

Business analytics

Faculty

Manjunatha B A

Assistant Professor

NMIT

- **Teaching Methodology:**

- Blackboard Teaching
- Power point presentation

- **Assessment Methods**

- Group Discussion for 10 Marks.
- Case study for 10 Marks.
- Three internals, 30 Marks each will be conducted and the Average of best of two will be taken.
- Final examination, of 100 Marks will be conducted and will be evaluated for 50 Marks.

Data Explosion

- About seven billion shares change hand in US equity markets everyday.
- About 350 million photos are uploaded every day in the Facebook.
- Amount of credit card debt in US: \$890.91 billion.
- Total amount of credit card fraud worldwide: \$5.5 billion.
- Number of bankruptcies filed in US in 2014 is 910, 090.
- Percentage of US credit card holders who have been victims of credit card fraud:
 - 10%
- Every week, about 260 million customers visit Walmart stores.

Business analytics

- Business analytics (BA) refers to the tools, techniques and processes for continuous exploration and investigation of past data to gain insights and help in decision making.
- Business Analytics is an integration between science, technology and business context that assist data driven decision making.
- Today several products and solutions are driven by analytics.

Applications in various domains:

- Amazon
- Agriculture Business Analytics.
- Network analytics
- Stock Marketing.
- Finance Marketing.
- Manufacturing Industry.
- Medical Methodology.
- Customer Relation Management.
- etc

Introduction

- Corporate Decision Making: The HIPPO Algorithm
- Highest Paid Person's Opinion
- Opinion based decision making-leads to incorrect decisions.
- BA is to improve the quality of decision making using data analysis.

Theory of bounded relation

- Theory of bounded relation proposed by Herbert Simon (1972)
- Increasing complexity of business problems, existence of several alternative solutions with limited time.
- Decision making difficult due to Uncertainty, incomplete information, lack of knowledge, effect relationship between parameters of importance etc.

Problems of e-commerce companies

- Forecasting demand for products directly sold by the company
- Cancellation of orders placed by customers before their delivery
- Fraudulent transactions resulting in financial loss
- Predicting delivery time
- Predicting customer to buy in future

Data-driven decision making process

- Identify the problem or opportunity
- Identify sources of data
- Pre-process the data
- Divide the dataset training and testing
- Build analytical model and identify the best model/s
- Implement solution/decision/develop product etc.

Flow diagram

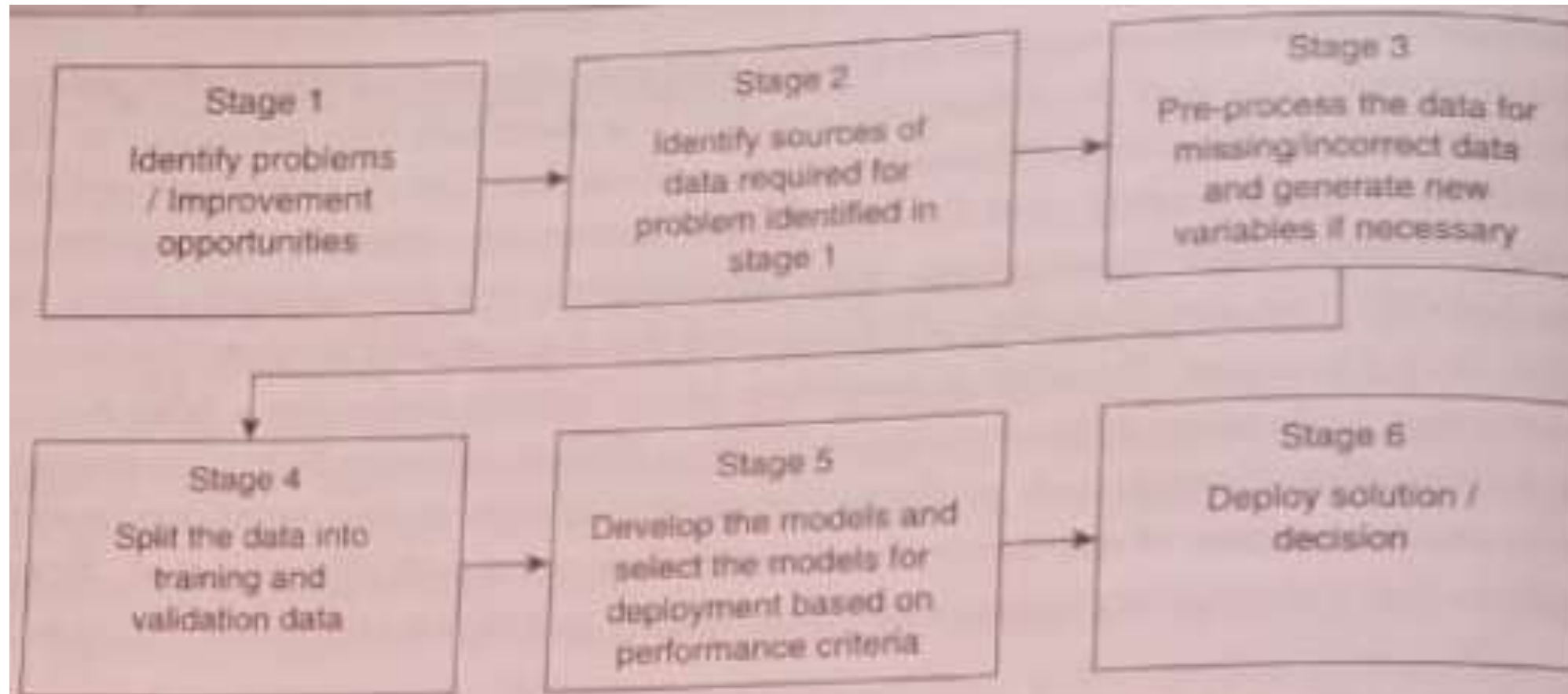
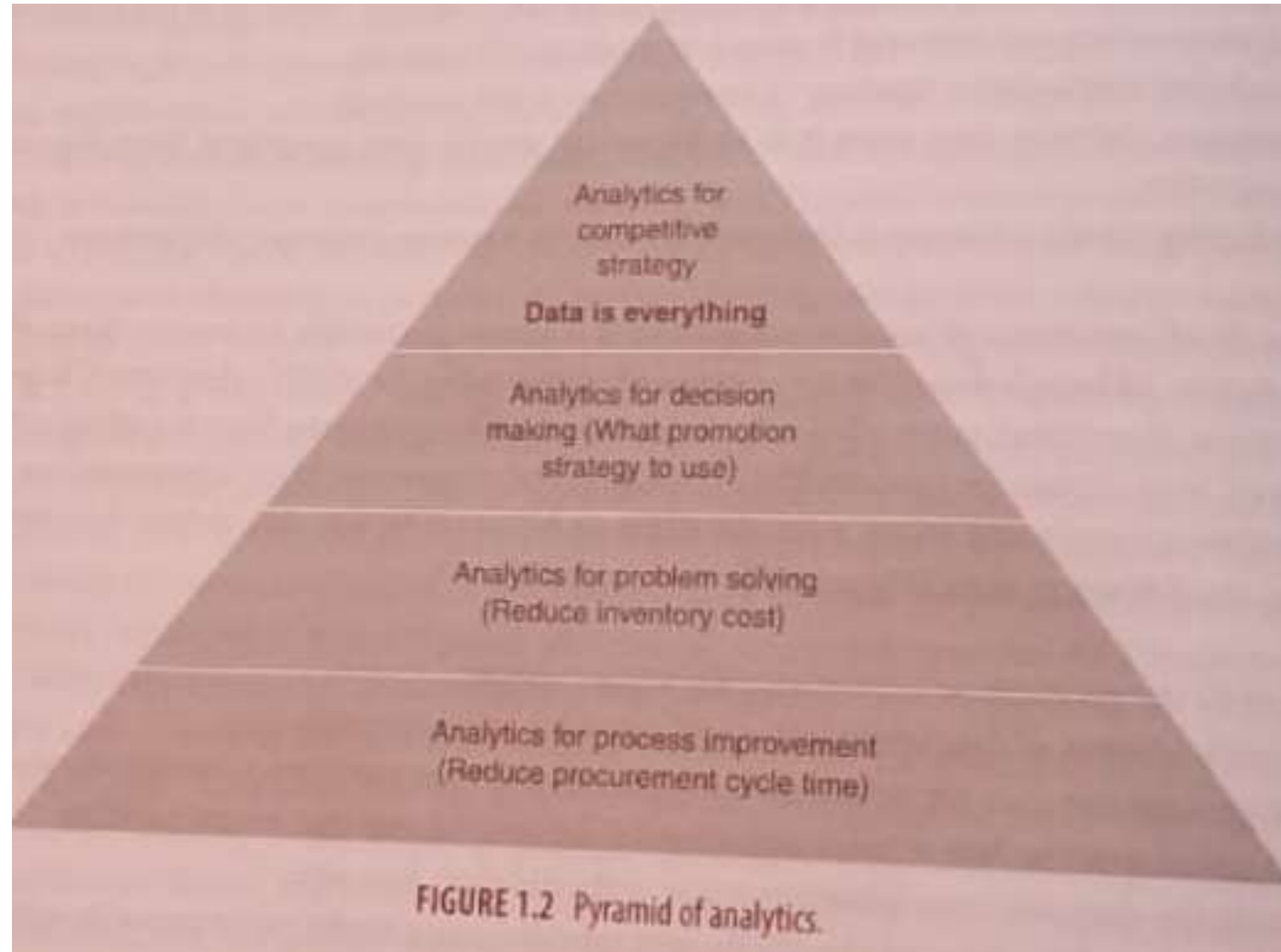


