Cognitive Computing and Big Data Analytics
By Judith Hurwitz, Marcia Kaufman and Adrian Bowles

CHAPTER

6

Applying Advanced Analytics to Cognitive Computing

Advanced analytics refers to a collection of techniques and algorithms for identifying patterns in large, complex, or high-velocity data sets with varying degrees of structure. It includes sophisticated statistical models, predictive analytics, machine learning, neural networks, text analytics, and other advanced data mining techniques. Some of the specific statistical techniques used in advanced analytics include decision tree analysis, linear and logistic regression analysis, social network analysis, and time series analysis. These analytical processes help discover patterns and anomalies in large volumes of data that can anticipate and predict business outcomes. Accordingly, advanced analytics is a critical element in creating long-term success with a cognitive system that can ask for the right answers to complex questions and predict outcomes. This chapter explores the technologies behind advanced analytics and how they can be leveraged in a knowledge-driven cognitive environment. With the right level of advanced analytics, you can gain deeper insights and predict outcomes in a more accurate and insightful manner.

Advanced Analytics Is on a Path to Cognitive Computing

The role of analytics in an organization's operational processes has changed significantly over the past 30 years. As illustrated in Table 6-1, companies are experiencing a progression in analytics maturity levels, ranging from descriptive analytics to

predictive analytics to machine learning and cognitive computing. Companies have been successful at using analytics to understand both where they have been and how they can learn from the past to anticipate the future. They can describe how various actions and events will impact outcomes. Although the knowledge from this analysis can be used to make predictions, typically these predictions are made through a lens of preconceived expectations. Data scientists and business analysts have been constrained to make predictions based on analytical models that are based on historical data. However, there are always unknown factors that can have a significant impact on future outcomes. Companies need a way to build predictive models that can react and change when there are changes to the business environment.

Table 6-1: Analytics Maturity Levels

ANALYTICS TYPE	DESCRIPTION	EXAMPLES OF QUESTIONS ANSWERED
Descriptive Analytics	Understand what happens when using analytic techniques on historical and current data.	Which product styles are selling better this quarter as compared to last quarter? Which regions are exhibiting the highest/lowest growth? What factors are impacting growth in different regions?
Predictive Analytics	Understand what might happen when using statistical predictive modeling capabilities, including data mining and machine learning. Predictive models use historical and current/real-time data to predict future outcomes. Models look for trends, clusters of behavior, and events. Models identify outliers.	What are the predictions for next quarter's sales by product and region? How does this impact raw material purchases, inventory management, and human resource management?
Prescriptive Analytics	Use to create a framework for making a decision about what to do or not do in the future. The "predictive" element should be addressed in prescriptive analytics to help identify the relative consequences of your actions. Use an iterative process so that your model can learn from the relationship between actions and outcomes.	What is the best mix of products for each region? How will customers in each region react to advertising promotions and offers? What type of offer should be made to each customer to build loyalty and increase sales?
Machine Learning and Cognitive Computing	Collaboration between humans and machines to solve complex problems. Assimilate and analyze multiple sources of information to predict outcomes. Need depends on the problems you are trying to solve. Improve effectiveness of problem solving and reduce errors in predicting outcomes.	How secure is the city environment? Are there any alerts from the vast amount of information streaming from monitoring devices (video, audio, and sensing devices for smoke or poisonous gases)? Which combination of drugs will provide the best outcome for this cancer patient based on the specific characteristics of the tumor and genetic sequencing?

The next frontier, which comes with opportunities for enormous change, includes big data analytics and incorporates the technologies of machine learning and cognitive computing. As shown in Figure 6-1, there is a convergence of technologies cutting across analytics and artificial intelligence. One major push for this convergence is the change in the timing and immediacy of data. Today's applications often require planning and operational changes at a fast rate for businesses to remain competitive. Waiting 24 hours or longer for results of a predictive model is no longer acceptable. For example, a customer relationship management application may require an iterative analytics process that incorporates current information from customer interactions and provides outcomes to support split-second decision making, ensuring that customers are satisfied. In addition, data sources are more complex and diverse. Therefore, analytic models need to incorporate large data sets including structured, unstructured, and streaming data to improve predictive capabilities. The multitude of data sources that companies need to evaluate to improve model accuracy includes operational databases, social media, customer relationship systems, web logs, sensors, and videos.

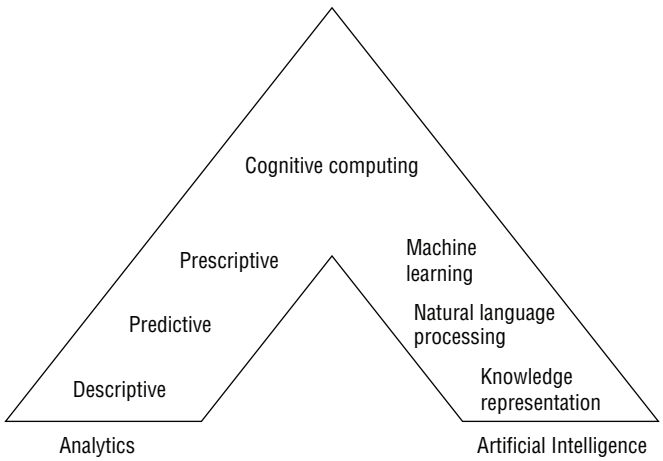

Figure 6-1: Converging technologies: analytics and artificial intelligence

Increasingly, advanced analytics is deployed in high-risk situations such as patient health management, machine performance, and threat and theft management. In these use cases the ability to predict outcomes with a high degree of accuracy can mean lives are saved and major crises are averted. In addition, advanced analytics and machine learning are used in situations in which the large volume and fast speed of data that must be processed demands automation to provide a competitive advantage. Typically, human decision makers use

the results of predictive models to support their decision-making capabilities and help them to take the right action.

