SUB NAME: **WEB AND SOCIAL MEDIA ANALYTICS**

SUB CODE : **D87PC23**

web and social media analytics play a crucial role in shaping

Marketing strategies, **business decisions**, and **customer engagement**.

"Social Media Web Analytics" focuses on leveraging analytical tools and techniques to extract meaningful insights from social media platforms and web interactions.

This course prepares us to ***interpret data, measure performance, and optimize strategies*** to achieve organizational goals.

- **Key Learning Objectives**

➢ **Understand the Fundamentals of Analytics**: Learn how social media and web analytics work, including key metrics and methodologies.

➢ **Master Analytical Tools**: Gain hands-on experience with tools like Google Analytics, Facebook Insights, Hootsuite, and others.

➢ **Analyze and Interpret Data**: Develop the ability to extract actionable insights from social media and website data.

- **Optimize Digital Marketing Campaigns:** Use analytics to evaluate the performance of campaigns and recommend improvements.
- **Support Business Decision-Making:** Use data-driven insights to guide business strategy, product development, and customer experience

- **<u>Core Topics Covered</u>**

The course includes theoretical knowledge, practical applications, and case studies, covering a wide range of topics:

1. Introduction to Social Media and Web Analytics

- Definition and scope of analytics.
- Importance of analytics in digital marketing and business strategy.
- Key concepts: Traffic, conversions, engagement, and retention.

2. Understanding Social Media Platforms

- Overview of popular platforms like Facebook, Instagram, Twitter, LinkedIn, YouTube, and TikTok.
- Platform-specific metrics (e.g., reach, impressions, shares, likes).
- Differences between organic and paid metrics.

3. Web Analytics Fundamentals

- Components of web analytics: Audience, acquisition, behavior, and conversion.
- Understanding user behavior through metrics like bounce rate, time on site, and click-through rate (CTR).
- Basics of cookies, tracking pixels, and tags.

4. Key Performance Indicators (KPIs)

- Identifying relevant KPIs for different industries and goals.
- Setting benchmarks and measuring success.
- Examples: Engagement rate, customer lifetime value (CLV), cost per click (CPC).

5. Tools and Technologies

- **Social Media Analytics Tools**:
  - Facebook Insights, Twitter Analytics, Instagram Insights.
  - Third-party tools like Hootsuite, Buffer, and Sprout Social.
- **Web Analytics Tools**:
  - Google Analytics, Adobe Analytics, Matomo.
  - Heatmapping tools like Crazy Egg and Hotjar

6. Data Collection and Tracking

- Methods for tracking website and social media activity.
- UTM parameters and campaign tracking.
- Best practices for ensuring data privacy and compliance (e.g., GDPR, CCPA).

7. Data Visualization and Reporting

- Creating dashboards and visualizations to communicate insights.
- Tools: Google Data Studio, Tableau, Power BI.
- Building reports for different stakeholders.

8. Measuring Campaign Effectiveness

- Evaluating ROI of digital marketing campaigns.
- A/B testing and multivariate testing.
- Attribution models: First-touch, last-touch, and multi-touch attribution.

9. Sentiment Analysis and Social Listening

- Using tools like Brandwatch, Sprinklr, and Mention to gauge public sentiment.
- Understanding customer opinions and trends from social media discussions.

10. Predictive Analytics and Forecasting

- Leveraging historical data to predict future trends.
- Using machine learning algorithms for advanced analytics.

11. Challenges in Analytics

- Dealing with data overload and determining what is relevant.
- Addressing platform algorithm changes and data discrepancies.
- Ethical considerations in data collection and usage.

12. Emerging Trends

- AI in social media and web analytics.
- Real-time analytics for dynamic decision-making.
- Integration of social commerce data with analytics.

- **Skills Acquired**

Upon completing this course, we will have:

Analytical Thinking: Ability to interpret complex data sets and identify trends.

Tool Proficiency: Expertise in social media and web analytics tools.

Strategic Planning: Capability to use data to optimize campaigns and business strategies.

Reporting Skills: Proficiency in presenting data insights through dashboards and visualizations.

Problem-Solving: Skills to address real-world challenges in analytics and marketing.

Career Opportunities

- Social media and web analytics skills are in high demand, leading to various career paths:
- **Social Media Analyst**: Analyzing platform performance and recommending strategies.
- **Digital Marketing Manager**: Overseeing campaigns and measuring their effectiveness.
- **Web Analyst**: Monitoring website performance and suggesting optimizations.
- **SEO/SEM Specialist**:SEO (Search Engine Optimization)/SEM (Search Engine Marketing) Optimizing search engine rankings and paid ad performance.
- **Content Strategist**: Using analytics to plan and refine content strategies.
- **Data Analyst**: Applying analytics beyond marketing, in broader business contexts.

- **What is WSMA in simple definition**:

Web and Social Media Analytics (WSMA) is the study of ***collecting***,

measuring, and

analyzing data from:Websites (like page visits, clicks)Social media platforms (likes, shares, comments)

- To understand user behavior and improve business decisions.

- **<u>WHY do we study WSMA?</u>**

To know how users interact with websites/apps to improve marketing strategies to increase traffic, engagement, and sales to track performance using KPIs (Key Performance Indicators)

Example:If a website has high visitors but low sales → WSMA helps find the problem.

- **<u>HOW does WSMA work?</u>**

WSMA follows these steps:

***Data Collection**

Tools like Google Analytics, Facebook Insights

*Data Processing

Cleaning and organizing data

*Data Analysis

Finding patterns (e.g., most visited page, peak time)

***Visualization**

Graphs, dashboards

***Decision Making**

Improve content, ads, UI/UX (User Interface/User Experience)

- **WHERE is WSMA used?**

➢ E-commerce websites (Amazon, Flipkart)

➢ Social media platforms (Instagram, Twitter)

➢ Digital marketing companies

➢ News/media websites

➢ Educational platforms

- **WHEN is WSMA used?**

➢ During marketing campaigns

➢ After launching a website/app

➢ While analyzing user engagement trends

➢ For continuous performance improvement.

Core concepts in WSMA.

Types of Analytics

- Descriptive Analytics → What happened? (reports, dashboards)
- Diagnostic Analytics → Why did it happen?
- Predictive Analytics → What will happen? (uses ML)
- Prescriptive Analytics → What should we do?

Analytical Techniques

- ***Sentiment Analysis*** → detects positive/negative opinions
- ***Text Mining*** → analyzing comments/reviews
- ***Network Analysis*** → studying user connections
- ***A/B Testing*** → comparing two versions of a webpage

Purpose of WSMA Subject

- The main purpose of Web & Social Media Analytics (WSMA) is:

To analyze online data (websites + social media) and convert it into useful insights for decision-making.

In simple words***:"Understand user behavior on the internet and use that knowledge to solve real-world problems."***

- **UNIT–I: Business Intelligence & Analytics**

Purpose:

- To understand how data helps in decision making.
- What is Business Intelligence (BI)
- Types of analytics:
- Descriptive → What happened
- Diagnostic Analytics → Why did it happen?
- Predictive → What may happen
- Prescriptive → What action to take Decision Support Systems (DSS)

Outcome:You learn how companies use data to make smart decisions

- **UNIT–II: Text Analytics & Text Mining**

Purpose:

To extract useful information from text data.

- Basics of Natural Language Processing (NLP)
- Text preprocessing:
  - *****Tokenization
  - *****Stop-word removal
  - *****Stemming
- Tools for text mining

Outcome: we learn how to analyze text like reviews, comments, messages

- **UNIT–III**: **Sentiment Analysis**

Purpose:

To understand ***people's opinions and emotions***

- Opinion mining
- Sentiment classification:

*Positive / Negative / Neutral

Techniques:

*Machine learning Lexicon-based

Outcome: we can analyze public opinion from social media

- **UNIT–IV: Web Analytics & Web Mining**
- **Purpose:**
- To analyze website data and user behavior

Web mining types:

*Content mining

*Structure mining

*Usage mining

*SEO (Search Engine Optimization)

*Tools like Google Analytics

Outcome: we learn how websites track:

- Visitors
- Clicks
- Traffic sources

- **UNIT–V: Social Media Analytics**
- **Purpose:**

To analyze social networks and trends

- Social network concepts:

    ***Nodes (users)

    ***Edge(connections)

- Metrics:

    ***Centrality

    ***Engagement

Trend analysis & user behavior

Outcome:we understand how platforms like Instagram and Twitter analyze users and trends

UNIT-1 : OVERVIEW OF BUSINESS INTELLIGENCE(BI), ANALYTICS & DECISION SUPPORT SYSTEMS.

a) Introduction to **Business Intelligence(BI).**

Business intelligence (BI) is a ***set of technologies, processes, and tools that transform raw data into actionable insights for decision-making***.

It enables organizations to
monitor performance,
Identify market trends, and
improve operational efficiency through
data visualization, analytics, and reporting.

Core Components of Business Intelligence(BI)

- **Data Gathering & Storage**: Collecting data from various sources into a centralized location, such as a data warehouse.
- **Data Analysis**: Using analytical tools to find trends, patterns, and relationships in historical and current data.
- **Data Visualization**: Presenting findings through interactive dashboards, graphs, and charts to make data easy to understand.
- **Actionable Insights:** Translating data into practical information that allows leaders to make informed, data-driven decisions

Key Benefits of BI

- **Improved Decision Making**: Decisions are based on data rather than intuition.
- **Competitive Advantage**: Helps organizations identify new business opportunities and react quickly to market changes.
- **Increased Efficiency**: Automates data analysis to save time and streamline operations.
- **Performance Monitoring**: Tracks key performance indicators (KPIs) against company goals.

- **Common BI Tools**

Modern BI solutions prioritize self-service analytics and data governance. Popular BI tools include:

***Microsoft Power BI**

***Tableau**

***Qlik**

***Google Looker**

- **Applications of BI**
- **Retail:** Analyzing purchasing patterns and managing inventory levels.
- **Finance:** Detecting fraudulent activities and assessing risk.
- **Healthcare:** Improving patient outcomes and optimizing resource allocation

b) WHAT IS ANALYTICS?

In Web and Social Media Analytics (WSMA), analytics refers to the **systematic, computational analysis of data collected from websites and social media platforms to gain actionable insights, optimize user experience, and drive informed business decisions**.It involves monitoring, collecting, and analyzing data patterns to optimize performance, such as tracking visitors to improve website effectiveness.

Core Components: It typically covers four types of data analysis:

- **Descriptive:** What happened?
- **Diagnostic:** Why did it happen?.
- **Predictive:** What is likely to happen?
- **Prescriptive:** What should we do about it?.

C) DECISION SUPPORT SYSTEMS(DSS)

A Decision Support System (DSS) in Web and Social Media Analytics (WSMA) is an interactive, computer based system that helps managers use data, models, and analytical tools to solve unstructured or semi structured problems.

It provides insights for decision making rather than automating decisions entirely.

Key Aspects of DSS in WSMA:

- **Definition & Purpose:** DSS combines data, models, and user expertise to assist in complex decision-making scenarios. It aids in planning and analyzing alternatives, often at the management control and strategic levels.

Components of a DSS:

- **Database Management System (DBMS):** Stores and retrieves relevant data (internal/external).
- **Model Management System:** Provides analytical models, such as forecasting or optimization.
- **User Interface:** Allows interaction between the user and the system.
- **Knowledge Management Subsystem:** Supports the other components with intelligent insights.

- **Types of DSS:** Data-driven (large datasets), Model-driven (mathematical models), and Knowledge-driven (AI/expert systems).

Functions & Capabilities:

- **What-if Analysis:** Evaluates the impact of changing variables.
- **Sensitivity Analysis:** Determines how different variables affect outcomes.
- **Goal Seeking:** Identifies necessary actions to achieve a specific target.
- **Exception Reporting:** Monitors and alerts on variables outside defined ranges

- **Role in WSMA:** In the context of WSMA in this DSS aids in
- sentiment analysis,

    predictive modeling, and

    analyzing social media data

to aid in marketing or strategic decisions.

- **Benefits:** Improves decision quality by offering

detailed, timely, and analytical data, reducing the time required for decision-making.

- TOPIC-1

TOPIC-1: VACCINE SUPPLY CHAIN ANALYTICS

The **Vaccine Supply Chain** is frequently used as a primary case study **to illustrate how real-time data, analytics, and IoT devices are used to prevent, predict, and prescribe solutions for maintaining critical cold chain integrity.**

Supply chain analytics is the use of data analysis, modeling, and technology to evaluate, optimize, and manage the end-to-end supply chain—from procurement and production to distribution. It transforms massive operational datasets into actionable insights, improving efficiency, reducing costs, forecasting demand, and enhancing decision-making regarding inventory and logistics.

The **vaccine supply chain** usually refers to the **cold chain system** used to safely store, transport, and distribute vaccines from manufacturers to hospitals and patients.

A vaccine supply chain includes:

- **Manufacturing** – vaccines are produced in factories.
- **Storage** – vaccines are stored in temperature-controlled refrigerators/freezers.
- **Transportation** – vaccines are moved through cold-storage vehicles and insulated containers.
- **Distribution** – vaccines reach hospitals, clinics, and health centers.
- **Administration** – vaccines are finally given to patients.

Leveraging Analytics for Effective and Safe Vaccine Supply Chain Management

1. Demand Forecasting & Planning
2. Inventory Optimization
3. Cold Chain Monitoring (Safety-Critical)
4. Distribution & Route Optimization
5. Traceability & Anti-Counterfeiting
6. Risk Management & Resilience
7. Performance Monitoring & KPIs
8. AI & Advanced Analytics Trends

1. Demand Forecasting & Planning

- Predicting vaccine demand using historical data, seasonal trends, population data, and disease outbreaks.
- Helps avoid vaccine shortages and overstock situations.
- **Uses:**
  - AI/ML forecasting models
  - Population analytics
  - Disease spread prediction
- **Benefits**
- Better production planning
- Reduced wastage
- Timely vaccine availability

2. Inventory Optimization

Managing stock levels efficiently across warehouses and hospitals.

Ensures the right quantity of vaccines is available at the right location.

Techniques

- EOQ (Economic Order Quantity)
- Safety stock analysis
- Automated inventory tracking

Benefits

- Reduced expired vaccines
- Lower storage cost
- Improved availability

3. Cold Chain Monitoring (Safety-Critical)

Vaccines must be stored within specific temperature ranges.

Analytics Used

- IoT sensors
- Real-time temperature monitoring
- Predictive alerts

Example

- If temperature exceeds safe limits, alerts are generated immediately.

Benefits

- Prevents vaccine spoilage
- Maintains vaccine effectiveness
- Ensures patient safety

4. Distribution & Route Optimization

Finding the fastest and safest routes for vaccine transportation.

Technologies

- GPS tracking
- Route optimization algorithms
- Traffic and weather analytics

Benefits

- Faster delivery
- Reduced transportation cost
- Better rural area coverag

5. Traceability & Anti-Counterfeiting

Tracking vaccines from manufacturer to patient.

- **Methods**
- Barcode systems
- RFID tags
- Blockchain technology

Benefits

- Prevents fake vaccines
- Improves transparency
- Easy product recall management

6. Risk Management & Resilience

- Identifying and reducing supply chain risks.

Risks

- Transportation delays
- Equipment failure
- Pandemic disruptions
- Cybersecurity threats

Analytics Role

- Risk prediction models
- Scenario simulation
- Backup supply planning

Benefits

- Improved emergency preparedness
- Faster recovery from disruptions

7. Performance Monitoring & KPIs

KPIs (Key Performance Indicators) are used to measure supply chain efficiency.

Common KPIs

- Delivery time
- Vaccine wastage rate
- Inventory turnover
- Cold chain compliance rate
- Order fulfillment accuracy

Benefits

- Continuous improvement
- Better decision-making
- Increased operational efficiency

8. AI & Advanced Analytics Trends

Modern vaccine supply chains increasingly use AI and advanced analytics.

Emerging Trends

- Machine Learning forecasting
- Digital twins
- Predictive maintenance
- Real-time dashboards
- Blockchain integration

Benefits

- Smart decision-making
- Automation
- Higher accuracy and safety

Types of supply chains:

There are **four main supply chain models** in use today:

- **the continuous-flow model**, **(This model is used when demand is stable and products are produced continuously in large quantities.)**
- **fast model,** **(This model focuses on quickly responding to changing market trends and customer demands.)**
- **efficient model**, **(This model is designed to maximize efficiency and reduce waste in highly competitive markets)**
- **custom-configured model.** **(This model emphasizes flexibility and quick adaptation to unpredictable demand changes)**

Each model plays a specific role in managing and optimizing the flow of a business's products or services

Types of data analyst…

The **four types of data analytics** are

descriptive (**<u>what</u>** happened),

diagnostic (**<u>why</u>** it happened),

predictive (**<u>what</u>** will happen), and

prescriptive (**<u>what to do</u>** about it)

Each type answers a different business question and builds on the others to move from insight to action.

1. Descriptive Analytics

Definition

- Descriptive analytics is the process of analyzing historical and current data to understand **what has happened** in the vaccine supply chain.
- **It converts raw data into meaningful information through reports, dashboards, graphs, and summaries**.

Purpose

The main purpose of descriptive analytics is:

- To monitor supply chain activities
- To track vaccine movement
- To measure operational performance
- To provide visibility into inventory and distribution

Data Used

- Inventory records
- Shipment details
- Storage temperatures
- Vaccination rates
- Delivery schedules
- Wastage reports

Applications in Vaccine Supply Chain

1. Inventory Monitoring

- Tracks:
- Number of vaccine doses available
- Vaccines used daily
- Remaining stock levels

2. Cold Chain Monitoring

- Monitors storage temperature history of vaccines in:
- Refrigerators
- Freezers
- Transport containers

3. Delivery Tracking

- Records:
- Delivery times
- Shipment locations
- Delayed deliveries
- **4. Wastage Analysis**
- Measures:
- Expired vaccines
- Damaged vaccines
- Unused doses

Tools Used

- Dashboards
- Reports
- Bar charts
- Data visualization tools
- Spreadsheet software

Advantages

- Improves operational visibility
- Helps monitor vaccine availability
- Detects inefficiencies
- Supports reporting and compliance

Limitations

- Only explains past events
- Does not predict future problems
- Cannot recommend solutions

- **Example**
- A health department dashboard displays:
- Total vaccines distributed
- Daily vaccinations
- Number of expired doses
- Regions with low stock levels
- This helps administrators understand the current status of vaccine distribution.

2. Diagnostic Analytics

Definition

- Diagnostic analytics examines data to determine **why something happened**.
- **It identifies root causes of problems and operational failures**.

Purpose

- The objectives are:
- To investigate issues
- To identify causes of disruptions
- To improve system performance

Data Used

- Sensor logs
- Temperature records
- Delivery records
- Equipment maintenance data
- Transportation reports

Applications in Vaccine Supply Chain

1. Cold Chain Failure Analysis

- Determines why vaccines were spoiled by checking:
- Power outages
- Refrigerator malfunction
- Transportation delays

2. Delivery Delay Investigation

- Analyzes causes such as:
- Traffic congestion
- Vehicle breakdowns
- Poor route planning

3. Stock-Out Analysis

- Finds reasons for shortages:
- Demand underestimation
- Poor inventory planning
- Supplier delays

- **4. High Wastage Investigation**
- Examines:
- Improper storage
- Incorrect handling
- Expired inventory
- **Techniques Used**
- Root cause analysis
- Correlation analysis
- Data mining
- Drill-down analysis

Advantages

- Identifies operational weaknesses
- Helps reduce future errors
- Improves vaccine safety
- Enhances process efficiency

Limitations

- Requires detailed historical data
- Time-consuming analysis
- May not prevent future uncertainties

Example

- A batch of vaccines became ineffective. Diagnostic analytics reveals:
- The freezer temperature increased above safe limits due to power failure.
- Backup generators failed to operate.
- This information helps prevent future cold-chain failures.

