

Using data mining to detect fraud



Every government agency that exchanges money with citizens, service providers or vendors risks exposure to fraud and abuse.

Agencies around the world lose more and more money through fraud every year. They need to recoup this lost money so they can continue providing superior services for their citizens. These agencies can identify fraudulent activity by mining their existing data.

Typically, agency auditors use past experience and intuition to create profiles describing fraudulently filed claims. But these unproven theories waste time and miss opportunities as auditors unknowingly review legitimate claims while failing to catch fraudulent ones.

To ensure auditors target claims that have the greatest likelihood of adjustment, many social service agencies have incorporated data mining into their investigating and auditing processes. Data mining combines powerful analytical techniques with your business knowledge to turn data you've already acquired into the insight you need to identify probable instances of fraud and abuse.

Discover how to recoup more money

How does your agency determine which of its thousands or millions of claims are legitimate? Perhaps, your auditors rely on hunches and intuition to determine which claims or payment requests might be fraudulent. Do these less-than-precise methods cause your auditors to waste time reviewing claims or payments that have little or no chance of being recouped? Or maybe your auditors tend to target claims or payment requests that represent inconsequential adjustments, while missing the ones that offer significant amounts of money to recoup. With data mining, your auditors can focus on recovering money so much-needed programs receive the funds required to effectively serve citizens.

What if you could:

- Isolate the factors that indicate a claim or payment request has a high probability of fraudulence?
- Develop rules and use them to flag only those claims or requests most likely to be fraudulent?
- Ensure your auditors could review claims or requests that are not only likely to be fraudulent but also have the greatest adjustment potential?

If your agency could accomplish these goals, you could use your resources more efficiently and more effectively manage your department. Then, your department could recoup a substantial amount of money and reinvest it in the programs your citizens need.

Capitalize on existing data

Your previously audited claims hold the key to recouping money in the future. By creating models from historical information, you can accurately pinpoint fraudulent claims out of the millions of claims you receive each year. These data mining models lower the cost of fraud and abuse while saving your auditors' time.

Data mining empowers a variety of government agencies with the ability to predict which

claims are fraudulent so they can effectively target their resources and recoup significant amounts of money.

The following case study shows how one government agency — in this instance a U.S. Medicare office — used data mining to build models based on previously audited claims to identify potentially fraudulent claims. With these models in place, the department's claim audit selection will be more exact, generate more money through claim adjustments and save time and manpower.

Building models to find fraudulent claims

A Medicare fraud detection office needs to accurately determine which claims are fraudulently filed so it can concentrate on recouping revenue to reinvest into its programs.

Over the years, this office collected audit results on Medicare claims. The department never used historical records to identify probable fraudulent claims in the future. In fact, it generated audits based solely on auditors' experiences or intuitions. This approach caused the department to miss opportunities to collect money and to waste auditors' valuable time reviewing legitimate claims.

Data mining now enables this department to predict which Medicare claims are likely to be fraudulent. This gives auditors the power to determine what returns they should target, thereby, recouping millions of dollars otherwise lost and saving auditors many hours of valuable time.

The Medicare fraud detection office used Clementine, the leading data mining workbench, to get results. Clementine examines each line entry on claims, compares the line entries against the amount of fraud dollars detected, ranks claims in the order of likely fraudulence and displays the results back.

The following steps describe how the department created models and predicted which claims might be fraudulent.

Step 1: Understand your data

Clementine's visual programming interface makes examining and modeling the audit records straightforward. Clementine displays various data management, visualization and modeling procedures as nodes or icons and groups them by function at the bottom of the screen. You route data from one node to another by clicking and dragging the nodes into the screen's palette, and then connecting them with an arrow. This first screen shows a very simple data path (called a "stream" in Clementine). This stream routes data from an SPSS data source into a table via a type node, which automatically sets the appropriate variable types for use further down the stream. Note: Clementine can read any ODBC-compliant data source file.

