commit literary theft, to present as new and original an idea or product derived from an existing source. It involves both stealing someone else's work and lying about it afterward (definition adopted from www.plagiarism.org).

In the end, fraudulent activities are intended to result in gains or benefits for the fraudster, as emphasized by the definition of fraud provided by the *Oxford Dictionary*. The potential, usually monetary, gain or benefit forms in the large majority of cases the basic driver for committing fraud.

The so-called fraud triangle as depicted in Figure 1.1 provides a more elaborate explanation for the underlying motives or drivers for

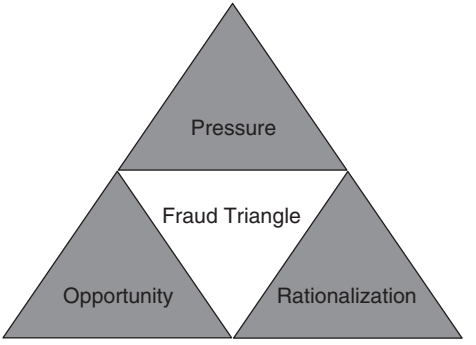


Figure 1.1 Fraud Triangle

committing occupational fraud. The fraud triangle originates from a hypothesis formulated by Donald R. Cressey in his 1953 book *Other People's Money: A Study of the Social Psychology of Embezzlement*:

> *Trusted persons become trust violators when they conceive of themselves as having a financial problem which is non-shareable, are aware this problem can be secretly resolved by violation of the position of financial trust, and are able to apply to their own conduct in that situation verbalizations which enable them to adjust their conceptions of themselves as trusted persons with their conceptions of themselves as users of the entrusted funds or property.*

This basic conceptual model explains the factors that together cause or explain the drivers for an individual to commit *occupational* fraud, yet provides a useful insight in the fraud phenomenon from a broader point of view as well. The model has three legs that together institute fraudulent behavior:

1. ***Pressure*** is the first leg and concerns the main motivation for committing fraud. An individual will commit fraud because a pressure or a problem is experienced of financial, social, or any other nature, and it cannot be resolved or relieved in an authorized manner.
2. ***Opportunity*** is the second leg of the model, and concerns the precondition for an individual to be able to commit fraud. Fraudulent activities can only be committed when the opportunity exists for the individual to resolve or relieve the experienced pressure or problem in an unauthorized but concealed or hidden manner.
3. ***Rationalization*** is the psychological mechanism that explains why fraudsters do not refrain from committing fraud and think of their conduct as acceptable.

An essay by Duffield and Grabosky (2001) further explores the motivational basis of fraud from a psychological perspective.

It concludes that a number of psychological factors may be present in those persons who commit fraud, but that these factors are also associated with entirely legitimate forms of human endeavor. And so fraudsters cannot be distinguished from nonfraudsters purely based on psychological characteristics or patterns.

Fraud is a social phenomenon in the sense that the potential benefits for the fraudsters come at the expense of the victims. These victims are individuals, enterprises, or the government, and as such society as a whole. Some recent numbers give an indication of the estimated *size* and the financial impact of fraud:

- A typical organization loses 5 percent of its revenues to fraud each year (www.acfe.com).
- The total cost of insurance fraud (non–health insurance) in the United States is estimated to be more than $40 billion per year (www.fbi.gov).
- Fraud is costing the United Kingdom £73 billion a year (National Fraud Authority).
- Credit card companies "lose approximately seven cents per every hundred dollars of transactions due to fraud" (Andrew Schrage, Money Crashers Personal Finance, 2012).
- The average size of the informal economy, as a percent of official GNI in the year 2000, in developing countries is 41 percent, in transition countries 38 percent, and in OECD countries 18 percent (Schneider 2002).

Even though these numbers are rough estimates rather than exact measurements, they are based on evidence and do indicate the importance and impact of the phenomenon, and therefore as well the need for organizations and governments to actively fight and prevent fraud with all means they have at their disposal. As will be further elaborated in the final chapter, these numbers also indicate that it is likely worthwhile to invest in fraud-detection and fraud-prevention systems, since a significant financial return on investment can be made.

The importance and need for effective fraud-detection and fraud-prevention systems is furthermore highlighted by the many different

forms or types of fraud of which a number have been summarized in Table 1.1, which is not exhaustive but, rather, indicative, and which illustrates the widespread occurrence across different industries and product and service segments. The broad fraud categories enlisted and briefly defined in Table 1.1 can be further subdivided into more specific subtypes, which, although interesting, would lead us too far into the particularities of each of these forms of fraud. One may refer to the further reading sections at the end of each chapter of this book, providing selected references to specialized literature on different forms of fraud. A number of particular fraud types will also be further elaborated in real-life case studies throughout the book.

FRAUD DETECTION AND PREVENTION

Two components that are essential parts of almost any effective strategy to fight fraud concern *fraud detection* and *fraud prevention*. Fraud detection refers to the ability to recognize or discover fraudulent activities, whereas fraud prevention refers to measures that can be taken to avoid or reduce fraud. The difference between both is clear-cut; the former is an ex post approach whereas the latter an ex ante approach. Both tools may and likely should be used in a complementary manner to pursue the shared objective, *fraud reduction*.

However, as will be discussed in more detail further on, preventive actions will change fraud strategies and consequently impact detection power. Installing a detection system will cause fraudsters to adapt and change their behavior, and so the detection system itself will impair eventually its own detection power. So although complementary, fraud detection and prevention are not independent and therefore should be aligned and considered a whole.

The classic approach to fraud detection is an *expert-based approach*, meaning that it builds on the experience, intuition, and business or domain knowledge of the fraud analyst. Such an expert-based approach typically involves a manual investigation of a suspicious case, which may have been signaled, for instance, by a customer complaining of being charged for transactions he did not do. Such a disputed transaction may indicate a new *fraud mechanism* to have been discovered or developed by fraudsters, and therefore requires

a detailed investigation for the organization to understand and subsequently address the new mechanism.

Comprehension of the fraud mechanism or pattern allows extending the fraud detection and prevention mechanism that is often implemented as a rule base or engine, meaning in the form of a set of If-Then rules, by adding rules that describe the newly detected fraud mechanism. These rules, together with rules describing previously detected fraud patterns, are applied to future *cases* or transactions and trigger an alert or signal when fraud is or may be committed by use of this mechanism. A simple, yet possibly very effective, example of a fraud detection rule in an insurance claim fraud setting goes as follows:

IF:

- Amount of claim is above threshold OR
- Severe accident, but no police report OR
- Severe injury, but no doctor report OR
- Claimant has multiple versions of the accident OR
- Multiple receipts submitted

THEN:

- Flag claim as suspicious AND
- Alert fraud investigation officer.

Such an expert approach suffers from a number of disadvantages. Rule bases or engines are typically expensive to build, since they require advanced manual input by the fraud experts, and often turn out to be difficult to maintain and manage. Rules have to be kept up to date and only or mostly trigger real fraudulent cases, since every signaled case requires human follow-up and investigation. Therefore, the main challenge concerns keeping the rule base lean and effective—in other words, deciding when and which rules to add, remove, update, or merge.