3. Predictive Analytics

Definition

- Predictive analytics uses **historical data, statistical models, and machine learning techniques to predict future events and trends.**

Purpose

- The goals are:
- To forecast demand
- To predict risks
- To improve preparedness
- To optimize resource allocation

Data Used

- Historical vaccination records
- Population health data
- Disease outbreak patterns
- Seasonal trends
- Transportation data

Applications in Vaccine Supply Chain

1. Demand Forecasting

- Predicts:
- Number of vaccine doses required
- Regional demand variations
- Seasonal vaccination needs

2. Outbreak Prediction

- Forecasts possible disease outbreaks based on:
- Infection rates
- Population movement
- Environmental conditions

3. Inventory Forecasting

- Estimates future stock requirements to avoid:
- Shortages
- Overstocking

4. Transportation Risk Prediction

- Predicts:
- Delivery delays
- Weather-related disruptions
- Traffic congestion

Techniques Used

- Machine learning
- Statistical forecasting
- Regression analysis
- Time-series analysis

Advantages

- Improves planning accuracy
- Reduces vaccine shortages
- Minimizes wastage
- Enhances emergency preparedness

Limitations

- Predictions may not always be accurate
- Requires quality historical data
- Complex models may be expensive

Example

- During influenza season, predictive analytics estimates vaccine demand in different cities based on previous years' data. Authorities distribute vaccines accordingly before shortages occur.

4. Prescriptive Analytics

Definition

- Prescriptive analytics recommends **what actions should be taken** to achieve desired outcomes.
- **It combines predictive analytics with optimization and decision-making models**.

Purpose

- The objectives are:
- To support decision-making
- To optimize supply chain operations
- To improve efficiency and safety

Data Used

- Predictive models
- Inventory data
- Transportation data
- Cost information
- Real-time sensor data

Applications in Vaccine Supply Chain

1. Route Optimization

- Suggests:
- Fastest delivery routes
- Fuel-efficient transportation plans
- Safe cold-chain paths

2. Inventory Optimization

- Recommends:
- Ideal stock levels
- Reorder quantities
- Emergency reserve stock

3. Vaccine Allocation

- Determines:
- Which regions should receive vaccines first
- Priority areas during outbreaks

4. Resource Scheduling

- Optimizes:
- Workforce allocation
- Vehicle scheduling
- Storage capacity usage

Techniques Used

- Optimization algorithms
- Simulation models
- Decision analysis
- Artificial intelligence

Advantages
- Improves operational efficiency
- Reduces transportation costs
- Enhances vaccine availability
- Supports quick decision-making

Limitations
- Requires accurate predictive data
- Complex implementation
- High computational requirements

Example
- If an outbreak occurs in a district, the system recommends:
- Sending additional vaccine doses immediately
- Using the fastest refrigerated transportation route
- Increasing storage capacity temporarily

5. Real-Time Analytics

Definition

- Real-time analytics **analyzes data immediately as it is generated.**

Purpose

- To monitor supply chain operations continuously and respond instantly to problems.

Applications

1. Temperature Monitoring

- Sensors continuously check vaccine storage temperatures.

2. GPS Tracking

- Tracks shipment location in real time.

3. Alert Systems

- Sends alerts when:
- Temperature exceeds safe range
- Deliveries are delayed
- Equipment malfunctions

Advantages

- Immediate response to issues
- Prevents vaccine spoilage
- Improves transparency

Example

- A refrigerated truck carrying vaccines experiences temperature rise. The system instantly sends alerts to operators for corrective action.

6. Risk Analytics

Definition

- Risk analytics **identifies and evaluates potential risks that may affect the vaccine supply chain**.

Purpose

- To reduce disruptions and improve preparedness.

Applications

1. Supplier Risk Assessment

- Evaluates supplier reliability.

2. Transportation Risk Analysis

- Analyzes risks from:
- Weather
- Road conditions
- Political instability

3. Storage Risk Assessment

- Identifies vulnerable storage facilities.

4. Cybersecurity Risk Analysis

- Protects digital healthcare systems from cyber attacks.

Advantages

- Improves supply chain resilience
- Enhances emergency planning
- Reduces operational disruptions

Example

- Risk analytics identifies a region prone to power failures. Authorities install backup generators in vaccine storage centers.

Benefits of Analytics in Vaccine Supply Chain

- Analytics helps to:
- Ensure vaccine safety and quality
- Maintain cold-chain integrity
- Reduce vaccine wastage
- Improve inventory management
- Enhance delivery efficiency
- Predict future demand
- Support public health planning
- Enable faster decision-making

TOPIC-2 Changing Business Environments and Decision Making

- Modern businesses operate in a rapidly changing environment influenced by **technology**, **globalization**, **competition**, **customer expectations**, **government regulations**, and **economic conditions**.
- To survive and grow, organizations must make timely and effective decisions.
- Computerized **Decision Support Systems (DSS)** help managers ***analyze information, evaluate alternatives, and make better decisions in dynamic business environments.***

[Changing Business Environments and Decision Support Systems]

- [Changing Business Environments]
- [Resource-based View]
- [Information and Knowledge – Organization's Strategic Resources]
- [Knowledge Management]
- [Creative Employees and Creative Organizations]
- [Information Communication Technology (ICT) as a Driver of Change and Innovation]
- [Information Communication Technology (ICT) in Business Value Creation]
- [Computer Decision Support Systems]
- [Designing Decision Support Systems]

Changing Business Environments

Meaning of Business Environment

- A business environment consists of **all internal and external factors that affect the operations, performance, and decision-making of an organization.**
- These factors continuously change, creating opportunities and challenges for businesses.

Definition

- Business environment refers to the ***totality of economic, social, technological, political, legal, and competitive forces that influence business activities***.

Key Types of Changes

- **Technological changes**
  New tools, AI, automation, digital platforms (e.g., online banking, e-commerce)
- **Economic changes**
  Inflation, recession, interest rates, income levels
- **Political & legal changes**
  New government policies, taxation rules, labor laws
- **Social & cultural changes**
  Lifestyle changes, consumer preferences, education levels
- **Globalization**
  Increased global competition and international markets

Impact on Business

- Forces companies to adapt quickly
- Creates both opportunities and threats
- Requires innovation and flexibility
- Makes long-term planning more uncertain

Example

- Retail businesses shifting from **physical stores to online platforms (Amazon, Flipkart)** due to digital transformation.

Resource-Based View (RBV)

Meaning

- The **Resource-Based View (RBV)** is a strategic management theory that says:
- A firm's competitive advantage comes from its internal resources and capabilities, not just external market conditions.
- The **Resource-Based View (RBV)** is a strategic management theory that explains ***how firms achieve and sustain competitive advantage through the unique resources and capabilities they control***. It emphasizes internal factors—rather than external market conditions—as primary determinants of performance differences among firms.

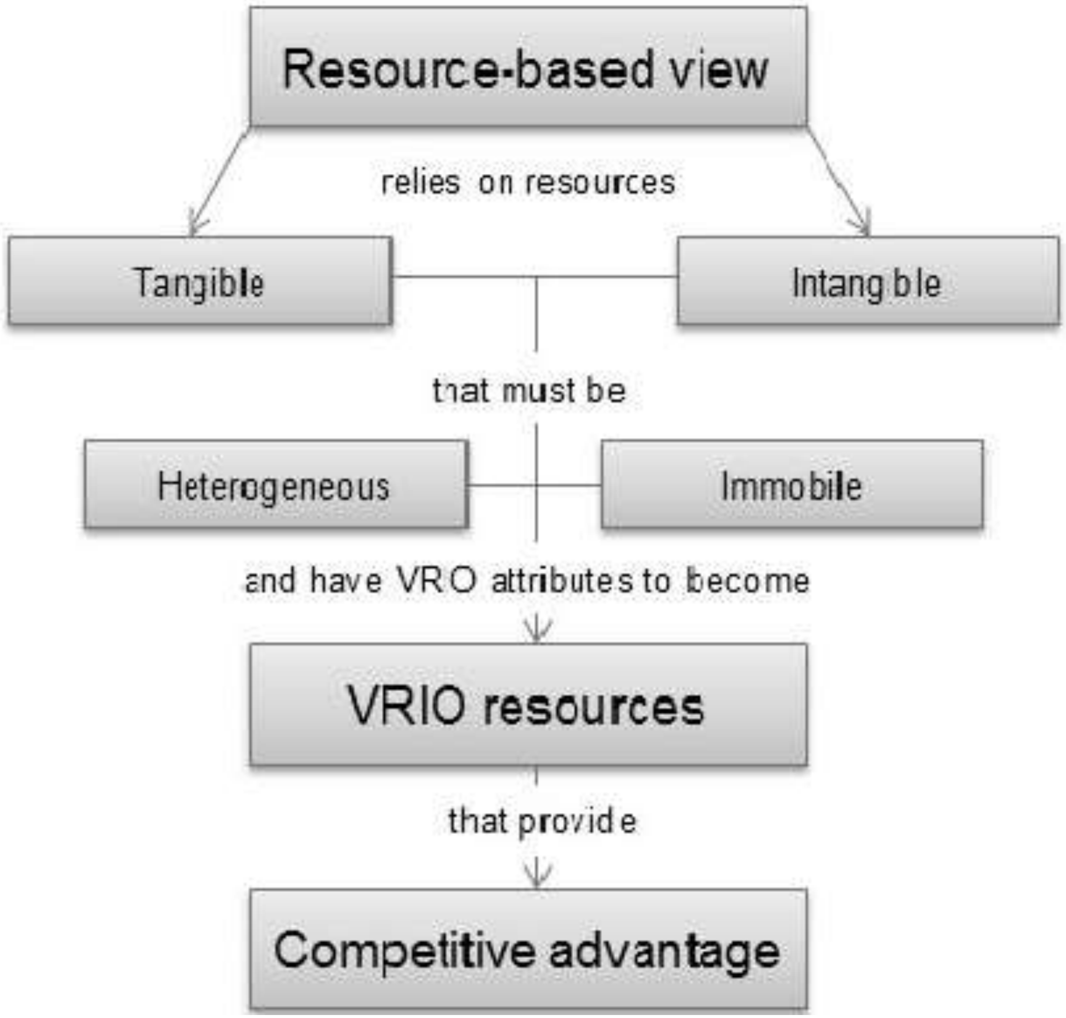
Key Idea

Instead of focusing only on competition outside the firm, RBV focuses on:

- **What the company already has**
- **How well it uses those resources**

The resource-based view (RBV) is a way of ***viewing the firm and in turn of approaching strategy.*** Fundamentally, this theory formulates the firm to be a bundle of resources.

- The following model explains RBV and emphasizes the key points of it.



- According to RBV proponents, it is much more feasible to exploit external opportunities using existing resources in a new way rather than trying to acquire new skills for each different opportunity. In RBV model, resources are given the major role in helping companies to achieve higher organizational performance.
- There are two types of resources: tangible and intangible

Types of Resources

- **Tangible resources**
  - Machines, buildings, capital, technology
- **Intangible resources**
  - Brand reputation, patents, goodwill, knowledge
- **Human resources**
  - Employee skills, experience, leadership

- **Tangible assets** are **physical things**. ***Land, buildings, machinery, equipment and capital – all these assets are tangible. Physical resources can easily be bought in the market so they confer little advantage to the companies in the long run because rivals can soon acquire the identical assets.***
- **Intangible assets** **are everything** ***else that has no physical presence but can still be owned by the company***. **Brand reputation**, **trademarks**, **intellectual property** are all intangible assets. ***Unlike physical resources, brand reputation is built over a long time and is something that other companies cannot buy from the market. Intangible resources usually stay within a company and are the main source of sustainable competitive advantage***.

The two critical assumptions of RBV are that resources must also be heterogeneous and immobile.

- **Heterogeneous**. The first assumption is that skills, capabilities and other resources that organizations possess differ from one company to another. If organizations would have the same amount and mix of resources, they could not employ different strategies to outcompete each other. What one company would do, the other could simply follow and no competitive advantage could be achieved. This is the scenario of perfect competition, yet real world markets are far from perfectly competitive and some companies, which are exposed to the same external and competitive forces (same external conditions), are able to implement different strategies and outperform each other. Therefore, RBV assumes that companies achieve competitive advantage by using their different bundles of resources

- The competition between Apple Inc. and Samsung Electronics is a good example of how two companies that operate in the same industry and thus, are exposed to the same external forces, can achieve different organizational performance due to the difference in resources. Apple competes with Samsung in tablets and smartphones markets, where Apple sells its products at much higher prices and, as a result, reaps higher profit margins. Why Samsung does not follow the same strategy? Simply because Samsung does not have the same brand reputation or is capable to design user-friendly products like Apple does. (heterogeneous resources)

- **Immobile**. The second assumption of RBV is that resources are **not mobile and do not move from company to company, at least in short-run**. Due to this immobility, companies cannot replicate rivals' resources and implement the same strategies. Intangible resources, such as brand equity, processes, knowledge or intellectual property are usually immobile

- Resources of the **firm** can include all assets, capabilities, organizational processes, firm attributes, information and knowledge.
- In short resources can be considered as inputs that facilitate the organization to perform its activities.
- All resources that an organization has may not have strategic relevance. Only certain resources are capable of being an input to a value creating strategy which put the organization in a position of competitive advantage.
- An **organization's resource** should have **four attributes to provide the potential for competitive advantage**. These form the VRIN characteristics

VRIN/VRIO Framework (Very Important)

A resource gives competitive advantage if it is:

- **V – Valuable** (helps exploit opportunities / reduce threats)
- **R – Rare** (not commonly owned by competitors)
- **I – Inimitable** (difficult to copy)
- **N – Non-substitutable** (cannot be replaced easily)

VRIO

VRIO is a later development of VRIN:

- **V – Valuable**
- **R – Rare**
- **I – Inimitable**
- **O – Organized**: Is the company organized and capable of fully utilizing the resource?

Definition:

- **VRIN/VRIO** is a framework used in the Resource-Based View to assess whether a firm's resources and capabilities can create and sustain a competitive advantage by being Valuable, Rare, Inimitable, and Non-substitutable (or supported by Organization in VRIO).

Information and Knowledge – Organization's Strategic Resources & Knowledge Management

Information and **knowledge** are recognized as an organization's most valuable strategic resources. Knowledge Management (KM) is the systematic process of **identifying**, **capturing**, **organizing**, **sharing**, and **utilizing** this collective intellectual capital to drive innovation, improve operational efficiency, and maintain a sustainable competitive advantage in dynamic business environments

- Information and Knowledge are valuable strategic resources of an organization. In today's competitive environment, organizations use information and knowledge to improve decision-making, innovation, productivity, customer satisfaction, and competitive advantage.

- Knowledge Management (KM) is the process of creating, capturing, storing, sharing, and utilizing knowledge effectively within an organization to achieve its goals.

Characteristics of Good Information

- Accurate
- Relevant
- Timely
- Complete
- Reliable
- Accessible

Examples

- Sales reports
- Customer feedback
- Market research reports
- Financial statements

2. Knowledge as a Strategic Resource

Definition

- Knowledge is information combined with experience, skills, understanding, and judgment.

Importance of Knowledge

- Improves organizational learning.
- Encourages innovation.
- Increases employee productivity.
- Helps solve complex problems.
- Provides sustainable competitive advantage.

Types of Knowledge

a) Explicit Knowledge

- Knowledge that can be documented and shared easily.
- **Examples:**
- Manuals
- Reports
- Policies
- Databases

b) Tacit Knowledge

Personal knowledge gained through experience.

Examples:

- Problem-solving skills
- Leadership abilities
- Technical expertise

3. Organization’s Strategic Resources

Strategic resources are assets that help an organization achieve long-term success.

Major Strategic Resources

- Human Resources
- Financial Resources
- Physical Resources
- Technological Resources
- Information Resources
- Knowledge Resources

4. Knowledge Management (KM)

Definition

- Knowledge Management is the systematic process of acquiring, organizing, storing, sharing, and applying knowledge to improve organizational performance.

Objectives of Knowledge Management

- Preserve organizational knowledge.
- Improve decision-making.
- Promote knowledge sharing.
- Enhance innovation.
- Increase productivity.
- Reduce duplication of work.

5. Knowledge Management Process

Knowledge Creation

(Generation of new ideas and insights.)

↓

Knowledge Capture

(Collecting knowledge from employees and systems.)

↓

Knowledge Storage

(Storing knowledge in databases, repositories, and documents.)

↓

Knowledge Sharing

(Distributing knowledge among employees.)

↓

Knowledge Application

(Using knowledge to solve problems and make decisions.)

↓

Organizational Learning

6. Components of Knowledge Management

People

- Employees
- Experts
- Managers

Processes

- Procedures for creating and sharing knowledge

Technology

- Databases
- Intranets
- Knowledge repositories
- Collaboration tools

Creative Employees and Creative Organizations (WSMA)

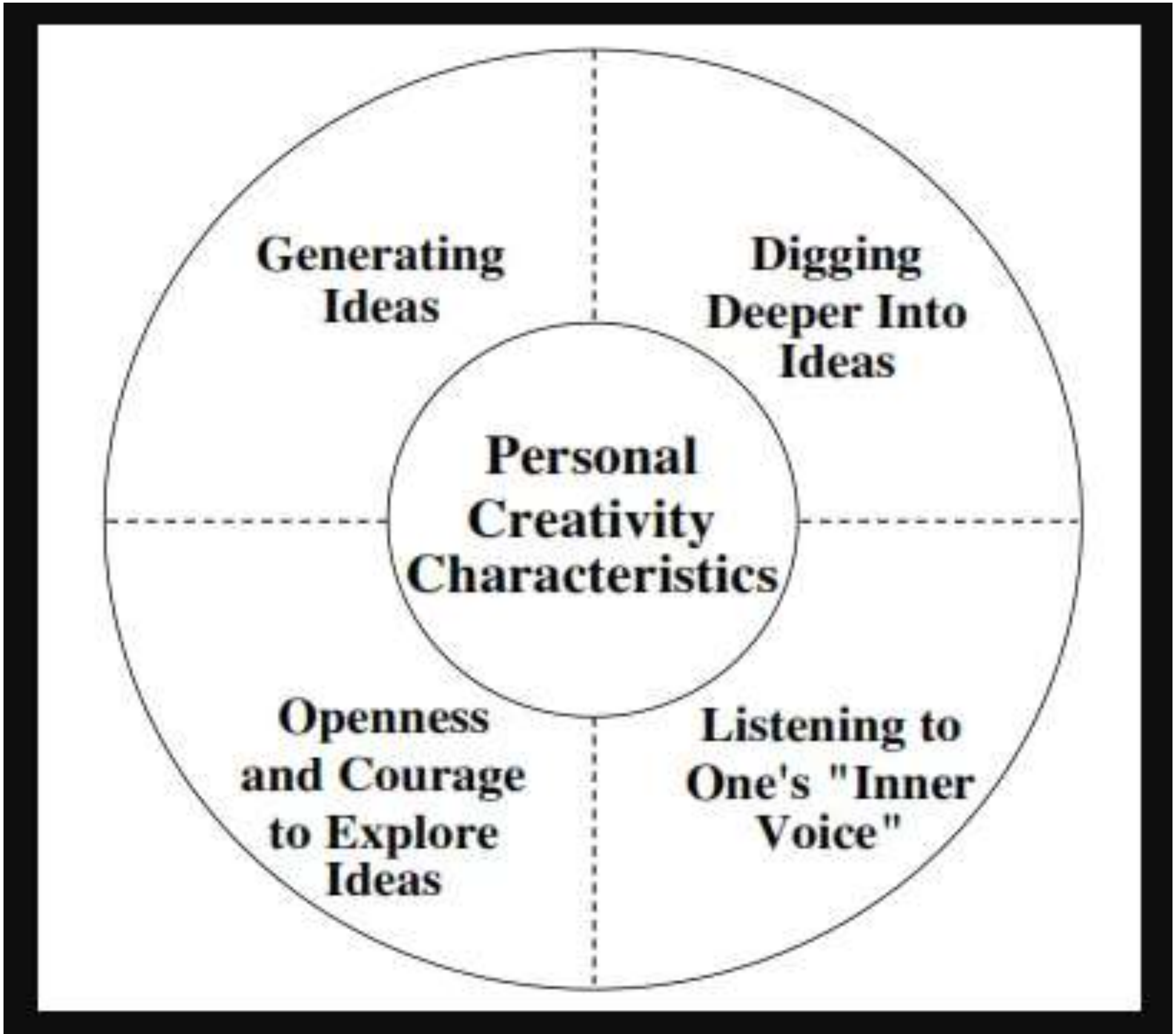
- In today's knowledge-driven economy, creativity has become one of the most valuable organizational resources.
- Organizations need employees who can think differently, generate innovative ideas, solve complex problems, and adapt to changing environments.
- Creative employees contribute to organizational growth, while creative organizations provide an environment that nurtures innovation and continuous improvement.