There are situations, however, in which pattern recognition and analytic processes lead to action without any human intervention. For example, investment banks and institutional traders use electronic platforms for automated or algorithmic trading of stocks. Statistical algorithms are created to execute orders for trades based on pre-established policies without humans stepping in to approve or manage the trades. Automated trading platforms use machine learning algorithms that combine historical data and current data that may have an impact on the price of the stock. For example, a trading algorithm may be designed to automatically adjust based on social media news feeds. This approach can provide rapid insight as large volumes of current data are processed at incredibly fast speeds. Although taking action based on this early (and unverified) information may improve trading performance, the lack of human interaction can also lead to erroneous actions. For example, automated trading algorithms have responded to fake or misleading social media news feeds leading to a rapid fall in the stock market. A human would have hopefully taken the time to check the facts.

Meeting business requirements for speed and accuracy of predictions with traditional approaches to analytics has become challenging. With the use of machine learning and cognitive computing, you can develop predictive models that account for relationships, patterns, and expectations that you may have never thought of before. You can move from describing what you see to impacting what you will see.

The following two examples illustrate how companies are using machine learning and analytics to improve predictive capabilities and optimize business results.

- **Analytics and machine learning predict trending customer issues.** The speed of social media can accelerate a small customer issue and grow it into a major complication before a company has time to react. Some companies decrease the time it takes to react to customer concerns by leveraging SaaS offerings that use machine learning to look for trends in social media conversations. The software compares the social media data to historical patterns and then continuously updates the results based on how the predicted pattern compares to actual results. This form of machine learning provides at least 72 hours of advanced warning on trending issues before they are picked up by mainstream media. As a result, marketing and public relations teams can take early action to protect the company's brand and mitigate customer concerns. Media buyers use the service to quickly identify changing customer purchasing trends, so they can determine where they should place their ads in mobile applications and web environments.

- **Analytics and machine learning speed the analysis of performance to improve service level agreements (SLA).** Many companies find it hard to monitor IT performance at fast enough intervals to identify and fix small problems before they escalate and negatively impact SLAs. Using machine learning algorithms, companies can identify patterns of IT behavior and orchestrate its systems and operational processes to become more prescriptive. These systems can learn to adapt to changing customer expectations and requirements. For example, telecommunications companies need to anticipate and prevent network slowdowns or outages so that they can keep the network operating at the speeds required by their customers. However, it can be nearly impossible to identify and correct for interruptions in bandwidth if network monitoring is not done at a sufficiently granular level. For example, with a new machine learning solution offered by Hitachi, telecoms can analyze large streams of data in real time. Hitachi's customers can combine analysis of historical data and real-time analysis of social media data to identify patterns in the data and make corrections in the performance of the network. There are many situations in which this would be helpful to customers. For example, if a streaming video application shows a popular sporting event and the game goes into overtime, an adaptive system could automatically add an additional 15 minutes of bandwidth support, so end users are provided with consistent high-quality service. Machine learning can help the system adapt to a variety of changes and unusual occurrences to maintain quality of performance.

Key Capabilities in Advanced Analytics

You can't develop a cognitive system without using some combination of predictive analytics, text analytics, or machine learning. It is through the application of components of advanced analytics that data scientists can identify and understand the meaning of patterns and anomalies in massive amounts of structured and unstructured data. These patterns are used to develop the models and algorithms that help determine the right course of action for decision makers. The analytics process helps you understand the relationships that exist among data elements and the context of the data. Machine learning is applied to improve the accuracy of the models and make better predictions. It is an essential technology for advanced analytics, particularly because of the need to analyze big data sources that are primarily unstructured in nature. In addition to machine learning, advanced analytics capabilities including predictive analytics, text analytics, image analytics, and speech analytics are described later in the chapter.

The Relationship Between Statistics, Data Mining, and Machine Learning

Statistics, data mining, and machine learning are all included in advanced analytics. Each of these disciplines has a role in understanding data, describing the characteristics of a data set, finding relationships and patterns in that data, building a model, and making predictions. There is a great deal of overlap in how the various techniques and tools are applied to solving business problems. Many of the widely used data mining and machine learning algorithms are rooted in classical statistical analysis. The following highlights how these capabilities relate to each other. Machine learning algorithms are covered in the next section in greater detail due to the importance of this discipline to advanced analytics and cognitive computing.