It is important to realize that fraudsters can, for instance by trial and error, learn the business rules that block or expose them and will devise inventive workarounds. Since the rules in the rule-based detection system are based on past experience, new emerging fraud patterns are not automatically flagged or signaled. Fraud is a dynamic

phenomenon, as will be discussed below in more detail, and therefore needs to be traced continuously. Consequently, a fraud detection and prevention system also needs to be continuously monitored, improved, and updated to remain effective.

An expert-based fraud-detection system relies on human expert input, evaluation, and monitoring, and as such involves a great deal of labor intense human interventions. An automated approach to build and maintain a fraud-detection system, requiring less human involvement, could lead to a more efficient and effective system for detecting fraud. The next section in this chapter will introduce several alternative approaches to expert systems that leverage the massive amounts of data that nowadays can be gathered and processed at very low cost, in order to develop, monitor, and update a high-performing fraud-detection system in a more automated and efficient manner. These alternative approaches still require and build on expert knowledge and input, which remains crucial in order to build an effective system.

EXAMPLE CASE

EXAMPLE CASE: EXPERT-BASED APPROACH TO INTERNAL FRAUD DETECTION IN AN INSURANCE CLAIM-HANDLING PROCESS

An example expert-based detection and prevention system to signal potential fraud committed by claim handling officers concerns the business process depicted in Figure 1.2, illustrating the handling of fire incident claims without any form of bodily injury (including death) (Caron et al. 2013). The process involves the following types of activities:

- Administrative activities
- Evaluation-related activities
- In-depth assessment by internal and external experts
- Approval activities
- Leniency-related activities
- Fraud investigation activities

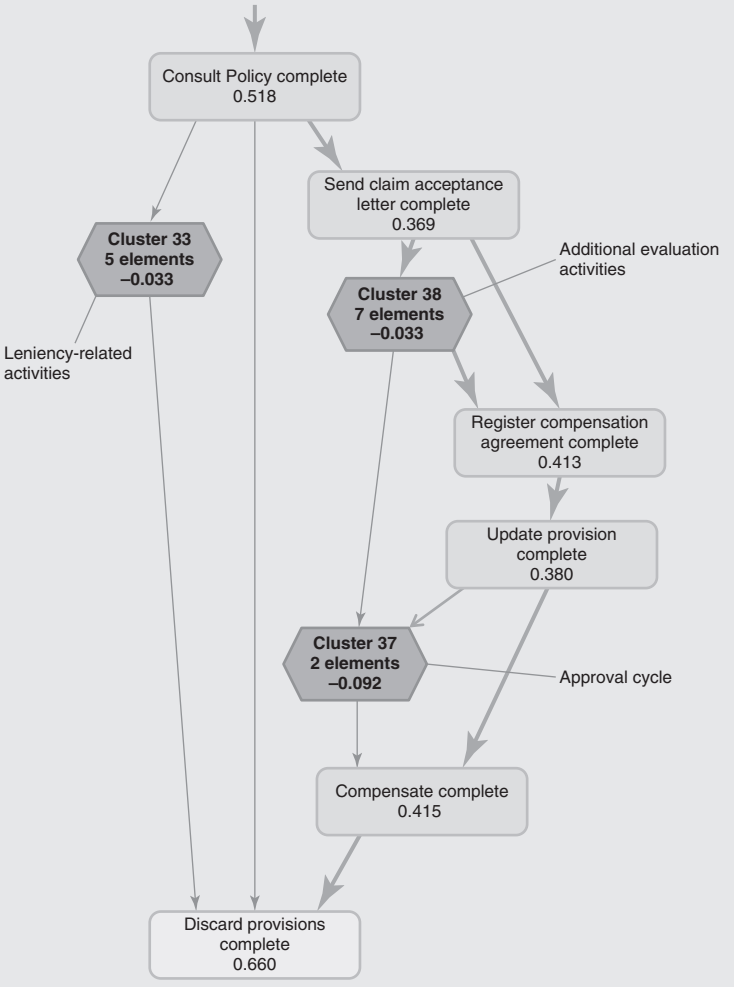


Figure 1.2 Fire Incident Claim-Handling Process

A number of harmful process deviations and related risks can be identified regarding these activities:

- Forgetting to discard provisions
- **Multiple partial compensations (exceeding limit)**

- **Collusion between administrator and experts**
- **Lack of approval cycle**
- Suboptimal task allocation
- Fraud investigation activities
- **Processing rejected claim**
- Forced claim acceptance, absence of a timely primary evaluation

Deviations marked in bold may relate to and therefore indicate fraud. By adopting business policies as a governance instrument and prescribing procedures and guidelines, the insurer may reduce the risks involved in processing the insurance claims. For instance:

Business policy excerpt 1 (customer relationship management related): If the insured requires immediate assistance (e.g., to prevent the development of additional damage), arrangements will be made for a single partial advanced compensation (maximum *x*% of expected covered loss).

- Potential risk: The expected (covered) loss could be exceeded through partial advanced compensations.

Business policy excerpt 2 (avoid financial loss): Settlements need to be approved.

- Potential risk: Collusion between the drafter of the settlement and the insured

Business policy excerpt 3 (avoid financial loss): The proposal of a settlement and its approval must be performed by different actors.

- Potential risk: A person might hold both the team-leader and the expert role in the information system.

Business policy excerpt 4 (avoid financial loss): After approval of the decision (settlement or claim rejection) no changes may occur.

- Potential risk: The modifier and the insured might collude.
- Potential risk: A rejected claim could undergo further processing.

Since detecting fraud based on specified business rules requires prior knowledge of the fraud scheme, the existence of fraud issues will be the direct result of either:

- An inadequate internal control system (controls fail to prevent fraud); or
- Risks accepted by the management (no preventive or corrective controls are in place).

Some examples of fraud-detection rules that can be derived from these business policy excerpts and process deviations, and that may be added to the fraud-detection rule engine are as follows:

Business policy excerpt 1:

- IF multiple advanced payments for one claim, THEN suspicious case.

Business policy excerpt 2:

- IF settlement was not approved before it was paid, THEN suspicious case.

Business policy excerpt 3:

- IF settlement is proposed AND approved by the same person, THEN suspicious case.

Business policy excerpt 4:

- IF settlement is approved AND changed afterward, THEN suspicious case.
- IF claim is rejected AND processed afterward (e.g., look for a settlement proposal, payment,… activity), THEN suspicious case.

BIG DATA FOR FRAUD DETECTION

When fraudulent activities have been detected and confirmed to effectively concern fraud, two types of measures are typically taken:

1. *Corrective measures,* that aim to resolve the fraud and correct the wrongful consequences—for instance by means of pursuing

restitution or compensation for the incurred losses. These corrective measures might also include actions to retrospectively detect and subsequently address similar fraud cases that made use of the same mechanism or loopholes in the fraud detection and prevention system the organization has in place.

