TUTORIAL T3
BUSINESS INTELLIGENCE, PREDICTIVE ANALYTICS, AND
DATA MINING

Content
T3.1 Introduction and Definition
T3.2 The Essentials of Business Intelligence
T3.3 Data Warehouses and Data Marts
T3.4 Business and Predictive Analytics
T3.5 Big Data and Intelligent Systems
T3.6 Knowledge Discovery: Data, Text, and Web Mining
T3.7 Resources for the Tutorial

T3.1 INTRODUCTION AND DEFINITION
Appropriate EC support activities (e.g., justification, market research, product pricing, allocation of resources including budget, or project management) require quantitative analysis. Sometimes such analysis needs to be done quickly and/or in real time involving large amounts of ever changing data. Also, EC management needs quick and easy access to different types of information as well as answers to queries. An increasingly popular set of tools to do just that is business intelligence and its major components: business analytics, predictive analytics, visualization, big data and data mining.

DEFINITION OF BI
Business intelligence (BI) is an umbrella term that includes computer-based architectures, tools, databases, analytical tools, applications, and methodologies. It is a content-free expression, so it means different things to different people. Part of the confusion about BI lies in the flurry of acronyms and buzzwords that are associated with its components. BI’s major objective is to enable interactive access to data, to enable
manipulation and processing of data, and to give business managers and analysts the ability to conduct appropriate analysis. By analyzing historical and current data, situations, and performances, decision makers can acquire valuable insights that enable them to make more informed, timely, and consequently better decisions. A BI predecessor for many years was the executive information system (EIS). Many BI software systems evolved from EIS.

**THE WIKIPEDIA COVERAGE**

Wikipedia has a major entry at en.wikipedia.org/wiki/Business_intelligence that is very comprehensive. The relevant content of the site in May 2013 was:

- Business intelligence (BI) and data warehousing
- Business intelligence and business analytics
- Applications in an enterprise
- Requirements gathering (preparation, identify the interview team, research the organization, and interview planning)
- Prioritization of business intelligence projects
- Success factors of implementation
- User aspect
- BI portals
- BI marketplaces
- Semi-structured or unstructured data and their treatment
- Future of BI

Wikipedia also covers in detail all the major components of BI (see en.wikipedia.org/wiki/Predictive_analytics). Also look at the entries: B1 2.0 and 3.0, business activity monitoring, mobile B1, and Big Data.

**T3.2 THE ESSENTIALS OF BUSINESS INTELLIGENCE**
Data for EC projects can be viewed as either transactional or analytical. Transactional data are those pieces of information that are collected in traditional transaction processing systems, are organized mainly in a hierarchical structure, and are centrally processed. Newer systems that contain transactional data are usually Web based; in medium to large organizations they may be part of an ERP system or order fulfillment, or sales records. These are known as operational systems, and the processing results are mainly summaries and reports. For comprehensive coverage, see Sharda, et al. (2014) or Sabherwal and Becerra-Fernandez (2011).

Today, the most successful companies are those that can respond quickly and flexibly to market changes and opportunities (i.e., they are agile). The key to this response is the effective and efficient use of data and information. EC transactions frequently must be done online in real time. This is done not only via transaction processing but also through the supplementary activity of analytical processing, which involves analysis of accumulated data, mainly by end users. Analytical processing includes Web applications, market research, predictive analytics, data mining, CRM activities, and decision-support systems. Placing strategic information in the hands of decision makers aids productivity, empowers users to make better decisions, and improves customer service, leading to greater competitive advantage and profitability.

**COLLECTING, ORGANIZING, AND STORING DATA FOR ANALYTICAL PROCESSING**

Analytical processing basically can be done in two ways. One is to work directly with the operational systems (the "let's use what we have" approach), using software tools and components known as front-end tools and middleware. This option can be optimal for companies that do not have a large number of end users running queries and conducting analyses against the operating systems. Since the mid-1990s, a wave of front-end
tools that allow end users to conduct queries and report on data stored in *operational databases* have become available. The problem with this approach, however, is that the tools are effective only with end users who have a medium- to high-degree of knowledge about databases and possess good computing skills.

These limitations call for a second, improved, option of analytical processing, which involves these concepts:

1. A business representation of data for end users
2. A user-friendly, Web-based environment that gives the customers and corporate employees query, reporting, and analysis capabilities
3. A single, server-based data repository—a **data warehouse (DW)**—that allows centralized analysis, security, and control over the data

These three concepts constitute the backbone of BI. In addition, BI includes data access, data retrieval, data analysis, and the results of the analyses with their visualization.