The resulting table shows the data that we will use in our example: Medicare billing records with detailed information such as recipient/provider codes and county of residence, diagnostic codes, admission source, length of stay and total charges claimed.

The screenshot shows the Clementine software interface. A data stream is visible, starting with an SPSS data source node (labeled 'table1002.dat') connected to a 'Table' node. The 'Table' node displays a preview of the data, which is a table of Medicare billing records. The table has the following columns and data rows:

RECFIDR	RECDCTY	PRVNOB	PRVLCY	TITWRDS	LOS	DRGCODE	AMCDSD	CDRDESC
45515786.0	67	22348023	67	46N.0	14	311	2 - Urgent	4
45278360.0	52	12861704	60	1476.8	12	311	2 - Urgent	4
40803060.0	17	14518718	85	1673.0	4	311	2 - Urgent	4
36633660.0	38	12861704	60	844.0	9	311	2 - Urgent	4
31101060.0	68	12861704	60	708.0	6	311	2 - Urgent	4
16744000.8	17	14518718	85	494.0	10	311	3 - Elective	4
16354400.8	62	12861704	60	1384.6	12	486	1 - Emergency	7
16321100.8	65	14518718	85	234.0	7	486	1 - Emergency	7
15868000.8	65	14518718	85	1672.0	4	486	1 - Emergency	7
15318700.8	34	14518718	85	878.0	11	486	1 - Emergency	7

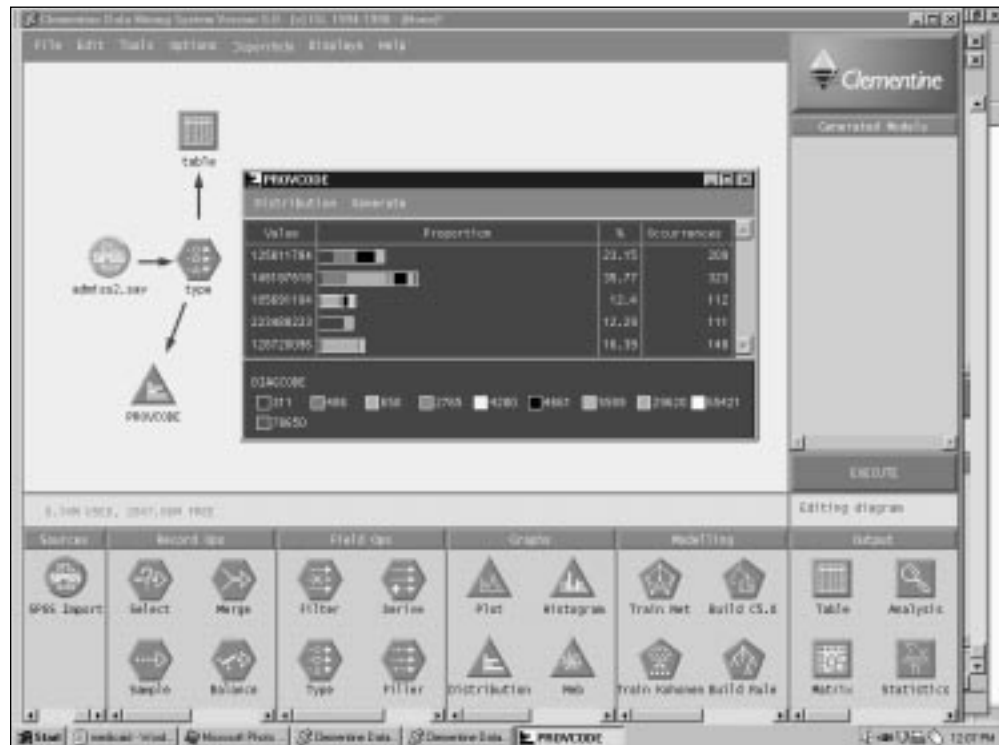
The interface also shows a palette at the bottom with various nodes categorized into Sources, Record Ops, Field Ops, Graphs, Model Ops, and Output. The 'Table' node is currently selected in the Output category.