2. *Preventive measures,* which may both include actions that aim at preventing future fraud by the caught fraudster (e.g., by terminating a contractual agreement with a customer, as well as actions that aim at preventing fraud of the same type by other individuals). When an expert-based approach is adopted, an example preventive measure is to extend the rule engine by incorporating additional rules that allow detecting and preventing the uncovered fraud mechanism to be applied in the future. A fraud case must be investigated thoroughly so the underlying mechanism can be unraveled, extending the available expert knowledge and allowing it to prevent the fraud mechanism to be used again in the future by making the organization more robust and less vulnerable to fraud by adjusting the detection and prevention system.

Typically, the sooner corrective measures are taken and therefore the sooner fraud is detected, the more effective such measures may be and the more losses can be avoided or recompensed. On the other hand, fraud becomes easier to detect the more time has passed, for a number of particular reasons.

When a fraud mechanism or *path* exists—meaning a loophole in the detection and prevention system of an organization—the number of times this path will be followed (i.e., the fraud mechanism used) grows in time and therefore as well the number of occurrences of this particular type of fraud. The more a fraud path is taken the more apparent it becomes and typically, in fact statistically, the easier to detect. The number of occurrences of a particular type of fraud can be expected to grow since many fraudsters appear to be repeat offenders. As the expression goes, "*Once a thief, always a thief*." Moreover, a fraud mechanism may well be discovered by several individuals or the knowledge shared between fraudsters. As will be shown in Chapter 5 on social network analytics for fraud detection, certainly some types of fraud tend to

spread virally and display what are called social network effects, indicating that fraudsters share their knowledge on how to commit fraud. This effect, too, leads to a growing number of occurrences and, therefore, a higher risk or chance, depending on one's perspective, of detection.

Once a case of a particular type of fraud has been revealed, this will lead to the exposition of similar fraud cases that were committed in the past and made use of the same mechanism. Typically, a retrospective screening is performed to assess the size or impact of the newly detected type of fraud, as well as to resolve (by means of corrective measures, cf. supra) as much as possible fraud cases. As such, fraud becomes easier to detect the more time has passed, since more similar fraud cases will occur in time, increasing the probability that the particular fraud type will be uncovered, as well as because fraudsters committing repeated fraud will increase their individual risk of being exposed. The individual risk will increase the more fraud a fraudster commits for the same basic reason: The chances of getting noticed get larger.

A final reason why fraud becomes easier to detect the more time has passed is because better detection techniques are being developed, are getting readily available, and are being implemented and applied by a growing amount of organizations. An important driver for improvements with respect to detection techniques is *growing data availability*. The informatization and digitalization of almost every aspect of society and daily life leads to an abundance of available data. This so-called big data can be explored and exploited for a range of purposes including fraud detection (Baesens 2014), at a very low cost.

DATA-DRIVEN FRAUD DETECTION

Although classic, expert-based fraud-detection approaches as discussed before are still in widespread use and definitely represent a good starting point and complementary tool for an organization to develop an effective fraud-detection and prevention system, a shift is taking place toward data-driven or statistically based fraud-detection methodologies for three apparent reasons:

1. *Precision*. Statistically based fraud-detection methodologies offer an increased detection power compared to classic approaches.

By processing massive volumes of information, fraud patterns may be uncovered that are not sufficiently apparent to the human eye. It is important to notice that the improved power of data-driven approaches over human processing can be observed in similar applications such as credit scoring or customer churn prediction. Most organizations only have a limited capacity to have cases checked by an inspector to confirm whether or not the case effectively concerns fraud. The goal of a fraud-detection system may be to make the most optimal use of the limited available inspection capacity, or in other words to maximize the fraction of fraudulent cases among the inspected cases (and possibly in addition, the detected amount of fraud). A system with higher precision, as delivered by data-based methodologies, directly translates in a higher fraction of fraudulent inspected cases.

2. *Operational efficiency*. In certain settings, there is an increasing amount of cases to be analyzed, requiring an automated process as offered by data-driven fraud-detection methodologies. Moreover, in several applications, operational requirements exist, imposing time constraints on the processing of a case. For instance, when evaluating a transaction with a credit card, an almost immediate decision is required with respect to approve or block the transaction because of suspicion of fraud. Another example concerns fraud detection for customs in a harbor, where a decision has to be made within a confined time window whether to let a container pass and be shipped inland, or whether to further inspect it, possibly causing delays. Automated data-driven approaches offer such functionality and are able to comply with stringent operational requirements.
3. *Cost efficiency*. As already mentioned in the previous section, developing and maintaining an effective and lean expert-based fraud-detection system is both challenging and labor intensive. A more automated and, as such, more efficient approach to develop and maintain a fraud-detection system, as offered by data-driven methodologies, is preferred. Chapters 6 and 7 discuss the cost efficiency and return on investment of data-driven fraud-detection models.

An additional driver for the development of improved fraud-detection technologies concerns the growing amount of interest that fraud detection is attracting from the general public, the media, governments, and enterprises. This increasing awareness and attention for fraud is likely due to its large negative social as well as financial impact, and leads to growing investments and research into the matter, both from academia, industry, and government.

Although fraud-detection approaches have gained significant power over the past years by adopting potent statistically based methodologies and by analyzing massive amounts of data in order to discover fraud patterns and mechanisms, still fraud remains hard to detect. It appears the Pareto principle holds with respect to the required effort and difficulty of detecting fraud: It appears the principle of decreasing returns holds with respect to the required effort and so forth. In order to explain the *hardness* and complexity of the problem, it is important to acknowledge the fact that fraud is a dynamic phenomenon, meaning that its nature changes in time. Not only fraud-detection mechanisms evolve, but also fraudsters adapt their approaches and are inventive in finding more refined and less apparent ways to commit fraud without being exposed. Fraudsters probe fraud-detection and prevention systems to understand their functioning and to discover their weaknesses, allowing them to adapt their methods and strategies.

FRAUD-DETECTION TECHNIQUES

Indeed, fraudsters develop advanced strategies to cleverly cover their tracks in order to avoid being uncovered. Fraudsters tend to try and blend in as much as possible into the surroundings. Such an approach reminds of camouflage techniques as used by the military or by animals such as chameleons and stick insects. This is clearly no fraud by opportunity, but rather, is carefully planned, leading to a need for new techniques that are able to detect and address patterns that initially seem to comply with normal behavior, but in reality instigate fraudulent activities.

Detection mechanisms based on unsupervised learning techniques or descriptive analytics, as discussed in Chapter 3, typically aim at

finding behavior that *deviates* from *normal* behavior, or in other words at detecting anomalies. These techniques learn from historical observations, and are called unsupervised since they do not require these observations to be labeled as either a fraudulent or a nonfraudulent example case. An example of behavior that does not comply with *normal behavior* in a telecommunications subscription fraud setting is provided by the transaction data set with call detail records of a particular subscriber shown in Table 1.2 (Fawcett and Provost 1997). Remark that the calls found to be fraudulent (last column in the table indicating *bandit*) are not suspicious by themselves; however, they are deviating from normal behavior for this particular subscriber.