**THE BUSINESS INTELLIGENCE PROCESS**

The process of building and using BI is shown in Exhibit T3.1. The organization's data are stored in operational systems (left side of the figure). They are the input to the DW. Not all data are transferred to the data warehouse; frequently, only a summary of the data is transferred in a process of *extraction, transformation, and load* (ETL). The data that are transferred are organized within the warehouse as a relational database so that it is easy for end users to access. Also, the data are organized by subject, such as by product, customer segment, or business partner. EC data also may be organized according to a business process, such as ordering, shipping, or available inventory. Data access for analysis is provided through Web browsers via middleware software. On the right side of the figure are various analytical applications that may use the data.
The activities conducted during much of the process described in Exhibit T3.1 are generally referred to as business intelligence. The major reason for the name is that these activities not only collect and process data, they also enable analysis that results in useful—intelligent—solutions to business problems.

The various benefits offered by BI are that by its use an organization can improve its business knowledge, provide competitive advantage, enhance customer service and satisfaction, facilitate decision making, and help in streamlining business processes and supply chain.
T3.3 DATA WAREHOUSES AND MARTS

The purpose of a data warehouse is to establish a repository that makes operational data accessible in a form readily acceptable for analytical processing activities, such as EC applications, decision support, Q&A, and other end-user applications.

Data warehouses provide for the storage of metadata, which are data about data. Metadata include software programs about data, rules for organizing data, and data summaries that are easy to index and search, especially with Web tools.

CHARACTERISTICS OF DATA WAREHOUSING

The major characteristics of data warehousing include the following:

- **Organization.** Data are organized by detailed subject (e.g., by customer, vendor, product, price level, and region) and only contain information relevant for decision support.

- **Consistency.** Data in different operational databases may be encoded differently. For example, gender data may be encoded 0 and 1 in one operational system and "m" and "f" in another. They will be coded in a consistent manner within each warehouse.

- **Time variant.** The data are kept for 5 to 10 years so that they can be used for trends, forecasting, and comparisons over time.

- **Nonvolatile.** Once entered into the warehouse, data are not updated. However, new, related data may replace or supplement old data.

- **Relational.** The data warehouse typically uses a relational structure (organized into tables of rows and columns).

BENEFITS OF DATA WAREHOUSES

The major benefits of data warehouses are (1) the ability of users to reach data quickly, because data are located in one place and organized properly, and (2) the ability to reach data easily, frequently by end users themselves,
using Web browsers. Another benefit is that a data warehouse provides a consolidated view of corporate data, which is better than providing many smaller (and differently formatted) views. For example, separate EC systems may track sales and online coupon distribution. Combining data from these two systems may yield insights into the cost efficiency of online coupon promotions that would not be immediately evident from the output data of either system alone. Integrated within a data warehouse, however, such information can be easily extracted.

Data warehouses allow information processing to be off-loaded from expensive operational systems onto low-cost servers (or processed by cloud computing and application service providers, ASPs). Once this is done, end-user tools can handle a significant number of end-user information requests (e.g., using Microsoft Excel, Microsoft Office 365). Furthermore, some operational system reporting requirements can be moved to Web-based decision-support systems, thus making information processing more efficient.

**SUITABILITY**

Data warehousing is most appropriate for organizations in which some of the following apply:

- Large amounts of data need to be accessed by end users (i.e., Big Data).
- The operational data are stored in different systems and places.
- An information-based approach to management is in use.
- The company has a large, diverse *customer base* (such as in a utility company or a bank).
- The same data are represented differently in different systems (e.g., a company’s departments).
- Data are stored in highly technical formats that are difficult to decipher.
- Extensive end-user computing is performed (many end users performing many activities).
Hundreds of successful applications have been reported (e.g., see client success stories and case studies at websites of vendors such as Oracle, MicroStrategy, SAP/Business Objects, IBH/Cognos Corp., Information Builders, Teradata, NCR Corp., and Software A&G). For further discussion, see Sharda, et al. (2014). Also visit The Data Warehouse Institute (tdwi.org).

Although data warehouses offer substantial benefits, the cost of a data warehouse can be very high, both to build and to maintain. Furthermore, it may be difficult and expensive to incorporate data from obsolete legacy systems. Finally, there may be a lack of incentives among departments within a company to share data. Therefore, a careful feasibility study must be undertaken before a commitment is made to data warehousing. Alternatively, one or more data marts or data stores can be used.