FIGURE 1.1 Business analytics – Data-driven decision-making flow diagram.

Pyramid of analytics



Theory of Firm to minimize the transaction cost

- Production cost
- Implementation cost
- Success/Failure cost

Example : Travelling salesman problem

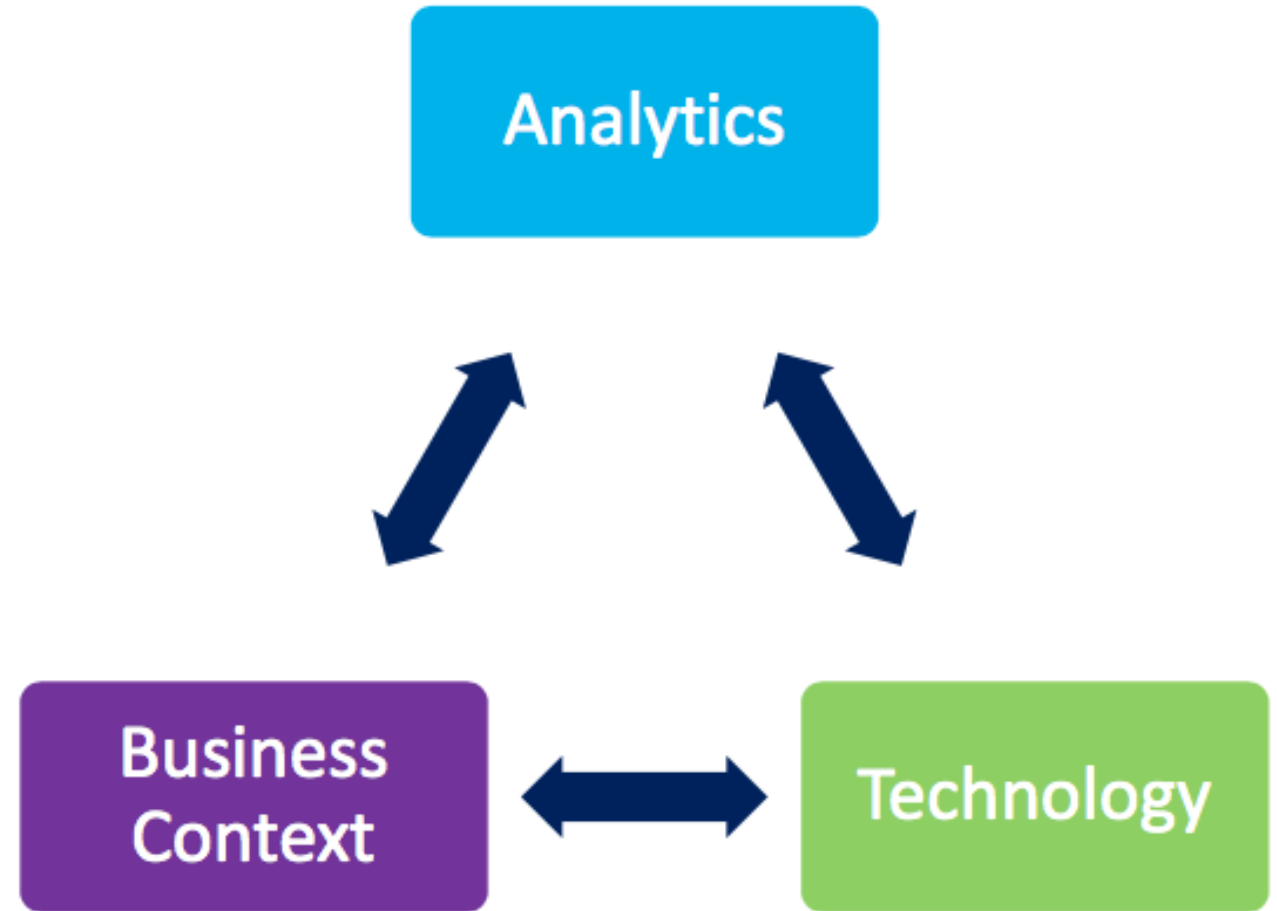


Business analytics: The science of data driven decision making

- Many companies collect data manually
- Lack of data
- Partially solved the problem of non-availability of the data
- Enterprise resource planning(ERP) system were not able to build analytics models
- To fill the gap BA can broken into 3 components

Components of BA

- 1 Business context
- 2 Technology
- 3 Data science (analytics)



Business context

- Targets pregnancy test/prediction
- Smart basket features @bigbasket.com

Technology

- Examples data has to be capture
- Data storage (structured, unstructured, semi structured)
- Data preparation
- Data analysis (R, python, SAS, SPSS, Tableau etc)
- Data share

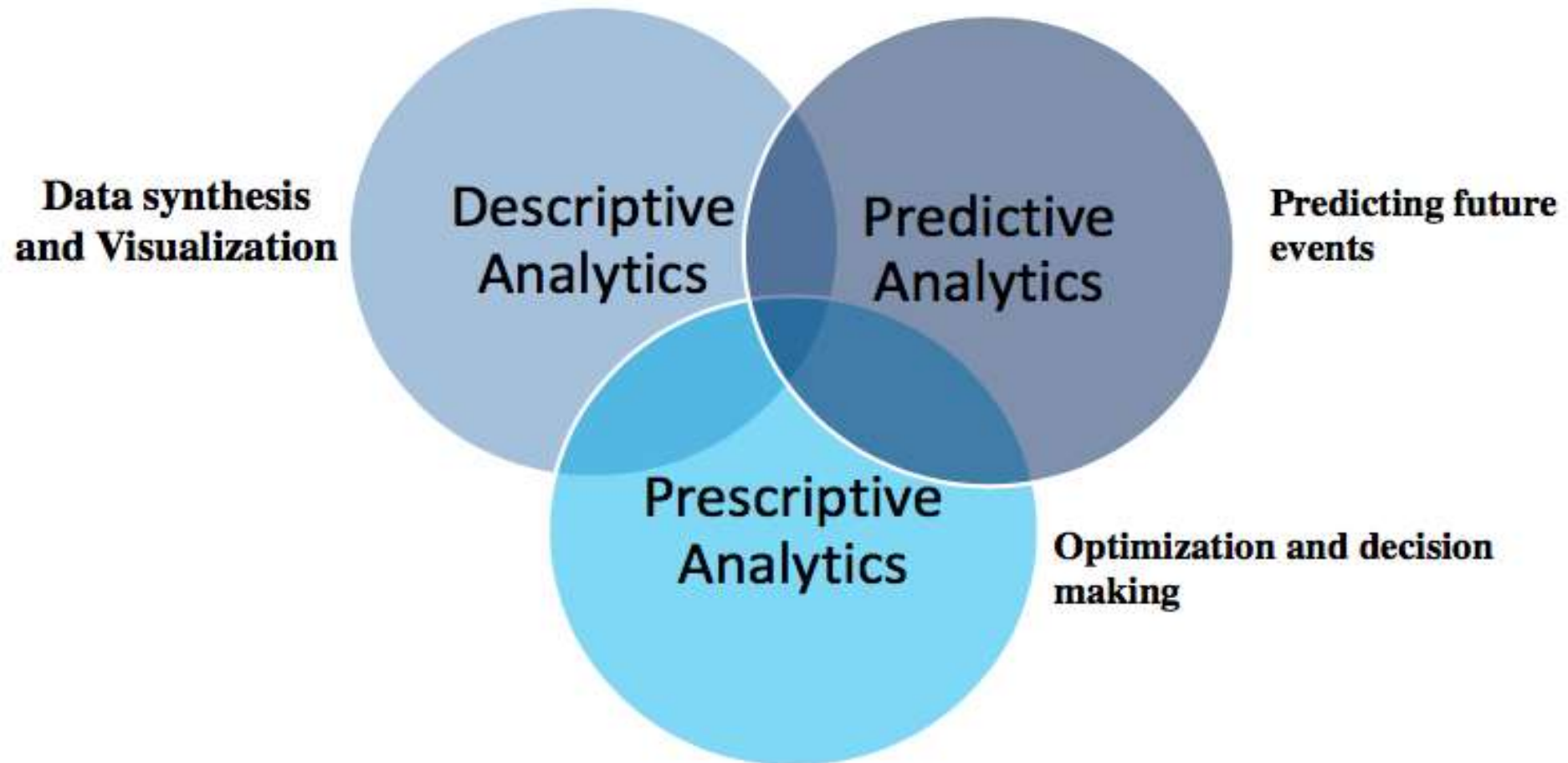
Data science (analytics)