Outlier-detection techniques have great value and allow detecting a significant fraction of fraudulent cases. In particular, they might allow detecting fraud that is different in nature from historical fraud, or in other words fraud that makes use of new, unknown mechanisms resulting in a *novel fraud pattern*. These new patterns are not discovered by expert systems, and as such descriptive analytics may be a first

Table 1.2 Call Detail Records of a Customer with Outliers Indicating Suspicious Activity (deviating behavior starting at a certain moment in time) at the Customer Subscription (Fawcett and Provost 1997)

Date (m/d)	Time	Day	Duration	Origin	Destination	Fraud
1/01	10:05:01	Mon	13 mins	Brooklyn, NY	Stamford, CT	
1/05	14:53:27	Fri	5 mins	Brooklyn, NY	Greenwich, CT	
1/08	09:42:01	Mon	3 mins	Bronx, NY	White Plains, NY	
1/08	15:01:24	Mon	9 mins	Brooklyn, NY	Brooklyn, NY	
1/09	15:06:09	Tue	5 mins	Manhattan, NY	Stamford, CT	
1/09	16:28:50	Tue	53 sec	Brooklyn, NY	Brooklyn, NY	
1/10	01:45:36	Wed	35 sec	Boston, MA	Chelsea, MA	Bandit
1/10	01:46:29	Wed	34 sec	Boston, MA	Yonkers, MA	Bandit
1/10	01:50:54	Wed	39 sec	Boston, MA	Chelsea, MA	Bandit
1/10	11:23:28	Wed	24 sec	White Plains, NY	Congers, NY	
1/11	22:00:28	Thu	37 sec	Boston, MA	East Boston, MA	Bandit
1/11	22:04:01	Thu	37 sec	Boston, MA	East Boston, MA	Bandit

complementary tool to be adopted by an organization in order to improve its expert rule–based fraud-detection system.

Descriptive techniques however show to be prone to deception, exactly by the camouflage-like fraud strategies already discussed. Therefore, the detection system can be further improved by complementing it by a tool that is able to unmask fraudsters adopting a camouflage-like technique.

Therefore, in Chapter 4, a second type of techniques is introduced. Supervised learning techniques or predictive analytics aim to learn from historical information or observations in order to retrieve patterns that allow differentiating between normal and fraudulent behavior. These techniques exactly aim at finding silent alarms, the parts of their tracks that fraudsters cannot cover up. Supervised learners can be applied to predict or detect fraud as well as to estimate the amount of fraud.

Predictive analytics has limitations as well, probably the most important one being that they need historical examples to learn from (i.e., a labeled data set of historically observed fraud behavior). This reduces their detection power with respect to drastically different fraud types making use of new mechanisms or methods, and which have not been detected thus far and are therefore not included in the historical database of fraud cases from which the predictive model was learned. As already discussed, descriptive analytics may perform better with respect to detecting such new fraud mechanisms, at least if a new fraud mechanism leads to detectable deviations from *normality*. This illustrates the complementarity of supervised and unsupervised methods and motivates the use of both types of methods as complementary tools in developing a powerful fraud-detection and prevention system.

A third type of complementary tool concerns social network analysis, which further extends the abilities of the fraud-detection system by learning and detecting characteristics of fraudulent behavior in a network of linked entities. Social network analytics is the newest tool in our toolbox to fight fraud, and proofs to be a very powerful means as will appear from the discussion and presented case study

in Chapter 5. Social network analytics allows including an extra source of information in the analysis, being the relationships between entities, and as such may contribute in uncovering particular patterns indicating fraud.

It is important to stress that these three different types of techniques may complement each other since they focus on different aspects of fraud and are not to be considered as exclusive alternatives. An effective fraud-detection and prevention system will make use of and combine these different tools, which have different possibilities and limitations and therefore reinforce each other when applied in a combined setup. When developing a fraud-detection system, an organization will likely follow the order in which the different tools have been introduced; as a first step an expert-based rule engine may be developed, which in a second step may be complemented by descriptive analytics, and subsequently by predictive and social network analytics. Developing a fraud-detection system in this order allows the organization to gain expertise and insight in a stepwise manner, hereby facilitating each next step. However, the exact order of adopting the different techniques may depend on the characteristics of the type of fraud an organization is faced with.

FRAUD CYCLE

Figure 1.3 introduces the *fraud cycle*, and depicts four essential activities:

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

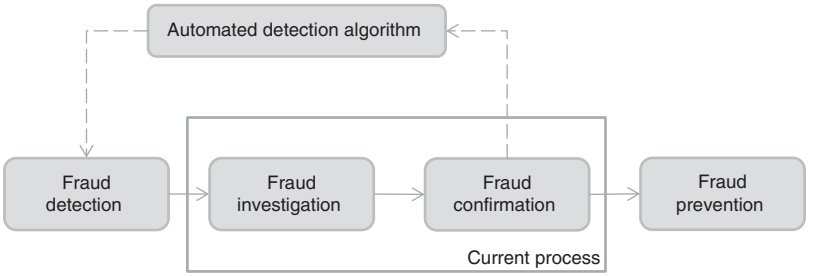


Figure 1.3 The Fraud Cycle

premise of the 1956 science fiction short story *Minority Report* by Philip K. Dick.

Remark the feedback loop in Figure 1.3 from the fraud confirmation activity toward the fraud-detection activity. Newly detected cases should be added (as soon as possible!) to the database of historical fraud cases, which is used to learn or induce the detection model. The fraud-detection model may not be retrained every time a new case is confirmed; however, a regular update of the model is recommendable given the dynamic nature of fraud and the importance of detecting fraud as soon as possible. The required frequency of retraining or updating the detection model depends on several factors:

- The volatility of the fraud behavior
- The detection power of the current model, which is related to the volatility of the fraud behavior
- The amount of (similar) confirmed cases already available in the database
- The rate at which new cases are being confirmed
- The required effort to retrain the model

Depending on the emerging need for retraining as determined by these factors, as well as possible additional factors, an automated approach such as reinforcement learning may be considered which continuously updates the detection model by learning from the newest observations.

EXAMPLE CASE

EXAMPLE CASE: SUPERVISED AND UNSUPERVISED LEARNING FOR DETECTING CREDIT CARD FRAUD

In order to fight fraud and given the abundant data availability, credit card companies have been among the early adopters of big data approaches to develop effective fraud-detection and prevention systems. A typical credit card transaction is registered in the systems of the credit card company by logging up to a hundred or more characteristics describing the details of a transaction. Table 1.3 provides for illustrative purposes a number of such characteristics or variables that are being captured (Hand 2007).

Table 1.3 Example Credit Card Transaction Data Fields

Transaction ID	Transaction type	Date of transaction
Time of transaction	Amount	Currency
Local currency amount	Merchant ID	Merchant category
Card issuer ID	ATM ID	Cheque account prefix

By logging this information over a period of time, a dataset is being created that allows applying descriptive analytics. This includes *outlier detection techniques,* which allow detecting abnormal or anomalous behavior and/or characteristics in a data set. So-called outliers may indicate suspicious activities, and may occur at the data item level or the data set level.