Route data from one node to another by clicking and dragging the nodes into the screen's palette, and then connecting them with an arrow.

Step 2: Determine your population makeup

In this step, we route the data to a distribution node, PROVCOD, to display the frequency that the different providers appear in our dataset. Then, we color-code each frequency bar with the diagnostic codes associated with each provider.

We want to ensure that our data doesn't disproportionately represent any provider or exclusively associate any provider with one particular diagnostic code. Neither appears to be the case.



Examine frequencies to gain a better understanding of the data.

Step 3: Discover relationships in your data

Compute and add new variables that might prove useful in our analysis by inserting a derive node into our stream. Our dataset contains the county of both the provider and recipient in each of the billing records. Since this information might be useful, we derive a new variable called “county.” This variable will carry the value T (true) if the recipient’s county is the same as the provider’s county, and F (false) if otherwise. Clementine adds the new variable to each record in our dataset.

We then proceed to route the data (now with the additional variable, “county”) to a node that will build a Web graphic. Specifically, we want to examine how frequently (and for which diagnostic codes) each provider filed claims for out-of-county services. Darkening lines connecting each variable value indicate frequency. For instance, providers 125611784 and 145197518 were more frequently associated with the value of F for county (provider county not the same as recipient county) than were any other providers. This information might prove useful further along in our analysis.

The screenshot displays the SPSS Clementine interface. At the top, a 'Field' dialog box is open, configuring a new field named 'county'. The 'Type' is set to 'Flag', with 'True value' as 'T' and 'False value' as 'F'. The 'True if' condition is 'PRCP_CTY = PRV_CTY'. Below the dialog, a workflow diagram shows a 'table' node connected to a 'type' node, which then branches into 'PRVCOB' and 'COUNTY' nodes. The 'COUNTY' node is connected to a 'PRVCOB x DIAGCODE' node. A 'Web' graphic node is also present, displaying a network visualization. The network shows nodes for provider IDs (125611784, 145197518, 12572239, 12548112) and 'county' values (T, F). The lines connecting these nodes are darkened to represent the frequency of associations. The bottom of the interface shows a toolbar with various data mining nodes like 'SPSS Input', 'Select', 'Merge', 'Filter', 'Join', 'Split', 'Balance', 'Type', and 'Filter'.

Visually discover if a relationship among the variables exists. Examine associations using a Web graphic.

Step 4: Build a model

In this step, we first model the total charges on a Medicare claim, using the admissions source, length of stay and diagnostic codes as inputs. We chose a modeling procedure called rule induction because it is easy to understand. When we browse the model, we discover that a diagnostic code of 4280 will somewhat consistently produce charges of \$3,960.30, whereas, the charges incurred for diagnostic codes 4661 and 486 vary based on a patient's length of stay.

The model itself is represented by the gold, diamond-shaped node labeled TOTCHGS. When we insert the model into the stream, the model will read the inputs (admissions source, length of stay and diagnostic code) for each record, and then produce a projected value for total charges, labeled \$R-TOTCHGS. We'll use this new value later on.

The screenshot shows the SPSS Data Mining interface. The main workspace contains a flow diagram with several nodes: 'admission', 'type', 'count', 'PROCEDURE', and 'TOTCHGS'. A pop-up window titled 'workbuilt Rule browser 1 for totchgs' displays the following rules:

```

BILCODE 4280 => 3960.3
BILCODE 4661 [len: 4280.0], Effect: +134.2810
LOS 1-5 => 3960.30
LOS 6-5 [len: 4280.0], Effect: +666.00
LOS 1-18 => 4141.4
LOS 19-19 [len: 4280.0], Effect: +1809.40
LOS 4-22 => 3280.0
LOS 4-23 => 3280.0
BILCODE 486 [len: 486.0], Effect: +101.8731
LOS 4-4 => 101.87
LOS 4-4 [len: 486.0], Effect: +480.10
LOS 1-3 [len: 486.0], Effect: +621.06
LOS 1-5 => 3322.0
LOS 6-5 [len: 486.0], Effect: +941.064
LOS 6-8 [len: 486.0], Effect: +605.30
LOS 1-9 => 8791.0
LOS 6-9 => 8041.0
  