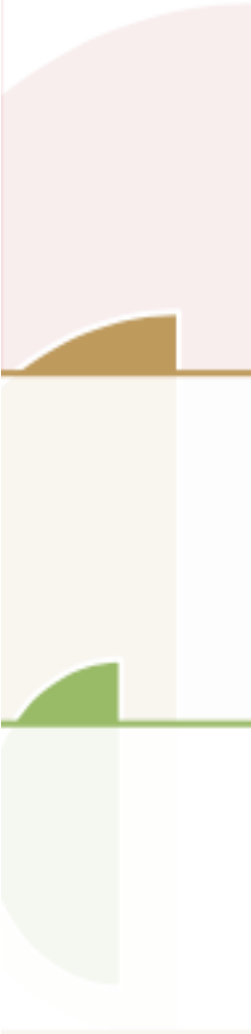
- It consists of statistics & operation research techniques
 - ML,
 - Deep learning
 - Data mining
-
- Identify appropriate statistical model/ML algorithms
 - Classification problems: logistic regression, RF, AdaBoost, Neural network etc.

BA grouped into three types

Analytics

FIGURE 11





Descriptive analytics

- Communicates the hidden facts and trends in the data
- Simple analysis of data can lead to business practices that result in financial rewards
- Helps SMEs uncover inefficiencies and eliminate them

Predictive analytics

- Predicts the probability of occurrence of a future event
- Helps organizations to plan their future course of action
- Most frequently used type of analytics across several industries

Prescriptive analytics

- Assists users in finding the optimal solution to a problem
- In most cases, provides an optimal solution/decision to the problem
- Inventory management is one of the problems that are most frequently addressed

Facebook Relationship Breakups

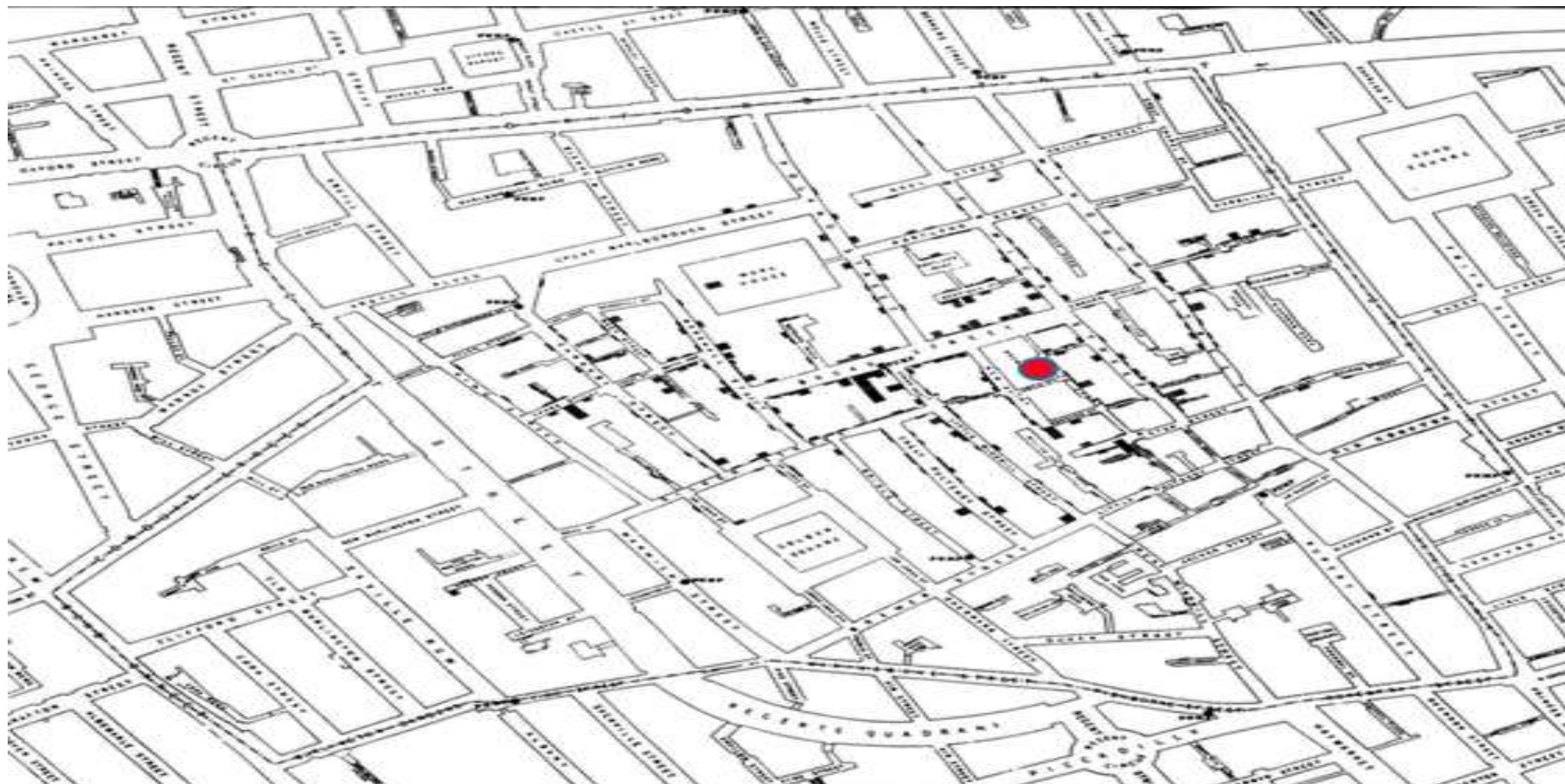


London Cholera Outbreak - 1854

- Severe outbreak of cholera that occurred near Broad Street (now Broad wick street) in Soho district of London in 1854.
- More than 500 people died within 10 days of the outbreak, the mortality rate in some parts of the city was as high as 12.8%.
- Prepared a spot map



Spot Map



Predictive Analytics

- Which product the customer is likely to buy in his next purchase ? (recommender system).
- Which customer is likely to default in his / her loan payment ? (credit risk).
- Who is likely to cancel the product that was ordered through e-commerce portal ?