Figure 1.4 provides an illustration of outliers at the data item level, in this example transactions that deviate from the *normal* behavior by a customer. The scatter plot clearly shows three clusters of regular, frequently occurring *types* as characterized by the time and place dimension of transactions for one particular customer, as well as two deviating transactions marked in black. These outliers are suspicious and possibly concern fraudulent transactions, and therefore may be flagged for further human investigation.

An outlier at the data set level means that the behavior of a person or instance does not comply with the overall behavior. Figure 1.5 plots the age and income characteristics of customers as provided when applying for a credit card. The two outliers marked in black in the plot may indicate so-called subscription fraud (cf. Table 1.1, definition of credit card fraud), since these combinations of age and income strongly deviate from the *normal behavior*.

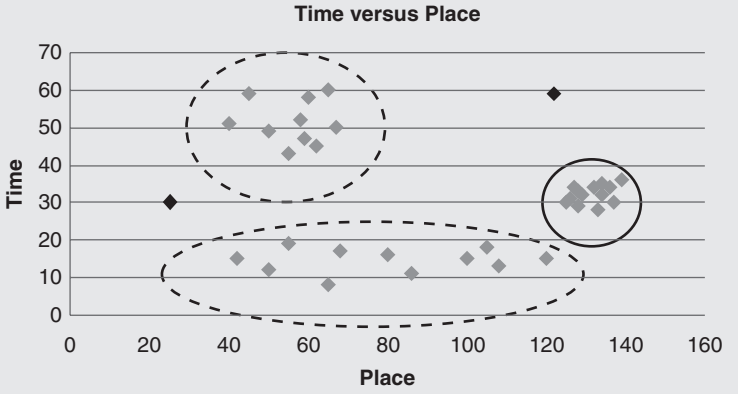


Figure 1.4 Outlier Detection at the Data Item Level

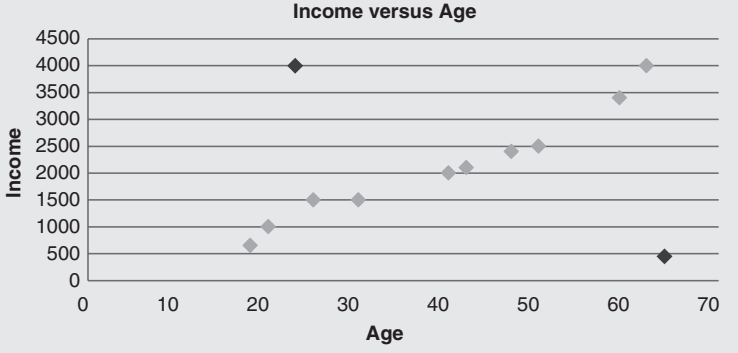


Figure 1.5 Outlier Detection at the Data Set Level

The addition of a field in the available transaction data set that indicates whether a transaction was fraudulent allows predictive analytics to be applied to yield a model predicting or classifying an instance as being fraudulent or not. As will be discussed in the next section as well as in Chapter 4, such models may be interpreted to understand the underlying credit card fraud behavior patterns that lead the model to predict whether a transaction might be fraudulent. Such patterns may be:

- Small purchase followed by a big one
- Large number of online purchases in a short period

- Spending as much as possible as quickly as possible
- Spending smaller amounts, spread across time

Such pattern may be harder to detect and concern advanced methods adopted by fraudsters and developed exactly to avoid detection.

THE FRAUD ANALYTICS PROCESS MODEL

Figure 1.6 provides a high-level overview of the analytics process model (Han and Kamber 2011; Hand, Mannila, and Smyth 2001; Tan, Steinbach, and Kumar 2005). As a first step, a thorough definition of the business problem is needed to be solved with analytics. Next, all source data must be identified that could be of potential interest. This is a very important step, as data are the key ingredient to any analytical exercise and the selection of data will have a deterministic impact on the analytical models that will be built in a subsequent step. All data will then be gathered in a staging area that could be a data mart or data warehouse. Some basic exploratory analysis can be considered here using for instance OLAP (online analytical processing, see Chapter 3) facilities for multidimensional data analysis (e.g., roll-up, drill down, slicing and dicing). This will be followed by a data-cleaning step to

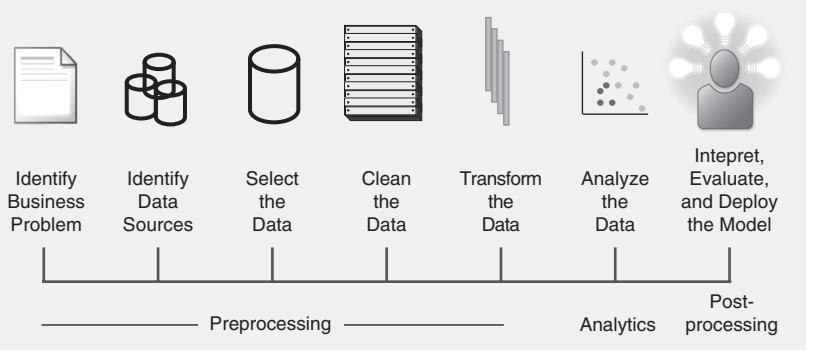


Figure 1.6 The Fraud Analytics Process Model

get rid of all inconsistencies, such as missing values and duplicate data. Additional transformations may also be considered, such as binning, alphanumeric to numeric coding, geographical aggregation, and so on. In the analytics step, an analytical model will be estimated on the preprocessed and transformed data. In this stage, the actual fraud-detection model is built. Finally, once the model has been built, it will be interpreted and evaluated by the fraud experts.

Trivial patterns that may be detected by the model, for instance similar to expert rules, are interesting as they provide some validation of the model. But of course, the key issue is to find the unknown yet interesting and actionable patterns (sometimes also referred to as knowledge diamonds) that can provide added insight and detection power. Once the analytical model has been appropriately validated and approved, it can be put into production as an analytics application (e.g., decision support system, scoring engine). Important to consider here is how to represent the model output in a user-friendly way, how to integrate it with other applications (e.g., detection and prevention system, risk engines), and how to make sure the analytical model can be appropriately monitored and backtested on an ongoing basis. These post-processing steps will be discussed in detail in Chapter 6.

It is important to note that the process model outlined in Figure 1.6 is iterative in nature in the sense that one may have to go back to previous steps during the exercise. For example, during the analytics step, the need for additional data may be identified, which may necessitate additional cleaning, transformation, and so on. The most time-consuming step typically is the data selection and preprocessing step, which usually takes around 80 percent of the total efforts needed to build an analytical model.

A fraud-detection model must be thoroughly evaluated before being adopted. Depending on the exact setting and usage of the model, different aspects may need to be assessed during evaluation in order to ensure the model to be acceptable for implementation. Table 1.4 reviews several key characteristics of *successful* fraud analytics models that may or may not apply, depending on the exact application.