```

The bottom of the screenshot shows a toolbar with various data mining operations like 'Select', 'Merge', 'Filter', 'Derive', 'Plot', 'Histogram', 'Train Net', 'Build CS', 'Table', and 'Analysis'.

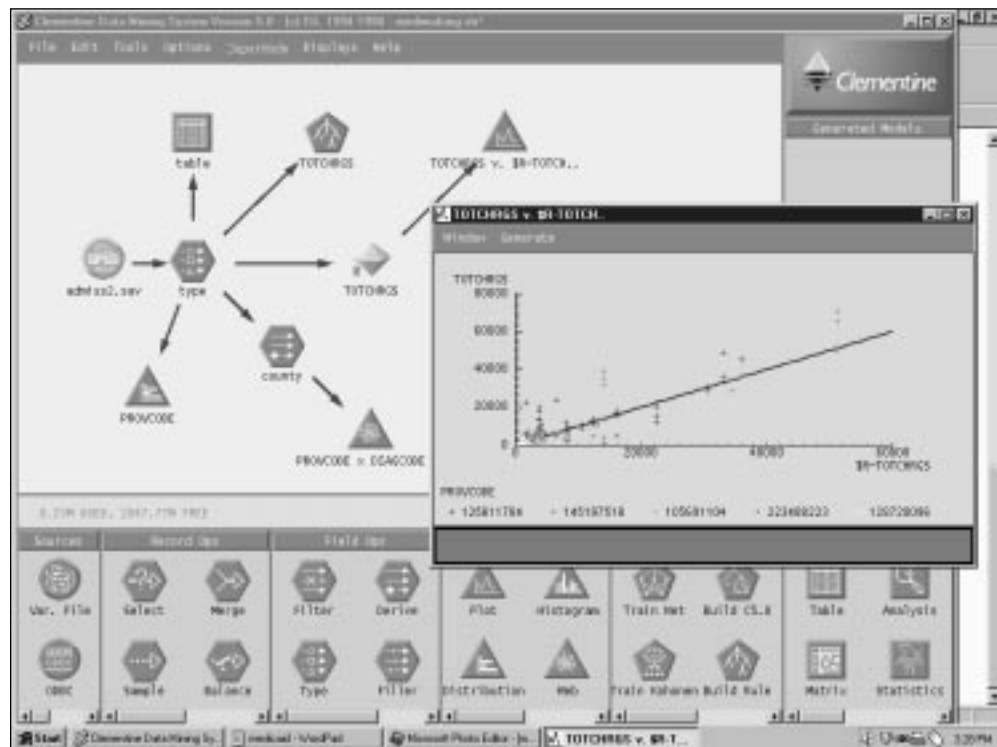
This model helps us determine a typical charge for a specific billing code.

Step 5: Use the model against actual records

To examine the difference between the actual charges recorded on each Medicare claim and the charges that our model projected, we will graph one against the other in a plot graphic. We color-code each record with its diagnostic code. (See the graph below. Each record in our data is represented by a plus sign.)

As part of the graphic, we also added a graph of the line, $y=x$. If the actual charges equal the projected charges on a particular record, that record's point should fall on this line. On the other hand, if the record's actual charges were greater than what the model projected, the record's point would be above this line.

It appears that the points above the line belong to two particular providers, 223488223 and 104269120.