- **Statistics** is the science of learning from data. Classical or conventional statistics is inferential in nature, meaning it is used to reach conclusions about the data (various parameters). Although statistical modeling can be used for making predictions, the focus is primarily on making inferences and understanding the characteristics of the variables. The practice of statistics requires that you test your theory or hypothesis by looking at the errors around the data structure. You test the model assumptions to understand what may have led to the errors with techniques such as normality, independence, and constant variances. The goal is to have constant variances around your model. In addition, statistics requires you to do estimation using confidence values and significance testing—test a null hypothesis and determine the significance of the results, called p–values.
- **Data mining**, which is based on the principles of statistics, is the process of exploring and analyzing large amounts of data to discover patterns in that data. Algorithms are used to find relationships and patterns in the data and then this information about the patterns is used to make forecasts and predictions. Data mining is used to solve a range of business problems such as fraud detection, market basket analysis, and customer churn analysis. Traditionally, organizations have used data mining tools on large volumes of structured data such as customer relationship management databases or aircraft parts inventories. Some analytics vendors provide software solutions that enable data mining of a combination of structured and unstructured data. Generally, the goal of data mining is to extract data from a larger data set for the purposes of classification or prediction. In classification, the idea is to sort data into groups. For example, a marketer might be interested in the characteristics of people who responded to a promotional offer versus those who didn't respond to

the promotion. In this example, data mining would be used to extract the data according to the two different classes and analyze the characteristics of each class. A marketer might be interested in predicting those who will respond to a promotion. Data mining tools are intended to support the human decision-making process.

- **Machine learning** uses some of the same algorithms that are used in data mining. One of the key differences in machine learning as compared to other mathematical approaches is the focus on using iterative methods to reduce the errors. Machine learning provides a way for systems to learn and thereby improve both the models and the results of those models. It is an automated approach that provides new ways of searching for data and enables many iterations of a model to occur, quickly improving accuracy. Machine learning algorithms have been used as "black box" algorithms that make predictions for large data sets without requiring a causal interpretation of the fitted model.

Using Machine Learning in the Analytics Process

Machine learning is essential to improving the accuracy of predictive models in a cognitive environment. These predictive models have a large number of attributes across many observations. The data sets are likely to be unstructured, massive in size, and subject to frequent change. Machine learning enables the models to learn from the data and enhance the knowledge base for a cognitive system. Hundreds or thousands of iterations of a model take place very quickly, leading to an improvement in the types of associations that are made between data elements. Due to their complexity and size, these patterns and associations could have easily been overlooked by human observation. In addition, complex algorithms can be automatically adjusted based on rapid changes in variables such as sensor data, time, weather data, and customer sentiment metrics. The improvements in accuracy are a result of the training process and automation. Machine learning refines the models and algorithms by continuously processing new data in real time and training the system to adapt to changing patterns and associations in the data.

Increasingly companies are incorporating machine learning to understand the context of various predictive attributes and how these variables relate to each other. This improved understanding of context leads to greater accuracy in the predictions. Companies are applying machine-learning technology to improve predictive analytics processes that have been in place for many years. For example, the telecommunications industry has used analytics to analyze historical customer information such as demographics, usage, trouble tickets, and products purchased to help predict and reduce churn. Over time the industry

progressed from a focus on data mining of structured customer information to include text analytics of historical unstructured information such as comments on customer surveys and notes from call center interactions. Currently, an advanced analytic approach followed by some telecoms brings the unstructured and structured information together to develop a more complete profile of an individual customer. In addition, historical information can be combined with the most current information sourced through social media applications. Machine learning technology is used to train systems to quickly identify those customers that the company is most at risk of losing and develop a strategy to improve retention. Machine learning is applied in many industries including healthcare, robotics, telecommunications, retail, and manufacturing.

Supervised and unsupervised machine learning algorithms are used in a variety of analytics applications. The machine learning algorithm chosen depends on the type of problem being solved and the type and volume of the data required to solve the problem. Typically, supervised learning techniques use labeled data to train a model, whereas unsupervised learning uses unlabeled data in the training process. Machine learning models that are trained on labeled data can then use this training to predict an accurate label for unlabeled data. *Labeled data* refers to the identification or tag that provides some information about the data. For example, unstructured data such as voice recordings could be "labeled" or "tagged" with a name of the speaker or some information about the topic on the recording. Humans often provide the labels to the data as part of the training. Unlabeled data does not include tags, other identifiers, or metadata. For example, unstructured data such as videos, social media data, voice recordings, or digital images would be considered unlabeled if it exists in its raw form without any preconceived human judgments about the data.

Supervised Learning

Supervised learning typically begins with an established set of data and a certain understanding of how that data is classified. Humans are involved to provide the training, and the analytical model is fit with data that is tagged or labeled. The algorithms are trained using preprocessed examples and then the performance of the algorithms is evaluated with test data. Occasionally, patterns that are identified in a subset of the data can't be detected in the larger population of data. If you fit the model to patterns that exist only in the training subset, then you create a problem called *overfitting*. To protect against overfitting, testing needs to be done against both labeled data and unlabeled data. Using unlabeled data for the test set can help to evaluate the accuracy of the model in predicting outcomes and results. Some applications of supervised learning include speech recognition, risk analysis, fraud detection, and recommendation systems.

