Understanding the Process Of Data Mining



As in the Stone Age, Bronze Age, and Industrial Age before it, the Information Age has produced a range of technologies that have irrevocably changed the way we live our lives.

One of the most prominent, and important, of these technologies is Big Data.

With every purchase made, every social media profile created, and nearly every type of media available for download, review, and offline consumption or analysis, organizations and individuals generate vast amounts of information every second of every day.

Hidden within all this information are useful insights organizations can use to drive continuous improvement, develop innovations, and achieve competitive advantage.

To reveal and use them, you need to understand the data mining process, best practices for using data mining in your own organization, and the tools you can use to ensure optimal process management and implementation.

What is the Data Mining Process?

If, like a lot of other businesses, you're aware of Big Data's potential and want to harness it, you need a system in place, supported by formal policies and practices for data collection, processing, integration, and transformation through data analysis.

Like miners boring into the earth in search of the motherlode, companies use the data mining process—also known as *knowledge discovery*— to chew through vast amounts of information to unearth valuable actionable insights.

Most data mining systems collect data from a wide range of sources, from internal sales and financial databases to social media to vendor compliance records to data warehouses and more.

Using advanced data analysis tools like machine learning, the raw data is cleaned, prepared, sorted, integrated, and refined to extract relevant information and identify interesting patterns that can be further analyzed to improve decision making, achieve process improvement, or enhance the accuracy and completeness of forecasts and reporting.

Like refining ore, data mining is an *iterative process*.

The same data may be refined multiple times using data analysis in order to improve data quality and the results generated.

No two businesses will approach the data mining process in identical ways, but in general the process will look something like this:

1. A company identifies a business requirement to be satisfied.
2. Sources of potential raw data, and their sources, are identified.
3. A data model is built based on the available data.
4. A data structure, based on the data model, is built.
5. The data structure is mined for useful information, interesting patterns, etc.

Because the process is iterative, steps 2-4 may be repeated multiple times as new data sources are added or updated.

Like miners boring into the earth in search of the motherlode, companies use

the data mining process—also known as knowledge discovery— to chew through vast amounts of information to unearth valuable actionable insights.

Data Mining Process: Step by Step

You've identified your business requirements, chosen your sources, and are ready to get mining. But before you can move through the five-step data mining process, you need to ensure you're using the best possible available data.

That's why the data mining process itself is actually *two* processes: *data preprocessing*, followed by the actual data mining.

Data preprocessing was developed to ensure the data being mined is on TRACC:

- Timely
- Relevant
- Accurate
- Complete
- Consistent

Data that meets the desired standards will prove much more useful than unrefined information.

Data preprocessing involves four steps; data mining, three. The total process spans seven distinct steps:

STEP PERFORMED	PROCESS TYPE
Data Cleaning	Data Preprocessing
Data Integration	Data Preprocessing
Data Reduction	Data Preprocessing
Data Transformation	Data Preprocessing
Data Mining	Data Mining
Pattern Evaluation	Data Mining
Knowledge Representation	Data Mining

1. **Data Cleaning** removes inaccurate, incomplete, or otherwise "dirty" data

(i.e., erroneous data) from your sources. Data is cleaned by either restoring missing data or removing the dirty data.

Missing data can be added manually, replaced with a calculated mean or average, or simply replaced with the most probable value as calculated by your team.

Noisy or dirty data can be removed using *binning,* a process that sorts data sets into virtual "bins" (also called buckets), which are then analyzed to find the median value based on the minimum and maximum value for each bin.

Alternatively, values can be replaced with the nearest boundary value (either the minimum or the maximum).

2. **Data Integration** collects and combines all your assorted data sources into a single source suitable for analysis and manipulation. Integrating all your available data in this way improves both the speed and accuracy of the actual data mining.

Data preparation and integration are essential, because different sources often have different names for similar or identical variables, or express them in different ways, creating redundant entries that must be parsed by your data mining tools.

Using data integration tools such as Online Analytical Processing (OLAP), Online Transaction Processing (OLTP) can help bridge the gap between (for example) two sources using data warehousing and databases, respectively.

3. **Data Reduction** refers to "slicing and dicing" data to obtain the most relevant information from the larger whole, without disrupting the overall integrity of the data sources or the samples taken.

Data reduction relies on a number of data analysis techniques, including:

- **Data compression**, which provides a compressed "thumbnail" of the source data.
- **Decision trees**, a type of algorithmic tool used to follow multiple potential paths to a desired goal and then identify the most effective one.
- **Neural networks**, which combine the power of machine learning and deep learning (specific applications of artificial intelligence) to combine multiple algorithms into a single entity meant to emulate a human brain's

ability to parse information and identify patterns.
   - **Dimensional and Numerosity Reduction,** two refinement tools that seek to separate the virtual wheat from the digital chaff by streamlining data sets.

4. **Data Transformation** Cleaned, reduced, and optimized data is further transformed in this stage. Data is refined even more thoroughly, smoothing away outliers, summarizing data sets where applicable, replacing raw data values with ranges (i.e., discretization), etc.

5. **Data Mining**, using the five-step, iterative process to the clean and optimized data.