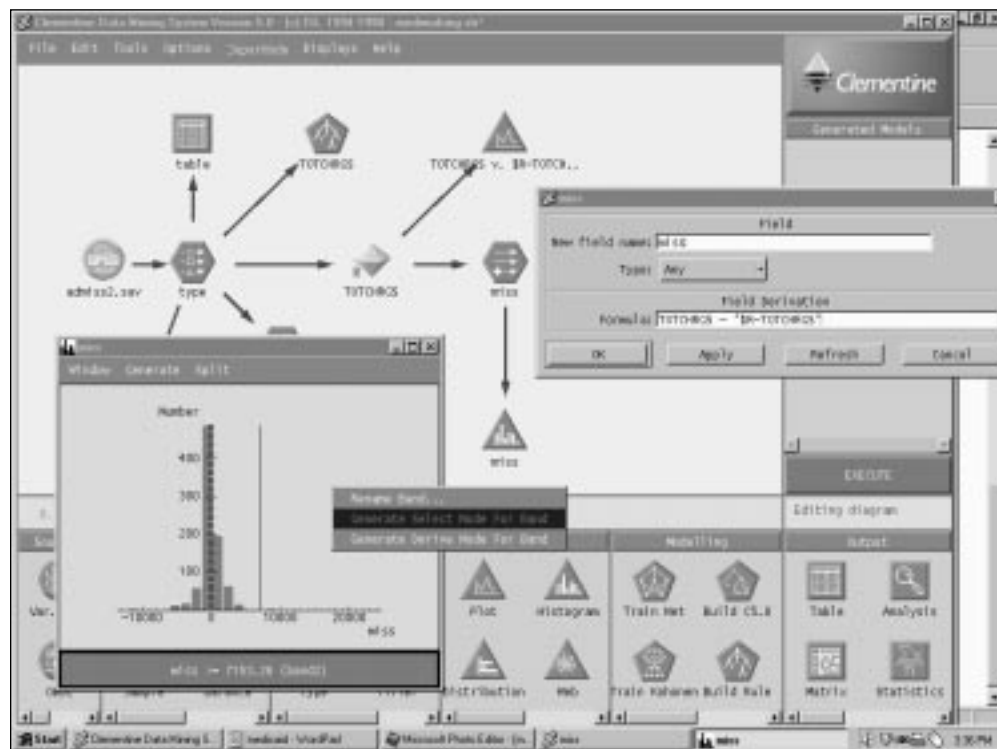


Using the model, we can compare actual charges on Medicare claims against charges the model provided.

Step 6: Segment your data

To further drill down on the differences between actual and projected charges, we derive a new variable to add to our data. This new variable, “miss,” is then graphed in a histogram. It shows that the majority of records in our data had a miss value clustered very close to \$0. However, few records have a miss value that extends to \$10,000 and beyond.

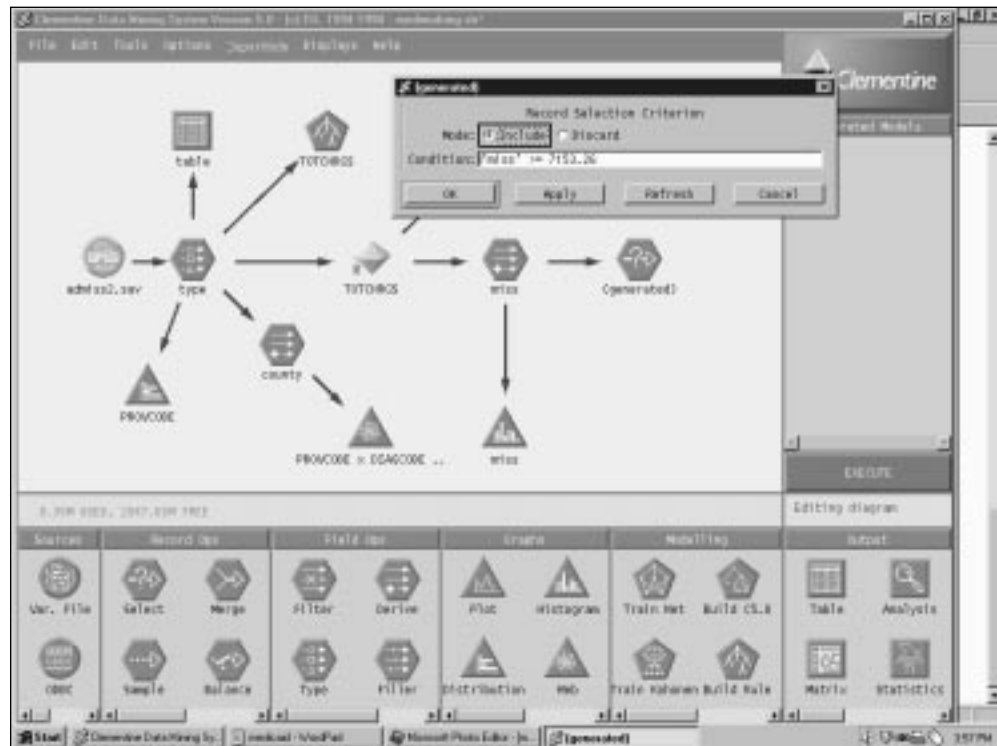
Let’s now narrow our attention to only those latter records. Select a point on the horizontal axis of the histogram, and then request a select node. It automatically generates and selects only those records whose miss value falls to the right of the \$10,000 point.