TABLE 1.2 List of predictive analytics applications

| Organization | Predictive Analytics Model |
|------------------------|---|
| Polyphonic HMI | Predicts whether a song will be a hit using machine learning algorithms. Their product 'Hit Song Science' uses mathematical and statistical techniques to predict the success of a song on a scale of 1 to 10 (Anon, 2003). |
| Okcupid | Predicts which online dating message is likely to get a response from the opposite sex (Siegel, 2013). |
| Amazon.com | Uses predictive analytics to recommend products to their customers. It is reported that 35% of Amazon's sales is achieved through their recommender system (Siegel, 2013, MacKinzie <i>et al.</i> , 2013). |
| Hewlett Packard (HP) | Developed a flight risk score for its employees to predict who is likely to leave the company (Siegel, 2013). |
| University of Maryland | Claimed that dreams can predict whether one's spouse will cheat (Whitelocks, 2013). |
| Flight Caster | Predicts flight delays 6 hours before the airline's alerts. |
| Netflix | Predicts which movie their customer is likely to watch next (Greene, 2006). 75% of what customer watch at Netflix is from product recommendations (MacKinzie <i>et al.</i> , 2013). |
| Capital One Bank | Predicts the most profitable customer (Davenport, 2007). |
| Google | Predicted the spread of H1N1 flu using the query terms (Carneiro and Mylonakis, 2010). |
| Farecast | Developed a model to predict airfare, whether it is likely to increase or decrease, and the amount of increase/decrease. ^a |

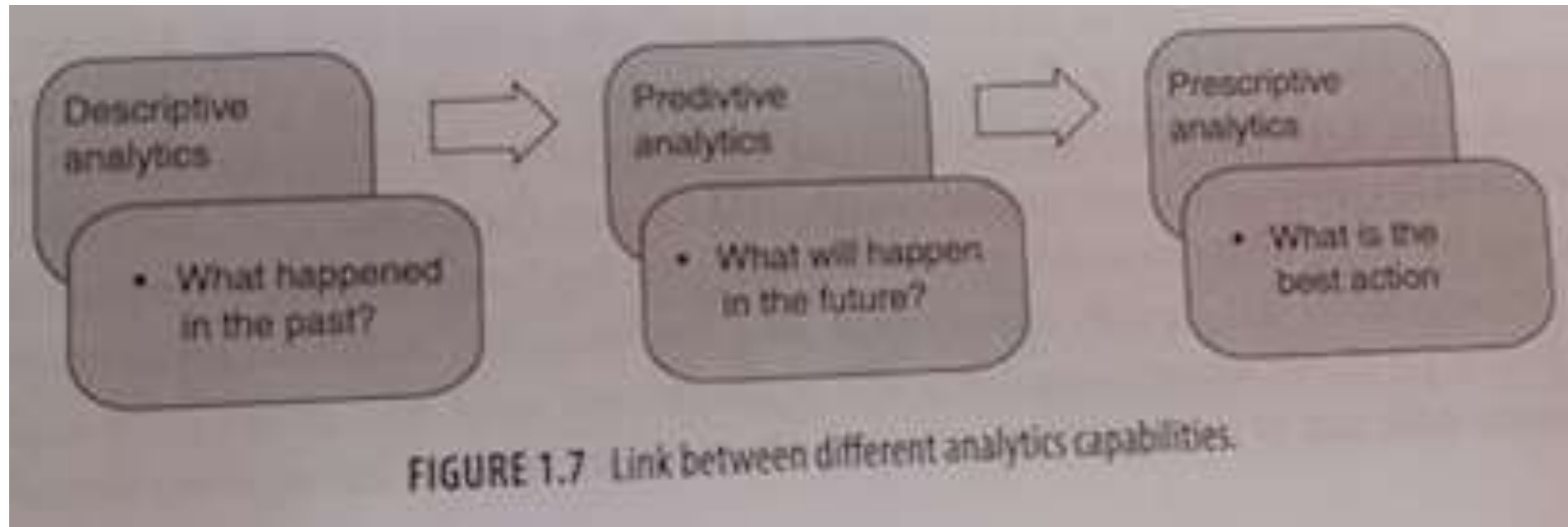
Prescriptive analytics

- Finding an optimal solution to a problem
- Making the right choice/decisions among several alternatives