A number of particular challenges may present themselves when developing and implementing a fraud-detection model, possibly leading to difficulties in meeting the objectives as expressed by the

Table 1.4 Key Characteristics of Successful Fraud Analytics Models

Statistical accuracy	Refers to the detection power and the correctness of the statistical model in flagging cases as being suspicious. Several statistical evaluation criteria exist and may be applied to evaluate this aspect, such as the hit rate, lift curves, AUC, etc. A number of suitable measures will be discussed in detail in Chapter 4. Statistical accuracy may also refer to statistical significance, meaning that the patterns that have been found in the data have to be real and not the consequence of coincidence. In other words, we need to make sure that the model generalizes well and is not overfitted to the historical data set.
Interpretability	When a deeper understanding of the detected fraud patterns is required, for instance to validate the model before it is adopted for use, a fraud-detection model may have to be interpretable. This aspect involves a certain degree of subjectivism, since interpretability may depend on the user's knowledge. The interpretability of a model depends on its format, which, in turn, is determined by the adopted analytical technique. Models that allow the user to understand the underlying reasons why the model signals a case to be suspicious are called white-box models, whereas complex incomprehensible mathematical models are often referred to as black-box models. It may well be in a fraud-detection setting that black-box models are acceptable, although in most settings some level of understanding and in fact validation which is facilitated by interpretability is required for the management to have confidence and allow the effective operationalization of the model.
Operational efficiency	Operational efficiency refers to the time that is required to evaluate the model, or in other words, the time required to evaluate whether a case is suspicious or not. When cases need to be evaluated in real time, for instance to signal possible credit card fraud, operational efficiency is crucial and is a main concern during model performance assessment. Operational efficiency also entails the efforts needed to collect and preprocess the data, evaluate the model, monitor and backtest the model, and reestimate it when necessary.
Economical cost	Developing and implementing a fraud-detection model involves a significant cost to an organization. The total cost includes the costs to gather, preprocess, and analyze the data, and the costs to put the resulting analytical models into production. In addition, the software costs, as well as human and computing resources, should be taken into account. Possibly also external data has to be bought to enrich the available in-house data. Clearly, it is important to perform a thorough cost-benefit analysis at the start of the project, and to gain insight in the constituent factors of the returns on investment of building an advanced fraud-detection system.
Regulatory compliance	Depending on the context there may be internal or organization-specific and external regulation and legislation that applies to the development and application of a model. Clearly, a fraud-detection model should be in line and comply with all applicable regulation and legislation, for instance with respect to privacy, the use of cookies in a web-browser, etc.

characteristics discussed in Table 1.4. A first key challenge concerns the dynamic nature of fraud. Fraudsters constantly try to beat detection and prevention systems by developing new strategies and methods. Therefore, adaptive analytical models and detection and prevention systems are required, in order to detect and resolve fraud as soon as possible. Detecting fraud as early as possible is crucial, as discussed before.

Clearly, it is also crucial to detect fraud as accurately as possible, and not to miss out on too many fraud cases, especially on fraud cases involving a large amount or financial impact. The cost of missing a fraudulent case or a fraud mechanism may be significant. Related to having good detection power is the requirement of having at the same time a low false alarm rate, since we also want to avoid harassing good customers and prevent accounts or transactions to be blocked unnecessarily.

In developing analytical models with good detection power and low false alarm rate, an additional difficulty concerns the skewedness of the data, meaning that we typically have plenty of historical examples of nonfraudulent cases, but only a limited number of fraudulent cases. For instance, in a credit card fraud setting, typically less than 0.5 percent of transactions are fraudulent. Such a problem is commonly referred to as a needle-in-a-haystack problem and might cause an analytical technique to experience difficulties in learning an accurate model. A number of approaches to address the skewedness of the data will be discussed in Chapter 4.

Depending on the exact application, also operational efficiency may be a key requirement, meaning that the fraud-detection system might only have a limited amount of time available to reach a decision and let a transaction pass or not. As an example, in a credit card fraud-detection setting the decision time has to be typically less than eight seconds. Such a requirement clearly impacts the design of the operational IT systems, but also the design of the analytical model. The analytical model should not take too long to be evaluated, and the information or the variables that are used by the model should not take too long to be gathered or calculated. Calculating trend variables in real time, for instance, may not be feasible from an operational perspective, since this is taking too much valuable time. This also

relates to the final challenge of dealing with the massive volumes of data that are available and need to be processed.

FRAUD DATA SCIENTISTS

Whereas in the previous section we discussed the characteristics of a good fraud-detection model, in this paragraph we will elaborate on the key characteristics of a good fraud data scientist from the perspective of the hiring manager. It is based on our consulting and research experience, having collaborated with many companies worldwide on the topic of big data, analytics, and fraud detection.

A Fraud Data Scientist Should Have Solid Quantitative Skills

Obviously, a fraud data scientist should have a thorough background in statistics, machine learning and/or data mining. The distinction between these various disciplines is getting more and more blurred and is actually not that relevant. They all provide a set of quantitative techniques to analyze data and find business relevant patterns within a particular context such as fraud detection. A data scientist should be aware of which technique can be applied when and how. He/she should not focus too much on the underlying mathematical (e.g., optimization) details but, rather, have a good understanding of what analytical problem a technique solves, and how its results should be interpreted. In this context, the education of engineers in computer science and/or business/industrial engineering should aim at an integrated, multidisciplinary view, with graduates formed in both the use of the techniques, and with the business acumen necessary to bring new endeavors to fruition. Also important is to spend enough time validating the analytical results obtained so as to avoid situations often referred to as data massage and/or data torture whereby data is (intentionally) misrepresented and/or too much focus is spent discussing spurious correlations. When selecting the optimal quantitative technique, the fraud data scientist should

take into account the specificities of the context and the problem or fraud-detection application at hand. Typical requirements for fraud-detection models have been discussed in the previous section and the fraud data scientist should have a basic understanding and feeling for all of those. Based on a combination of these requirements, the data scientist should be capable of selecting the best analytical technique to solve the particular business problem.

A Fraud Data Scientist Should Be a Good Programmer

As per definition, data scientists work with data. This involves plenty of activities such as sampling and preprocessing of data, model estimation and post-processing (e.g., sensitivity analysis, model deployment, backtesting, model validation). Although many user-friendly software tools are on the market nowadays to automate and support these tasks, every analytical exercise requires tailored steps to tackle the specificities of a particular business problem and setting. In order to successfully perform these steps, programming needs to be done. Hence, a good data scientist should possess sound programming skills (e.g., SAS, R, Python, etc.). The programming language itself is not that important as such, as long as he/she is familiar with the basic concepts of programming and knows how to use these to automate repetitive tasks or perform specific routines.

A Fraud Data Scientist Should Excel in Communication and Visualization Skills

Like it or not, analytics is a technical exercise. At this moment, there is a huge gap between the analytical models and the business users. To bridge this gap, communication and visualization facilities are key. Hence, data scientists should know how to represent analytical models and their accompanying statistics and reports in user-friendly ways using traffic-light approaches, OLAP (online analytical processing) facilities, If-then business rules, and so on. They should be capable of communicating the right amount of information without

getting lost into complex (e.g., statistical) details, which will inhibit a model's successful deployment. By doing so, business users will better understand the characteristics and behavior in their (big) data which will improve their attitude toward and acceptance of the resulting analytical models. Educational institutions must learn to balance, since many academic degrees form students who are skewed to either too much analytical or too much practical knowledge.