Let’s see the extent of differences between the actual costs and the projected costs.

Step 7: Pinpoint records that have a higher chance of being fraudulent

After generating this select node, we can insert it into the stream. In doing so, we now examine only the records that meet this selection criterion.



Easily segment your data to use only those records where the difference between actual and projected costs exceeds \$10,000.

Step 8: Compare your subset to the entire population

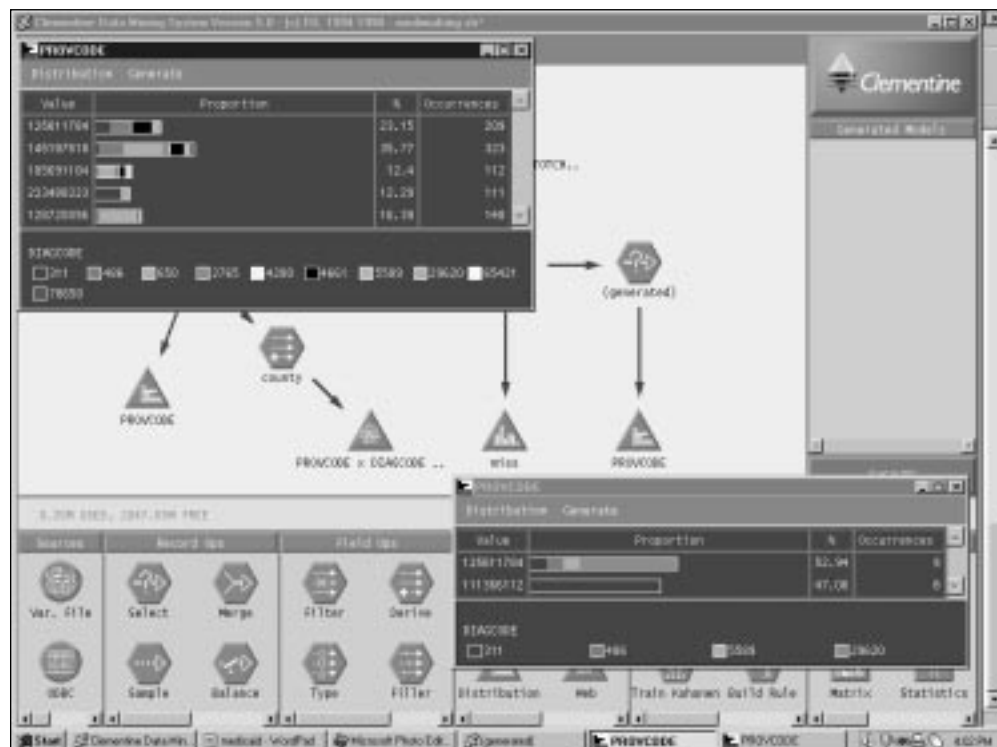
We now direct the selected records to a distribution graphic node and examine the frequencies with which each provider appears in the records. The graphic appears in the lower right part of this screen. Next, we compare the subset of the data against the makeup of the original file of all records to see if any providers are disproportionately represented with records that exceed the projected \$10,000. This graphic appears in the upper left-hand side of the screen.