Techniques

- OR
- ML
- Metaheuristics
- Advanced statistical models

Link b/w different analytics



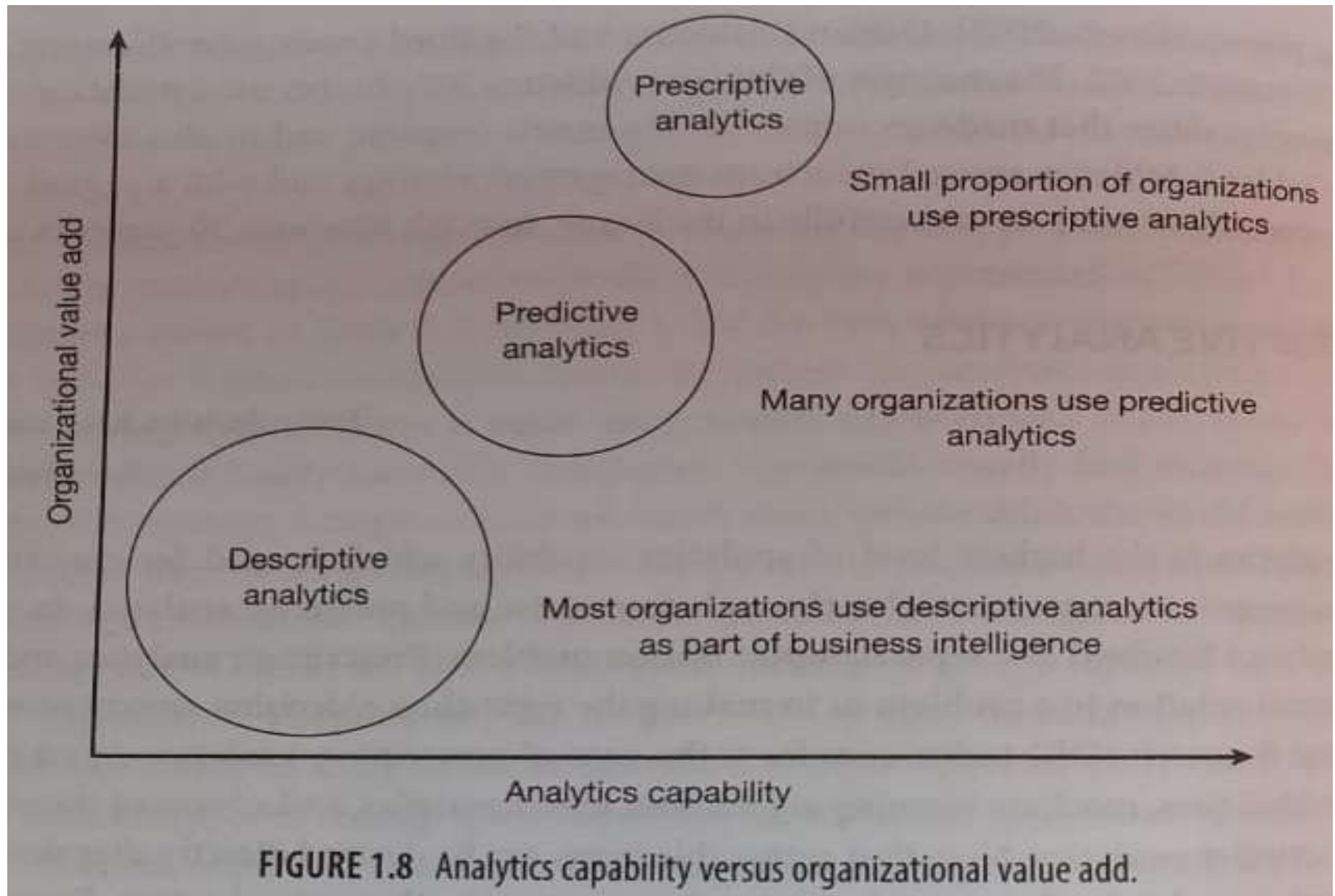


TABLE 1.3 Predictive and prescriptive analytics techniques

| Analytics Techniques | Applications |
|-------------------------------------|--|
| Regression | Regression is the most frequently used predictive analytics tool. It is a supervised learning algorithm. In management and social sciences, almost all hypotheses are validated using regression models. In business, irrespective of the sector, the decision maker would like to know how the key performance indicators (KPIs) of the business are related to macro-economic parameters and other internal process parameters. Regression is an excellent tool for establishing the existence of an association relationship between a response variable (KPI) and other explanatory variables. Unfortunately, regression is one of the most highly misused techniques in analytics. |
| Logistic and Multinomial Regression | Logistic and multinomial logistic regression techniques are used to find the probability of occurrence of an event. Logistic regression is a supervised learning algorithm. Logistic regression is used for solving classification and discrete choice problems. Classification problems are common in many businesses. For example, banks and financial institutions would like to classify their customers into several risk categories. Companies would like to predict which customer is highly likely to churn in the next quarter. Marketers would like to know which brand a customer is likely to buy and whether promotions can make a customer change his/her brand loyalty. Credit scoring and fraud detection are other popular applications of logistic regression. |
| Decision Trees | Decision trees or classification trees are usually used for solving classification problems. There are several types of classification tree models. Chi-Squared Automatic Interaction Detection (CHAID) and Classification Trees (CART) are frequently used for solving classification problems. Although the decision trees are usually used for solving classification problems (in which the outcome variable is discrete), they can also be used when the outcome variable is continuous. |
| Markov Chains | Olle Haggstrom (2007) wrote an article stating that problem solving is often a matter of cooking up an appropriate Markov chain. One of the initial applications of Markov chains was implemented by the American retail giant Sears. They used a Markov Decision Process to decide the optimal mailing policy for their catalogues (Howard, 2002). Today, Markov chains are one of the key analytics tools in marketing, finance, operations, and supply chain management. |
| Random Forest | Random forest is one of the popular machine learning algorithms that uses ensemble approach to solve the problem by generating a large number of models. |
| Linear Programming | Since its origins during World War II, linear programming is one of the most frequently used techniques in prescriptive analytics. Problems such as resource allocation, product mix, cutting-stock problem, revenue management, and logistics optimisation are frequently solved using linear programming. |
| Integer Programming | Many optimization problems in real life may have variables that can take only integer values. When one or more variables in the problem can take only an integer solution, the model is called an integer programming model. Capital budgeting, scheduling, and set covering are a few problems that are solved using integer programming. |

Big Data Analytics

- Extremely large data sets that may be analysed computationally to reveal patterns, trends, associations, and interactions.

This is identified using 4 Vs

- Volume
- Velocity
- Varsity
- Veracity

Value



Clinically relevant data
Longitudinal studies

Volume



High-throughput technologies
Continuous monitoring of vital signs

Velocity



High-speed processing for fast clinical decision support
Increasing data generation rate by the health infrastructure

Variety



Heterogeneous and unstructured data sources
Differences in frequencies and taxonomies

Veracity



Data quality is unreliable
Data coming from uncontrolled environments

Variability



Seasonal health effects and disease evolution
Non-deterministic models of illness and health

Techniques to solve Big data

- Apache Hadoop
- Map reduce
- Spark
- Pig
- Hive
- etc

Machine learning

- Machine learning is an application of artificial intelligence (AI) that provides systems the ability to automatically learn and improve from experience without being explicitly programmed.

Three dimensions of ML

- Learning strategies used by the system
- Knowledge or skill acquired by the system
- Application domain for which the knowledge is obtained.

ML classified into four categories

- Supervised
- Unsupervised
- Reinforcement
- Evolutionary

Web and social media analytics

- What is the most effective social media tactics?
- What was the best way to engage the costumers with social media?
- How to calculate the returns on investment?
- What are the best social media management tools?
- How do you create a social media strategy for the organization?

Engage the costumers

- Potential reach in a wider audience and create viral impact in short duration.
- SM important for marketing products and services.
- Relationship between social media and box office collection.
- It is less expensive than conventional media.

Social media measurable

In terms of

- Impressions
- Visits
- Views
- Clicks
- Comments
- Shares
- Likes
- Followers
- Fans
- Subscribers etc.

Return on investment (ROI)

- $\text{ROI} = (\text{gain from SM marketing} - \text{cost of SM marketing}) / \text{cost of SM marketing}$

Way of calculate ROI are:

- Return on engagement (ROE)
- Return on Influence
- Anecdots
- Correlation
- Multivariate testing
- Linking and tagging
- Social commerce approach
- Share of conversation
- Sentimental analysis

Return on engagement (ROE)

- Facebook- $(\text{number of likes, shares}) / (\text{total number of face book page likes})$
- Twitter- $(\text{number of replies retweets}) / (\text{total number of follwers})$
- Youtube- $(\text{number of comments, ratings and likes}) / (\text{number of video views})$. or
- $(\text{number of comments, ratings and likes}) / (\text{number of subscribers})$

Framework for data-driven decision making

Problem or Opportunity Identification

- Domain knowledge is very important at this stage of the analytics project.
- This will be a major challenge for many companies who do not know the capabilities of analytics.

Collection of relevant data

- Once the problem is defined clearly, the project team should identify and collect the relevant data.
- This may be an interactive process since "relevant data" may not be known in advance in many analytics projects.
- The existence of ERP systems will be very useful at this stage.

Data Pre-processing

- Data preparation and data processing forms a significant proportion of any analytics project.
- This would include data imputation and the creation of additional variables such as interaction variables and dummy variables in the case of predictive analytics projects.

Model Building

- Analytics model building is an iterative process that aims to find the best model.
- Several analytical tools and solution procedures will be used to find the best analytical model in this stage.

Communication of the data analysis

- The communication of the analytics output to the top management and clients plays a crucial role.
- Innovative data visualization techniques may be used in this stage.

Analytics capability building

- Top management support
- Analytics talent
- Information technology
- Innovation