DATA MARTS
The high cost of data warehouses confines their use mostly to large companies. An alternative used by many other firms is the creation of a lower-cost, scaled-down version of a data warehouse called a data mart. A data mart is a small warehouse designed for a strategic business unit (SBU) or a department. A data mart can be fully dedicated to specific areas such as marketing, finance, or EC.

The advantages of data marts over data warehouses include the following:

- The cost is low (prices under $100,000 versus $1 million or more for large data warehouses).
- The lead time for implementation is significantly shorter, often less than 90 days.
- Data marts are controlled locally rather than centrally, conferring power on the using group.
- They contain less information than the data warehouse. Hence, they have
more rapid response and are more easily understood and navigated than an enterprisewide data warehouse.

- They allow an EC department to build its own decision support systems without relying on a centralized IS department.

Data marts are either replicated or stand-alone. *Replicated data marts* are those in which functional subsets of the data warehouse have been replicated (copied) into smaller data marts. The reason for using replicated data marts is that sometimes it is easier to work with a small subset of the data warehouse. Each of these replicated data marts is dedicated to a certain area. A stand-alone data mart may mix and match data from a DW and other sources, or can operate without the need for a DW.

**OPERATIONAL DATA STORES**

An *operational data store* is a database for transaction processing systems that uses data warehouse concepts to provide clean data. That is, it brings the concepts and benefits of the data warehouse to the operational portions of the business, often at a lower cost. It is used for short-term decisions involving mission-critical applications rather than for the medium- and long-term decisions associated with the regular data warehouse. Short-term decisions often require current information. For example, when a customer sends an e-mail query to a bank, the bank will quickly need to access all of the customer's current accounts. The operational data store can be viewed as situated between the operational data (legacy systems) and the data warehouse.

**SUCCESSES AND Failures OF DATA WAREHOUSING AND BI INITIATIVES**

Since their early inception, data warehouses and BI initiatives have produced many success stories. However, many failures have also occurred. Carbone (1999) identified several types of warehouse failures, which also exist today. For example the warehouse:
Did not meet the expectations of those involved
- Was completed, but went severely over budget, in relation to time, or both
- Failed one or more times, but eventually was completed at additional cost
- Failed and no effort was made to revive it

Carbone (1999) identified a number of specific reasons for failures (which are typical for BI in general). Some of these are included in Exhibit T3.2, which includes additional factors.

**Exhibit T3.2 Why Data Warehousing and BI Initiatives Fail**

<table>
<thead>
<tr>
<th>Data Problems</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Not enough summarization of data.</td>
</tr>
<tr>
<td>• Failure to align data marts and data warehouses.</td>
</tr>
<tr>
<td>• Poor data quality (e.g., omitted information).</td>
</tr>
<tr>
<td>• Incomplete user input makes data irrelevant.</td>
</tr>
<tr>
<td>• Incorrectly using data marts instead of data warehouses (and vice versa).</td>
</tr>
<tr>
<td>• Certain types of data are excluded or are not expressed properly.</td>
</tr>
<tr>
<td>• Insecure access to data manipulation.</td>
</tr>
<tr>
<td>• Data were not standardized properly.</td>
</tr>
<tr>
<td>• Poor upkeep of information (e.g., failure to keep information current).</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Technology Problems</th>
</tr>
</thead>
<tbody>
<tr>
<td>• The middleware is not working properly.</td>
</tr>
<tr>
<td>• Inappropriate architecture was used.</td>
</tr>
<tr>
<td>• Using the warehouse only for operational, not informational, purposes.</td>
</tr>
<tr>
<td>• Poor upkeep of technology and the use of obsolete technology (e.g., too slow).</td>
</tr>
<tr>
<td>• Inappropriate format of information—a single, standard format was not used.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Other Problems</th>
</tr>
</thead>
</table>
- Need long and expensive training.
- Vendors overselling capabilities of products used in the DW/BI.
- Lack or inappropriate training and support for users.
- Inexperienced/untrained/inadequate number of users and maintainers.
- The benefits do not justify the costs.
- Integration with other systems is poorly done.
- Unrealistic expectations—overly optimistic time schedule or underestimation of cost.
- Lack of coordination (or requiring too much coordination).
- Corporate and employee cultural issues were ignored.
- Improperly managing multiple users with various needs.
- Unclear business objectives; not knowing the information requirements of the users.
- Lack of effective project sponsorship.
- Corporate politics interfere.

Suggestions on how to avoid data warehouse failure can be found at tdwi.org and dmreview.com.