A Fraud Data Scientist Should Have a Solid Business Understanding

While this might be obvious, we have witnessed (too) many data science projects that failed because the respective analyst did not understand the business problem at hand. By "business" we refer to the respective application area. Several examples of such application areas of fraud-detection techniques were summarized in Table 1.1. Each of those fields has its own particularities that are important for a fraud data scientist to know and understand in order to be able to design and implement a customized fraud-detection system. The more aligned the detection system with the environment, the better its performance will be, as evaluated on each of the dimensions already discussed.

A Fraud Data Scientist Should Be Creative

A data scientist needs creativity on at least two levels. First, on a technical level, it is important to be creative with regard to feature selection, data transformation, and cleaning. These steps of the standard analytics process have to be adapted to each particular application, and often the "right guess" could make a big difference. Second, big data and analytics is a fast-evolving field. New problems, technologies, and corresponding challenges pop up on an ongoing basis. Moreover, also fraudsters are very creative and adapt their tactics and methods on an ongoing basis. Therefore, it is crucial that a fraud data scientist keeps up with these new evolutions and technologies and has enough creativity to see how they can create new opportunities.

Figure 1.7 summarizes the key characteristics and strengths constituting the ideal fraud data scientist profile.

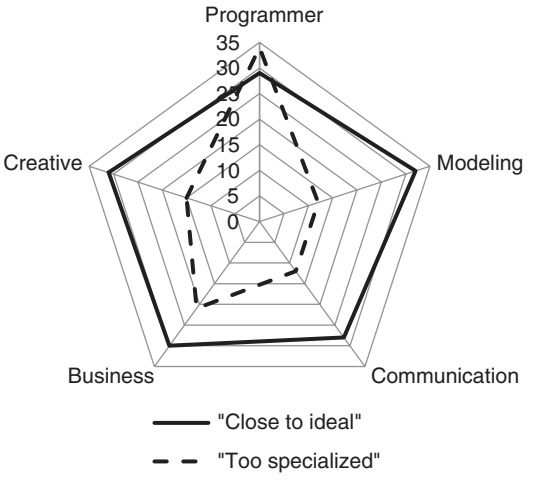


Figure 1.7 Profile of a Fraud Data Scientist

A SCIENTIFIC PERSPECTIVE ON FRAUD

To conclude this chapter, let's provide a scientific perspective about the research on fraud. Figure 1.8 shows a screenshot of the Web of Science statistics when querying all scientific publications between 1996 and 2014 for the key term *fraud*. It shows the total number of papers published each year, the number of citations and the top five most-cited papers.

A couple of conclusions can be drawn as follows:

- 6,174 scientific papers have been published on the topic of fraud during the period reported.
- The h-index is 44, implying that there are at least 44 papers with 44 citations on the topic of fraud.
- The number of publications is steadily increasing, which shows a growing interest from the academic community and research on the topic.
- The citations are exponentially growing, which is associated with the increasing number of publications.
- Two of the five papers mentioned study the use of analytics for fraud detection, clearly illustrating the growing attention in the field for data-driven approaches.

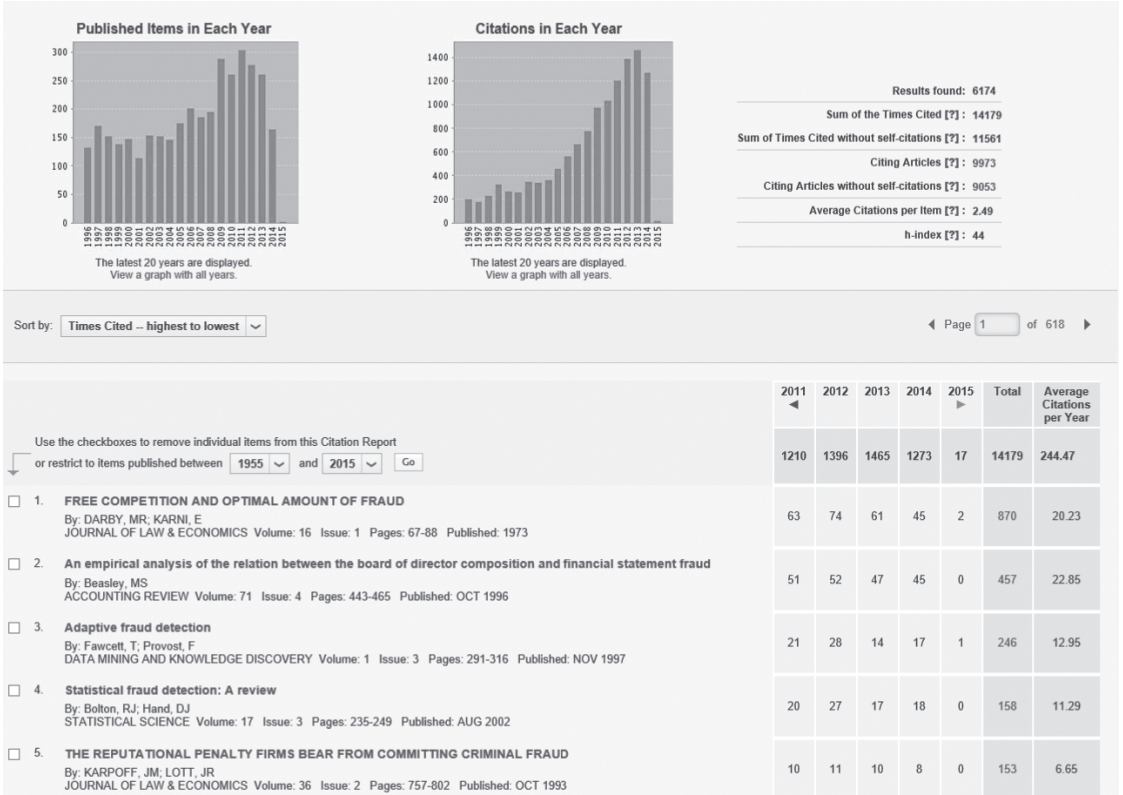


Figure 1.8 Screenshot of Web of Science Statistics for Scientific Publications on Fraud between 1996 and 2014

REFERENCES

Armstrong, J. S. (2001). Selecting Forecasting Methods. In J.S. Armstrong, ed. *Principles of Forecasting: A Handbook for Researchers and Practitioners*. New York: Springer Science + Business Media, pp. 365–386.

Baesens, B. (2014). *Analytics in a Big Data World: The Essential Guide to Data Science and Its Applications*. Hoboken, NJ: John Wiley & Sons.

Bolton, R. J., & Hand, D. J. (2002). Statistical Fraud Detection: A Review. *Statistical Science*, *17* (3): 235–249.

Caron, F., Vanden Broucke, S., Vanthienen, J., & Baesens, B. (2013). Advanced Rule-Based Process Analytics: Applications for Risk Response Decisions and Management Control Activities. *Expert Systems with Applications*, Submitted.