The following tools and techniques are often used to implement supervised learning algorithms.

- **Regression—**Regression models were developed in the statistical community. LASSO regression, Logistic regression, and Ridge regression can be used in machine learning. LASSO is a type of linear regression that minimizes the sum of squared errors. Logistic regression is a variant of standard regression but extends the concept to deal with classification. It measures the relationship between a categorical-dependent variable and one or many independent variables. Ridge regression is a technique used to analyze data with highly correlated independent variables (collinearity). Ridge regression introduces some bias to the estimates in order to reduce the standard errors or misleading variances that result from collinearity with least squares estimates.
- **Decision tree—**A decision tree is a representation or data structure that captures the relationships among a set of categories. Leaf or end nodes represent the categories, while all other nodes represent "decisions" or questions that refine the search through the tree. A variety of machine learning algorithms are based on traversing decision trees. For example, gradient boosting and random forest algorithms assume that the underlying data is stored as a decision tree. Gradient boosting is a technique for regression problems, which produces a prediction model in the form of an ensemble. Random forest is an algorithm that organizes data using classification and regression trees to look for outliers, anomalies, and patterns in the data. The algorithm builds a model by initially selecting predictors at random and then continuously repeats the process to build hundreds of trees. Random forest is a bagging tool (ensembles of regression trees fit to bootstrap samples) that leads to more accurate models by relying on an iterative approach to many alternative analyses. After the trees are grown, it is possible to identify clusters or segments in data and rank the importance of variables used in the model. Leo Breiman, Statistics Department at the University of California, Berkeley, created this algorithm and described it in a paper published in 2001. Random forest algorithms are used extensively in risk analytics.
- **Neural networks—**Neural network algorithms are designed to emulate human/animal brains. The network consists of input nodes, hidden layers, and output nodes. Each of the units is assigned a weight. Using an iterative approach, the algorithm continuously adjusts the weights until it reaches a specific stopping point. Errors identified in training data output are used to make adjustments to the algorithm and improve the accuracy of the analytic model. Neural networks are used in speech recognition, object recognition, image retrieval, and fraud detection. Neural networks can be used in recommendation systems like the Amazon.com system that makes selections for the buyer based on previous purchases and searches. Deep neural networks can be used to build models on unlabeled data.

As noted in Chapter 5, "Representing Knowledge in Taxonomies and Ontologies," two recent projects highlight the power of neural networks as discovery and classification engines. The Google Brain project used 16,000 processors in a neural network to discover patterns in images that converged as it learned to recognize cats, without a predefined template for cat images. Microsoft's Project Adam can identify dog breeds from photographs, using asynchronous neural networks.

Neural networks can be challenging to use in environments that require an audit trail or traceability, such as financial services trading systems. Because these systems are designed to learn autonomously, it could be cost-prohibitive to track changes made in a way that would satisfy the requirements of an outside examiner.

- **Support Vector Machine (SVM)**—SVM is a machine learning algorithm that works with labeled training data and output results to an optimal hyperplane. A *hyperplane* is a subspace of the dimension minus one (that is, a line in a plane). SVM is usually used when there are a small number of input features. The features are expanded into higher dimension space. SVM is not scalable to billions of elements of training data. An alternative algorithm for situations with extremely large volumes of training data would be logistic regression.
- **k-Nearest Neighbor (k-NN)**—k-NN is a supervised classification technique that identifies groups of similar records. The k-Nearest Neighbor technique calculates the distances between the record and points in the historical (training) data. It then assigns this record to the class of its nearest neighbor in a data set. k-NN is often selected when there is limited knowledge about the distribution of the data.

Unsupervised Learning

Unsupervised learning algorithms can solve problems that require large volumes of unlabeled data. As in supervised learning, these algorithms look for patterns in the data that enable an analytics process. For example, in social analytics, you may need to look at large volumes of Twitter messages (tweets), Instagram photos, and Facebook messages to collect adequate information and develop insight into the problem you want to solve. This data is not tagged; and given the large volume, it would take too much time and other resources to attempt to tag all this unstructured data. As a result, unsupervised learning algorithms would be the most likely choice for social media analytics.

Unsupervised learning means that the computer learns based on an iterative process of analyzing data without human intervention. Unsupervised learning algorithms segment data into groups of examples (clusters) or groups of features. The unlabeled data creates the parameter values and classification of

the data. Unsupervised learning can determine the outcome or solution to a problem, or it can be used as the first step that is then passed on to a supervised learning process.

The following tools and techniques are typically used in unsupervised learning.