T3.4 BUSINESS AND PREDICTIVE ANALYTICS

Once the data are in the data warehouse and/or data marts, they can be accessed by end users. Users can then conduct several types of business analytical activities with the data, using a set of tools and methodologies. In EC, analysis can also be done on Web data and on data stored in repositories other than DW or data marts.

BUSINESS ANALYTICS

Analytics is the science of analysis. Generally, it refers to analysis of data. There are many ways to conduct such analysis. Indeed, there are many methods and hundreds of software tools to conduct the analysis. (See IBM
Business analytics (BA) is a broad category of applications and techniques for gathering, storing, analyzing, and providing access to data to help enterprise users make better business and strategic decisions. BA is also known as analytical processing, BI tools, and BI applications. BA is becoming a major tool for most medium and large corporations. For a comprehensive description, see Davenport (2010) and Laursen (2010).

Marketing managers can run new data information through BA that forecast, for example, the probability of a customer placing its next order. A company then can use this information to determine marketing strategies to influence the customer to buy more of its products without spending more on marketing strategy than it has to.

A BA application allows for activities such as:

- Automating the thinking and, in most cases, a portion of the decision making of a human being.
- Typically using complex quantitative techniques, such as multivariate regression analysis, data mining, artificial intelligence, or nonlinear programming.

Examples: A BA application used for credit scoring for a loan applicant might: calculate a creditworthiness score, automatically accept or deny the loan application, determine the loan limit, select which credit card product to accept from buyers or which other type of loan approve.

Using BA software, the user can make queries, request ad hoc reports, or conduct analyses. For example, an analysis can be carried out by performing multilayer queries. Because all the databases are linked, you can search for what products are overstocked in a particular store. You can then determine which of those products commonly sell with popular items based on previous sales. After planning a promotion to move the excess stock along with the popular products (e.g., bundling them together), you can dig deeper into the
data to see where this promotion would be most popular and most profitable. The results of your request can be reports, predictions, alerts, and/or graphical presentations.

The major tools (techniques) used by BA packages are presented next.

**Reports and Queries**

Business analytics include static and dynamic reporting, all types of queries, discovery of information, multidimensional views, a drill down to details, and so on. *Routine reports* are generated automatically and distributed periodically to subscribers on mailing lists. Examples are weekly sales figures, units produced each day and each week, and monthly hours worked. Here is an example of how a report is used in BI: A store manager receives store performance reports generated weekly by the BI software. After a review of one such weekly report on store sales, the manager notices that sales for computer peripherals have dropped off significantly from previous weeks. She clicks on her report and immediately drills down to another enterprise report, which shows her that the three best-selling hard drives are surprisingly underperforming. Now the manager needs to investigate why. Drilling down further by individual day may reveal that bad weather caused the problem.

Reports can be generated directly from operational data (e.g., ERP, point-of-sales [POS] systems) and/or from a data warehouse.

**Ad Hoc (On-Demand) Queries and Reports**

*Ad hoc queries* allow users to request real-time information from the computer that is not available in periodic reports. Such answers are needed to expedite decision making. The system must be intelligent enough to understand what the user wants. Simple ad hoc query systems are based on menus. More intelligent systems use SQL (structured query language) and query-by-example approaches or Web-based applications.

**Web-Based Ad Hoc Query Tools.** Web-based ad hoc query tools
allow users to access, navigate, and explore relational data to make key business decisions in real time. For instance, users can gauge the success of a Web marketing campaign according to the number of Web hits received last month, last week, or even yesterday, in relation to products or services purchased. This insight helps companies better target marketing efforts and forge closer, more responsive relationships with customers. Several vendors offer such tools. For example, Cognos Corp., an IBM company, offers Web users powerful ad hoc exploration of corporate data assets, with little or no user training needed (see cognos.com/products/query.html).

Advanced query tools can be connected to intranets and extranets for B2B and CRM querying. Answers to queries can be supported by visualization tools.

**Online Analytical Processing and Multidimensional Analysis**

*Online analytical processing (OLAP)* refers to such end-user activities as simple as the use of spreadsheets and graphics to fairly complex modeling. OLAP enables the user to query the system, conduct an analysis, make a prediction, and so on, while the user is at a PC or a tablet.

A typical OLAP query might access a multigigabyte, multiyear sales database in order to find all product sales in each customer segment (female, male, young people, etc.). After reviewing the results, an analyst might further refine the query to find sales volume for each sales channel by hours of the day or by product type. As a last step, the analyst might want to perform year-to-year or quarter-to-quarter comparisons for each sales channel. This whole process must be carried out online with a rapid response time.