2) Creative Employees

Definition

A **creative employee** is an

- Creativity is not limited to artists or designers; it can be demonstrated by engineers, managers, teachers, researchers, software developers, and employees in any profession.


Generating Ideas
Digging Deeper Into Ideas
Personal Creativity Characteristics
Openness and Courage to Explore Ideas
Listening to One's "Inner Voice"

Characteristics of Creative Employees

1. Original Thinking

- Generate unique and innovative ideas.
- Look beyond traditional solutions.
- Think "outside the box."

2. Problem-Solving Ability

- Analyze problems from different perspectives.
- Develop effective and practical solutions.

3. Curiosity and Learning Orientation

- Continuously seek new knowledge.
- Ask questions and explore possibilities.

4. Risk-Taking Attitude

- Willing to experiment with new approaches.
- Accept failures as learning opportunities.

5. Adaptability

- Adjust quickly to changing technologies and business environments.
- Remain flexible in uncertain situations.

6. Self-Motivation

- Possess strong internal drive.
- Take initiative without constant supervision.

7. Open-Mindedness

- Welcome diverse opinions and suggestions.
- Consider alternative viewpoints.

8. Knowledge Sharing

- Share expertise with colleagues.
- Contribute to organizational learning.

Importance of Creative Employees

- **Innovation**

Creative employees generate ideas that lead to new products and services.

- **Competitive Advantage**

Unique ideas help organizations stay ahead of competitors.

- **Increased Productivity**

Creative solutions improve efficiency and reduce costs.

- **Better Decision Making**

Innovative thinking enhances problem-solving and strategic planning.

- **Organizational Growth**

New ideas contribute to long-term success and sustainability.

Factors Influencing Employee Creativity

Individual Factors

- Knowledge and expertise
- Skills and experience
- Personality traits
- Motivation

Organizational Factors

- Supportive leadership
- Positive work culture
- Training opportunities
- Reward systems

Environmental Factors

- Technological advancements
- Market competition
- Customer demands

Example of Creative Employees

- **Software Engineer**
- Develops a new algorithm that reduces processing time by 50%.
- **Teacher**
- Introduces interactive digital learning methods.
- **Manager**
- Creates innovative workflows that improve team productivity.

2) Creative Organizations

Definition

- A **creative organization** is an organization that encourages innovation, supports new ideas, promotes knowledge sharing, and continuously adapts to changing business environments.
- Creative organizations view creativity as a strategic resource for achieving long-term success.

Characteristics of Creative Organizations

1. Innovation-Oriented Culture

- Encourages experimentation and innovation.
- Supports idea generation.

2. Knowledge Sharing

- Employees freely exchange information and expertise.
- Learning becomes a continuous process.

3. Employee Empowerment

- Employees are given authority to make decisions.
- Encourages initiative and responsibility.

4. Open Communication

- Transparent communication across all levels.
- Ideas can flow freely within the organization.

5. Supportive Leadership

- Leaders encourage creativity and innovation.
- Provide guidance and resources.

6. Continuous Learning

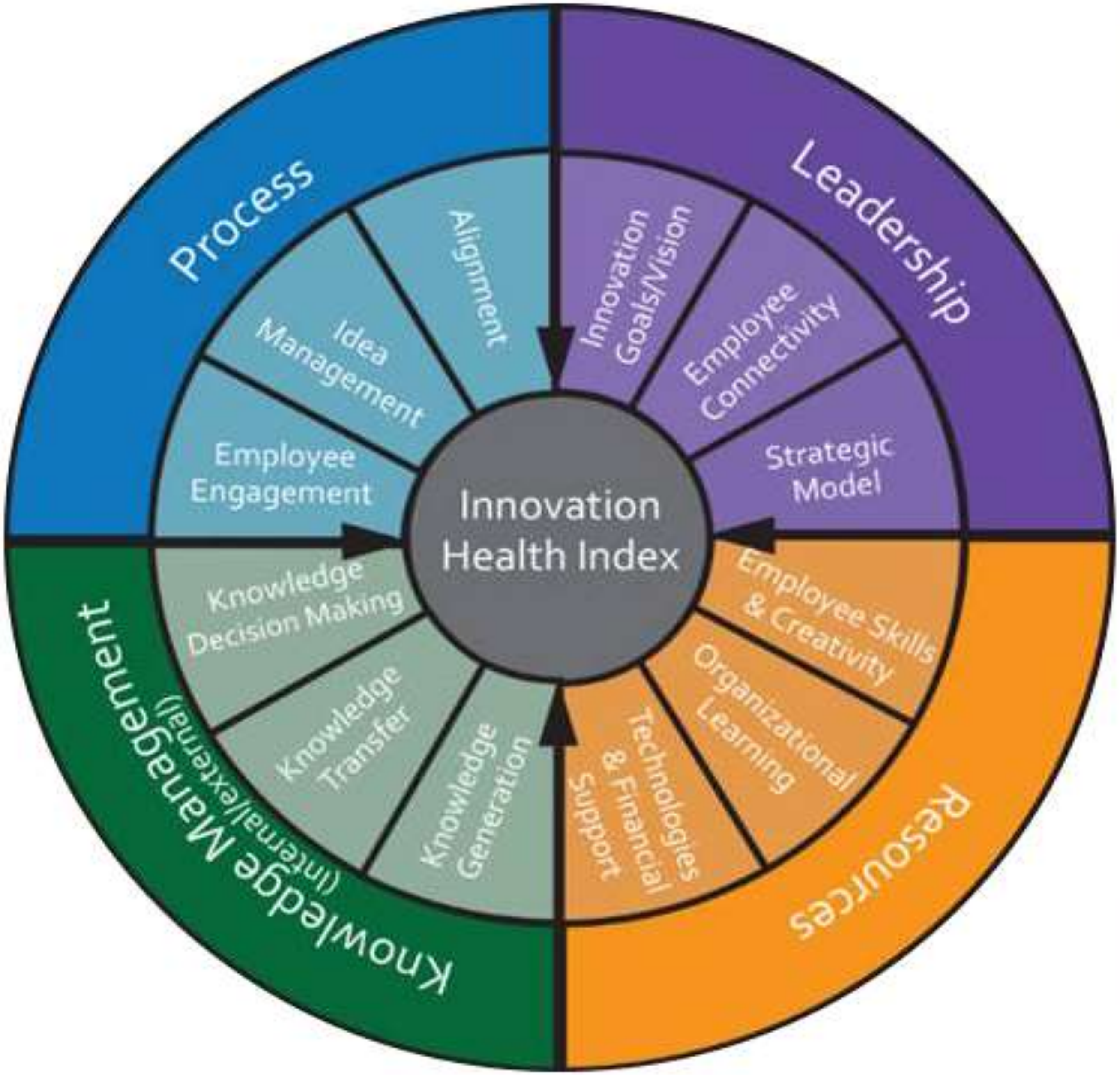
- Regular training and development programs.
- Encourages skill enhancement.

7. Collaboration and Teamwork

- Employees work together to solve problems.
- Diverse teams generate better ideas.

8. Tolerance for Failure

- Mistakes are treated as learning experiences.
- Encourages experimentation without fear.



Process
Leadership
Knowledge Management
(internal/external)
Resources
Alignment
Idea Management
Employee Engagement
Innovation Goals/Vision
Employee Connectivity
Strategic Model
Innovation Health Index
Knowledge Decision Making
Knowledge Transfer
Knowledge Generation
Employee Skills & Creativity
Organizational Learning
Technologies & Financial Support

Seven traits of highly innovative organizations that drive improved financial performance.

1-Executives embrace innovation, create an innovative vision and strategies, and communicate them relentlessly

2-Employee and external partner creativity is unleashed and learning is encouraged

3-A clear process exists to move ideas forward.

4-An innovation knowledge management system is in place.

5-Investment in resources, skills, time, advanced technology, digitization, and space are dedicated to supporting innovation

6-Innovation is measured

7-Performance management systems incent and encourage innovation

Importance of Creative Organizations

- **Sustainable Growth**
- Continuous innovation supports long-term development.

Improved Customer Satisfaction

- Innovative products and services meet changing customer needs.

Increased Competitive Advantage

- Organizations remain relevant in dynamic markets.

Better Employee Engagement

- Employees feel valued and motivated.

Effective Knowledge Management

- Knowledge is captured, stored, shared, and utilized effectively.

Building a Creative Organization

Encourage Knowledge Sharing

- Conduct workshops and seminars.
- Create knowledge repositories.

Promote Innovation

- Reward creative ideas.
- Organize innovation competitions.

Develop Learning Culture

- Offer training programs.
- Support professional development.

Use Technology

- Implement knowledge management systems.
- Utilize collaboration tools.

Foster Teamwork

- Encourage cross-functional collaboration.
- Build diverse teams.

Relationship Between Creative Employees and Creative Organizations

Creative Employees	Creative Organizations
Generate innovative ideas	Provide environment for innovation
Share knowledge	Facilitate knowledge sharing
Solve problems creatively	Support creative problem solving
Drive innovation	Encourage and reward innovation
Improve performance	Create systems for continuous improvement

NOTE:

Creative employees are the foundation of a creative organization, while creative organizations provide the culture and resources needed for employees to express and develop their creativity.

Diagram: Creative Employees and Creative Organizations

- Creative Employees

  ↓
- New Ideas

  ↓
- Knowledge Sharing

  ↓
- Innovation

  ↓
- Creative Organization

  ↓
- Improved Products & Services

  ↓
- Competitive Advantage

  ↓
- Organizational Growth

– Information Communication Technology (ICT) as a Driver of Change and Innovation.**(A)**

– Information Communication Technology (ICT) in Business Value Creation.**(B)**

Information Communication Technology (ICT) as a Driver of Change and Innovation**(A)**

Information Communication Technology (ICT) serves as the **critical backbone for modern structural evolution and corporate growth, transforming how organizations operate.**

By converging computing, enterprise software, and telecommunications, ICT functions simultaneously as a disruptive **driver of innovation** and a primary engine for sustainable competitive advantage and long-term organizational transformation

Components of ICT

The term information and communications technology (ICT) is generally accepted to mean all technologies that, combined, allow people and organizations to interact in the digital world.

- **ICT (information and communications technology)** is the infrastructure and components that enable modern computing. Among the goals of IC technologies, tools and systems is to improve the way humans create, process and share data or information with each other. Another is to help them improve their abilities in numerous areas, including business, education, medicine, real-world problem-solving, and even leisure activities related to sports, music and movies.
- There is no universal definition of ICT because the technologies, devices and even ideas related to ICT are constantly evolving. However, the term is generally accepted to mean all devices, networking components and applications.
- When combined, these help people and organizations interact in the digital world.

- *Organization* is broadly defined to encompass businesses, nonprofit agencies, governments and even criminal enterprises, so virtually anyone can benefit from the proper application and use of IC technologies, devices and innovations.

What technologies are included in ICT?

- ICT encompasses the internet, as well as mobile infrastructures powered by wireless networks. It includes antiquated technologies, such as landline telephone, radio, television and broadcast -- all of which remain widely used alongside today's cutting-edge ICT components, such as smartphones, artificial intelligence (AI) and robotics.

- The internet, internet of things (IoT), metaverse, virtual reality and social media are also part of ICT, as are cloud computing services, video conferencing and collaboration tools, unified communications systems and mobile communication networks. Newer technologies, such as 5G, 6G, Web3 and quantum computing, are also part of the ICT universe.
- Any technology, infrastructure, component or device that enables communications, data sharing and global connectivity between humans and between humans and machines may be considered as part of ICT.

➢ICT vs. IT

- ICT is sometimes used synonymously with **IT.**
- However, **ICT** ***generally represents a more comprehensive list of components related to computer systems and digital technologies***.
- **IT** ***is more about managing the technologies related to information and its various technical aspects, including*** software, hardware **and networking**.
- IT management **does not include considerations of** telecommunications **devices and technologies, while ICT does.**
- IT can be considered a subset of ICT

What are the components of ICT?

- **The list of ICT components is exhaustive and continues to grow.**
- Some components, such as **computers and telephones**, have **existed for decades**, while others, such as **smartphones**, **digital TVs and robots**, are more recent entries.

- ICT components include the following:

Devices (hardware).

Software.

Middleware.

Data.

Wired networks.

Wireless networks.

Computer networks.

Communication technologies.

Cloud entities and services.

Communications protocols and interfaces.

Information security and information governance policies

Note:

Middleware.

- Middleware is software that **bridges the gap between applications and operating systems (OSes) by providing a method for communication and data management.**
- This capability is useful for applications that otherwise do not have any way to exchange data with other software tools or databases

ICT as a Driver of Change

ICT brings significant changes in organizations by improving efficiency, communication, and decision-making processes.

1. Organizational Transformation

- ICT changes traditional business processes into digital processes.
- Organizations move from paper-based systems to automated systems.
- It supports digital transformation and modern business models.

Example: Banks offering online banking instead of manual transactions.

2. Improved Communication

- ICT enables instant communication through emails, video conferencing, messaging applications, and collaboration platforms.
- Employees can communicate effectively regardless of geographical location.

Example: Use of video conferencing tools for remote meetings

3. Better Decision Making

- ICT provides real-time information and data analytics.
- Managers can make informed decisions based on accurate data.

Example: Business Intelligence (BI) dashboards showing sales performance.

4. Increased Productivity

- Automation reduces manual work and saves time.
- Employees can focus on strategic activities rather than routine tasks.

Example: Automated payroll and attendance systems.

5. Global Connectivity

- ICT connects organizations with customers, suppliers, and partners worldwide.
- Businesses can operate globally with reduced communication costs.

Example: E-commerce companies selling products internationally.

ICT as a Driver of Innovation

Innovation refers to introducing new ideas, products, services, or processes. ICT plays a crucial role in fostering innovation.

1. Product Innovation

- ICT helps develop new and improved products.
- Technologies such as AI, IoT, and cloud computing enable smart products.

Example: Smart watches and smart home devices.

2. Process Innovation

- Organizations redesign business processes using ICT.
- Automation and digital workflows improve operational efficiency.

Example: Online ticket booking systems replacing manual reservations.

3. Service Innovation

- ICT enables organizations to offer innovative services.
- Customers receive faster and personalized services.

Example: Mobile banking and digital payment applications.

4. Business Model Innovation

- ICT creates entirely new ways of doing business.
- Digital platforms and subscription-based services emerge.

Example: Netflix, Amazon, and online education platforms.

5. Knowledge Sharing and Collaboration

- ICT facilitates collaboration among employees and teams.
- Knowledge management systems encourage idea generation and innovation.

Example: Collaborative project management tools.

Benefits of ICT in Driving Change and Innovation

- Faster communication
- Improved productivity
- Better customer service
- Enhanced decision making
- Cost reduction
- Increased competitiveness
- Global market access
- Continuous innovation
- Better resource utilization
- Greater organizational flexibility

Challenges of ICT Adoption

- High implementation cost
- Cybersecurity risks
- Resistance to change by employees
- Need for continuous training
- Data privacy concerns
- Rapid technological obsolescence

Conclusion as a **NOTE:**

- Information and Communication Technology (ICT) is a key enabler of organizational change and innovation.
- It transforms **business operations, enhances communication, improves decision-making, and supports the development of innovative products, services, and business models**.
- Organizations that effectively utilize ICT gain competitive advantages, improve efficiency, and achieve sustainable growth in the digital economy.

– Information Communication Technology (ICT) in Business Value Creation.**(B)**

Business Value Creation refers to the process of generating benefits for customers, employees, stakeholders, and the organization itself. ICT plays a vital role in creating value by improving efficiency, reducing costs, enhancing customer satisfaction, and supporting innovation.

Role of ICT in Business Value Creation

- **1. Operational Efficiency**
- ICT automates routine tasks and business processes.
- Reduces time and operational costs.
- Improves productivity and accuracy.

Example: ERP systems integrating various business functions.

2. Customer Value Creation

- ICT helps organizations understand customer needs.
- Enables personalized products and services.
- Improves customer experience and satisfaction.

Example: Customer Relationship Management (CRM) systems.

3. Strategic Decision Making

- ICT provides data analytics and business intelligence tools.
- Helps managers identify opportunities and risks.

Example: Data-driven marketing strategies.

4. Innovation and Competitive Advantage

- ICT enables organizations to develop innovative products and services.
- Supports differentiation from competitors.

Example: Digital payment systems and online shopping platforms.

5. Supply Chain Optimization

- ICT improves coordination among suppliers, manufacturers, and distributors.
- Reduces delays and inventory costs.

Example: Real-time inventory management systems.

6. Knowledge Management

- ICT facilitates the creation, storage, sharing, and utilization of organizational knowledge.
- Enhances learning and problem-solving capabilities.

Example: Knowledge repositories and intranet systems.

7. Revenue Generation

- ICT opens new business opportunities and markets.
- Enables online sales and digital services.

Example: E-commerce websites and mobile applications.

Areas of Business Value Creation through ICT

Area	Value Created
Operations	Efficiency and productivity
Marketing	Customer engagement and market reach
Finance	Cost control and accurate reporting
Human Resources	Better employee management
Customer Service	Improved satisfaction and loyalty
Supply Chain	Faster and efficient logistics
Innovation	New products and services

Business Benefits of ICT

- **Tangible Benefits**
- Increased revenue
- Reduced operational costs
- Improved productivity
- Higher profitability
- Faster transaction processing

Intangible Benefits

- Better customer satisfaction
- Enhanced organizational reputation
- Improved employee collaboration
- Better knowledge sharing
- Increased innovation capability

ICT Value Creation Model

- ICT Investment

  ↓
- Process Improvement

  ↓
- Operational Efficiency

  ↓
- Customer Satisfaction

  ↓
- Competitive Advantage

  ↓
- Business Value Creation

- **Conclusion as a NOTE:**
- ICT is a critical resource for business value creation. It enhances ***operational efficiency, supports innovation, improves customer relationships, and enables strategic decision-making***.
- By effectively leveraging ICT, ***organizations can achieve higher productivity, stronger competitive advantage, and long-term business success***.
- These benefits make ICT one of the most important drivers of value creation in the modern digital economy.

TOPIC-3 Brief Introduction to Big Data Analytics

- **Big Data Analytics** ***is the process of examining large, complex, and diverse datasets to discover hidden patterns, correlations, trends, and useful information that can support better decision making.***
- It combines **advanced analytical techniques**, **statistical methods, data mining**, **machine learning,** and **visualization tools** to extract meaningful insights from data.

- The term **Big Data** refers to **datasets that are too large, fast, or complex to be processed using traditional database management systems.**
- Organizations generate **massive amounts of data** from sources such as **social media**, **sensors, mobile devices, business transactions, websites**, and **IoT devices.**
- Big Data Analytics helps organizations transform this raw data into valuable knowledge.

Objectives of Big Data Analytics

- Improve decision-making processes.
- Identify market trends and customer preferences.
- Enhance operational efficiency.
- Detect fraud and security threats.
- Predict future outcomes through predictive analytics.
- Support innovation and business growth

Characteristics of Big Data

- Big Data is commonly described using the **5 Vs**:
- **Volume** – Huge amounts of data.
- **Velocity** – High speed of data generation and processing.
- **Variety** – Different types of data (structured, semi-structured, and unstructured).
- **Veracity** – Reliability and accuracy of data.
- **Value** – Useful insights derived from data.

Types of Big Data Analytics

- **Descriptive Analytics** – Explains what happened.
- **Diagnostic Analytics** – Explains why it happened.
- **Predictive Analytics** – Forecasts what may happen in the future.
- **Prescriptive Analytics** – Suggests actions to achieve desired outcomes.

Applications of Big Data Analytics

- **Healthcare:** Disease prediction and patient monitoring.
- **Banking**: Fraud detection and risk analysis.
- **Retail:** Customer behavior analysis and recommendation systems.
- **Education:** Student performance analysis.
- **Manufacturing:** Predictive maintenance and quality control.
- **Transportation:** Traffic management and route optimization.