- **Clustering techniques** are used to find clusters that exist in the data sample. Clustering categorizes the variables into groups based on certain criteria (all the variables with X or all the variables without X).
  - The **K-means algorithm** can estimate the unknown means based on the data. This is probably the most widely used unsupervised learning algorithm. It is a simple local optimization algorithm.
  - The **EM-Algorithm for clustering** can maximize the mixture density given the data.
- **Kernel density estimation (KDE)** estimates the probability distribution or the density of a data set. It measures the relationship between random variables. KDE can smooth the data when inferences are made from a finite data sample. KDE is used in analytics for risk management and financial modeling.
- **Nonnegative matrix factorization (NMF)** is useful in pattern recognition and to solve challenging machine learning problems in fields such as gene expression analysis and social network analysis. NMF factorizes a non-negative matrix into two non-negative matrixes of lower rank, and can be used as a clustering or classification tool. Used in one way, it is similar to K-means clustering. With another variation, NMF is similar to probabilistic latent semantic indexing—an unsupervised machine-learning approach for text analytics.
- **Principal Components Analysis (PCA)** is used for visualization and feature selection. PCA defines a linear projection where each of the projected dimensions is a linear combination of the original.
- **Singular Value Decomposition (SVD)** can help to eliminate redundant data to improve the speed and overall performance of the algorithm. SVD can help decide which variables are most important and which ones can be eliminated. For example, assume you have two variables that are highly correlated, such as "humidity index and probability of rain" and, therefore, do not add value to the model when used together. SVD can be used to determine which variable should be kept in the model. SVD is often used in recommendation engines.
- **Self Organizing Map (SOM)** is an unsupervised neural network model that was developed in 1982 by Tuevo Kohonen. SOM is a pattern recognition process. The patterns are learned without any external influences. It is an abstract mathematical model of topographic mapping from the

(visual) sensors to the cerebral cortex. It is used to understand how the brain recognizes and processes patterns. This understanding of how the brain works has been applied to machine learning pattern recognition. These techniques have been applied to manufacturing processes.

Predictive Analytics

Predictive analytics is a statistical or data mining solution consisting of algorithms and techniques that can predict future outcomes. Data mining, text mining, and machine learning can find hidden patterns, clusters, and outliers in both structured and unstructured data. These patterns form the basis of the answers and predictions made with a cognitive system. Predictive modeling uses the independent variables that were identified through data mining and other techniques to determine what is likely to occur under various future circumstances. Organizations use predictive analytics in many ways, including prediction, optimization, forecasting, and simulation. Predictive analytics can be applied to structured, unstructured, and semi-structured data. In predictive analytics, the algorithm you use applies some sort of objective function. For example, Amazon.com uses an algorithm that learns about your buying behavior and makes predictions regarding your interest in making additional purchases.

The focus on analyzing unstructured data represents a change for the use of predictive analytics. Traditionally, statistics and data mining technology have been applied to large databases of structured data. The internal operational systems of record at an organization are typically stored as structured data. However, it is the wide range of data types in unstructured data that represents the majority of data required to form a knowledge base for a cognitive system. These unstructured data sources include e-mails, log files, customer call center notes, social media, web content, video, and literature. Until recently, it has been more difficult for companies to extract, explore, and leverage unstructured data for decision making. Technology advancements such as Hadoop have improved the speed and performance of statistical analysis of unstructured data. The ability to analyze these unstructured data sources is key to the development of cognitive systems.

Business Value of Predictive Analytics

Companies use predictive analytics to solve many business challenges, including reducing customer churn, improving the overall understanding of customer priorities, and reducing fraud. Businesses can use predictive analytics to target customers that fit a certain profile, and segment customers according to prior purchases and current sentiment. By fine-tuning the models with iterative analytics and machine learning, predictive analytics can improve outcomes for businesses. Table 6-2 illustrates several examples of predictive analytics customer use cases.

Table 6-2: Predictive Analytics Use Cases

USE CASE	EXAMPLE	HOW PREDICTIVE ANALYTICS CHANGES OUTCOMES
Predicting consumer behavior	A manufacturer can identify patterns in consumer preferences that it could not recognize using traditional analysis of the data. Use of predictive analytics has improved supply chain management and the ability to react to consumer demand. This manufacturer can now predict customer orders 4 months in advance with an accuracy rate of approximately 98 percent.	The company deployed a real-time data warehouse to ensure that multiple sources of data could be well integrated and available at the right time for analytics. The company is building more accurate models using timely data and diverse data types. The models are designed to identify hidden patterns and create accurate forecasts.
Sales and inventory forecasting	A large multistore retailer uses advanced analytics to develop models at a faster pace than in the past using larger volumes of data. This company benefited by improving the accuracy of its sales forecasting models and reducing inventories. The company achieved 82 percent accuracy in its forecasting, a major improvement compared to traditional approaches.	This retailer implemented an analytics platform that standardizes and automates a portion of the predictive analytics process. Using this platform, the company can build 500 predictive models per month as compared to 1 model using traditional methods. The increased granularity in its models is yielding greater accuracy.
Predicting failures in machinery	A medical equipment manufacturer embeds sensors in its equipment to monitor performance. The recorded data is constantly streamed and analyzed to predict potential failures with enough lead time to make adjustments and avoid harm to patients.	Advanced analytics is used to build sophisticated algorithms that can uncover hidden patterns of failure and monitor sensitive equipment more accurately than traditional methods. The volume of data that needs to be analyzed is large and streaming.
Predicting and reducing fraud	An insurance company used advanced analytics to transform its approach to claims processing and improve fraud detection. The company improved its success rate in pursuing fraudulent claims from 50 percent to 90 percent and saved millions of dollars.	Predictive analytics is used to look at the whole claims process differently. Patterns of fraud are analyzed and used to rate the likelihood that each new claim may be fraudulent. Text mining is incorporated into the system to gain insight from analyzing the content of police reports and medical records.

