

# Mining the Mountain of Financial Discovery

By Jason Slick VP of Engineering - Cynic Inc

In today's global economy the importance of data organization and classification has become the cornerstone on which a financial investigation relies. Everything from converting digital data to searching information for classification has become an arduous task as fraud cases have become more and more complicated. To address this increase in data complexity, a few standardizations of techniques and technology can make a huge difference in the number of hours needed to take a fraud investigation from start to finish. In this article, I will discuss the various techniques and processes for building financial fraud cases using computer technology. From data entry and conversion to advanced search algorithms that aid in classification, managing a financial fraud case has never been easier to do, if you know the right techniques and tools to use.

Since this article is written for the average fraud investigator, I think it is important to clarify some of the more advanced technical words I am going to be using.

## Important Definitions

**Optical Character Recognition**, abbreviated as **OCR**, is the mechanical or electronic translation of images of handwritten, typewritten or printed text into machine-editable text.

**Data Object** is a digital representation of a real-world object's state and characteristics.

A **Method** provides a mechanism for accessing and manipulating the encapsulated data stored in an object.

**Data Mining** is the process of searching and sorting through large amounts of data and picking out relevant information.

An **Algorithm** is a sequence of instructions, often used for calculation and data processing.

Now I know you are saying to yourself, "oh no, I cannot learn advanced computer terminology" or "I better let my computer guy read this", but stick with me and I will try and clarify the importance of having a basic understanding of these terms. Even if you use some software that has these features hidden from you in a nice user-interface, it is always good to understand the concepts that make the computer work for the modern-day fraud investigator.

## Data Gathering, Entry and Conversion

With the advent of technology, the data entry process has become the most important step in building a financial fraud case that is both accurate and complete. Whether the data entry process is manual or computer-aided, the information being precise is of the utmost importance. In this section, I will discuss some of the various data entry techniques that are available to the modern financial investigation team.

Whether you are an old-pro or new to the business, everyone is familiar with the discovery that is important to any fraud case. Bank records, credit card statements and receipts make up most of the information and getting it translated into a form which is usable by the computer can be a daunting task. The computer is not smart enough yet to show it a piece of paper and have it recognize the type of document, then enter the data and organize it. This is where the importance of the data entry group and the eyeballs of humans play the most important role. You need to have people verify that any information that is inputted into the computer is accurate. There is no point for a fraud examiner to even analyze data that is wrong; the case is only as good as the information obtained. Having a person who can validate the integrity of data entered whether manual or using Optical Character Recognition is important. Humans and computers can mistakenly enter 5's for S's and commas for periods, this makes for errors in the case management process. Catching these problems is the job of the data entry group and there are many tools that can help.

The single most important part of your data entry group is obtaining the discovery for analysis. Whether your evidence comes in the form of boxes and boxes of paper records or digital documents, proper organization of data saves much as far as time and headaches are concerned. If you do not have a PDF or any other digital type of the document, you must first get the information into a digital image. Using a scanner is the best method. Most scanners work easily with programs such as Adobe Acrobat and Microsoft Word for making useable images of documents. The most important setting during scanning is quality of image; I set all my scanners at 300 dpi setting (dots per inch) for any document. The higher the resolution the better chance the software has for accurate translation, which also makes it easier for a human to read. But if you have your dpi settings too high your file sizes can be large, which in turn makes the amount of storage space for each document increase. The second most important thing when obtaining data is getting the good data. Bank and credit card statements are the best form of financial discovery; while receipts and invoices are good for data organization, actual statements from the banks and companies can be more easily verified for accuracy of the data. Once you get all of your documents, you will need to apply Bates Stamps to the entire digital discovery. There are many different options for software applications that can help with bates stamping, but Adobe Acrobat has a feature that allows you to add a bates stamp easily to the page headers and footers.