- Big Data for Business Intelligence
  - Era of Data-intensive Computing
  - Big Data Term – Interpretation
  - Big Data V Model
  - Techniques for Big Data Analyzing
  - Big Data Technologies
  - Big Data Management

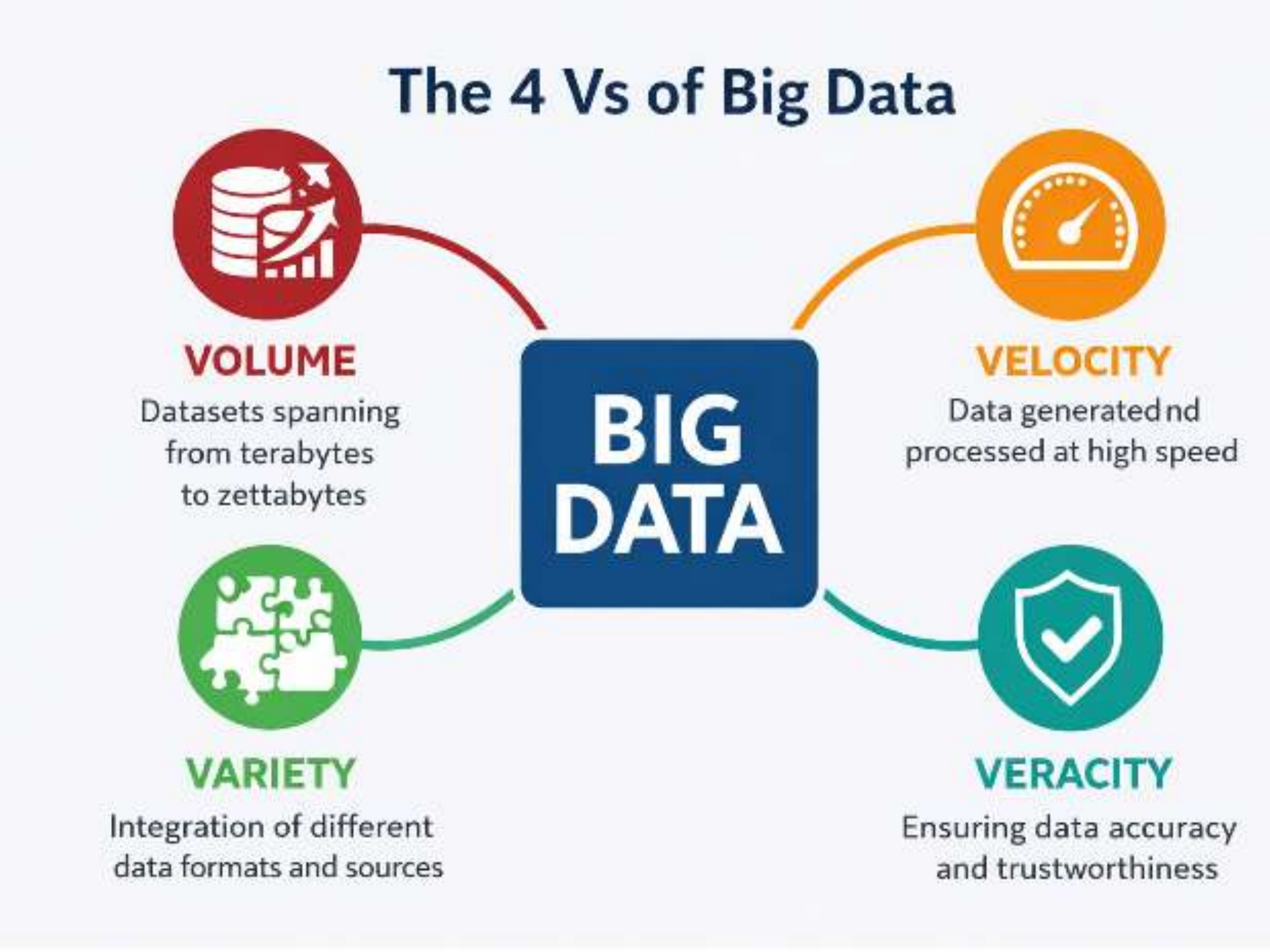
- Big Data for Business Intelligence
  - Era of Data-intensive Computing

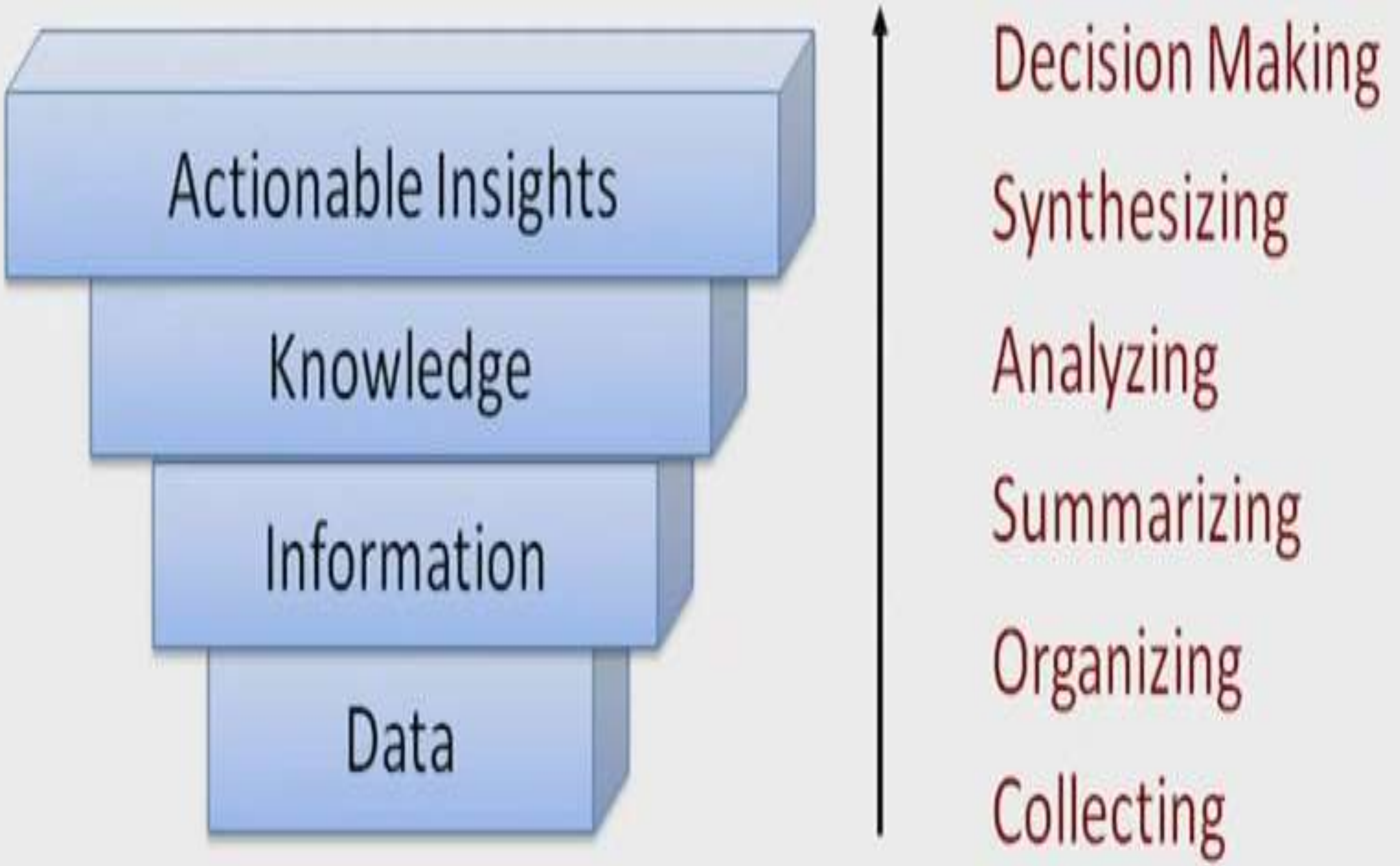
  - Big Data is the foundation of modern **Business Intelligence (BI)**, shifting organizations away from static spreadsheets and toward real-time, predictive decision-making.
  - Operating in this era of **data-intensive computing** means organizations can extract actionable insights from petabytes(1 Petabyte = 1,024 Terabytes = about 1 million Gigabytes) of structured and unstructured information across their global operations.

The Pillars of Big Data in BI

Big Data is traditionally defined by four core characteristics (the "4 Vs"):

- **Volume:** Datasets spanning from terabytes to zettabytes, requiring scalable distributed processing rather than local servers.
- **Velocity:** The speed at which data is generated, forcing systems to deliver **real-time insights** rather than batch reporting.
- **Variety:** The integration of different formats, including traditional financial spreadsheets, IoT sensor logs, and unstructured social media streams.
- **Veracity:** The process of filtering and cleaning raw information to ensure accuracy and trustworthiness before it hits an executive dashboard.

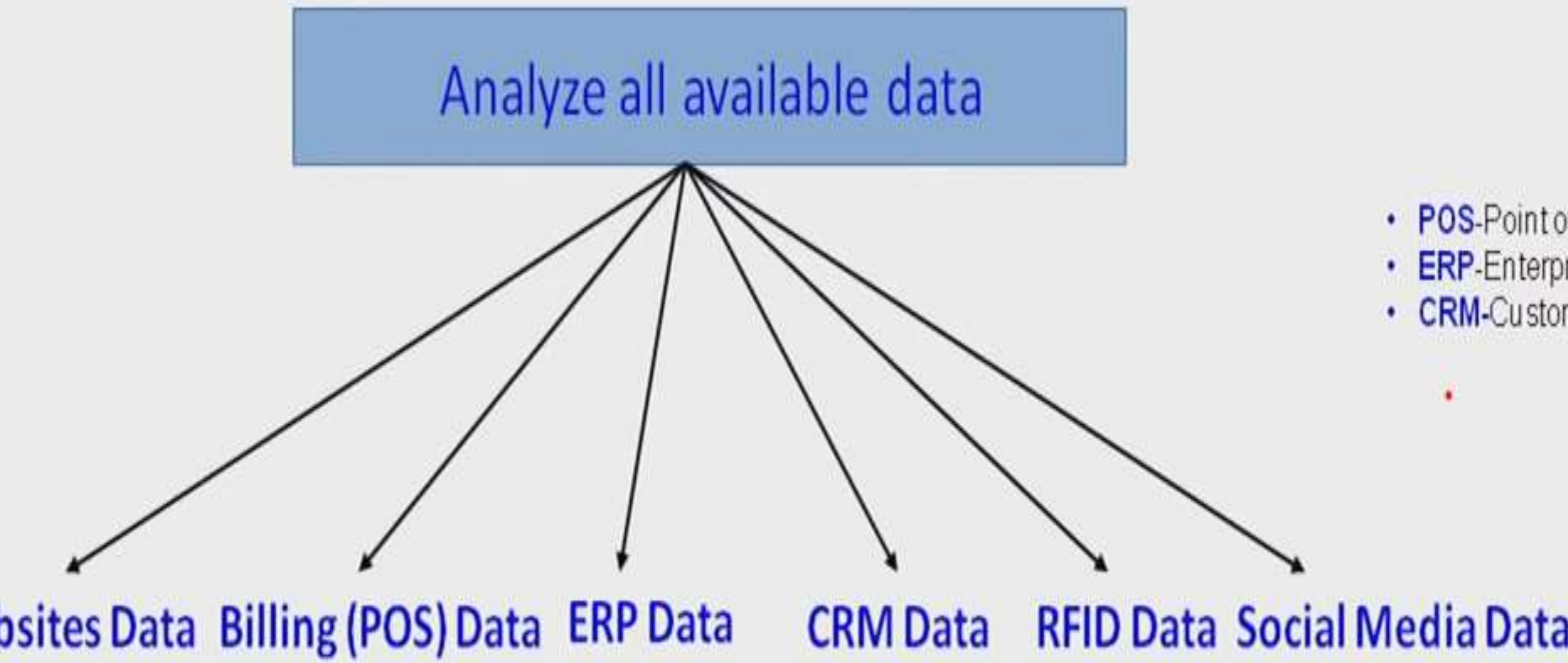

The 4 Vs of Big Data
BIG DATA
VOLUME
Datasets spanning from terabytes to zettabytes
VELOCITY
Data generated nd processed at high speed
VARIETY
Integration of different data formats and sources
VERACITY
Ensuring data accuracy and trustworthiness


Actionable Insights
Knowledge
Information
Data
Decision Making
Synthesizing
Analyzing
Summarizing
Organizing
Collecting

Big data Analytics

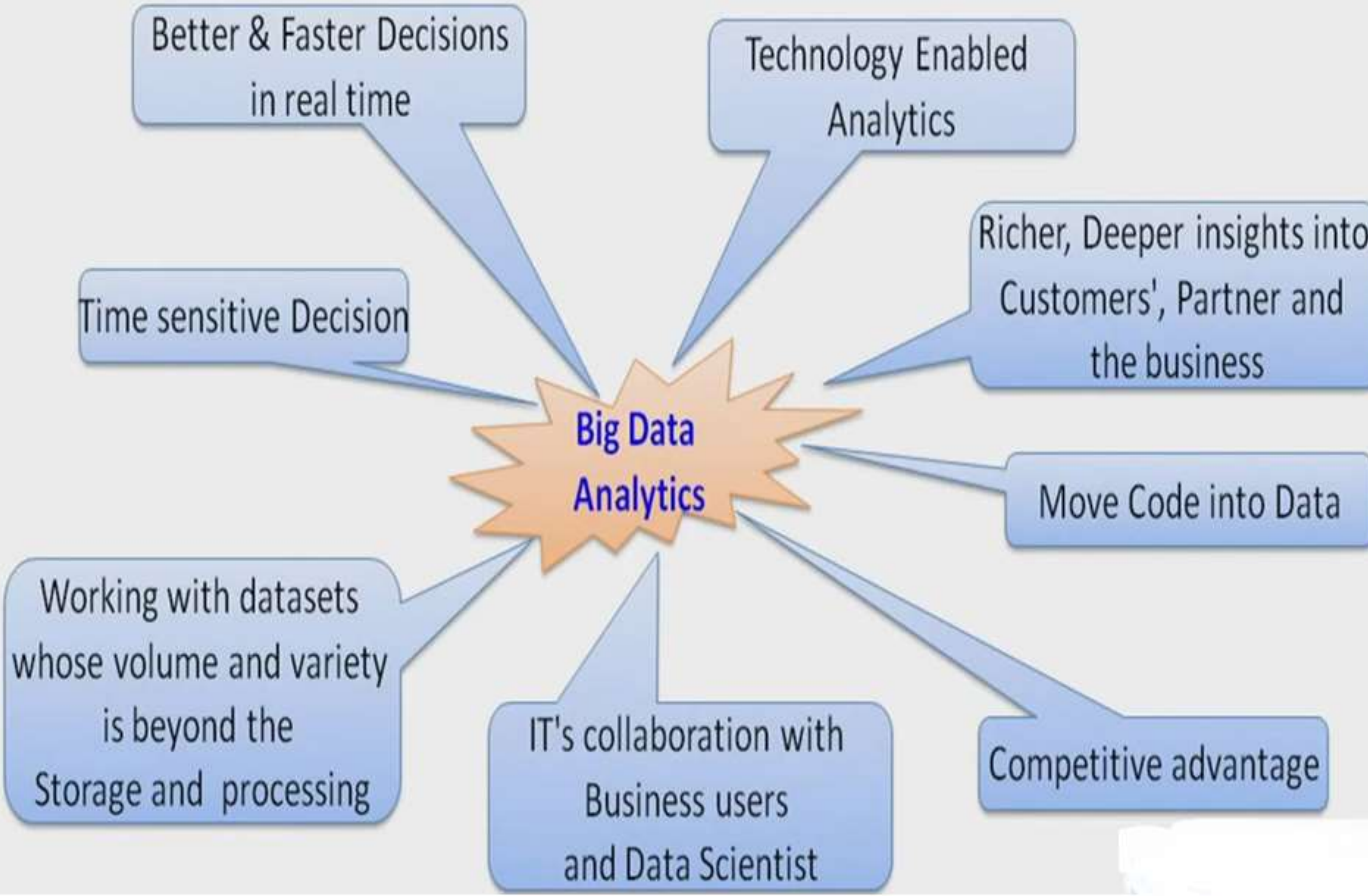
It is the process of examining big data to uncover patterns, unearth trends, a
find unknown correlations and other useful information to make faster and bet
decisions.

Analysis begin with analyzing all available data.



- **POS**-Point of Sale
- **ERP**-Enterprise Resource Pl
- **CRM**-Customer Relationship

What is Big Data Analytics?



2. Era of Data-Intensive Computing

Concept

- The **Era of Data-Intensive Computing** refers to the **modern computing age** where organizations generate and process enormous amounts of data every second.

In earlier days, computing focused on:

- Processing calculations
- Running applications
- Today, computing focuses on:
- Managing huge datasets
- Extracting knowledge from data
- Real-time analytics

Evolution of Computing

- Traditional Computing
- ↓
- Information Age
- ↓
- Internet Age
- ↓
- Big Data Age
- ↓
- Data-Intensive Computing Era

Characteristics

1. Massive Data Generation

- Organizations produce petabytes and exabytes of data.

2. High-Speed Processing

- Data must be analyzed in real time.

3. Distributed Computing

- Data is stored and processed across multiple computers.

4. Scalability

- Systems can easily expand as data grows.

5. Advanced Analytics

- Machine learning and AI techniques are used to discover patterns

- **Example**
- A bank processes:
- ATM transactions
- Credit card payments
- Mobile banking activities
- Online transfers
- Millions of transactions are analyzed instantly to detect fraud and improve customer services.

1. Big Data Term – Interpretation

Definition of Big Data

- **Big Data** refers to extremely large and complex datasets that cannot be efficiently processed, stored, or analyzed using traditional database management systems and software tools.
- Big Data includes:
- Structured data (tables, databases)
- Semi-structured data (XML, JSON)
- Unstructured data (text, images, videos, social media posts, emails)

- **Simple Interpretation**
- Big Data means collecting, storing, processing, and analyzing huge amounts of data generated every second from various sources to extract useful insights and support decision-making.

Sources of Big Data

- Social Media (Facebook, Instagram, X)
- E-commerce websites
- Banking transactions
- Sensors and IoT devices
- Mobile applications
- Healthcare records
- Scientific research
- Government databases

Examples

- Google processes billions of searches daily.
- Amazon analyzes customer purchase behavior.
- Netflix studies viewing habits to recommend movies.
- Hospitals analyze patient records for disease prediction.

- **2. Big Data V Model**
- The characteristics of Big Data are commonly explained through the **5V Model**.
- **Figure: Big Data V Model**

```
              BIG DATA
                  |
     ------------------------------------------
     |          |          |          |         |
  Volume Velocity Variety Veracity Value
```

1. Volume (Size of Data)

- Volume refers to the enormous amount of data generated every day.
- **Examples**
- Social media uploads
- Video streaming data
- **Challenges**
- Storage requirements
- Data processing complexity

2. Velocity (Speed of Data)

- Velocity refers to the speed at which data is generated, transmitted, and processed.
- **Examples**
- Online transactions
- Stock market updates
- Real-time sensor readings
- **Challenges**
- Real-time processing
- Instant decision-making

3. Variety (Different Types of Data)

- Variety refers to different formats of data.

Types

Structured Data

- Relational databases
- Excel sheets

Semi-Structured Data

- XML
- JSON

Unstructured Data

- Images
- Videos
- Emails
- Audio files

4. Veracity (Data Quality)

- Veracity refers to the reliability and accuracy of data.

Issues

- Missing values
- Duplicate records
- Incorrect information

Importance

- High-quality data leads to accurate analysis and decision-making.

5. Value (Business Benefit)

- Value refers to extracting useful information from Big Data.

Examples

- Customer recommendations
- Market trend prediction
- Healthcare diagnosis

3. Techniques for Big Data Analyzing

- Big Data Analytics involves examining large datasets to uncover patterns, trends, and insights.

A. Data Mining

- Process of discovering useful patterns from large datasets.
- **Applications**
- Market basket analysis
- Customer segmentation
- Fraud detection

B. Machine Learning

- Uses algorithms that learn from data and make predictions.