Text Analytics

Given the business value of text-based unstructured sources, text analytics is a critical element of cognitive systems. Text analytics is the process of categorizing unstructured text, extracting relevant information, transforming it into

structured information, and analyzing it in various ways. The analysis and extraction processes used in text analytics leverage techniques that originate in computational linguistics, natural language processing, statistics, and other computer science disciplines. The text can be extracted and transformed, and then analyzed iteratively to identify patterns or clusters and determine relationships and trends. In addition, the transformed information from text analytics can be combined with structured data and analyzed using various business intelligence or predictive and automated discovery techniques.

The need for businesses to make decisions based on real-time information makes text analytics an increasingly important capability. For example, a telecom provider wanting to understand which customers might be most likely to switch to the competition unless they receive the right incentive needs real-time data on customer sentiment. The accuracy of predictive models that rely on customer sentiment data requires rapid analysis of large volumes of unstructured data. Sentiment scoring and natural language processing engines can build more accurate models and improve the speed of analysis. Machine learning can improve the models' capability to react to sentiment data from social media as it is fed back into the model. Text analytics is widely used to help organizations increase customer satisfaction, build customer loyalty, and predict changes in customer behavior. Text analytics can also improve search capabilities in areas such as faceted navigation.

Business Value of Text Analytics

The business value of text analytics increases with an organization's capability to understand how to act or make decisions based on the content. Text analytics is used in areas such as marketing analysis, social media analytics, sentiment analysis, market basket analysis, sales forecasting, product selection, and inventory management (see Table 6-3). To take the right action, companies need to understand not only what a customer is saying, but also what the customer's intent might be. Text analytics can help companies listen to what its customers are saying both individually and as a group. Understanding what the customer intends to do next requires deep insight into sentiment at a granular level. This deep listening to a customer is often part of a voice of the customer (VOC) program. For example, by combining knowledge of prior purchases with an analysis of one customer's relationship with others, his buying behavior and high-priority issues, a company is in a better position to take the next best action in any customer interaction.

The goal of a VOC program is to understand customer pain and identify where you may have the greatest challenges with your customers. For example, do you have a new product introduction that is not meeting expectations, or are you having problems with defects? By incorporating text analytics into your VOC program, you can identify changes in customer sentiment quicker. These

sentiments may be found in e-mails, customer surveys, and social media. There is often a lot of noise in big data that may contain valuable information on consumer sentiment. Text analytics can reduce the noise by identifying patterns in large volumes of unstructured information, providing an early indicator of changes in customer behavior. In sentiment analytics, the input is text and the output is a sentiment score (scale ranging from positive to negative). The model computes the score with an algorithm. You can look at how sentiment changes over time or how customers view your products compared to competitors.

Table 6-3: Text Analytics Use Cases

Marketing	Churn analysis, voice of the customer, sentiment analysis, customer survey analysis, social media analysis, market research
Operations	Voice of the employee, document categorization, competitive intelligence
Legal/Risk and Compliance	Document categorization, risk analysis, fraud detection, warranty analysis, e-discovery

Image Analytics

The sources used to develop knowledge corpora for a cognitive system are likely to include videos, photos, or medical images. There has been an enormous increase in the volume of images created and managed by governments, organizations, and individuals. As a result, image analytics capabilities are important in cognitive computing. The ability to quickly identify clusters and patterns in these images can have a major impact on IT and physical security, healthcare, transportation logistics, and many other areas. For example, facial recognition technology is used to both verify and identify individuals as a means to help prevent fraud and solve crimes. Governments use facial and image analytics to anticipate and prevent terrorist activities. Facial recognition can be used in video indexing to label faces in the video and identify the speakers in the video. Although facial recognition is a significant part of image analytics, cognitive systems will demand the capability to identify content in many different types of images. Image analytics can index and search video events by classifying objects into different categories such as people, animals, and cars, or to look for anomalies in a medical digital image such as an X-ray or CT scan.

Facial recognition was one of the earliest research areas in the field of image analytics. The first system for face recognition was developed in the 1960s. It was only partially automated, however, and there were a lot of manual steps involved. In the late 1980s, Kirby and Sirvich developed a system called Principal Components Analysis (PCA), which compares digitized sections of photos (eigenfaces). Compression techniques eliminate data that is not going to help with the comparison. This research represented a significant advancement, enabling

greater automation and improvements in speed and accuracy. There is great deal of ongoing research focusing on this area of technology, and it continues to improve. For example, there are algorithms that focus on the unique skeletal and musculature features in a face. These features have an impact on facial expressions that are consistent over time even as person ages.