*Hint: Run your scanner with a small piece of paper, if the extra space of the document is a color other than white you need to find a way to make it white. Sometimes it is a setting in the scanning software or you have to actually cover a background in white. This will save you a lot of headaches with printing and reading documents later.*

The tried and true methods of manual data entry have been overtaken by the advances Optical Character Recognition, referred to as OCR from here out. While it can take up to two minutes for a human to enter one bank transaction, the computer can read a scanned document and translate the image to text in seconds, if not quicker. There are many applications that work well at character recognition and finding one that is useful for you

is important. How they work is simple. Once you have the image of the document, use your OCR application to translate or recognize the characters then save the data in whichever text or file format you plan on using. Most OCR programs can handle word, excel and other file types, even export the data as text using comma or tab delimited values. Using some sort of delimiter gives the data a consistent marker that aids with data separation and organization, while maintaining an application neutral data file. Open the file in something such as Notepad or other simple text editor; from here you can use the features of copy and paste to further organize the data. What you are looking for is consistency of information layout and one line of data per transaction. Make sure word wrap is turned off for the application and verify each transaction has the same number of entries.

Example of simple bank transactions that are tab delimited

```
02/01/2010 08:01:00 150.23
02/01/2010 09:30:25 -75.00
datetime3      amount3
```

In this example the transaction can be either debit or credit depending on whether the values is positive (credit) or negative (debit). Even though you cannot perceive it, there is hidden code, \t which means tab, which separates the datetime from the amount. The computer can now read the example and be instructed to separate the two values from each other using the tab as the delimiter. By creating patterned data the amount of manual data manipulation is decreased through consistency. In fact, there are many occasions where the time of OCRing just a few transactions is slower than manual data entry so being consistent with your delimiters is the key.

*Hint: Try and create delimiters that are unique and have almost no chance of being in the text you are working with*

*Example of simple bank transactions that use a custom text delimiter of %\$%*

```
datetime1%#%amount1
```

```
datetime2%#%amount2
```

Either way the data has to be verified as correct and laid out in such a way to be useful for the investigator. Using tools like Microsoft Excel or most any spreadsheet program, you can import the delimited text easily with all data going into their respective columns. Some spreadsheet applications, like Excel, allow the programming of macros and algorithms that can be used to calculate the running balance of transactions as they apply to the bank or credit card statements. There are also many features embedded into spreadsheet programs, like sum total, that can perform these calculations quickly. As an example a bank statement has a daily balance included on the statement, using the sum of the transactions a person can compare the digital data with what the statement says. If you are missing any information or there is an error in typing like extra numbers (50.00 instead of 5.00), the error will be noticeable because of the deviation from the statement or failure of the application to perform as intended.

## Data Objects and the Financial Investigator

Data Objects are not an easy concept to grasp for some, but I feel it is important to have a small knowledge of them to understand how computers handle data efficiently. From the previous definition given, a Data Object is a digital representation of a real-world item. What does this mean? In the case of bank account, this means that a Bank Account Data Object has certain properties and associations that are universal to every bank account. Bank accounts always have an account number and opened date. Bank Accounts also can contain other data objects within them, like bank transactions. Bank Transaction Data Objects have their own properties such as datetime of transaction, whether the type is a debit or credit, and the amount of the transaction. By creating data objects for every possible financial transaction, we can build a set of rules that the computer can interpret and aid the investigator in case management.

I know this concept is hard to understand, so I am going to try and clarify the Data Object concept. Let us look at an apple and orange, and create a fruit data object that can be used to store information about both. What we do when creating a data object is find ways to generally describe the properties of the items we are classifying. Whether it is the color or flavor, all fruit share similar characteristics that can be organized into an easy to manage data object.

### Fruit Data Object

Name

Description

Skin

- Color

- Flavor

- Edible

Meat

- Color

- Flavor

- Edible



Now that you can see what the Fruit Data Object looks like, imagine using the data object to classify all of the different types of fruit. This simple object covers quite a few of the attributes of the various types of fruit and makes it easier for us to explain to another person, or computer, what type of fruit we are talking about. Using the data object as a baseline, we can add more attributes to store more information or even add Methods that can be applied to the fruit object in general. Now let us apply an easy to understand Method to our Fruit Data Object, something like EatFruit. With a method like EatFruit, it is easy to apply general functionality to all fruit objects, like consume the meat of the fruit.