6. **Pattern Evaluation,** wherein the patterns uncovered during data mining are analyzed and converted to useful information understandable to end users, e.g. seasonal buying patterns that indicate an opportunity to capture additional sales during periods of peak demand.

7. **Knowledge Representation** converts the useful information into multimedia formats for further review, analysis, and presentation. The insights gleaned during pattern evaluation can be used, for example, to create a sales forecast, supply chain adjustments, new production schedules, etc.

Common Data Mining Models

Data mining requires the use of data models, which are distinct approaches developed to achieve specific data mining goals.

Two of the most common are the Cross-Industry Standard Process for Data Mining (CRISP-DM) and Sample, Explore, Modify, Model, and Assess (SEMMA).

CRISP-DM is cyclical, iterative, and versatile. Steps can be performed in any order, but must be completed to achieve the desired results.

Crisp-DM has six phases:

1. **Business Understanding** (Organizational goals established; steps necessary to meeting said goals documented).

2. **Data Understanding** (Data collection and population within data analysis toolset. Data is organized by source, location, acquisition method, and potential errors, then visualized for further review.)
3. **Data Preparation** (The most useful data is selected, cleaned, and integrated across multiple databases.)
4. **Data Modeling** (Data mining techniques chosen; data models built and tested; models are reviewed for completeness and utility.)
5. **Evaluation** (The data model is reviewed for utility, completeness, and ability to meet established business requirements.)
6. **Deployment** (A deployment plan is created; processes used to monitor data mining for utility and accuracy; process review is used to determine whether further refinements to the model are necessary or any stages need to be repeated to accomplish the desired business goals or accommodate new business requirements.)

SEMMA was created by the Statistical Analysis System (SAS) Institute and is designed for flexible exploration of data models of varying complexity.

Like CRISP-DM, it is an iterative data mining model, but takes a different approach to data collection, refinement, and analysis.

The five steps of SEMMA include:

1. **Sample** (A sample representing the entire dataset is extracted and used as a statistical synecdoche to reduce demand on the data analysis tools.)
2. **Explore** (Data is reviewed for broad patterns; any outliers or other anomalies are noted for additional insights into the nature of the data set.)
3. **Modify** (Data is organized into groups and subgroups, with a focus on the business goals pursued.)
4. **Model** (Models are built to clarify patterns uncovered in analysis.)
5. **Assess** (The constructed model is reviewed for utility, accuracy, and completeness, using real-world datasets to test the validity of the model itself.)

Everyday Applications for Data Mining Technologies

It probably won’t surprise you to learn data mining applications are in use around the world by a wide variety of organizations.

Some of the most common applications include:

- **Consumer Behavior.** Companies across the retail sector harvest and analyze shopping habits, trends, and feedback across media streams to improve customer service, create and recommend new products, and, of course, sell more goods and services.
- **Network Security.** Data mining tools can identify patterns that indicate potential threats to network resources and help stop distributed denial of service (DDoS) attacks, data breaches, and site hacks before they begin.
- **Financial Analysis.** Banks, investment firms, credit services, insurance companies, and other financial institutions all use data mining to decide where to invest, determine who receives lines of credit, and how best to protect and build value as well as profits.

Data Mining in Procurement

One of the most productive and valuable places to begin pursuing your own data mining goals is the procurement department, using procurement software with data analysis and mining capabilities.

A comprehensive procurement software solution like PLANERGY will help you optimize key processes like your procure-to-pay (P2P) workflows with artificial intelligence, centralized data collection and management, and process automation.

But more importantly, by capturing, collecting, and organizing all of your spend data, it provides an outstanding starting point for data mining and analysis in general. It also provides a central point of collection and integration with data flowing from different data sources, including marketing, sales, accounting, and legal.

Whether you're trying to improve vendor compliance, develop a more robust, agile, and resilient supply chain, or ferret out inefficiencies in your internal workflows, mining your procurement data can help you begin the process of capturing and analyzing *all* of the Big Data that flows in and out of your business.

Convert Big Data into Big Value with Data Mining

Like raw ore, all your available data isn't producing optimal value if you're not refining it.

Once you understand the potential, particulars, and limitations, you can develop your own data mining plan to extract valuable strategic insights and other useful information from your data sources.

And by putting your data mining techniques to work with help from a best-in-class software solution, you can be sure you're getting optimal data quality and analysis to produce data mining results that help you meet all your business objectives, no matter what they might be.

What's your goal today?

1. Use PLANERGY to manage purchasing and accounts payable

We've helped save billions of dollars for our clients through better spend management, process automation in purchasing and finance, and reducing financial risks. To discover how we can help grow your business:

- Read our case studies, client success stories, and testimonials.
- Visit our "Solutions" page to see the areas of your business we can help improve to see if we're a good fit for each other.
- Learn about us, and our long history of helping companies just like yours.

Book a Live Demo

2. Download our guide “Preparing Your AP Department For The Future”

Download a free copy of our guide to future proofing your accounts payable department. You’ll also be subscribed to our email newsletter and notified about new articles or if have something interesting to share.

download a free copy of our guide

3. Learn best practices for purchasing, finance, and more

Browse hundreds of articles, containing an amazing number of useful tools, techniques, and best practices. Many readers tell us they would have paid consultants for the advice in these articles.

Related Posts