Thus, OLAP queries are able to analyze the relationships between many types of business elements (e.g., sales, products, regions, and channels) involving aggregated data over time (e.g., sales volumes, budgeted dollars, and dollars spent, on a monthly, quarterly, or yearly basis). The ability to
present data in different perspectives involving complex calculations between data elements (e.g., expected profit calculated as a function of sales revenue for each type of sales channel in a particular region) enables users to pursue an analytical thought process without being stymied by the system.

**Multidimensionality**

Raw and summary data can be organized in different ways for analysis and presentation. An efficient way to do this is called *multidimensionality*. The major advantage of multidimensionality is that it enables data to be organized the way individual managers, rather than system analysts, like to see them. Different presentations of the same data can be arranged easily and quickly.

**Multidimensional Presentations.** Spreadsheet tables have two dimensions. Information with three or more dimensions can be presented by using a set of two-dimensional tables (i.e., several tables) or one fairly complex table. In decision support, an attempt is made to simplify information presentation and allow the user to easily and quickly change the structure of tables to make them more meaningful (e.g., by flipping columns and rows, aggregating several rows and columns—rollup, disaggregating a set of rows or columns).

Three factors are considered in multidimensionality: *dimensions*, *measures*, and *time*. Here are some examples:

1. **Dimensions.** Some examples of dimensions are products, salespeople, market segments, business units, geographical locations, distribution channels, countries, and industries.

2. **Measures.** Some examples of measures are money, sales volume, head count, inventory, and actual versus forecasted profit.

3. **Time.** Some examples of time are daily, weekly, monthly, quarterly, and yearly.
Example. IBM’s Cube Views automates the creation of OLAP metadata at the database level so that metadata can be shared among applications that access the database. Cube Views aggregates data into multidimensional charts, allowing users to access the data from different perspectives, and it returns answers to queries as XML-based Web services. Cube Views is supported by many BI vendors.

Although OLAP can be quite useful in reporting, querying, and simple analysis (e.g., trends, summaries), it is retrospective in nature and cannot provide the automated and prospective knowledge discovery that is done by predictive analytics and data mining techniques, which are referred to as advanced analytics (see Henschen 2010 and Henschen 2009 for details).

PREDICTIVE ANALYTICS

Prediction is commonly referred to as the act of telling about the future. It differs from simple guessing by taking into account the experiences, opinions, and other relevant information in conducting the task of foretelling.

Large numbers of applications are listed in en.wikipedia.org/wiki/Predictive_analytics.

Definition

Predictive analytics is an area of quantitative analysis that deals with extracting information from data and using it to predict future trends and behavior patterns. The core of predictive analytics relies on capturing relationships between explanatory variables and predicted variables from past occurrences and exploiting it to predict future outcomes. For an introduction, software, the relationship to social media, and more, see Vance (2010).

The Tools of Predictive Analytics

Depending on the vendors’ offering (e.g., SAS Analytics) you can find some or all of the following components in their product:

- Data mining
- Optimization models (3–8)
- Other operations research/management science models
- Forecasting models
- Statistics
- Text analytics and text mining
- Web analytics and Web mining
- Quality improvement
- Data visualization

For details, see predictiveanalyticsworld.com and visit its annual conference.

For tools of predictive analysis, see zementis.com.

**Predictive Analytics in E-Commerce**

Predictive analytics produces a predictive score for each customer or other organizational element. Predictive analytics optimizes e-marketing campaign and website behavior to increase customer responses, conversions, and clicks, and to decrease churn. Each customer’s predictive score informs actions to be taken with that customer. Predictive analysis is especially useful in analyzing the results of market research and marketing ad campaigns. It is also useful in conducting justification and preparing the business case for EC projects.

**T3.5 KNOWLEDGE DISCOVERY: DATA, TEXT, AND WEB MINING**

The process of extracting useful knowledge from volumes of data is known as knowledge discovery in databases (KDD), or just knowledge discovery (KD). The objective of KDD is to identify valid, novel, potentially useful, and ultimately understandable patterns in data. KDD is useful because it is supported by three technologies that are now sufficiently mature to produce meaningful data: massive data collection, powerful
multiprocessor computers, and data mining algorithms.

**DATA MINING**

*Data mining* derives its name from the similarities between searching for valuable business information in a large database and mining a mountain for a vein of valuable ore. Both processes require either sifting through an immense amount of material or intelligently probing it to find exactly where the value resides. In some cases the data are consolidated in a data warehouse and data marts; in others they are kept on the Internet and intranet servers.