Types

Supervised Learning

- Classification
- Regression

Unsupervised Learning

- Clustering
- Association Rules

Reinforcement Learning

- Learning through rewards and penalties

Applications

- Recommendation systems
- Spam filtering

C. Predictive Analytics

- Predicts future outcomes using historical data.

Examples

- Weather forecasting
- Sales forecasting
- Disease prediction

D. Text Analytics

- Extracting meaningful information from text data.

Sources

- Emails
- Reviews
- Social media posts

Applications

- Sentiment analysis
- Opinion mining

E. Statistical Analysis

- Uses mathematical techniques to analyze data.

Methods

- Mean
- Median
- Standard deviation
- Correlation analysis

F. Stream Analytics

- Analyzes real-time data streams.
- **Examples**
- Traffic monitoring
- Online fraud detection
- IoT sensor monitoring

G. Visualization Analytics

- Presents data graphically for easier understanding.

Tools

- Tableau
- Power BI
- QlikView

Benefits

- Easy interpretation
- Faster decision-making

UNIT - II

Text Analytics and Text Mining:

- Machine Versus Men on Jeopardy!
  The Story of Watson
- Text Analytics and Text Mining Concepts and
- Definitions,
- Natural Language Processing,
- Text Mining Applications,
- Text Mining Process,
- Text Mining Tools.

➢ Machine Versus Men on Jeopardy!: The Story of Watson

- The term ***"human champions in Jeopardy"*** primarily refers to the legendary 2011 ***"Man vs. Machine" exhibition*** match where **IBM's Watson** artificial intelligence defeated the show's ***two most dominant players***, **Ken Jennings** and **Brad Rutter**.

- One of the most remarkable achievements in the field of ***Text Analytics and Text Mining*** is the development of ***Watson***, **an intelligent computer system created by IBM.**
- Watson became famous after competing against human champions in the quiz show ***Jeopardy!*** in **2011.**
- This event demonstrated the power of artificial intelligence, natural language processing, and text analytics in understanding and answering complex human language questions.


TEXT SOURCES
Documents
Emails
Social Media
Research Papers
WATSON
TEXT ANALYTICS
Sentiment Analysis
Entity Recognition
Topic Analysis
TEXT MINING
Keyword Extraction
Pattern Discovery
Trend Analysis
INSIGHTS & DECISIONS
Business Intelligence

What is Watson?

Watson is a cognitive computing system developed by IBM to process and analyze large amounts of unstructured data.

Unlike traditional computers that rely on predefined rules, Watson can:

- **Understand natural language**
- **Analyze vast amounts of text**
- **Generate hypotheses**
- **Evaluate evidence**
- **Provide answers with confidence levels**
- **The system was named after Thomas J. Watson, one of IBM's influential leaders.**

What is Jeopardy?

- **Jeopardy! is a television quiz show, not a stage.**
- Jeopardy! is a famous American TV game show where contestants answer clues in the form of a question..
- Example:

Clue (Answer): This city is the capital of India.

Response (Question): What is New Delhi?

This unique answer-to-question format makes Jeopardy! challenging and requires strong language understanding, reasoning, and general knowledge

To succeed in Jeopardy!, a contestant must:

- **Understand language nuances**
- **Interpret puns and wordplay**
- **Analyze context**
- **Retrieve information quickly**
- **Make decisions under time pressure**

These requirements make Jeopardy! a challenging task for computers.


WATSON
$77,147
JEOPARDY!

- **The figure of Jeopardy with Watson** represents IBM's Watson AI system competing on the television quiz show Jeopardy! against human champions.
- It visually **symbolizes a milestone in artificial intelligence** — the moment when a machine demonstrated advanced natural language understanding and reasoning in a real-world challenge.

Key Indications-

Watson’s podium and logo show its identity as an AI contestant, not a human. -

- The score display **($77,147)** **highlights Watson’s victory, proving its ability to process complex questions faster and more accurately than humans**.
- - ***The Jeopardy! stage symbolizes the integration of AI into human knowledge domains — language, logic, and decision-making***.


DINK
PIENSE
想
$1,000
KEN
$2,000
WATSON
$400
BRAD

Event:

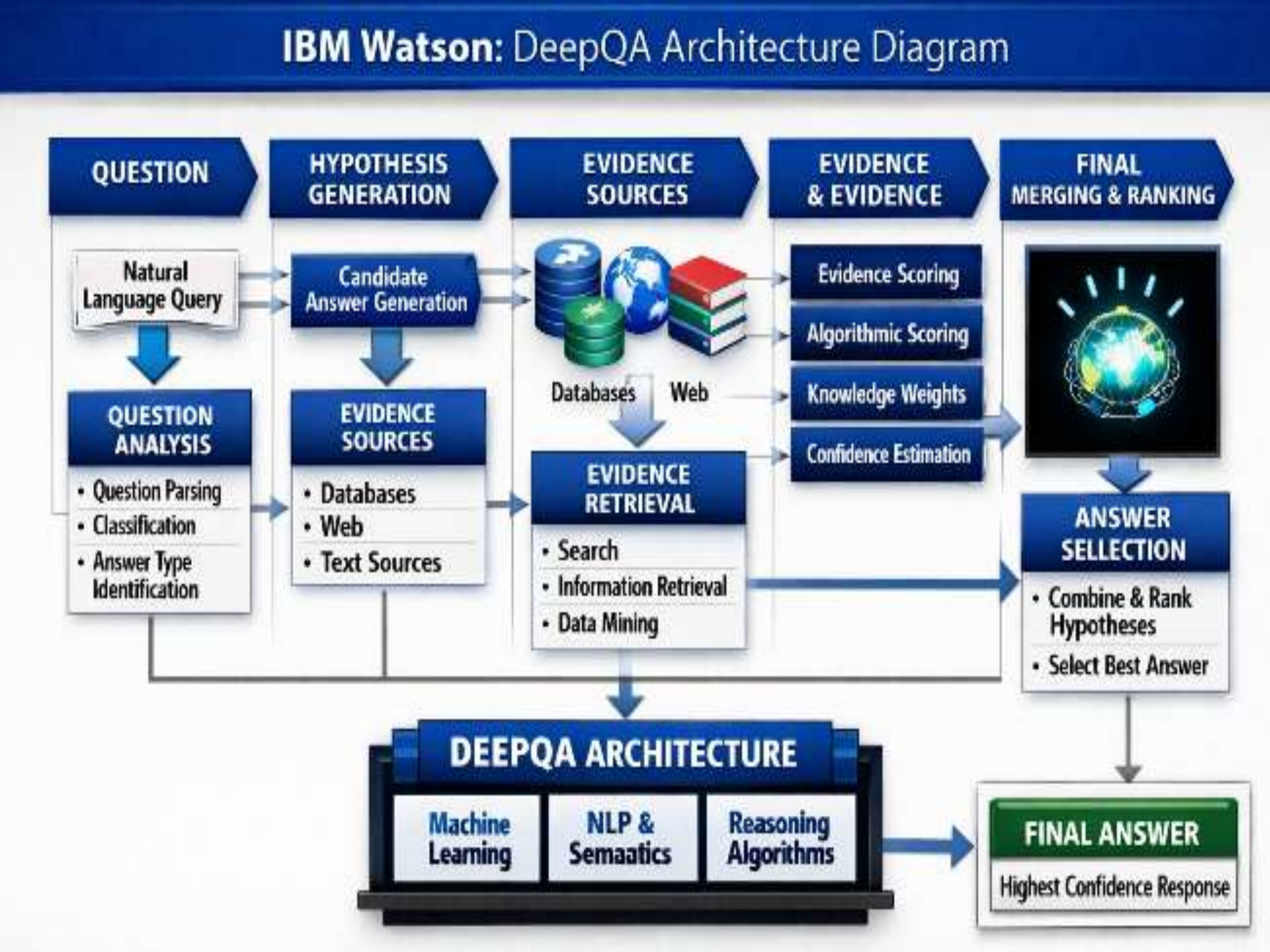
IBM Watson competed on Jeopardy! in 2011 against champions Ken Jennings and Brad Rutter.

Technology:

Based on the **DeepQA architecture** — a pipeline that performs question analysis, hypothesis generation, evidence scoring, and confidence estimation. -

Significance: Demonstrated how AI can interpret natural language, search vast databases, and reason probabilistically.

Outcome: Watson won with a high score, marking a breakthrough in AI natural language processing and machine learning applications.


IBM Watson: DeepQA Architecture Diagram
QUESTION
HYPOTHESIS GENERATION
EVIDENCE SOURCES
EVIDENCE & EVIDENCE
FINAL MERGING & RANKING
Natural Language Query
Candidate Answer Generation
Databases
Web
Evidence Scoring
Algorithmic Scoring
Knowledge Weights
Confidence Estimation
QUESTION ANALYSIS
• Question Parsing
• Classification
• Answer Type Identification
EVIDENCE SOURCES
• Databases
• Web
• Text Sources
EVIDENCE RETRIEVAL
• Search
• Information Retrieval
• Data Mining
ANSWER SELLECTION
• Combine & Rank Hypotheses
• Select Best Answer
DEEPQA ARCHITECTURE
Machine Learning
NLP & Semaatics
Reasoning Algorithms
FINAL ANSWER
Highest Confidence Response

Why Was Watson Developed?

- IBM researchers wanted to create a machine capable of understanding and processing human language similar to humans.

Objectives:

- **Advance Artificial Intelligence.**
- **Demonstrate machine understanding of natural language.**
- **Develop technologies useful in healthcare, finance, education, and business.**
- **Show the practical applications of text analytics and text mining.**

Technologies Behind Watson

Watson combines several advanced technologies.

A. Natural Language Processing (NLP)

- NLP helps Watson understand human language.
- Functions include:
- Parsing sentences
- Identifying keywords
- Understanding grammar
- Detecting relationships between words

B. Text Analytics

Text analytics converts unstructured text into meaningful information.

- Watson analyzes:
- Books
- Encyclopedias
- News articles
- Research papers
- Databases
- The system extracts facts, concepts, and relationships from text.

C. Text Mining

Text mining discovers patterns and knowledge hidden within large text collections.

Activities include:

- Information extraction
- Pattern recognition
- Classification
- Clustering
- Trend analysis

D. Machine Learning

- Watson continuously learns from examples and feedback.

Capabilities:

- Improves answer accuracy
- Learns new patterns
- Refines confidence estimates

How Watson Answers Questions

Watson follows a multi-step process.

Step 1: Question Analysis

The question is analyzed to identify:

- Keywords
- Context
- Question type

Step 2: Hypothesis Generation

Possible answers are generated.

Step 3: Evidence Gathering

- Watson searches millions of documents to gather supporting evidence for each answer.
- Sources include:
- Encyclopedias
- Dictionaries
- Literature databases
- Online resources

Step 4: Scoring

- Each candidate answer receives a confidence score based on evidence.

Step 5: Final Answer Selection

- The highest-scoring answer is selected and presented.

Watson's Architecture

- The architecture consists of several modules.

Input Layer

- Receives the question.

Analysis Layer

- Performs language processing and text analysis.

Knowledge Sources

Contains:

Books

Articles

Databases

Reference materials

Evidence Evaluation Layer

- Evaluates candidate answers.

Ranking Layer

- Ranks answers based on confidence.

Output Layer

- Provides the final answer.

Role of Text Analytics in Watson

- Text analytics enabled Watson to:

Information Extraction

- Identify relevant facts from documents.

Entity Recognition

- Recognize:
- People
- Places
- Organizations
- Events

Sentiment and Context Analysis

- Understand contextual meaning.

Semantic Analysis

- Interpret the intended meaning of words and phrases.

Role of Text Mining in Watson

- Text mining helped Watson:

  Discover Hidden Knowledge
- Extract facts from large text collections.
- **Pattern Identification**
- Identify relationships among concepts.
- **Classification**
- Categorize information.
- **Knowledge Discovery**
- Generate insights from unstructured data.

Applications of Watson

- **Healthcare**
- Disease diagnosis
- Clinical decision support
- Medical record analysis
- **Banking and Finance**
- Fraud detection
- Risk assessment
- Customer service
- **Education**
- Personalized learning
- Student assistance

- **Customer Support**
- Chatbots
- Virtual assistants
- **Research**
- Literature review
- Knowledge discovery

Advantages of Watson

- **Speed**
- Processes millions of documents within seconds.
- **Accuracy**
- Provides highly reliable answers.
- **Scalability**
- Handles massive volumes of data.
- **Learning Ability**
- Improves through experience.
- **Natural Language Understanding**
- Interacts using human language.

Limitations of Watson

- **Dependence on Data Quality**
- Poor-quality data leads to poor results.
- **High Cost**
- Requires substantial computational resources.
- **Lack of Human Common Sense**
- May misunderstand ambiguous situations.
- **Complexity**
- Building and maintaining such systems is challenging.

Significance in Text Analytics and Text Mining

Watson demonstrated that machines can:

- Understand natural language
- Analyze unstructured text
- Discover hidden information
- Support decision-making
- Compete with human experts in knowledge-intensive tasks
- It became a landmark achievement in Artificial Intelligence and proved the practical value of text analytics and text mining.

Relationship Between Watson and Text Analytics/Text Mining

The connection can be represented as:

- **Huge Text Collections**
  ↓
  Text Mining
  ↓
  Information Extraction
  ↓
  Text Analytics
  ↓
  Natural Language Understanding
  ↓
  Evidence Analysis
  ↓
  Watson
  ↓
  Correct Answer Generation

- Watson is included in the Text Analytics and Text Mining unit because it is a real-world application that demonstrates how machines can analyze and understand large volumes of unstructured text.
- Developed by IBM, Watson uses Natural Language Processing, Text Analytics, Text Mining, Machine Learning, and the DeepQA architecture to understand questions, extract relevant information from millions of documents, evaluate evidence, and generate accurate answers.
- Its victory in Jeopardy! proved that computers can process human language and discover knowledge from textual data, making it one of the most significant applications of Text Analytics and Text Mining.

- **Text Analytics and Text Mining Concepts and**
- **Definitions,**

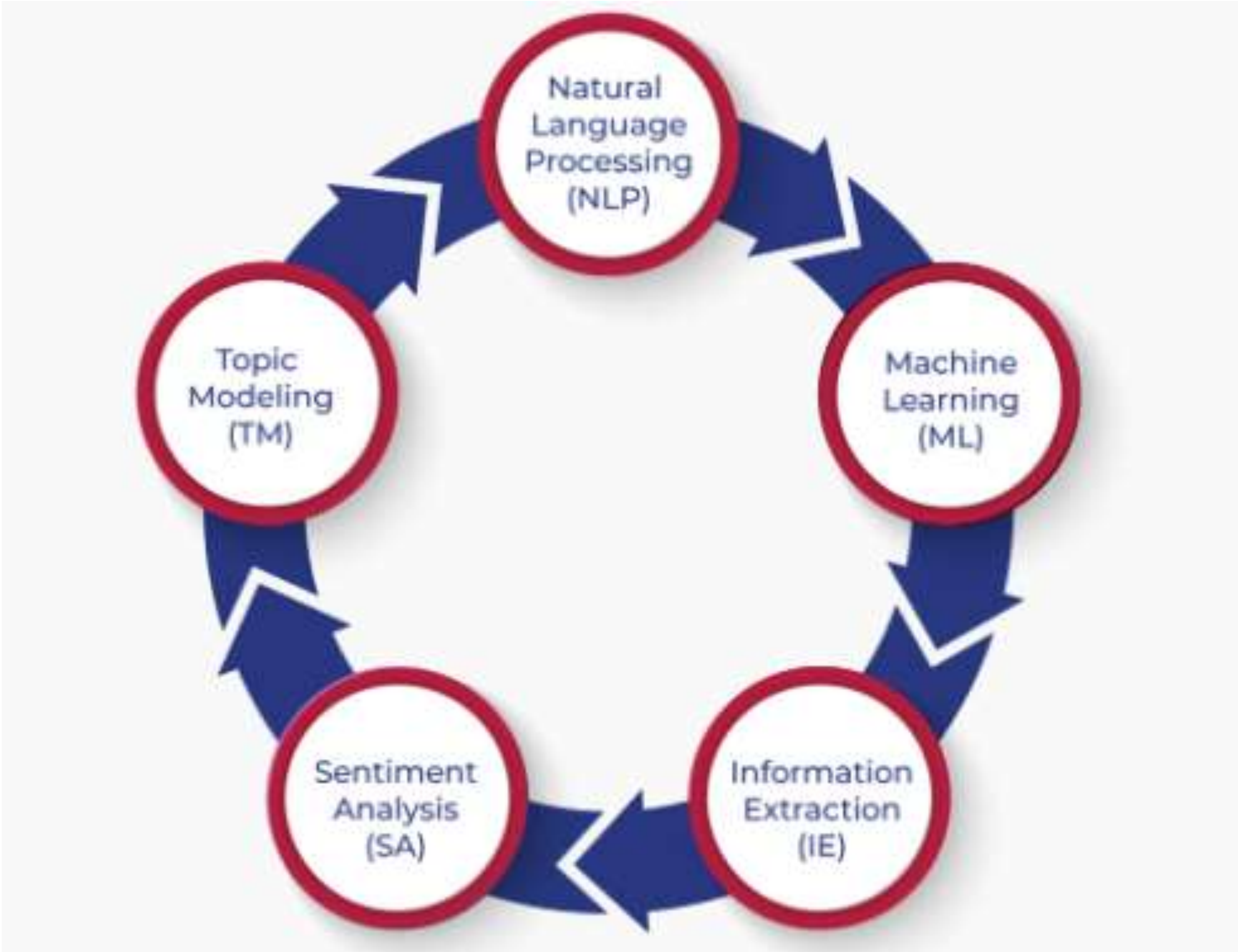
Text is everywhere. From customer reviews to social media posts and emails to documents, text data is one of the most abundant and valuable sources of information for businesses.

However, text data is also one of the most challenging and complex data types to analyze and extract insights from. That is where text analytics and text mining come in.

- **Text analytics** and **text mining** are powerful techniques that can help businesses transform ***unstructured text data into structured, actionable information.***
- By applying
- ***Natural Language Processing(NLP)***
- ***Machine Learning(ML)***
- ***Information Extraction,***
- ***Sentiment Analysis,***
- ***Topic Modeling,*** and other methods,

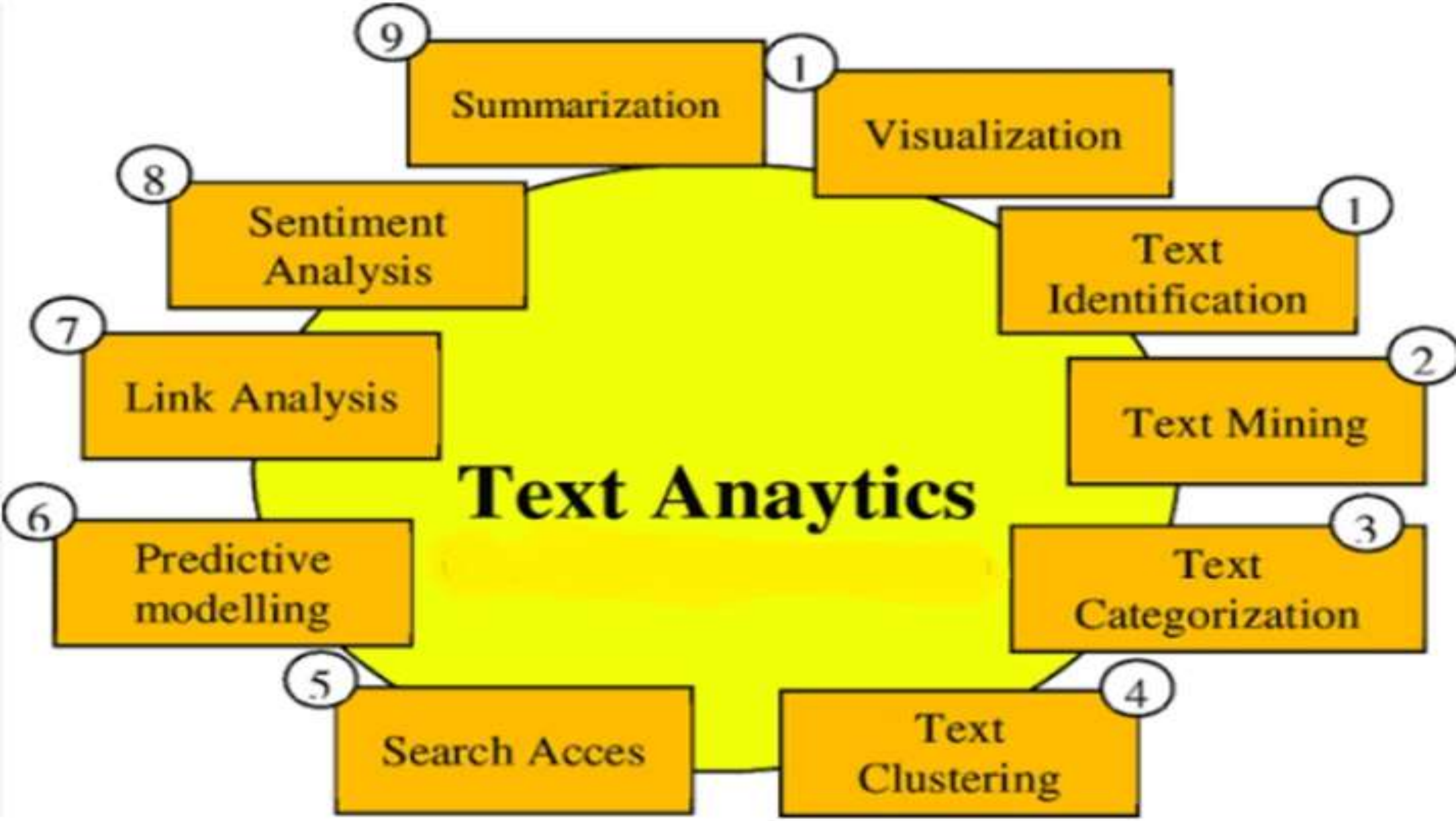
text analytics and text mining can help businesses better understand their customers, markets, products, competitors, risks, opportunities, and more.