Facial recognition is currently a big research area for many technology companies. For example, Facebook's research into facial recognition has resulted in software called DeepFace that is based on an advanced machine learning neural network. The machine learning algorithm analyzes a large number of human faces looking for recurring patterns in facial features such as eyebrows and lips. The learning process for DeepFace is based on a corpus of 4 million photos of faces. Facebook and other companies such as Google and Apple use facial recognition technology to enable users to identify and tag friends in photographs. Facebook's DeepFace project will be used to improve facial recognition capabilities on Facebook. Current test metrics show that DeepFace is almost as accurate as the human brain when comparing two photos to see if the face is the same. With this high level of accuracy there could be many other applications for DeepFace for marketing, sales, and security.

One key aspect of image analytics technology is the capability to detect the edge or the boundaries of objects in images. Edge detection algorithms look for discontinuities in brightness and can be used to segment the images. Some of the most common edge detection algorithms include Sobel, Canny, Prewitt, Roberts, and fuzzy logic. These algorithms are applied to all objects, not just faces. The process of facial recognition begins with finding the face in the image and identifying the facial features. Another aspect is based on determining color segmentation by looking at ratios of skin tone pixels. A face recognition algorithm using Eigenface-Fisher Linear Discriminant (EFLD) and Dynamic Fuzzy Neural Network (DFNN) helps with the dimension of features and classification. It reduces errors compared to previous algorithms. It works well on a face database with different expressions, poses, and illuminations. Machine learning frameworks can improve modeling and classification capabilities of large volumes of images.

Zintera, an emerging company based in San Diego, California, has developed a technology platform to enable image and video processing based on a biophysical neural network model. Zintera's technology requires very sparse training sets so that the neural network models can process images and videos quickly.

There are many potential applications for image analytics in healthcare. For example, IBM has a long-term, grand challenge project called Medical Sieve that incorporates image analytics. The goal of this project is to build a next-generation cognitive assistant with advanced multimodal analytics, clinical knowledge, and reasoning capabilities. Medical Sieve, an image-guided

informatics system, will be tuned to assist in clinical decision making in radiology and cardiology. Radiologists typically need to view thousands of images per day, leading to eyestrain and the possibility of a misdiagnosis. Medical Sieve uses sophisticated medical text and image processing, pattern recognition, and machine learning techniques guided by advanced clinical knowledge to process clinical data about the patient and identify anomalies in the images. Finally, it creates advanced summaries of imaging studies, capturing the most critical anomalies that are detected in various views of the images.

Speech Analytics

Text, image, and speech analytics can be used in a cognitive system to provide the right context to answer a question correctly or make an accurate prediction. Although text analytics is used to gain insight into sentiment, many emotions and attitudes can be easily masked in text. Images and speech can provide many more clues to a person's emotions and anticipated actions. *Speech analytics* is the process of analyzing recorded speech to extract information about the person speaking or the content of his words. Identifying the patterns of words and phrases that are good indicators of emotion and intent to act in a certain way can lead to improved accuracy of predictive models.

Speech analytics has been applied to call center processes for many years. There was significant research in the field of automatic speech recognition (ASR) as early as the 1950s. ASR systems needed to provide accurate information for people with vastly different speech patterns and regional accents without the need for training. Statistical models can create speech-clustering algorithms for different word and sound reference patterns. Although various statistical modeling techniques were applied to solving the problem of ASR, the hidden Markov model (HMM) and the stochastic language model became the most widely used techniques in the 1980s. As call centers developed in the 1990s as an important way to create a more efficient and cost-effective way to properly route calls, companies such as AT&T began using automatic speech recognition technology as part of the call center process.

Automated speech recognition (ASR) is one component of speech analytics. ASR can determine the words and phrases used in an individual's spoken language. This basic analysis can prioritize calls or get an initial understanding of the reason for the call. However, speech analytics is intended to go much deeper to gain more insight into context. Calls are categorized to identify patterns and anomalies. In the call center environment, speech analytics can answer many different types of questions. What is the subject matter discussed? What is the emotional tone of the speech? Is the speaker angry, impatient, or dissatisfied with a product? Are expectations for customer service agent performance met?

Using Advanced Analytics to Create Value

Ultimately, the goal of deploying advanced analytics processes and cognitive computing is to improve decision making. Companies are using analytics to differentiate from the competition by doing a better job of listening to customers, anticipating their needs, and making highly targeted offers. Government agencies use analytics to differentiate their cities by making them safer, more responsive to citizens' needs, and more ecologically sound. Healthcare organizations use analytics to improve physician training, eliminate unnecessary hospitalizations, and improve overall quality of care. Building the analytics models and cognitive computing environments that support these improvements in decision making requires more data, more accurate data, more refined data, and the ability to manage and interpret data from all input streams at fast speeds.

Figure 6-2 illustrates the trade-offs that companies often need to make to create business value from data. The ability to make faster and more execution-oriented decisions depends on reducing the "degree of difficulty" of interpreting the right data from all input streams. You need to manage the volume and complexity of the data. At the same time, you need to sample high-velocity data in a meaningful way. Raw data as captured by systems and sensors has potential value, but needs to be processed and analyzed to build business value. Referring to Figure 6-2, business value increases as volume, complexity, and speed are managed during the analytics process. The value to an organization comes from acting on the analysis of the data.