Roadmap for Analytics capability building

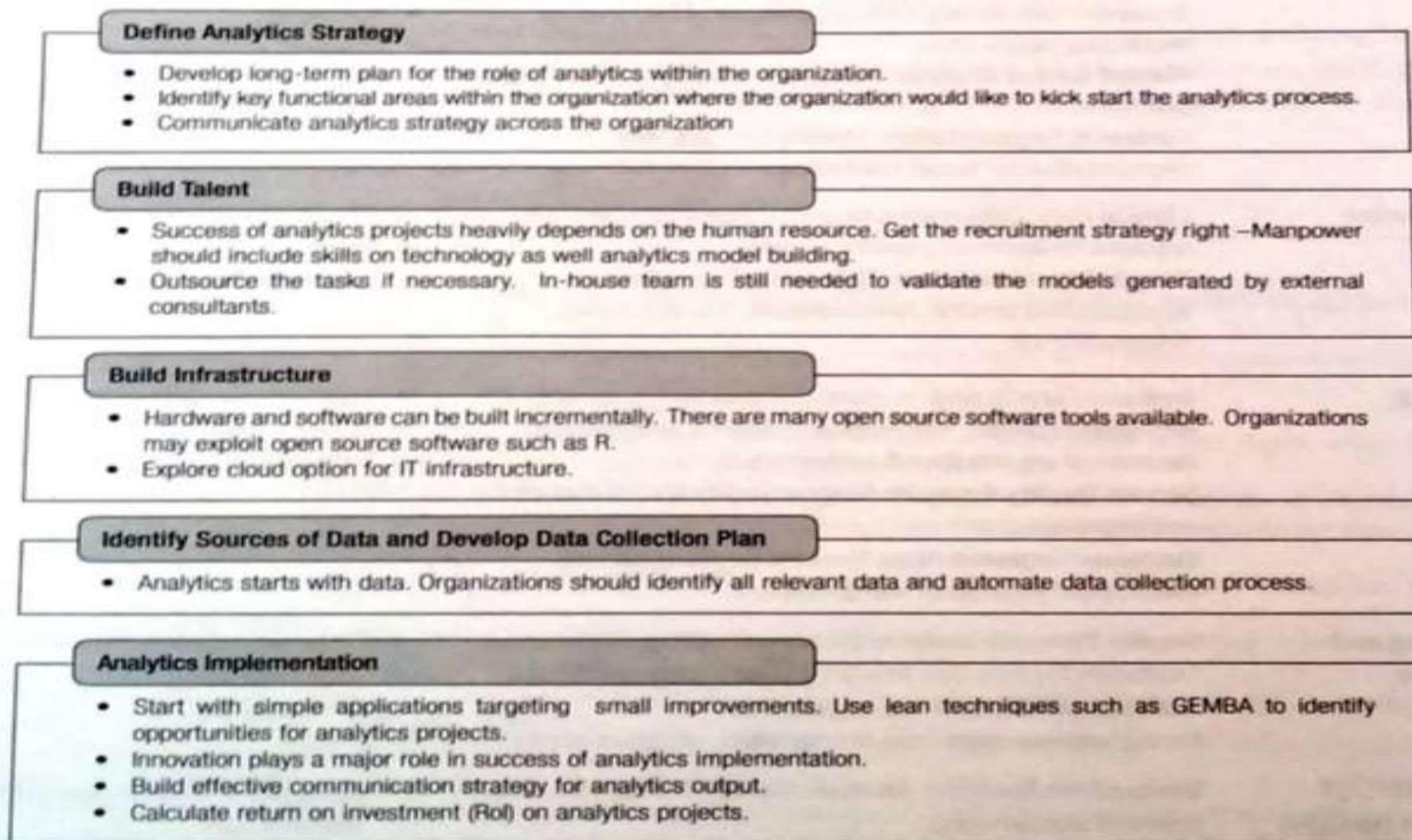


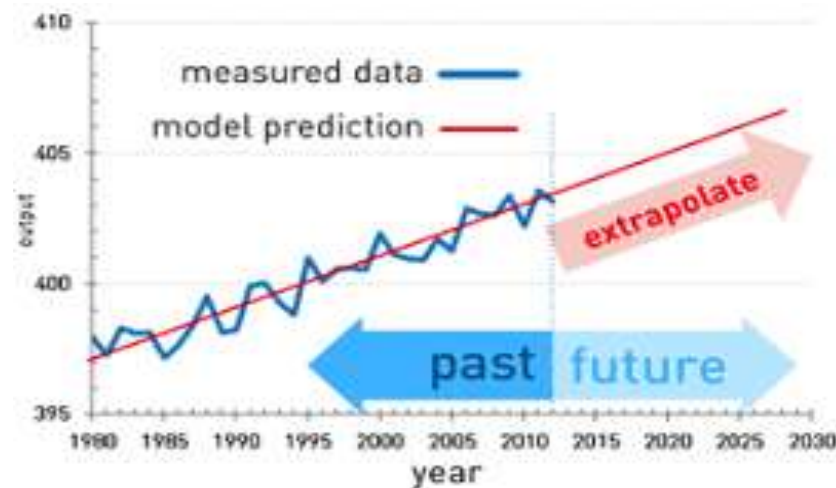
FIGURE 1.11 Roadmap for analytics capability building.

TABLE 1.4 Examples of industry-wise analytical problems and data resources

| Industry Sector | Sample Analytical Problems | Data Sources |
|--|---|--|
| Manufacturing | Supply Chain Analytics: Inventory management, procurement, vendor selection, distribution management Quality and Process Improvement: Product quality, manufacturing quality, process improvement Revenue and Cost Management: Revenue maximization and cost minimization. Warranty Analytics: Manage end customer warranty and after sales support. | <ul style="list-style-type: none"> ▪ Procurement, sales and production data. ▪ Warranty and after sales service data. ▪ Commodity price data ▪ Manufacturing data. ▪ Macroeconomic data. |
| Retail | Assortment Planning: Category and SKU (stock keeping unit) management that will maximize the revenue and improve loyalty. Promotion Planning: Decide promotion strategy such as temporary price cuts, markdowns, bundling, etc. Demand Forecasting: Forecast demand at SKU level for managing supply chain. Market Basket Analysis: Association among SKUs in customer purchase. Customer Segmentation: Identify the customer segmentation for target marketing. | <ul style="list-style-type: none"> ▪ Price data. ▪ Demand data at SKU and at category level. ▪ SKU level sales data with and without promotions. ▪ Planogram ▪ Customer demographics data. ▪ Point of Sales (PoS) data. ▪ Loyalty program data. |
| Healthcare | Clinical Care: Data related to clinical care and treatment required for improving quality of care. Hospitality related data: Data related to issues such as registration process, housekeeping, nursing, utility, diagnostics, etc. | <ul style="list-style-type: none"> ▪ All patient care related data ▪ Hospitality related data. ▪ Patient feedback data |
| Service | Demand Forecasting: Forecast demand for the service. NPS Optimization: Net Promoters Score is an important measure of organizational performance. Service Quality Analysis: Analyse quality for benchmarking and improvement. Customer Segmentation: Used for target marketing. Promotion: Promotion and its impact. | <ul style="list-style-type: none"> ▪ Transactional and feedback data ▪ Pricing and demand data ▪ Promotional data |
| Banking and Finance | Service Demand Analysis: Demand for different services. Customer Transaction Analysis: Used for many different analytics and decision-making insights. Credit Scoring: Important for managing different portfolios. | <ul style="list-style-type: none"> ▪ Customer transactional data ▪ Loan originating data ▪ Credit scoring data |
| IT and ITES (IT enables services) | Demand for Analytics Services: Identify demand for analytics products and services. Software Development Cycle Time: Cost and time reduction. | <ul style="list-style-type: none"> ▪ Customer interaction and market research data ▪ Internal product development data |

Descriptive Analytics

- Descriptive analytics consisted merely of the presentation of data in tables and charts; nowadays, it includes the summarization of data by means of numerical descriptions and graphs.
- **Predictive statistics** is an area of statistics that deals with extracting information from data and using it to predict trends and behavior patterns



Prescriptive analytics

- **Prescriptive analytics** is a process that analyzes data and provides instant recommendations on how to optimize business practices to suit multiple predicted

Data types and Scales

- Structured
- Unstructured
- Cross sectional
- Time series
- Panel data

Types of data measurement scales

- Nominal scale(Qualitative/categorical): example: married, unmarried
- Ordinal scale: feedback such as 1=poor, 2=fair, 3=good, 4=Vgood, 5=excellent
- Interval scale: temperature, score card such as A, B, C, D, E, F etc
- Ratio scale: These methods are generally implemented to compare two or more ordinal groups. Such as salarys

Two Basic Concepts—Population and Sample

- **Population:** is the set of all possible observations for a given context of the problem
- **unit:** A single entity, usually an object /records/subjects or person, whose characteristics are of interest.
- **Sample:** is the subset taken from a population.

| Population | Unit | Variables/Characteristics |
|--|------------|--|
| All students currently enrolled in school | student | GPA number of credits hours of work per week major right/left-handed |
| All printed circuit boards manufactured during a month | board | type of defects number of defects location of defects |
| All campus fast food restaurants | restaurant | number of employees seating capacity hiring/not hiring |
| All books in library | book | replacement cost frequency of checkout repairs needed |

Measures of central tendency

- Mean
- Median
- Mode
- Percentile
- Decile
- Quartile

Measures of Variation

- Range
- Boxplots(Inter quartile distance)
- Variance
- Standard deviation

BUSINESS INTELLIGENCE (BI)

- **Business intelligence (BI)** is an umbrella term that combines architectures, tools, data-bases, analytical tools, applications, and methodologies.
- **Business intelligence (BI)** combines **business** analytics, data mining, data visualization, data tools and infrastructure, and best practices to help organizations to make more data-driven decisions.
- **BI(Business Intelligence)** is a set of processes, architectures, and technologies that convert raw data into meaningful information

A Brief History of BI

- The term *BI* was coined by the Gartner Group in the mid-1990s.
- However, the concept is much older; it has its roots in the MIS reporting systems of the 1970s.
- During that period, reporting systems were static, two dimensional, and had no analytical capabilities.
- In the early 1980s, the concept of *executive information systems* (EIS) emerged.
- Today, a good BI-based enterprise information system contains all the information executives need. So, the original concept of EIS was transformed into BI.

Evolution of Business Intelligence (BI).