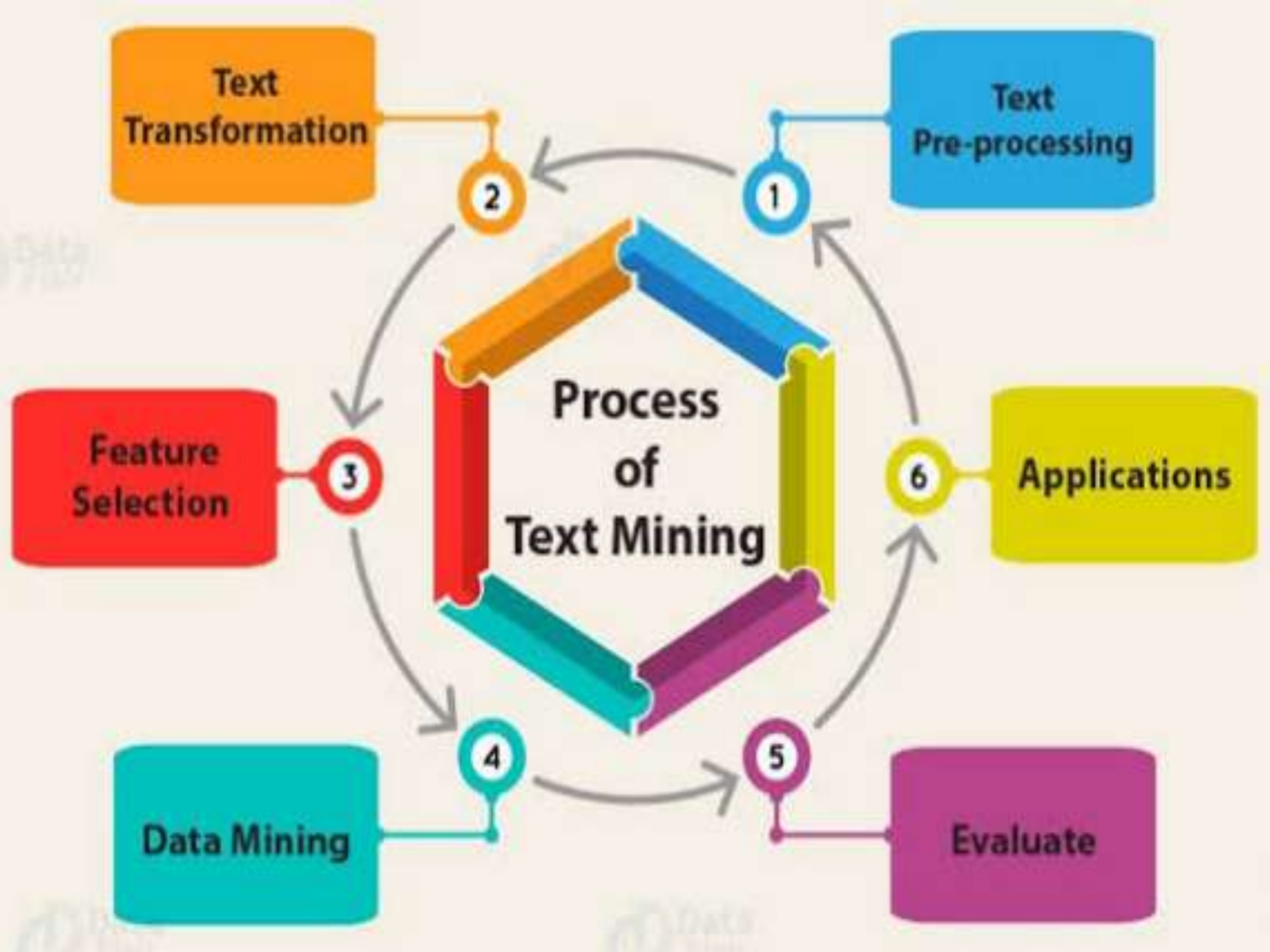
The ***most common and valuable techniques and methods used for text analytics and text mining*** are:


Natural Language Processing (NLP)
Machine Learning (ML)
Information Extraction (IE)
Sentiment Analysis (SA)
Topic Modeling (TM)

- **Text analytics** and **text mining** are often used interchangeably but have slight differences.

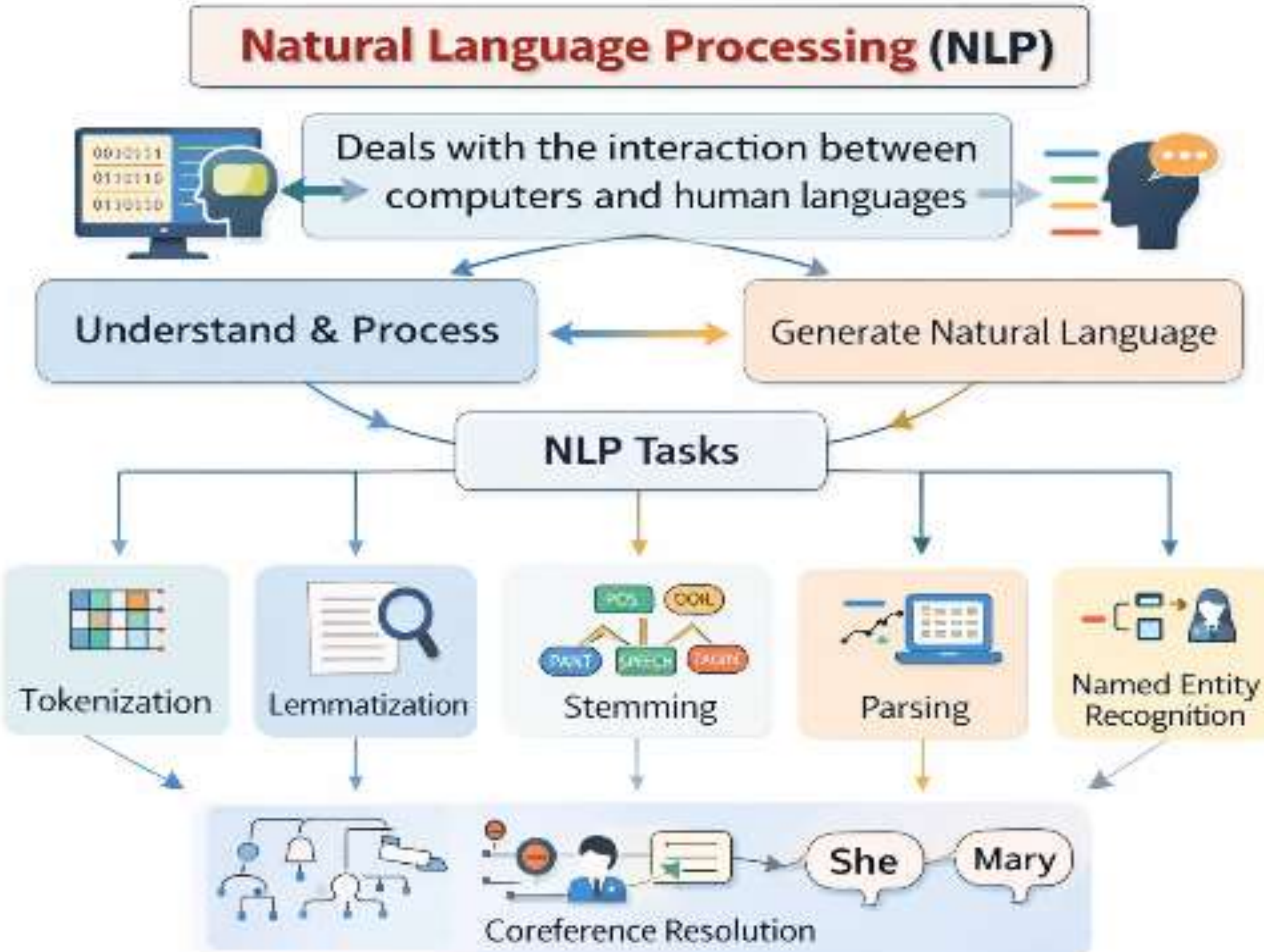
- **Text analytics** analyzes text data to derive meaning, insights, and patterns.
- **Text mining** is extracting relevant and valuable information from text data, such as entities, keywords, topics, sentiments, etc.
- **Text mining is a subset of text analytics**, and ***both techniques use similar methods and tools.***


Text Anaytics
1
Visualization
1
Text
Identification
2
Text Mining
3
Text
Categorization
4
Text
Clustering
5
Search Acces
6
Predictive
modelling
7
Link Analysis
8
Sentiment
Analysis
9
Summarization


Text Transformation
Text Pre-processing
Process of Text Mining
Feature Selection
Applications
Data Mining
Evaluate
1
2
3
4
5
6

➢ Natural Language Processing (NLP):

- ➢ This ***branch of artificial intelligence*** deals with the ***interaction between computers and human languages.***
- ➢ NLP enables computers to understand, process, and generate natural language texts, such as **speech** and **writing**.
- ➢ It involves various tasks, such as tokenization, lemmatization, stemming, part-of-speech tagging, parsing, named entity recognition, coreference resolution, etc.


Natural Language Processing (NLP)
Deals with the interaction between computers and human languages
Understand & Process
Generate Natural Language
NLP Tasks
Tokenization
Lemmatization
Stemming
Parsing
Named Entity Recognition
She
Mary
Coreference Resolution

➢ **<u>Machine Learning (ML):</u>**

➢ This ***<u>branch of artificial intelligence enables computers to learn from data and improve their performance without explicit programming</u>***.

➢ ML involves various **<u>algorithms</u>**, such as

<u>supervised learning</u>,

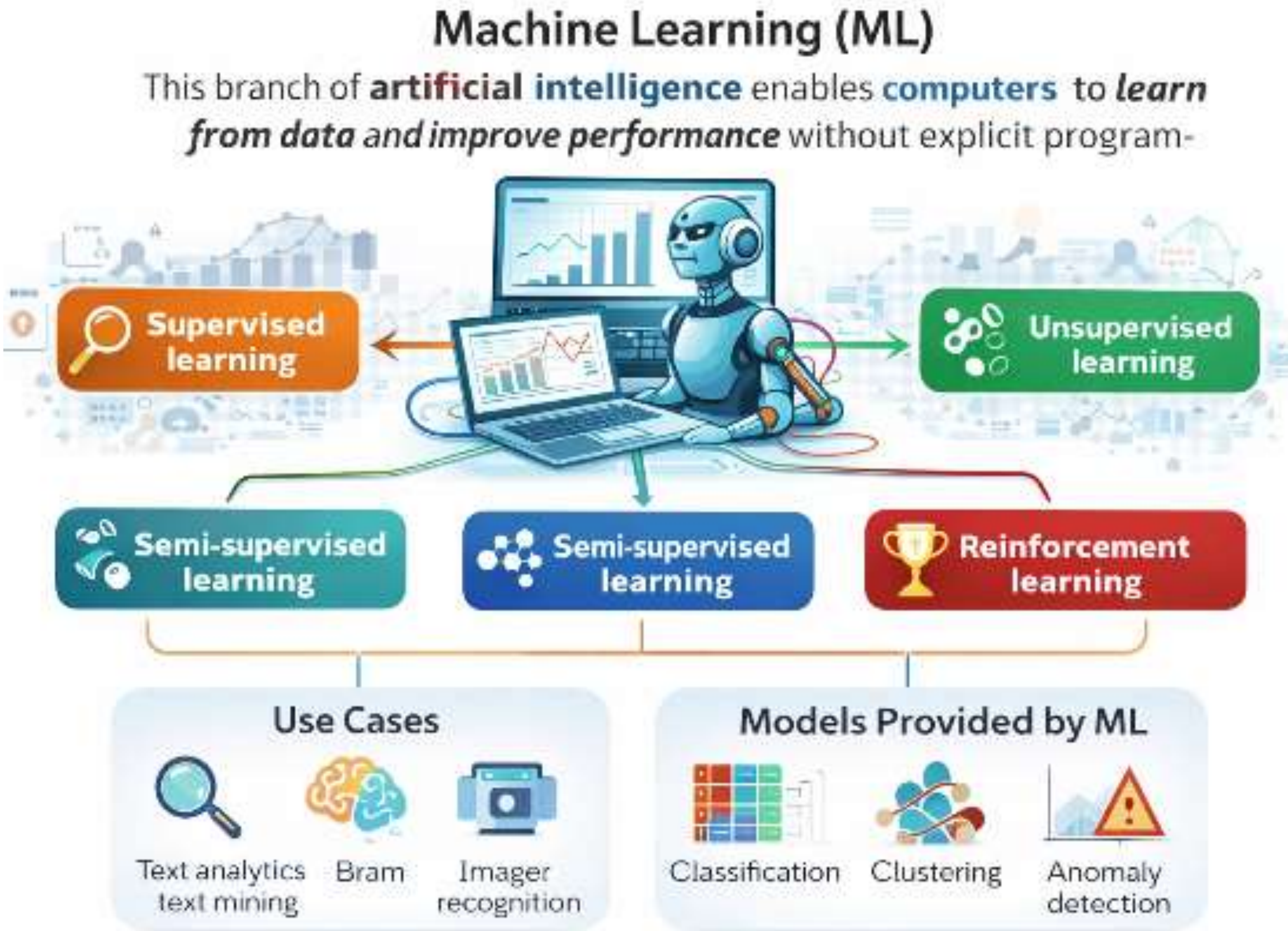
<u>unsupervised learning</u>,

<u>semi-supervised learning</u>,

<u>reinforcement learning</u>,

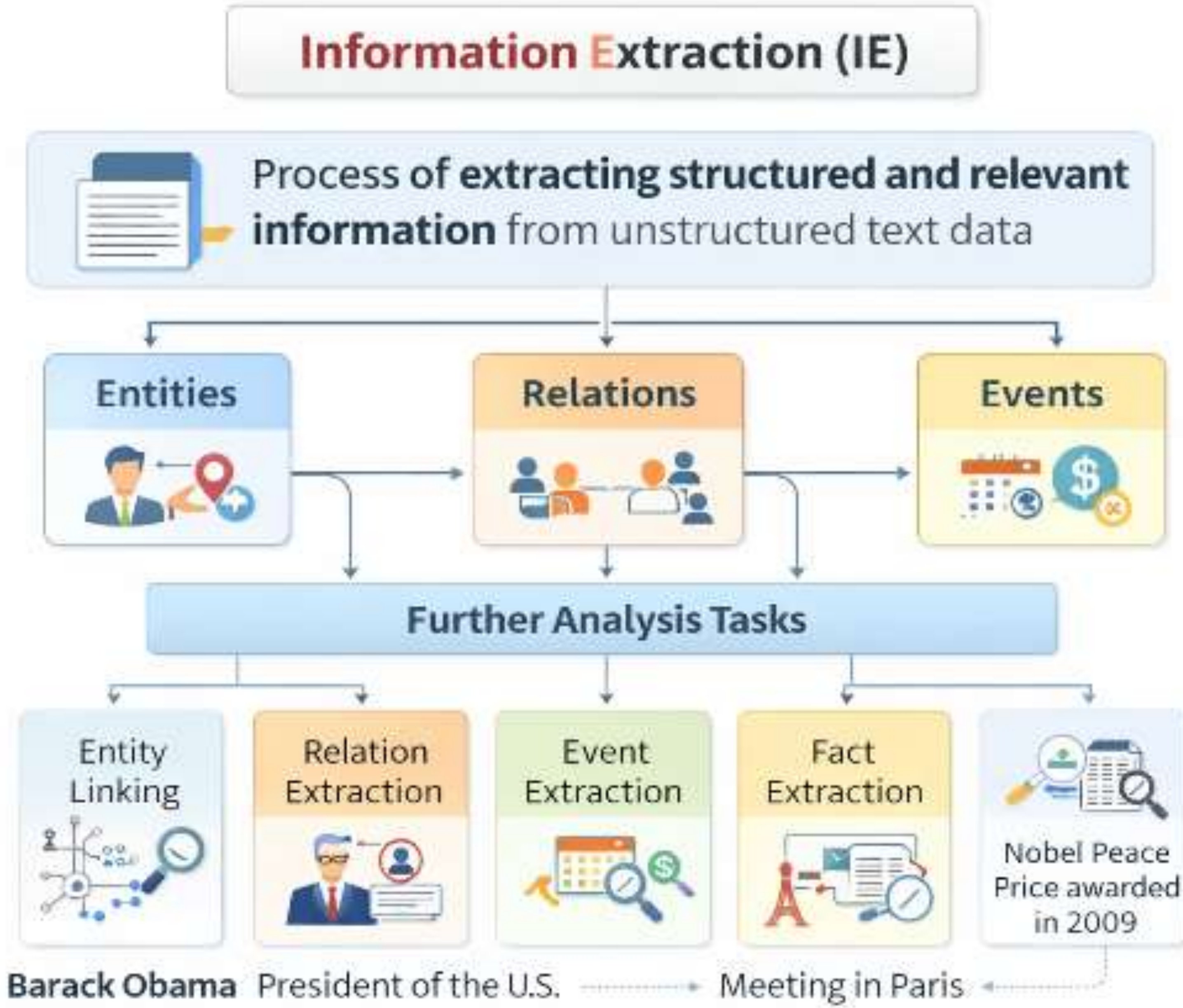
<u>deep learning,</u> etc.

It can help text analytics and text mining by providing models for classification, clustering, regression, anomaly detection, etc.


Machine Learning (ML)
This branch of artificial intelligence enables computers to learn from data and improve performance without explicit program-
Supervised learning
Unsupervised learning
Semi-supervised learning
Semi-supervised learning
Reinforcement learning
Use Cases
Text analytics text mining
Bram
Imager recognition
Models Provided by ML
Classification
Clustering
Anomaly detection

➢ <u>**Information Extraction (IE):**</u>

- This is <u>***the process of extracting structured and relevant information from unstructured text data, such as entities, relations, events, facts, etc.***</u>
- IE can help text analytics and text mining by <u>**providing data for further analysis**</u>, such as <u>**entity linking**</u>, <u>**relation extraction**</u>, <u>**event extraction**</u>, <u>**fact extraction**</u>, etc.