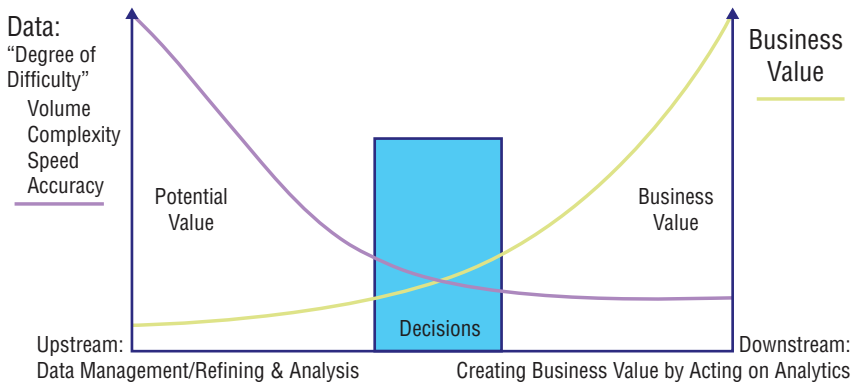

Figure 6-2: Refining raw data to create business value

Source: "An Executive Guide to Analytics Infrastructure," January 2014 by STORM Insights, Inc.

Building Value with In-memory Capabilities

Managing the dimensions of speed, complexity, and volume can be best addressed by an optimization strategy that includes both software and hardware. To scale appropriately to support big data analysis, many companies opt for hardware that is pre-integrated and optimized to run advanced analytics workloads. To achieve business value, the analytical models and predictions need to be fully integrated into operational business processes. All applications need access to the necessary data at the right time. Capabilities such as columnar alignments, graph databases, and in-memory computing can help support advanced analytics. To enhance the speed, shared memory, shared disk, high-speed networks, and optimized storage are valuable techniques.

Platforms that are designed for high-speed and volume analytics may rely on in-memory capabilities. Transaction processing, operational processing, analytics, and reporting and visualization can be integrated within one in-memory platform. It eliminates the time-consuming effort of data extraction and transformation. In-memory analytics provides a way to process large and complex analytic workloads quickly. It can improve application performance. These workloads can be chunked into smaller units and distributed across a parallel system. This approach is used primarily with structured data. In-memory analytics can help to overcome the challenges of trying to visualize and analyze big data at fast enough speeds. For example, with real-time streaming data, in-memory capabilities can ensure that computations are performed in RAM to increase processing times much faster than data access from disk.

Current approaches to model development, advanced analytics, and cognitive computing demand highly scalable architectures. The iterative processes used in machine learning yield more precise results, but at the same time require extreme speeds that can be provided by in-memory computing. With data viewed as the most important asset in advanced analytics, one benefit of using in-memory capabilities is that you eliminate the need to push the data to where the computations are taking place. Managing data sets in-memory means that the data can be used for transaction processing and analytics simultaneously. Organizing the documents for machine learning can be processing-intensive, so executing tasks such as creating tags or labels for documents can process faster in memory. Many companies find that they are building hundreds of models to develop more accurate and customized predictions. In addition, machine learning algorithms typically require complex technical computations on large volumes of data. Leveraging infrastructure with the right power and speed is critical to the success of these predictive models and cognitive systems.

Impact of Open Source Tools on Advanced Analytics

Open source analytics tools are having a major impact on the growth of predictive analytics at many organizations. The open source software environment and programming language, R, is fast becoming one of the primary tools for data scientists, statisticians, and other enterprise users. R, which is designed for computational statistics and data visualization, is the language of choice for graduate students doing research in advanced analytics and cognitive computing. Strong interest in R has led to a very active open source community. Members of the community share information on models, algorithms, and coding best practices. Users like the flexibility that a special-purpose programming language and environment offers for building custom applications. Some of the benefits of R include its flexibility and adaptability. R is actually an implementation of the statistical programming language S, developed at Bell Laboratories, as a higher-level alternative to using FORTRAN statistical subroutines.

Although R can be complicated to use unless you are an experienced data scientist or statistician, many vendors offer some sort of connection to R that makes it easier to use. Vendors are providing algorithms that are preset and ready to use in model development. The open source community has spawned and supports many projects that form the foundation for advanced analytics applications. For example, two important projects within the Apache Foundation are projects like Cassandra (distributed DBMS) and Spark (an analytics framework for cluster computing in the Hadoop space).

Summary

Advanced analytics helps the cognitive system gain insight from the corpora and ontologies. For example, machine learning algorithms and predictive modeling are applied to ensure that the cognitive system is constantly learning. The system needs to understand context, provide the right answers to questions, make accurate predictions, and apply the right information at the right time. The actual machine learning algorithms selected will depend on the goals of the analysis. For example, is the cognitive system applied to some aspect of healthcare? Are the goals related to improving medical diagnosis accuracy, reducing costs, reducing re-admission rates after patients are discharged from the hospital, or improving overall health for individuals and communities?

Machine learning algorithms will be applied to help discover the patterns that are important to building a cognitive system that is both accurate and fast. Algorithms for prediction, classification, segmentation, forecasting, sequence pattern discovery, association pattern discovery, geo-spatial and temporal discovery, or pattern detection may all be applied as needed to improve results of the system.