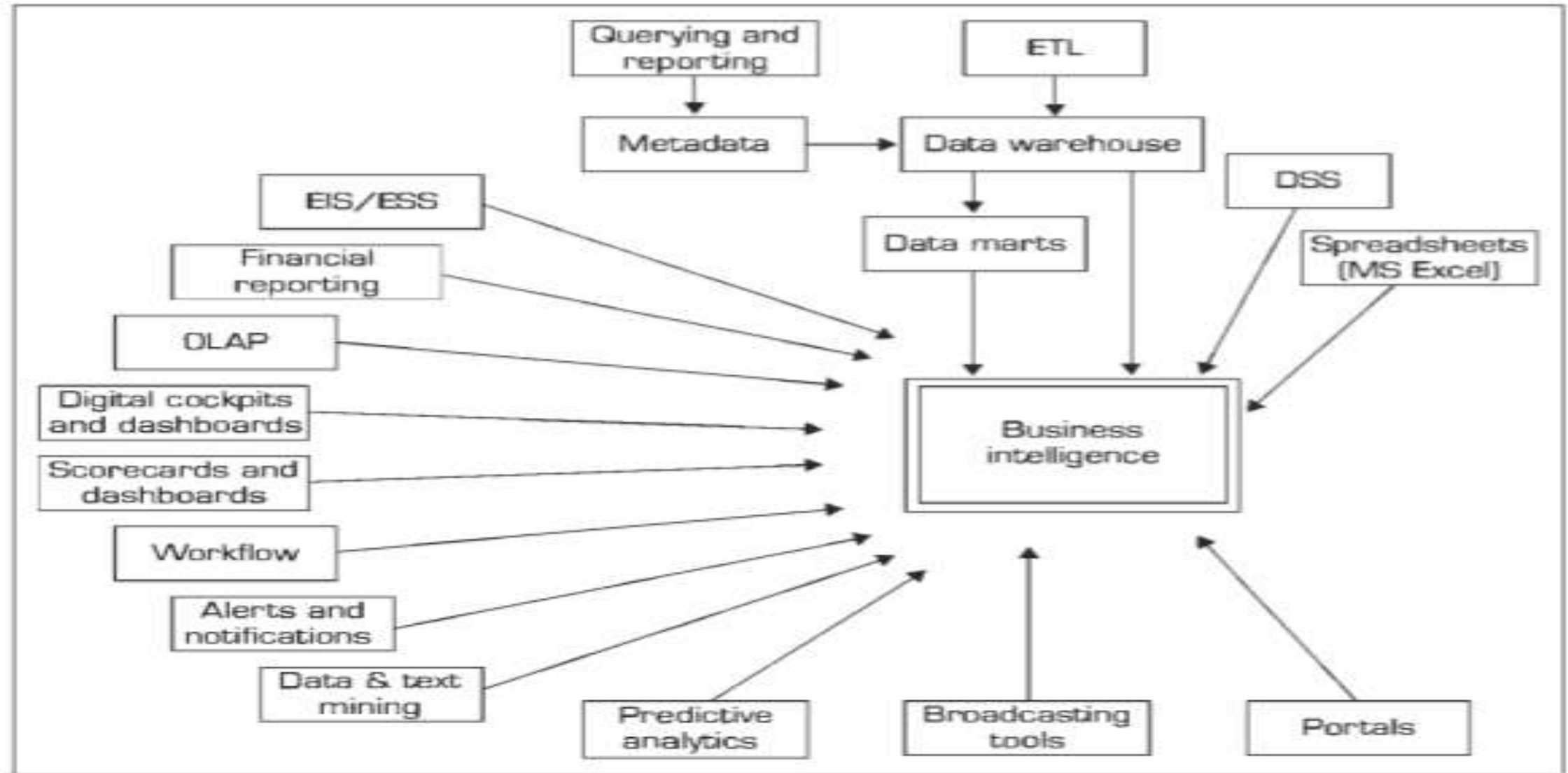
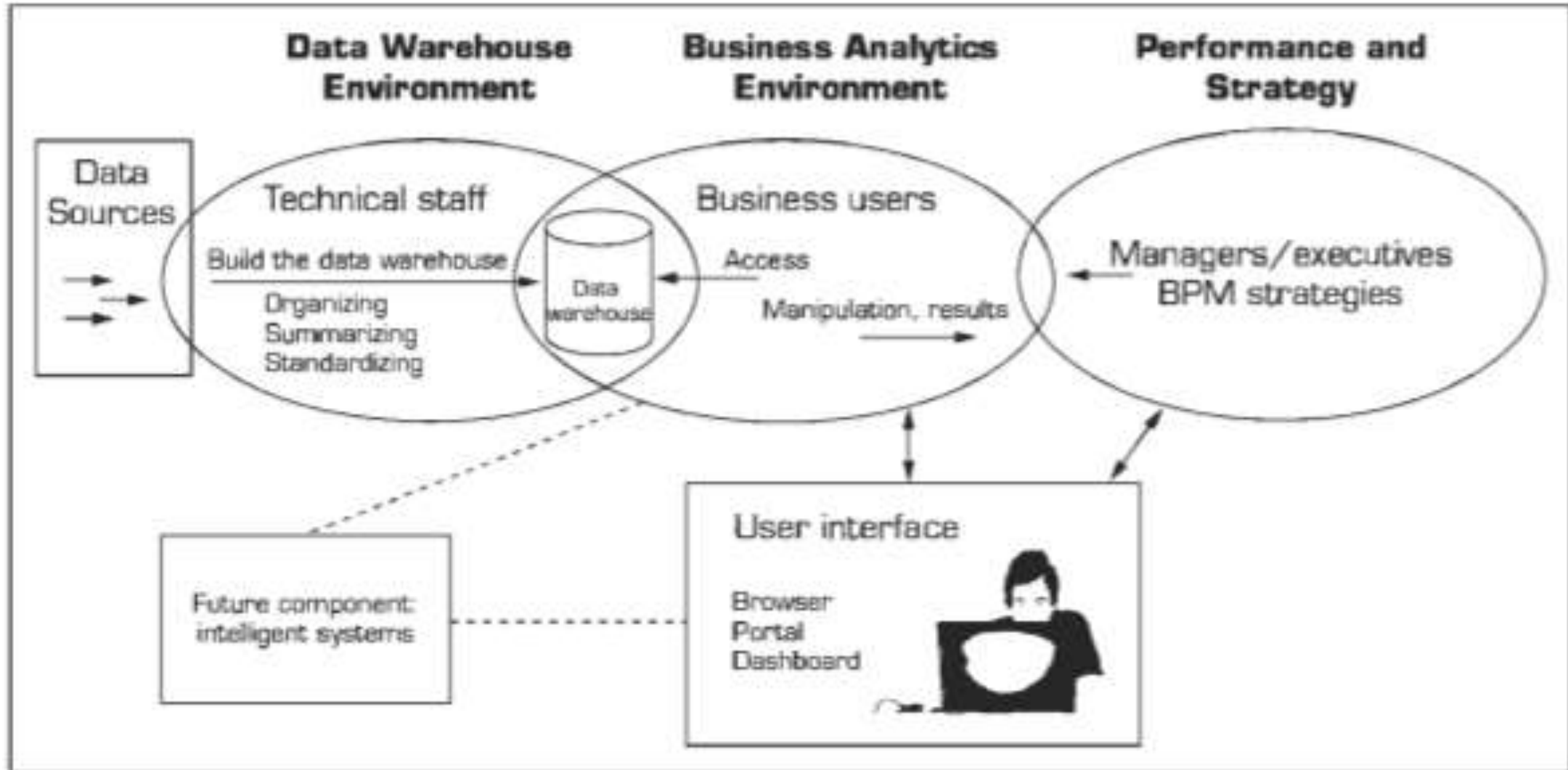


FIGURE 1.3 Evolution of Business Intelligence (BI).

The Architecture of BI



BI system has four major components:

- *A data warehouse*, with its source data;
- *Business analytics*, a collection of tools for manipulating, mining, and analyzing the data in the data warehouse;
- *Business performance management (BPM)* for monitoring and analyzing performance;
- *A user interface* (e.g., a dashboard).

Styles of BI

- The five styles are report delivery and alerting;
- enterprise reporting (using dashboards and scorecards);
- cube analysis (also known as slice-and-dice analysis);
- ad hoc queries;
- statistics and data mining.