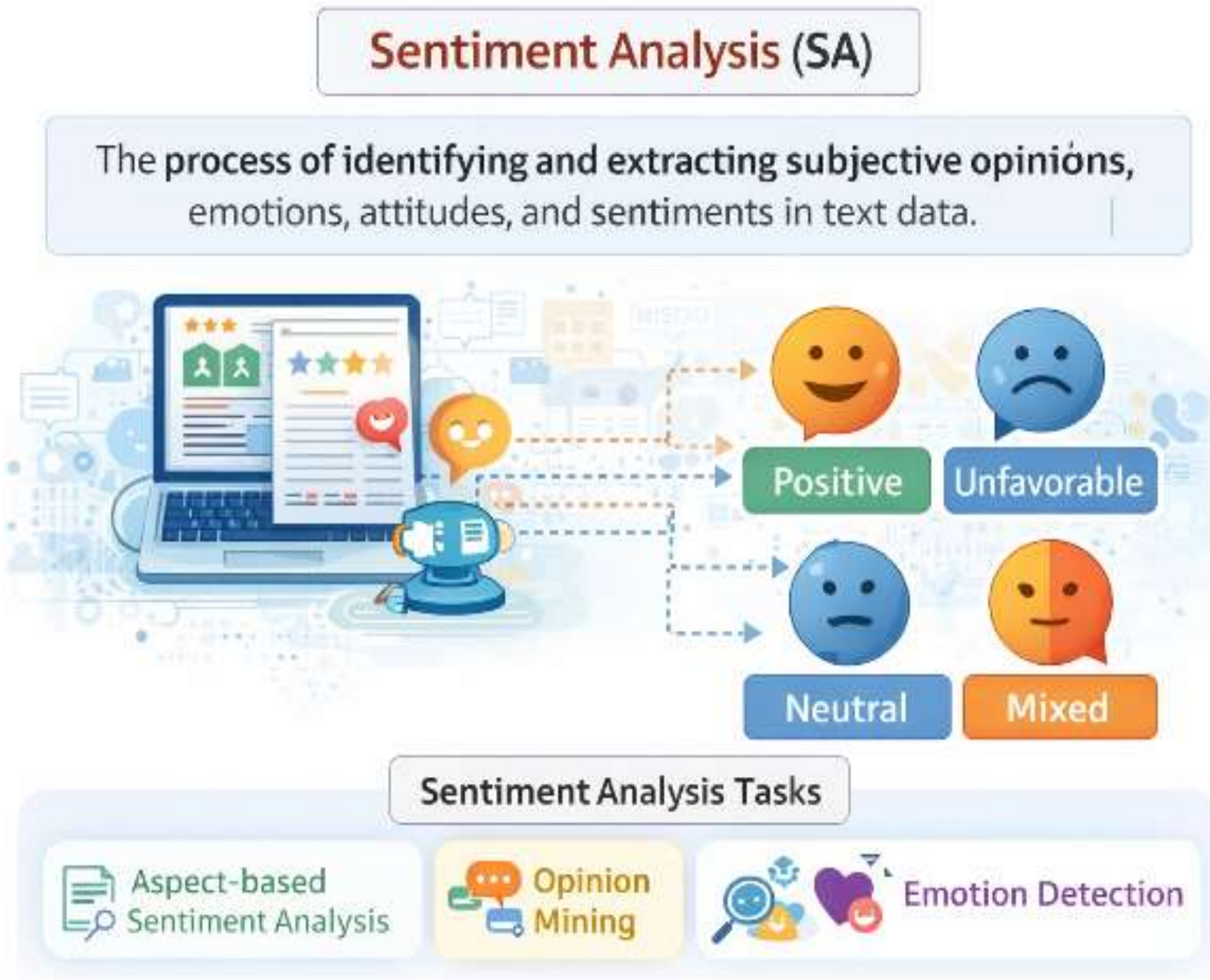

Information Extraction (IE)
Process of extracting structured and relevant information from unstructured text data
Entities
Relations
Events
Further Analysis Tasks
Entity Linking
Relation Extraction
Event Extraction
Fact Extraction
Nobel Peace Price awarded in 2009
Barack Obama
President of the U.S.
Meeting in Paris

➢ Sentiment Analysis (SA):

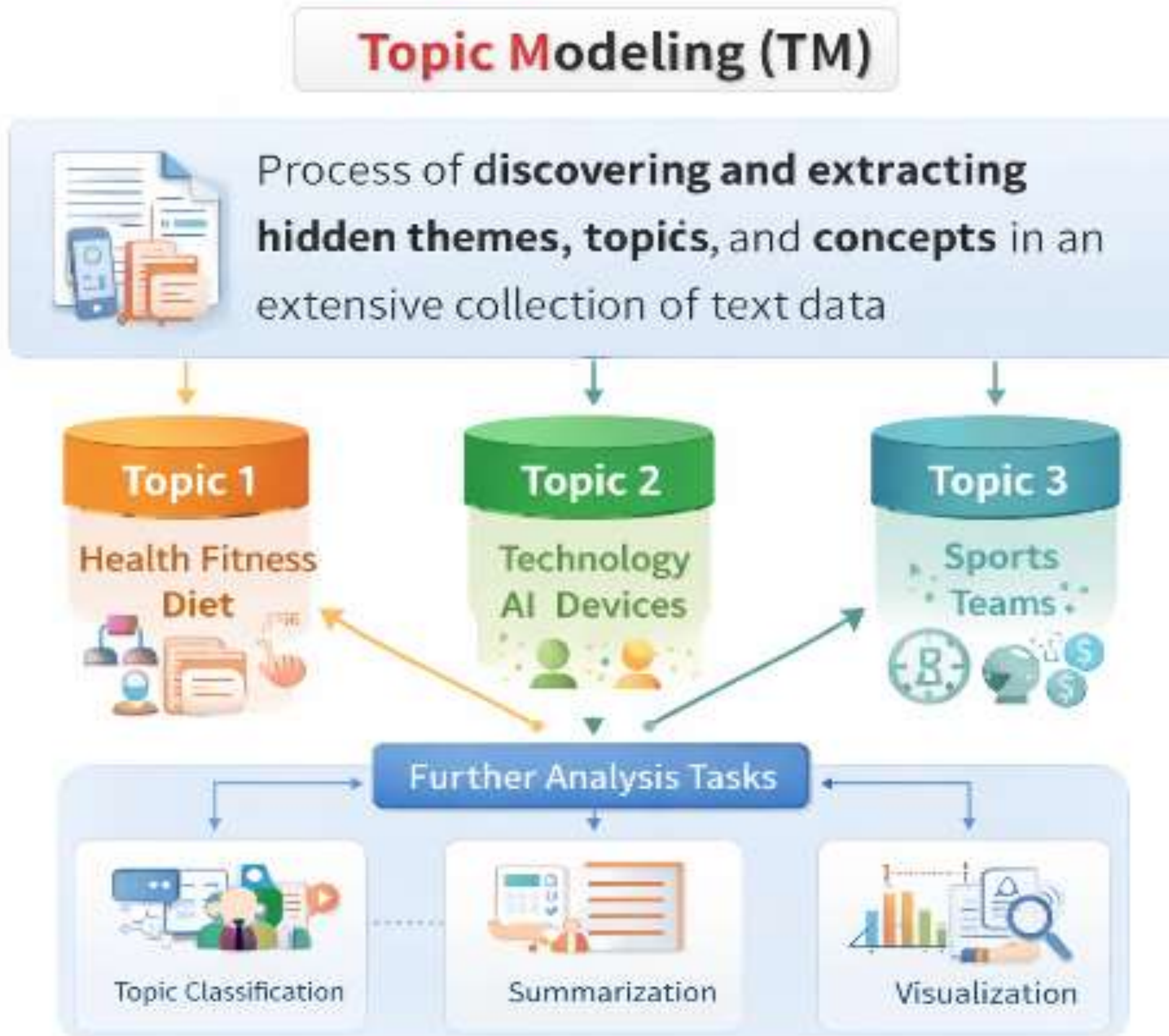
- This is ***the process of identifying and extracting the subjective opinions, emotions, attitudes, and sentiments expressed in text data, such as positive, unfavorable, neutral, mixed, etc***.
- SA can help text analytics and text mining by **providing data for further analysis**, such as **aspect-based sentiment analysis**, **opinion mining**, **emotion detection,** etc.

➢ Sentiment Analysis (SA):

- This is ***the process of identifying and extracting the subjective opinions, emotions, attitudes, and sentiments expressed in text data, such as positive, unfavorable, neutral, mixed, etc***.
- SA can help text analytics and text mining by **providing data for further analysis**, such as **aspect-based sentiment analysis**, **opinion mining**, **emotion detection,** etc.


Sentiment Analysis (SA)
The process of identifying and extracting subjective opinions, emotions, attitudes, and sentiments in text data.
Positive
Unfavorable
Neutral
Mixed
Sentiment Analysis Tasks
Aspect-based Sentiment Analysis
Opinion Mining
Emotion Detection

- **Topic Modeling (TM):**
- This is ***the process of discovering and extracting the hidden themes, topics, and concepts that are present in an extensive collection of text data, such as documents, articles, reviews, etc***.
- TM can help text analytics and text mining by **providing data for further analysis**, such as **topic classification**, **summarization**, and **visualization.**



Topic Modeling (TM)
Process of discovering and extracting hidden themes, topics, and concepts in an extensive collection of text data
Topic 1
Health Fitness
Diet
Topic 2
Technology
AI Devices
Topic 3
Sports
Teams
Further Analysis Tasks
Topic Classification
Summarization
Visualization

These are just some of the techniques and methods used. There are many more, such as

Text Summarization,
Text Generation,
Text Classification,
Text Clustering,
Text Visualization,
Text Similarity,
Text Search,
Text Recommendation, etc.

The choice of techniques and methods depends on the text data's type, size, and complexity, as well as the objectives and goals of the analysis and mining.

Text Analytics and **Text Mining** Applications

Text analytics and text mining have many applications and benefits for businesses across various industries and domains.

Some of the most common and practical applications for businesses are:


Customer service
Marketing
Product development
Risk management
Compliance

➢ **Customer service**:

- Text analytics and text mining can help businesses **improve customer service** by analyzing and extracting insights from customer feedback, such as reviews, surveys, emails, chats, calls, etc.
- They can help businesses identify and address customer issues, complaints, suggestions,preferences, satisfaction,loyalty, etc.
- They can also help companies provide automated and personalized responses and Solutions to customer queries and problems, such as chatbots,FAQs, knowledge bases, etc.

➢ **Marketing**:

- Text analytics and text mining can help businesses **improve their marketing** by analyzing and extracting insights from **customer behavior**, such as social media posts, online reviews, web searches, web browsing, etc.
- They can help businesses understand and segment their customer base, identify and target potential customers, monitor and measure their **brand reputation**, **sentiment**, and **awareness, and optimize and personalize their marketing campaigns, content, and offers.**

- **Product development**:

Text analytics and text mining can help businesses **improve their product development** by analyzing and extracting insights from product feedback, such as reviews, ratings, comments, suggestions, complaints, etc.

They can help businesses identify and address product issues, defects, bugs, errors, etc.

They can also help companies to discover and validate product features, functionalities, requirements, specifications,etc.

➢ **<u>Risk management</u>**:

Text analytics and text mining can help businesses **<u>improve risk management by analyzing and extracting insights from risk-related data, such as news articles, Reports, documents, regulations, policies, etc</u>**. They can help businesses identify and assess potential risks, threats, vulnerabilities, etc.

They can also help companies to monitor and mitigate existing risks, incidents, crises, etc.

- **Compliance:**

- Text analytics and text mining can **help businesses improve compliance by analyzing and extracting** insights from compliance-related data, such as ***laws***, **rules, standards**, **guidelines,** etc.
- They can help companies ensure and verify their compliance with various regulations, such as GDPR, HIPAA, KYC, AML, etc. They can also help businesses detect and prevent non-compliance, fraud, abuse, etc

- GDPR- **G**eneral **D**ata **P**rotection **R**egulation
- HIPAA- **H**ealth **I**nsurance **P**ortability and **A**ccountability **A**ct
- KYC- **K**now **Y**our **C**ustomer
- AML- **A**nti **M**oney **L**aundering

These are just some of the applications and benefits. There are many more, such as human resources, finance, healthcare, education, research, etc. The possibilities are endless, as text analytics and text mining can help businesses leverage the power of text data to achieve their goals and objectives.

- **Text Analytics** and **Text Mining** **Challenges and Solutions**

Text analytics and text mining are not without challenges and limitations. Some of the main challenges and solutions of text analytics and text mining are:

Data quality

Scalability

Privacy

Data governance

Conclusion

- Text analytics and text mining can unlock your business potential by leveraging the power of text data.. We are with **digital transformation company** with expertise in this field.

UNIT-3

SENTIMENT ANALYSIS

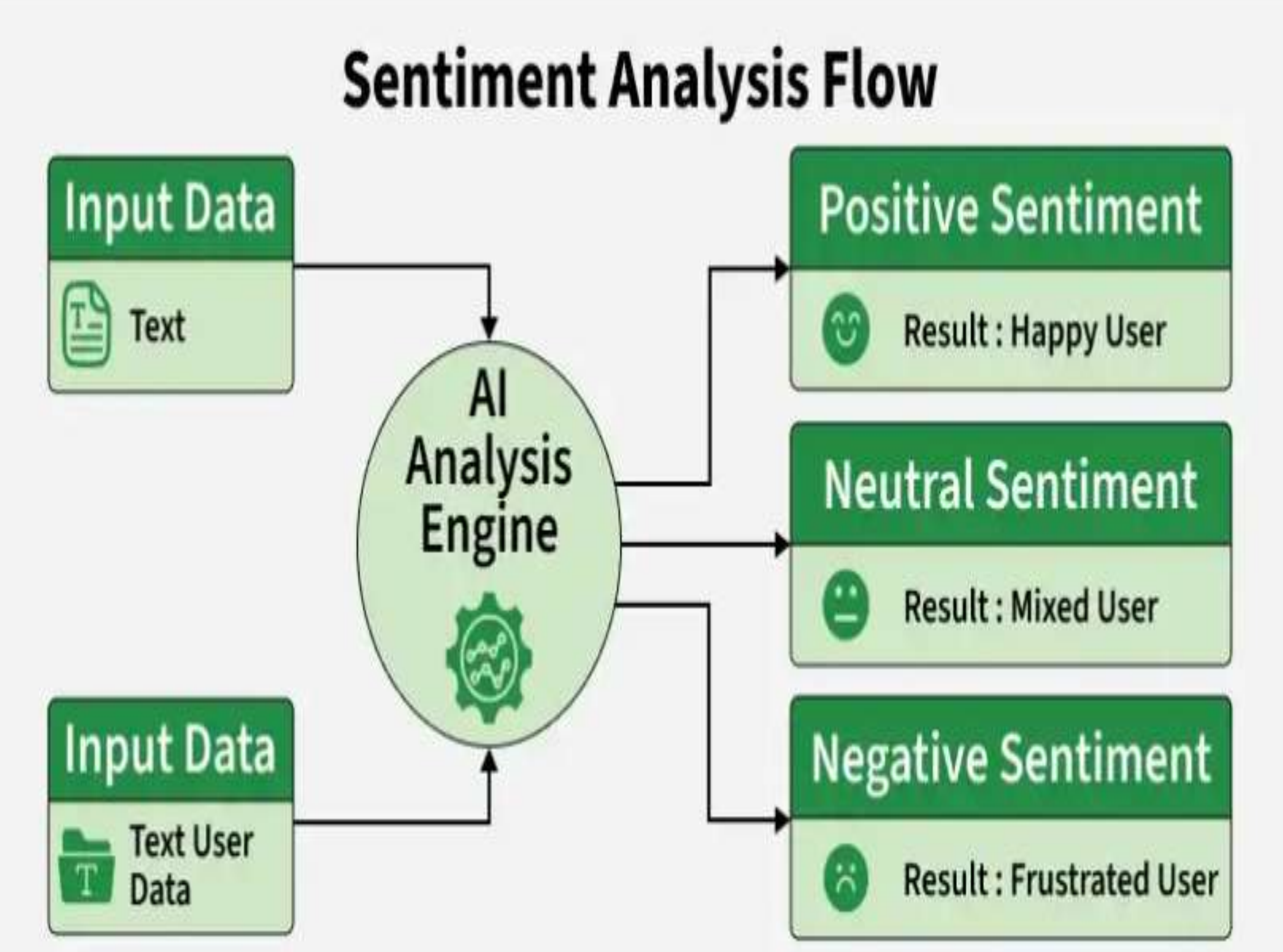
Sentiment analysis, also referred to as *opinion mining*, is an approach to natural language processing (NLP) that identifies the emotional tone behind a body of text. This is a popular way for organizations to determine and categorize opinions about a product, service or idea.

- Sentiment analysis involves the
  - use of **Data Mining**,

Machine Learning (ML),

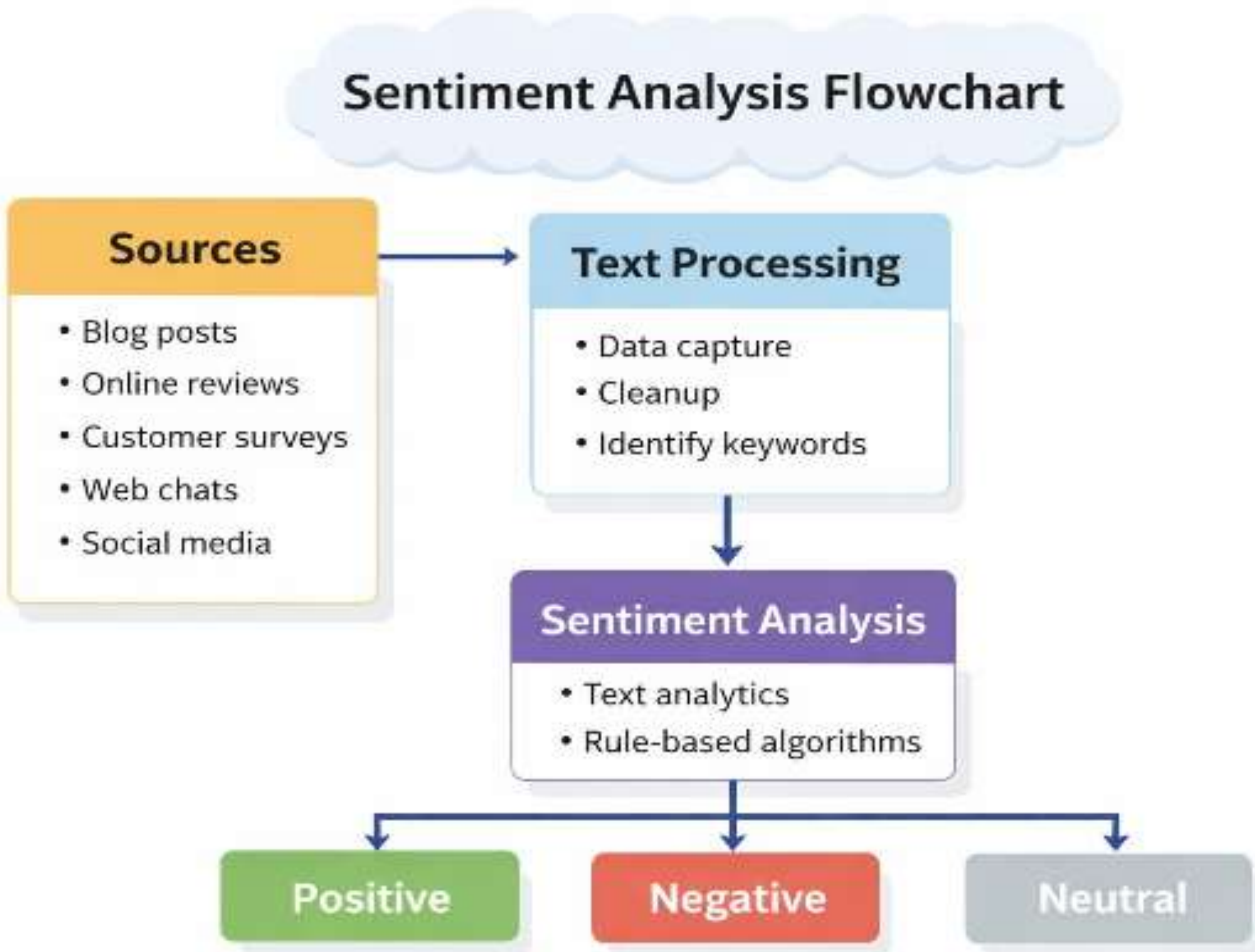
Artificial Intelligence (**AI**) and

Computational linguistics to mine text for sentiment and subjective information, such as whether it's expressing positive, negative or neutral feelings.