Now that we have a basic understanding of how data objects work, we can apply the same concepts to accounting principles, like create a bank account and the transactions in that account. With data objects the computer can determine whether a field must meet some type of parameter or criteria, like the date being in a particular format like mm/dd/yyyy or amount of transaction being of the money type. This allows for the computer to warn the user, or another computer application, of data entry errors that need to be addressed. This approach can guarantee that check amounts are money fields and

names need to exist for our case's subjects. By using data objects we can maintain minimum standards of data that guarantee our search algorithms and association techniques will function correctly, giving our case building and reporting more robust features.

## Data Mining and Classification

A very important feature that an investigator needs is the ability to search, sort and classify the data. Without searching and sorting features the classification of financial transactions can be a laborious. The investigator needs to be able to classify transactions as business expenses, income, investments, etc. efficiently and consistently. In this section I will discuss the various methodologies available to the investigator that help with data organization and classification.

The first thing a financial investigator needs to do with classification is to first organize the recurring items whether they are expenses or incomes. Such as a business might have an office with a rent payment, electricity, phones and internet that are due and paid every month. These transactions have a tendency to be paid from the same account or even in the same manner consistently, like always by check or using a debit card. These transactions usually always have the same information profile, like notations in the bank statements. An example of this is transactions for the cable tv/internet might always have a check being written to Cable Co; by using the name of the company in a simple search we can easily find every transaction that has a reference to that name. Simple search is just typing a word or words in exactly as you expect them to appear from within some text. Simple search and sorting techniques help by narrowing the data set of available financial objects that require more classification by categorizing repeating data more effectively. If there are on average ten checks written by a company every month and six of those checks are for company bills, we have limited the unorganized data set by 60%. This greatly reduces the size of the unorganized data and can effectively aid the investigator in more quickly recognizing patterns of fraud or uncharacteristic spending habits.

The second pass on the data needs to determine the movement or transfer of money between accounts. Sometimes the transfer of money is easy to spot because there is a direct one-to-one correlation between the values, like a \$1000 debit from one account becoming a \$1000 credit in another account. These transfers sometimes can be hard for an investigator to spot when working with multiple accounts. This is where a simple search on the amount fields can show just the transactions with that particular amount. Other times searching is more difficult, as an example some of the money was transferred to an asset like a car or even put in their pocket as cash. Anytime that money is taken out of a bank account and spent as cash, a transfer of money needs to be made to a virtual transaction in the subject's cash account. A virtual transaction is a representation of an assumed value or item that needs to be accounted for through reconciliation. An example of this is if the person says they paid cash for a \$2000 television, the case data needs to reflect the purchase from the cash account. Sometimes a transaction is split up, like a \$1000 debit going to pay a \$500 company expense while the other \$500 went into the person's pocket. Being able to handle the breaking up and joining of transaction data is a

cornerstone to the financial investigation. By sorting out all of these account transfers we have reduced the chance of tabulating the transactions more than once, because as example a deposit of money from another one of the subject's account is not income and should not be counted as such.

By time we reach the third pass we should have over 90% of all data organized and need to apply only few more advanced searching techniques to further organize the case. There are many different methods for searching that can be applied to financial objects; we are only going to discuss a few. The math and computer science behind these search techniques are a bit technical for this article but I feel it is important to mention them. Fuzzy searching is the name given to finding strings of text that has an approximate match, like finding ball when looking for bill. Soundex searching is looking for words that sound similar in the English language, such as pelt and belt. Wildcard searching uses a character that can be substituted for any subset of a string; wildcards are usually represented by the \* key (star). An example of this is searching for fra\* should reveal results like France, fray, Frank, frazzle, etc. Fuzzy and Soundex searching are good tools for finding data where typing errors have occurred. Just because we have built our data objects properly does not mean that errors in data entry did not occur. A data object value that is a text field will accept '5+3ve' just as it will 'Steve'; if you get lots of errors such as this there is a breakdown in your data entry process somewhere that needs addressed

The rest of the investigation after the third pass just comes down to solid fraud investigator work, mental data mining if you will. Remember what you might be looking for could have been categorized or sorted in the data mining process. Smart criminals try and hide fraud through fake businesses, employees and other methods, but if you have organized your data correctly when you make that connection every related transaction is already at your fingertips.