TABLE 1.3 Business Value of BI Analytical Applications

| Analytic Application | Business Question | Business Value |
|-----------------------------|---|--|
| Customer segmentation | What market segments do my customers fall into, and what are their characteristics? | Personalize customer relationships for higher satisfaction and retention. |
| Propensity to buy | Which customers are most likely to respond to my promotion? | Target customers based on their need to increase their loyalty to your product line. Also, increase campaign profitability by focusing on the most likely to buy. |
| Customer profitability | What is the lifetime profitability of my customer? | Make individual business interaction decisions based on the overall profitability of customers. |
| Fraud detection | How can I tell which transactions are likely to be fraudulent? | Quickly determine fraud and take immediate action to minimize cost. |
| Customer attrition | Which customer is at risk of leaving? | Prevent loss of high-value customers and let go of lower-value customers. |
| Channel optimization | What is the best channel to reach my customer in each segment? | Interact with customers based on their preference and your need to manage cost. |

Transaction Processing VERSUS Analytic Processing

- *Transaction processing* systems are constantly involved in handling updates to what we might call *operational databases*. For example, in an ATM withdrawal transaction, we need to reduce our bank balance accordingly; a bank deposit adds to an account.
- These **online transaction processing (OLTP)** systems handle a company's routine ongoing business.
- In contrast, a data warehouse is typically a distinct system that provides storage for data that will be made use of in *analysis*.
- This analysis is to give management the ability to scour data for information about the business, and it can be used to provide tactical or operational decision support.
- DWs are intended to work with informational data used for **online analytical processing (OLAP)** systems.

OLTP

- Online transaction processing applications have high throughput and are insert- or update-intensive in database management.
- These applications are used concurrently by hundreds of users.
- The key goals of OLTP applications are availability, speed, concurrency and recoverability.
- OLTP systems process all kinds of queries (read, insert, update and delete)
- Reduced paper trails and the faster, more accurate forecast for revenues and expenses are both examples of how OLTP makes things simpler for businesses.
- However, like many modern online information technology solutions, some systems require offline maintenance, which further affects the cost-benefit analysis of an online transaction processing system.

Data warehouses

- *Data warehouses* contain a wide variety of data that present a coherent picture of business conditions at a single point in time.
- The idea was to create a database infrastructure that is always online and contains all the information from the OLTP systems, including historical data, but reorganized and structured in such a way that it was fast and efficient for querying, analysis, and decision support.
- Separating the OLTP from analysis and decision support enables the benefits of BI that were described earlier and provides for competitive intelligence and advantage

OLAP

- A **cube** in OLAP is a multidimensional data structure (actual or virtual) that allows fast analysis of data.
- It can also be defined as the capability of efficiently manipulating and analyzing data from multiple perspectives.
- The arrangement of data into cubes aims to overcome a limitation of relational databases: Relational databases are not well suited for near instantaneous analysis of large amounts of data.
- Using OLAP, an analyst can navigate through the database and screen for a particular subset of the data (and its progression over time) by changing the data's orientations and defining analytical calculations.
- These types of user-initiated navigation of data through the specification of slices (via rotations) and drill down /up (via aggregation and disaggregation) is sometimes called "slice and dice."
- Commonly used OLAP operations include slice and dice, drill down, roll up, and pivot

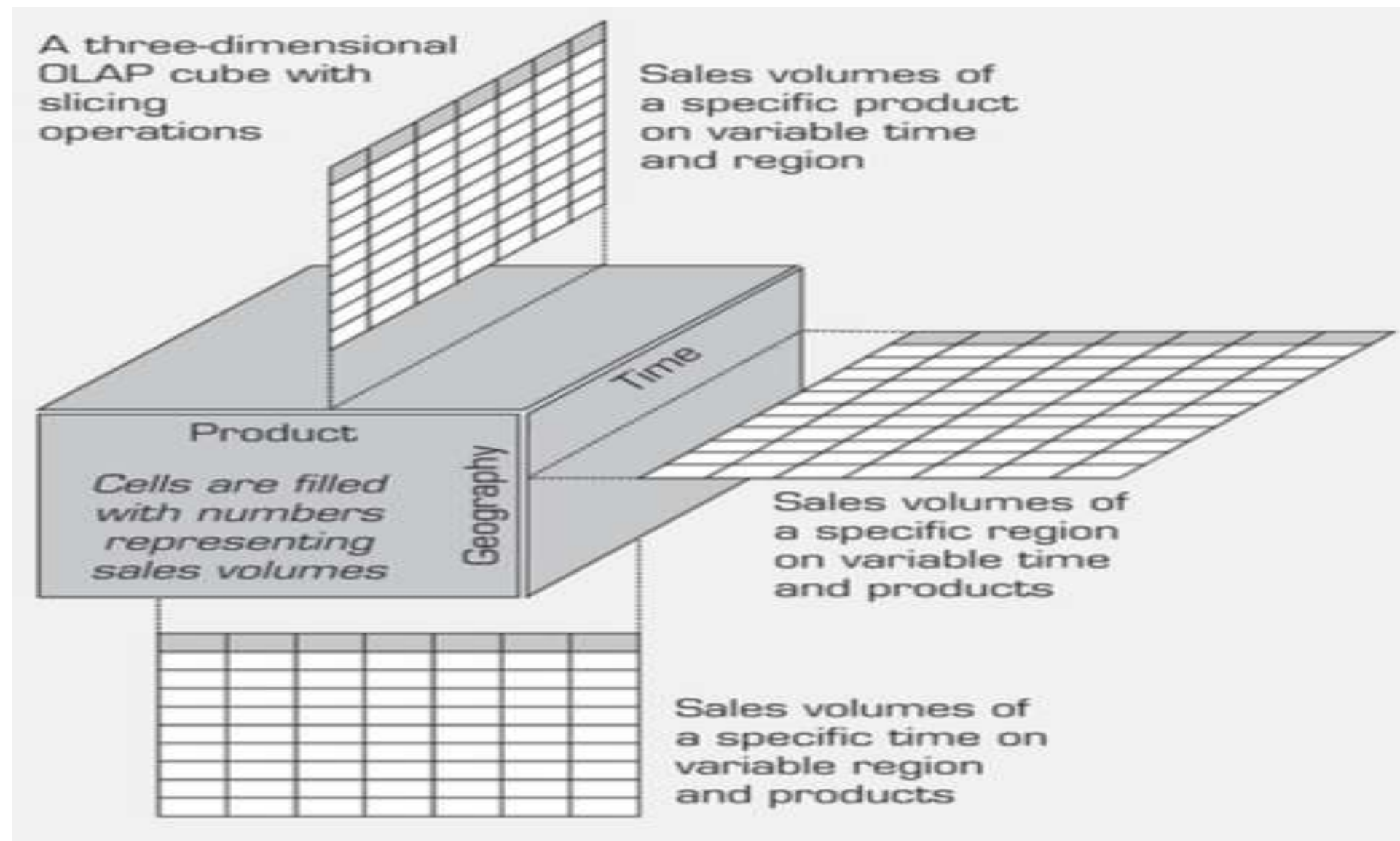


Figure 2.11 Slicing Operations on a Simple Three-Dimensional Data Cube.

OLAP operations

- Slice. A slice is a subset of a multidimensional array (usually a two-dimensional representation) corresponding to a single value set for one (or more) of the dimensions not in the subset. A simple slicing operation on a three-dimensional cube.
- Dice. The dice operation is a slice on more than two dimensions of a data cube.
- Drill Down/Up. Drilling down or up is a specific OLAP technique whereby the user navigates among levels of data ranging from the most summarized (up) to the most detailed (down).
- Roll-up. A roll-up involves computing all of the data relationships for one or more dimensions. To do this, a computational relationship or formula might be defined.
- Pivot. This is used to change the dimensional orientation of a report or ad hoc query-page display.

OLTP vs OLAP

| Criteria | OLTP | OLAP |
|-----------------------|---|--|
| Purpose | To carry out day-to-day business functions | To support decision making and provide answers to business and management queries |
| Data source | Transaction database (a normalized data repository primarily focused on efficiency and consistency) | Data warehouse or data mart (a nonnormalized data repository primarily focused on accuracy and completeness) |
| Reporting | Routine, periodic, narrowly focused reports | Ad hoc, multidimensional, broadly focused reports and queries |
| Resource requirements | Ordinary relational databases | Multiprocessor, large-capacity, specialized databases |
| Execution speed | Fast (recording of business transactions and routine reports) | Slow (resource intensive, complex, large-scale queries) |