Chakraborty, G., Murali, P., & Satish, G. (2013). *Text Mining and Analysis: Practical Methods, Examples, and Case Studies Using SAS*. Cary, NC: SAS Institute.

Cressey, D. R. (1953). *Other People's Money; A Study of the Social Psychology of Embezzlement*. New York: Free Press.

Duffield, G., & Grabosky, P. (2001). The Psychology of Fraud. In *Trends and Issues in Crime and Criminal Justice, Australian Institute of Criminology* (199).

Elder IV,, J., & Thomas, H. (2012). *Practical Text Mining and Statistical Analysis for Non-Structured Text Data Applications*. New York: Academic Press.

Fawcett, T., & Provost, F. (1997). Adaptive Fraud Detection. *Data Mining and Knowledge Discovery 1–3* (3): 291–316.

Grabosky, P., & Duffield, G. (2001). Red Flags of Fraud. *Trends and Issues in Crime and Criminal Justice, Australian Institute of Criminology* (200).

Han, J., & Kamber, M. (2011). Data Mining: Concepts and Techniques, Third Edition: Morgan Kaufmann.

Hand, D. (2007, September). *Statistical Techniques for Fraud Detection, Prevention, and Evaluation*. Paper presented at the NATO ASI: Mining Massive Data sets for Security, London, England.

Hand, D. J., Mannila, H., & Smyth, P. (2001). *Principles of Data Mining*. Cambridge, MA: Bradford.

Jamain, A. (2001). *Benford's Law*. London: Imperial College.

Junqué de Fortuny, E., Martens, D., & Provost, F. (2013). Predictive Modeling with Big Data: Is Bigger Really Better? *Big Data 1* (4): 215–226.

Little, R. J. A., & Rubin, D. B. (2002). *Statistical Analysis with Missing Data* (2nd ed.). Hoboken, NJ: Wiley.

Maydanchik, A. (2007). *Data Quality Assessment*. Bradley Beach, NC: Technics Publications.

Navarette, E. (2006). Practical Calculation of Expected and Unexpected Losses in Operational Risk by Simulation Methods (Banca & Finanzas: Documentos de Trabajo, *1* (1): pp. 1–12).

Petropoulos, F., Makridakis, S., Assimakopoulos, V., & Nikolopoulos, K. (2014). "Horses for courses" in demand forecasting. *European Journal of Operational Research*, *237* (1): 152–163.

Schneider, F. (2002). Size and Measurement of the Informal Economy in 110 Countries around the World. In *Workshop of Australian National Tax Centre, ANU,* Canberra, Australia.

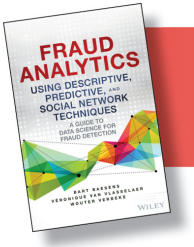
Tan, P.-N. N., Steinbach, M., & Kumar, V. (2005). *Introduction to Data Mining*. Boston: Addison Wesley.

Van Gestel, T., & Baesens, B. (2009). *Credit Risk Management: Basic Concepts: Financial Risk Components, Rating Analysis, Models, Economic and Regulatory Capital*. Oxford: Oxford University Press.

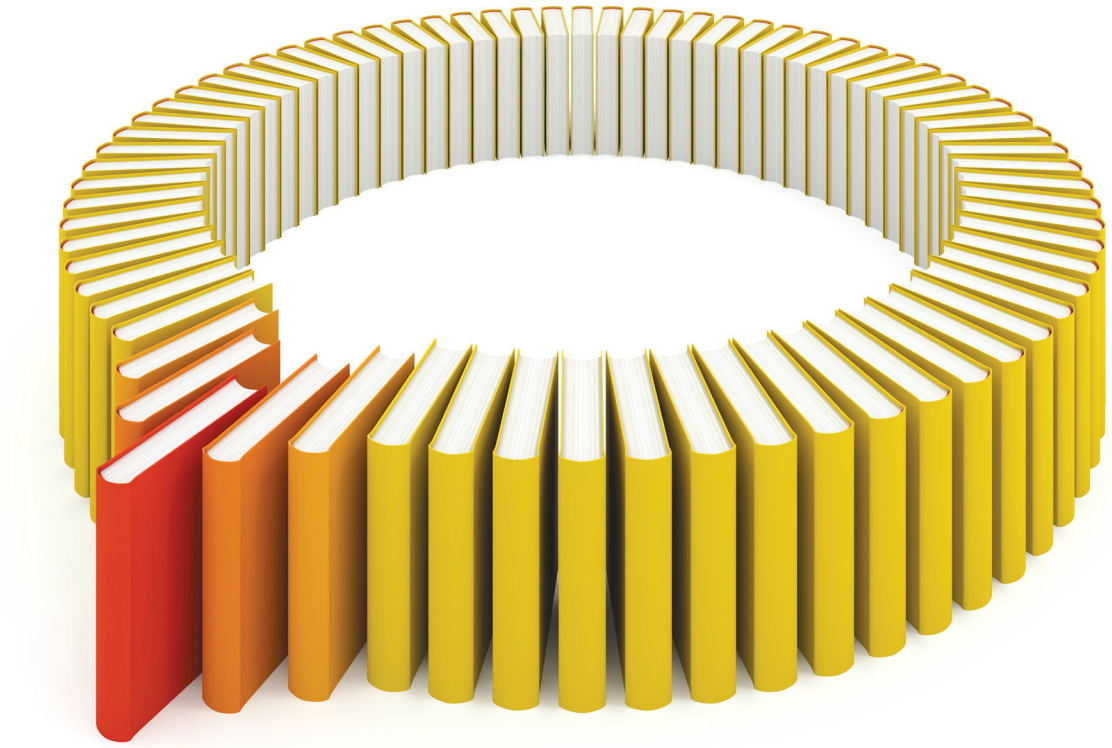
Van Vlasselaer, V., Eliassi-Rad, T., Akoglu, L., Snoeck, M., & Baesens, B. (2015). Gotcha! Network-based Fraud Detection for Social Security Fraud. *Management Science*, Submitted.

Verbeke, W., Dejaeger, K., Martens, D., Hur, J., & Baesens, B. (2012). New Insights into Churn Prediction in the Telecommunication Sector: A Profit Driven Data Mining Approach. *European Journal of Operational Research* 218: 211–229.

From *Fraud Analytics Using Descriptive, Predictive, and Social Network Techniques: A Guide to Data Science for Fraud Detection*, by Bart Baesens, Véronique Van Vlasselaer, and Wouter Verbeke. Copyright © 2015, SAS Institute Inc., Cary, North Carolina, USA. ALL RIGHTS RESERVED.



From *Fraud Analytics Using Descriptive, Predictive, and Social Network Techniques.*
Full book available for purchase here.



Gain Greater Insight into Your SAS® Software with SAS Books.

Discover all that you need on your journey to knowledge and empowerment.

support.sas.com/bookstore
for additional books and resources.

sas
THE POWER TO KNOW®

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration. Other brand and product names are trademarks of their respective companies. © 2013 SAS Institute Inc. All rights reserved. S107969US.0413