**Data Mining Capabilities**

Given databases of sufficient size and quality, data mining technology can generate new business opportunities by providing the following capabilities.

*Automated Prediction of Trends and Behaviors.* Data mining automates the process of finding predictive information in large databases. Questions that traditionally required extensive hands-on analysis can now be answered directly and quickly from the data. A typical example of a predictive problem is targeted marketing. Data mining can use data on past promotional mailings to identify the targets most likely to respond favorably to future mailings. Other predictive examples include forecasting the success of online campaigns and identifying segments of a population likely to respond similarly to given events.

*Automated Discovery of Previously Unknown Patterns.* Data mining tools identify previously hidden patterns in one step. An example of pattern discovery is the analysis of retail sales data to identify seemingly unrelated products that are often purchased together, such as baby diapers and beer. Other pattern discovery problems include detecting fraudulent credit card transactions and identifying *invalid (anomalous) data* that may represent data entry keying errors. This is especially useful for checking the validity of credit cards in e-shopping.
Data mining also can be conducted by nonprogrammers. The "miner" is often an end user, empowered by "data drills" and other power query tools to ask ad hoc questions and get answers quickly, with little or no programming skill. Data mining tools can be combined with spreadsheets and other end-user software development tools, making it relatively easy to analyze and process the mined data. Data mining appears under different names, such as knowledge extraction, data dipping, data archeology, data exploration, data pattern processing, data dredging, and information harvesting. "Striking it rich" in data mining often involves finding unexpected, valuable results.

**Types of Information Resulting from Data Mining Analysis**

Data mining yields five major types of information:

1. **Association.** Relationships between events that occur at one time (e.g., the contents of a shopping cart, such as orange juice and cough medicine)
2. **Sequences.** Relationships that exist over a period of time (e.g., repeat visits to a supermarket)
3. **Classifications.** The defining characteristics of a certain group (e.g., customers who have been lost to competitors)
4. **Clusters.** Groups of items that share a particular characteristic that was not known in advance of the data mining
5. **Forecasting.** Future values based on patterns within large sets of data (e.g., demand forecasting)

Data miners use several tools and techniques: case-based reasoning (using historical cases to recognize patterns); neural computing (a machine-learning approach by which historical data can be examined for patterns through massive parallel processing); association analysis (using a specialized set of algorithms to sort through data sets and express statistical rules among items); and intelligent agents (expert or knowledge-based
software embedded in information systems). For details, see Vercellis (2009).

A Sampler of Data Mining Applications in E-Commerce

Some applications of data mining in e-commerce follow.

- **Retailing and sales distribution.** Predicting sales, determining correct inventory levels and distribution schedules among outlets in B2B, B2C, and e-CRM.

- **Banking.** Forecasting levels of bad loans and fraudulent credit card use, predicting credit card spending by new customers, predicting which kinds of customers will best respond to (and qualify for) new loan offers (used in e-banking).

- **Manufacturing and production.** Predicting machinery failures, finding key factors that control optimization of manufacturing capacity (important in mass-customization).

- **Brokerage and securities trading.** Predicting when bond prices will change, forecasting the range of stock fluctuations for particular issues and the overall market, determining when to buy or sell stocks (can be used in e-stock trading and investment).

- **Computer hardware and software.** Predicting disk-drive failures, forecasting how long it will take to create new chips, predicting potential security violations (used by Google, Akami, and other support services of EC).

- **Travel industry (airlines, hotels/resorts, rental car companies).** Data mining has a variety of uses in the travel industry. It is successfully used to (1) predict sales of different services (seat types in airplanes, room types in hotels/resorts, car types in rental companies) in order to optimally price services to maximize revenues as a function of time-varying transactions (commonly referred to as yield management); (2) forecast demand at different locations to better allocate limited
organizational resources; (3) identify the most profitable customers and provide them with personalized services to maintain their repeat business; and (4) retain valuable employees by identifying and acting on the root causes for attrition.

- **Health care.** Data mining has a number of health care applications. It can be used to (1) identify people without health insurance and the factors underlying this undesired phenomenon; (2) identify novel cost–benefit relationships between different treatments to develop more effective strategies, (3) forecast the level and the time of demand at different service locations to optimally allocate organizational resources; and (4) understand the underlying reasons for customer and employee attrition.