The records showing a large disparity between actual charges claimed and projected charges belong to two particular providers — neither of whom were disproportionately represented in the complete dataset.

Further, since we've color-coded each of these two providers' frequencies, we can easily tell how often they file specific diagnostic codes. One of the providers filed exclusively a particular code; the other submitted a variety of codes.

Data mining proved useful for two reasons. First, it yielded valuable information about potential cases of fraud in the records currently on file. Although by no means an open-and-shut case of fraudulent claims submittal, the evidence we gathered can now be passed to investigators. They can focus on claims that may yield larger adjustments.

Plus, the model we built can be applied to future claims. The model will compute total projected charges on incoming claims. These projections can be compared against actual charges, and the system will “flag” questionable claims for investigation.



Only two providers claimed charges at least \$10,000 more than the model projected they would file.

Although this Medicare fraud detection office used data mining for provider fraud, you could also use it for eligibility fraud.

And, because circumstances change over time, you can periodically review the models and update them so they continue to be effective and deliver the best results.

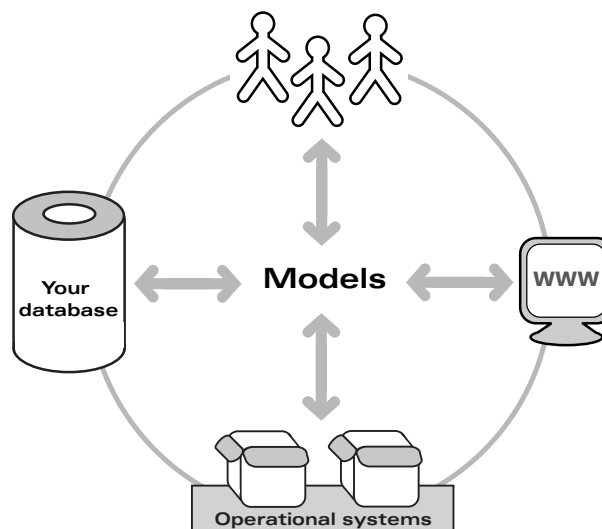
Strategically deploy your data mining results for optimum success

Once you have models that predict fraudulence, you need to strategically deploy your results to the people who can use that information to eradicate fraud and recoup money. Only by using your results can you meet your data mining objective.

Strategic deployment means integrating models into your agency's daily operations. When you implement deployment in your agency, a front-line manager, for example, can use the deployed models to feed new information into the models and get results. Strategic deployment empowers you to put timely, consistent information into the right hands. Everyone in your organization is on the same page and can act more quickly to recoup the most money for your agency.

SPSS deployment technology enables you to score values against new claims and payment requests, then deploy your model and distribute the results (for example, a list of claims most likely to be non-compliant). Depending on your needs, you can display these results over an intranet, through e-mail, or via hard-copy reports.

For agencies that have local or branch offices, strategic data mining deployment provides an additional benefit. The central office can store and mine data for the entire organization and deploy the data mining results to local offices, which are often charged with stopping fraud and abuse. Deploying data mining results to local offices can stretch scarce resources, empowering you to consistently share information throughout the organization.



You can maximize the value of your data mining investment by deploying data mining models to decision makers, virtual decision makers, databases and operational systems.

Solve your complex business problems

You can track and solve critical business problems using the best practice approach to data mining, Cross-Industry Standard Process for Data Mining (CRISP-DM). CRISP-DM is a comprehensive data mining methodology and process model that makes large data mining projects faster, more efficient and less costly. SPSS subscribes to this best practice approach to data mining — which it co-authored with several other leading companies — and brings it to your organization to deliver actionable results. For more information about CRISP-DM, see www.crisp-dm.org.

Data mining makes the difference™

Discover patterns that indicate which claims have a high probability of fraudulence when you apply sophisticated data mining techniques to your past claims data.

SPSS provides sound, objective solutions to help you use data and technology better and improve your ability to recoup significant amounts of money. SPSS teams with you to merge deep analytical and technical knowledge with your business expertise. Along the way, SPSS experts educate your staff, provide recommendations and build a repeatable process so your organization can apply the skills and tools to easily proceed on your own.

Analyze your data using a variety of techniques, from simple reports to advanced methods for predicting provider, client or vendor behavior. It's important to use multiple analytical methods so you can work with many types of data in many applications and always get answers that lead to the best chance to realize substantial claim adjustments.

SPSS has over 30 years experience providing analytical solutions to governments and industries, and offers you the analytical depth, breadth and expertise to ensure you get accurate results. SPSS works with you to make sure the data you have is ready for analysis and may also collect additional data that will improve your results.

Data mining makes the difference

SPSS Inc. enables organizations to develop more profitable customer relationships by providing analytical solutions that discover what customers want and predict what they will do. The company delivers analytical solutions at the intersection of customer relationship management and business intelligence. SPSS analytical solutions integrate and analyze market, customer and operational data and deliver results in key vertical markets worldwide including: telecommunications, health care, banking, finance, insurance, manufacturing, retail, consumer packaged goods, market research and the public sector. For more information, visit www.spss.com.

Contacting SPSS

To place an order or to get more information, call your nearest SPSS office or visit our World Wide Web site at www.spss.com

SPSS Inc.	+1.312.651.3000	SPSS Israel	+972.9.9526700
Toll-free	+1.800.543.2185	SPSS Italia	+800.437300
SPSS Argentina	+5411.4814.5030	SPSS Japan	+81.3.5466.5511
SPSS Asia Pacific	+65.245.9110	SPSS Korea	+82.2.3446.7651
SPSS Australasia	+61.2.9954.5660	SPSS Latin America	+1.312.651.3539
Toll-free	+1.800.024.836	SPSS Malaysia	+603.7873.6477
SPSS Belgium	+32.16.317070	SPSS Mexico	+52.5.682.87.68
SPSS Benelux	+31.183.651.777	SPSS Miami	+1.305.627.5700
SPSS Brasil	+55.11.5505.3644	SPSS Norway	+47.22.40.20.60
SPSS Czech Republic	+420.2.24813839	SPSS Polska	+48.12.6369680
SPSS Danmark	+45.45.46.02.00	SPSS Russia	+7.095.125.0069
SPSS East Africa	+254.2.577.262	SPSS Schweiz	+41.1.266.90.30
SPSS Federal Systems (U.S.)	+1.703.527.6777	SPSS Singapore	+65.324.5150
Toll-free	+1.800.860.5762	SPSS South Africa	+27.11.807.3189
SPSS Finland	+358.9.4355.920	SPSS South Asia	+91.80.2088069
SPSS France	+01.55.35.27.00	SPSS Sweden	+46.8.506.105.50
SPSS Germany	+49.89.4890740	SPSS Taiwan	+886.2.25771100
SPSS Hellas	+30.1.72.51.925	SPSS Thailand	+66.2.260.7070
SPSS Hispanoportuguesa	+34.91.447.37.00	SPSS UK	+44.1483.719200
SPSS Hong Kong	+852.2.811.9662		
SPSS Ireland	+353.1.415.0234		

SPSS is a registered trademark and the other SPSS products named are trademarks of SPSS Inc. All other names are trademarks of their respective owners.

Printed in the U.S.A © Copyright 2000 SPSS Inc. DMFWP-1000