- **Entertainment industry.** Data mining is successfully used by the entertainment industry to (1) analyze viewer data to decide what programs to show during prime time and how to maximize returns by knowing where to insert advertisements; (2) predict the financial success of movies before they are produced to make investment decisions and to optimize the returns; (3) forecast the demand at different locations and different times to better schedule entertainment events and to optimally allocate resources; and (4) develop optimal pricing policies to maximize revenues.

- **Marketing.** Classifying customer demographics that can be used to predict which customers will respond to a mailing or buy a particular product (popular in market research).

**TEXT MINING**

**Text mining** is the application of data mining to nonstructured or less-structured text files. Data mining takes advantage of the infrastructure of stored data to extract predictive information. For example, by mining a customer database, an analyst may discover that everyone who buys
product A also buys products B and C, but does so 6 months later. Text mining, however, operates with less-structured information. Documents rarely have strong internal infrastructure, and when they do, it is frequently focused on document format rather than document content.

Text mining helps organizations find the "hidden" content of documents, as well as additional useful relationships. It also helps them group documents by common themes (e.g., identify all the customers of an insurance firm who have similar complaints). For details, see Sharda, et al. (2014).

**Web Mining**

The previous discussion of data mining refers to data that usually are stored in a data warehouse. However, to analyze a large amount of data on the Web, one needs different mining tools. **Web mining** is the application of data mining and similar techniques to discover meaningful patterns, profiles, and trends from websites. The term Web mining is used to describe two different types of information mining. The first, Web content mining, is the process of discovering information from millions of Web documents. The second, Web usage mining, is the process of analyzing what customers are doing on the Web—that is, analyzing clickstream data.

In Web mining, the data are clickstream data, usually stored in a special clickstream data warehouse (see Sweiger, et al. 2002) or in a data mart. The strategies used may be the same in both. Several companies provide tools for Web mining; for example, iOpus (iopus.com), KDNuggets (kdnuggets.com), Megaputer (megaputer.com), and SPSS (spss.com). For details, see Sharda, et al. (2014).

**T3.6 BIG DATA**

According to Wikipedia, **big data** refers to “a collection of data sets so large and complex that it becomes difficult to process using on-hand database
management tools or traditional data processing applications. The challenges include capture, curation, storage, search, sharing, transfer analysis, and visualization.” Big data contain large data sets. This allows for correlations to be found, in order to make predictions and find complex statistical associations.

**BIG DATA VS. BUSINESS INTELLIGENCE**

What is considered big data varies depending on the nature of the data and on the capabilities of the existing BI tools. Big data analysis yields better results than analyzing individually the same sets of data. Since existing BI software and hardware cannot process big data sets, within time and cost requirements, new platforms are being developed under the umbrella of "big Data", for conducting the various data manipulations. Specifically, according to Wikipedia, “Big data” is difficult to work with using most relational database management systems and desktop statistics and visualization packages. What is needed is massively parallel software running on tens, hundreds, or even thousands of servers.

According to Wikipedia, Business Intelligence uses descriptive statistics with data with high information density to measure things, detect trends, etc., while "Big Data" uses inductive statistics with data with low information density whose huge volume allow to infer business relationships (e.g., via statistical regressions and associations), and thus giving (with limits of interference reasoning) to "Big Data" strong predictive capabilities. To help companies to understand big data and how to turn them to competitive advantage, IBM wrote an ebook (see Zikopoulos, et al. 2013) which is provided for free.

Of special interest is Gartner’s big data definition (see Sincular 2013) which is based on three “Vs” (Volume, Velocity, and Variety). For an overview see the video by Hengeveld (2013).

**Usage in E-Commerce**
According to Hurwitz et al. (2013), EC data sets such as customer transactions for a mega-retailer, or social network activity can quickly outpace the capacity of traditional data management (BI) tools. The authors define, explain and guide users through the big data concept. The following EC applications are described in Akerkar (2013): Big social data analysis (e.g., those collected by monitoring conversations), knowledge management, smart grids, and smart buildings. Other EC applications include health care patients’ data, financial transactions, transportation, customers’ transactions, retail sales and global trade analysis. Predicting of shopping patterns is described by Mayer-Schonberger and Cukier (2013).

**Using Intelligent Systems for Big Data Manipulation**

There are several methodologies and tools to manipulate big data. Several of these are described by Akerkar (2013) and by Minelli et al. (2013). Examples are PureSystems from IBM (see Groenfeldt 2013), and IBM’s Big Data analytic platform.

**T3.7 RESOURCES FOR THE TUTORIAL**

**VIDEOS**

1. Business Intelligence and Data Warehousing (6:13 minutes)
   [youtube.com/watch?v=mRW7gimydFE&feature=related](https://youtube.com/watch?v=mRW7gimydFE&feature=related)
2. Business Intelligence Demonstration (27:01 minutes)
   [youtube.com/watch?v=-j5J7lXav7Y&feature=related](https://youtube.com/watch?v=-j5J7lXav7Y&feature=related)
3. Data Warehousing BI DW Introduction (10:56 minutes)
   [youtube.com/watch?v=zDk4yPL6Adc&feature=related](https://youtube.com/watch?v=zDk4yPL6Adc&feature=related)
4. Google Analytics for Business Intelligence (8:48 minutes)
   [youtube.com/watch?v=Of3uc2Aum-o&feature=related](https://youtube.com/watch?v=Of3uc2Aum-o&feature=related)
5. Introduction to Data Mining (5:28 minutes)
   [youtube.com/watch?v=p5jD5clrIhQ](https://youtube.com/watch?v=p5jD5clrIhQ)
7. The Value of Business Intelligence (6:00 minutes)
   youtube.com/watch?v=ArOFILzblHo
8. What Is Business Intelligence (2 minutes)
   youtube.com/watch?v=od2BlvlNkBY
9. What Is Data Mining? (6:06 minutes)
   youtube.com/watch?v=wqpMyQMio0to&feature=related
10. IBM SPSS Predictive Analytics for ISVs (5:03 minutes)
    youtube.com/watch?v=INbDzaXW5no
11. Predictive Analytics + Business Rules = Enhanced Decisioning (9:25 minutes)
    youtube.com/watch?v=jDsSVmsYRh8
12. How to Use Predixion Predictive Analytics: Detect Categories (4:50 minutes)
    metacafe.com/watch/5115349/how_to_use_predixion_predictive_analytics_detect_categories

BOOKS
   (accessed May 2013).


**JOURNALS**

1. *The International Journal of Artificial Intelligence, Neural Networks, and Complex Problem-Solving Technologies*  
   springer.com/computer/ai/journal/10489

2. *Business Intelligence Journal* businessintel.org

3. *International Journal of Business Intelligence and Data Mining*  
   inderscience.com/browse/index.php?journalCODE=ijbidm

4. Data Mining and Knowledge Discovery  
   springer.com/computer/database+management+%26+information+retrieval/journal/10618

5. SAS. “Real Analytics for Real Business.” White paper.  

**REPRESENTATIVE PROFESSIONAL ASSOCIATIONS**

1. Society of Competitive Intelligence Professionals scip.org

2. Certified Computing Professionals iccp.org

3. Data Management Association dama.org

4. The Data Warehouse Institute (TDWI) tdwi.org

5. Business Intelligence Network b-eye-network.com

6. DSS Resources dssresources.com

7. Teradata University teradatauniversitynetwork.com

8. Predictive Analytics predictiveanalytics.org

9. Data Mining Group dmg.org

**REFERENCES**

mitre.org/work/tech_papers/tech_papers_00/d-warehouse_presentation/problems_with_dws_990503.pdf (no longer available online).


GLOSSARY

**big data** “A collection of data sets so large and complex that it becomes difficult to process using on-hand database management tools or traditional data processing applications. The challenges include capture, curation, storage, search, sharing, transfer analysis, and visualization” (per Wikipedia).

**business analytics (BA)** A broad category of applications and techniques for gathering, storing, analyzing, and providing access to data to help enterprise users make better business and strategic decisions.

**business intelligence (BI)** Activities that not only collect and process data, but also make possible analysis that results in useful—intelligent—solutions to business problems.

**data mart** A small data warehouse designed for a strategic business unit (SBU) or department.

**data mining** The process of searching a large database to discover previously unknown patterns; automates the process of finding predictive information.

**data warehouse (DW)** A single, server-based data repository that allows centralized analysis, security, and control over data.

**knowledge discovery in databases (KDD)/knowledge discovery (KD)** The process of extracting useful knowledge from volumes of data.

**metadata** Data about data, including software programs about data, rules for organizing data, and data summaries.

**online analytical processing (OLAP)** End-user analytical activities, such as DSS modeling using spreadsheets and graphics, that are done online.

**operational data store** A database for use in transaction processing
(operational) systems that uses data warehouse concepts to provide clean data.

**predictive analytics** A collection of tools in BI that support data analysis including making predictions on events and people behavior.

**text mining** The application of data mining to nonstructured or less-structured text files.

**Web mining** The application of data mining techniques to discover meaningful patterns, profiles, and trends from both the content and usage of websites.