Sentiment Analysis Flow
Input Data
Text
Input Data
Text User Data
AI Analysis Engine
Positive Sentiment
Result : Happy User
Neutral Sentiment
Result : Mixed User
Negative Sentiment
Result : Frustrated User

Sentiment analysis systems help organizations gather insights into real-time customer sentiment, customer experience and brand reputation. Generally, these tools use text analytics to analyze online sources, such as emails, blog posts, online reviews, customer support tickets, news articles, survey responses, case studies, web chats, tweets, forums and comments. Algorithms are used to implement rule-based, automatic or hybrid methods of scoring whether the customer is expressing positive, negative words or neutral words.

- Sentiment analysis can also extract the polarity or the amount of positivity and negativity, as well as the subject and opinion holder within the text.
- This approach is used to analyze various parts of text, such as a full document or a paragraph, sentence or subsentence.
-


Sentiment Analysis Flowchart
Sources
• Blog posts
• Online reviews
• Customer surveys
• Web chats
• Social media
Text Processing
• Data capture
• Cleanup
• Identify keywords
Sentiment Analysis
• Text analytics
• Rule-based algorithms
Positive
Negative
Neutral

Sentiment analysis generally follows these steps:

- **Collect data.**
- **Data preprocessing.**
  - **Tokenization**
  - **lemmatization**
- **Keyword analysis**
- **Text scoring.**
- **Sentiment classification.**

Sentiment Analysis Workflow

Sentiment analysis is typically conducted through a series of steps: collecting data, **preprocessing the text** (including tokenization and **lemmatization**), analyzing **keywords**, **scoring the text**, and classifying the sentiment as positive, negative, or neutral.

- **Collect data.**

The text being analyzed is identified and collected. This involves using a ***web scraping bot*** or a ***scraping application programming interface(SAPI)***

- **Data preprocessing.**

In this stage, the data is processed to identify keywords that highlight the core meaning of the text. Other preprocessing steps include the following:

➢ ***Web scraping bot***

Writing your own code to pull data directly from websites.

➢ ***Scraping API(SAPI)***

Using a ready-made service that gives you structured data without coding.

In short:

Web scraping = DIY(Do it Yourself),(In this context, it refers to writing your own code or program to collect data directly from websites, instead of relying on a ready-made service.)

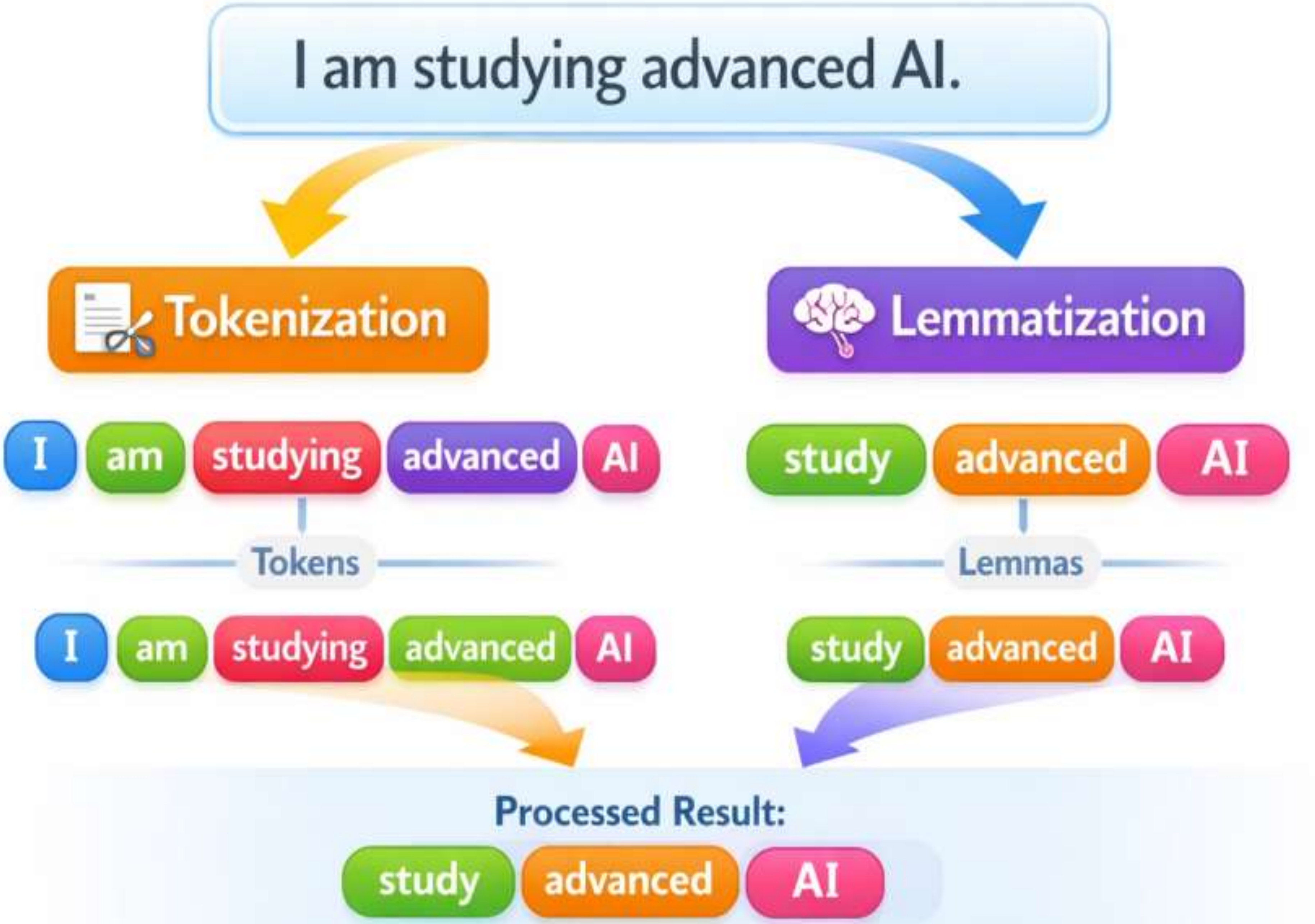
coding, flexible but harder.

- ***Scraping API(SAPI)*** = prebuilt service, easier but limited

A Scraping API is a ready-made service that collects data from websites and gives it to you in a clean, structured format (like JSON or CSV). Unlike web scraping bots where you write your own code, here the API does the heavy lifting — you just send a request and get the data back.

- Tokenization is used to break a sentence down into multiple elements, called tokens.
- Stop-word removal is performed to remove parts of speech that don't have meaning relevant to the sentiment of the text. This includes contractions, such as *I'm*, and words that have little information such as *is*, articles such as *the*, punctuation, URLs, special characters and capital letters.
- Lemmatization then converts keywords into their root form.

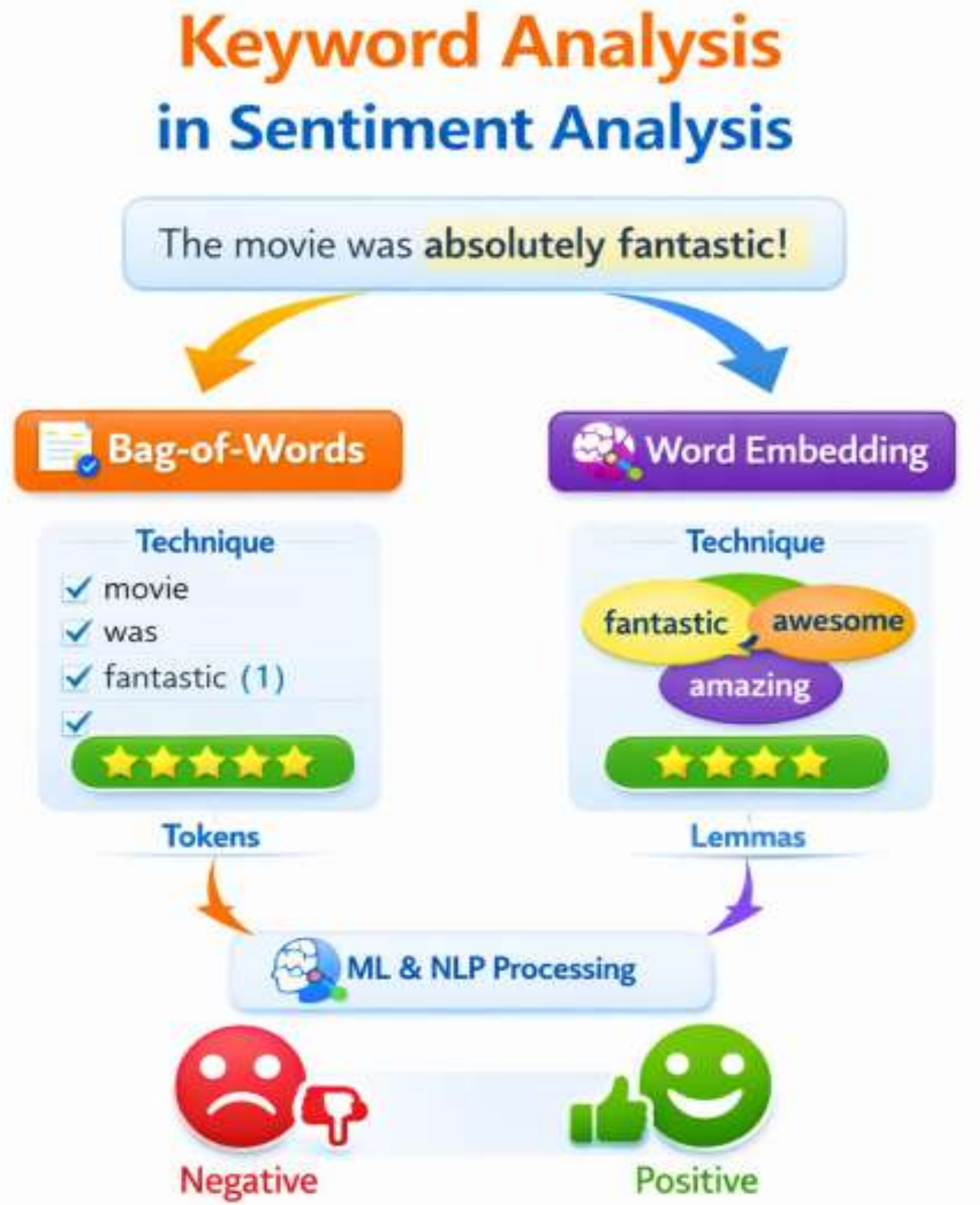
Tokenization and Lemmatization



- **Tokenization** → Breaks a sentence into smaller parts (tokens).
- **Lemmatization** → Converts words to their root/base form

- **Keyword analysis.**

ML and NLP algorithms automatically extract text features to identify negative or positive sentiment. MLapproaches used include the bag-of-words Technique that tracks the occurrence of words in text and the more nuanced word-embedding technique that uses neural networks to analyze words with similar meanings.



Keyword Analysis
in Sentiment Analysis
The movie was absolutely fantastic!
Bag-of-Words
Word Embedding
Technique
movie
was
fantastic (1)
Technique
fantastic
awesome
amazing
Tokens
Lemmas
ML & NLP Processing
Negative
Positive

➢ **Text scoring.**

A sentiment analysis tool scores the text using a rule based, automatic or hybrid ML model. Rule-based systems perform sentiment analysis based on predefined, lexicon-based rules and are often used in domains such as law and medicine, where a high degree of precision and human control is needed. Automatic systems use ML and deep learning techniques to learn from data sets. A hybrid model combines both approaches and is generally considered the most accurate model. These modelsoffer different approaches to assigning sentiment scores to pieces of text.

- **Sentiment classification.**

Once a model is picked and used to analyze a piece of text, it assigns a sentiment score to the text, including positive, negative or neutral. Organizations can also decide to view the results of their analysis at different levels, including document level, which pertains mostly to professional reviews and coverage; sentence level for comments and customer reviews; and sub-sentence level, which identifies phrases or clauses within sentences.

-

What is Sentiment Analysis? How does it relate to Text Mining?

Sentiment Analysis

- Sentiment Analysis, also known as **Opinion Mining**, is the process of using **Natural Language Processing (NLP), Machine Learning (ML), and Text Analytics** to identify, extract, and classify emotions, opinions, attitudes, or sentiments expressed in textual data.
- The primary objective is to determine whether a piece of text expresses:

Positive sentiment (e.g., "The product is excellent.")

Negative sentiment (e.g., "The service was terrible.")

Neutral sentiment (e.g., "The package arrived today.")

- Some advanced sentiment analysis systems can also identify emotions such as **joy, anger, sadness, surprise, fear, and disgust**.

Example

- **Customer Review:**

"The laptop performs amazingly, but the battery life is disappointing."

Sentiment Analysis identifies:

Laptop performance → Positive

Battery life → Negative

This is called **Aspect-Based Sentiment Analysis**, where different aspects of the same product receive different sentiments.

- **Relation between Sentiment Analysis and Text Mining**
- Text Mining is the broader process of extracting useful information, patterns, and knowledge from unstructured text data.
- Sentiment Analysis is a **specialized application of Text Mining** that focuses specifically on identifying opinions and emotions.

Text Mining	Sentiment Analysis
Extracts knowledge from text	Extracts opinions and emotions
Covers many tasks	One important application of text mining
Finds patterns, keywords, entities, topics	Determines positive, negative, or neutral opinions
Uses NLP and data mining techniques	Uses NLP, machine learning, and text analytics

- **How They Work Together**
- Text Mining performs tasks such as:
- Text preprocessing
- Tokenization
- Stop-word removal
- Stemming
- Feature extraction
- Sentiment Analysis uses these processed texts to classify the sentiment expressed.

- **Applications**
- Product review analysis
- Brand reputation monitoring
- Customer feedback analysis
- Social media monitoring
- Political opinion analysis
- Movie review analysis
- Hotel and restaurant review analysis

What are the Sources of Data for Sentiment Analysis?

Sentiment analysis requires large amounts of textual data from various online and offline sources.

1. **Social Media Platforms**
2. **Customer Reviews**
3. **Surveys and Feedback Forms**
4. **Blogs and Discussion Forums**
5. **News Articles**
6. **Emails**
7. **Chat Messages**
8. **Call Center Transcripts**
9. **Online Question-and-Answer Websites**
10. **Government and Public Feedback Portals**

Uses:

- Brand monitoring
- Election analysis
- Public opinion
- Example:
- "This mobile phone camera is amazing!"

What are the Common Challenges that Sentiment Analysis has to Deal With?

- Although sentiment analysis has become highly effective, it still faces several challenges due to the complexity of human language

1. Sarcasm and Irony

People often express the opposite of what they actually mean.

Example:

"Wonderful! Another software crash."

The word "Wonderful" is positive, but the actual sentiment is negative.

2. Context Dependency

The meaning of words changes depending on context.

Example:

"The battery is light."

Positive for a laptop.

"The exam was light."

Means easy.

The same word has different meanings.

3. Negation Handling

Negative words completely change the sentiment.

Example:

"The phone is good."

Positive.

"The phone is not good."

Negative.

The system must correctly identify negation words.

4. Mixed Sentiments

A single sentence may contain both positive and negative opinions.

Example:

"The display is excellent, but the battery is poor."

Positive toward display.

Negative toward battery.

5.Ambiguous Words

Some words have multiple meanings.

Example:

- The word **"Cool"** is ambiguous because it can carry different meanings depending on the context.
- Examples of "Cool"- Traditional meaning (temperature): - "The weather is cool today." → Neutral (describes low temperature). - Slang meaning (positive): - "That idea is really cool." → Positive (impressive, great).

6. Domain Dependency

A word may carry different sentiments in different domains.

Example: **Sharp**

Medical domain: - "The patient felt sharp pain." - Sentiment → Negative (pain is unpleasant).

Business/Intellectual domain: - "he has a sharp mind." Sentiment → Positive (intelligent, quick thinker).

Physical/Descriptive domain: - "The knife is sharp." - Sentiment → Neutral (factual description, no emotional tone).

The word "Sharp" flips sentiment depending on the domain: negative in medicine, positive in business, neutral in physical description.

7. Emojis and Emoticons

Users often express emotions using emojis instead of words.

Example:

Positive

Negative

Happiness

Proper interpretation is essential.

8. Spelling Mistakes and Informal Language

Social media contains:

- Short forms
- Misspellings
- Slang
- Internet abbreviations

Example:

"Gud product“

"Gr8 service“

“Cool”, “Lit”, “salty”

“LOL”, “OMG”, “TYSM”

Systems must normalize such text.

9. Multilingual Text

Users mix multiple languages in one sentence.

Example:

"Movie chaala bagundi."

English + Telugu

This makes sentiment detection more difficult.

10. Fake Reviews and Spam

Some reviews are intentionally misleading.

Examples:

Paid positive reviews

Competitor's fake negative reviews

Such reviews reduce analysis accuracy.

11. Lack of Training Data

Machine learning models require large, labeled datasets.

Problems:

Expensive annotation

Limited data for regional languages

Domain-specific datasets may be unavailable

12. Aspect-Level Sentiment Detection

Different aspects of the same product may have different sentiments.

Example:

"Camera is excellent but battery is poor."

The system should identify:

Camera → Positive

Battery → Negative

instead of assigning a single overall sentiment.

SENTIMENTAL ANALYSIS APPLICATIONS

- Social Media Monitoring
- Customer Experience (CX) & Support
- Market Research & Product Development
- Financial & Stock Market Prediction
- Employee Satisfaction
- Political Analysis

WEB AND SOCIAL MEDIA ANALYTICS

UNIT-IV

Web Analytics & Web Mining

Topics:

1. Security First Insurance Deepens Connection with Policyholders

2. Web Mining Overview

3. Web Content and Web Structure Mining

4. Search Engines

5. Search Engine Optimization

6. Web Usage Mining (Web Analytics)

7. Web Analytics Maturity Model and Web Analytics Tools

1. Security First Insurance Deepens Connection with Policyholders

Security First Insurance is one of the largest homeowners' insurance companies in Florida. Headquarters in Ormond Beach, it employs more than 80 insurance professionals to serve its nearly 190,000 customers.

CHALLENGE

Being There for Customers Storm After Storm, Year After Year

Florida has more property and people exposed to hurricanes than any state in the country. Each year, the Atlantic Ocean averages 12 named storms and nine named hurricanes. Security First is one of a few Florida homeowners' insurance companies that has the financial strength to withstand multiple natural disasters. "One of our promises is to be there for our customers, storm after storm, year after year," says Werner Kruck, chief operating officer for Security First.

During a typical month, Security First processes 700 claims. However, in the aftermath of a hurricane, that number can swell to tens of thousands within days. It can be a challenge for the company to quickly scale up to handle the influx of customers trying to file post-storm insurance claims for damaged property and possessions. In the past, customers submitted claims primarily by phone and sometimes email. Today, policyholders use as many means available to connect with an agent or claims representative, including posting a question or comment on the company's Facebook page or Twitter account.

Although Security First provides ongoing monitoring of its Facebook and Twitter accounts, as well as its multiple email addresses and call centers, the company knew that the communication volume after a major storm required a more aggressive approach. "We were concerned that if a massive number of customers contacted us through email or social media after a hurricane, we would be unable to respond quickly and appropriately," Kruck says. "We need to be available to our customers in whatever way they want to contact us. "In addition, Security First recognized the need to integrate its social media responses into the claims process and document those responses to comply with industry regulations.

SOLUTION

Providing Responsive Service, No Matter How Customers Get in Touch

Security First contacted IBM Business Partner Integrity for help with harnessing social media to improve the customer experience. Integrity configured a solution built on key IBM Enterprise Content Management software components, featuring IBM Content Analytics with Enterprise Search, IBM Content Collector for Email and IBM® FileNet® Content Manager software. Called Social Media Capture (SMC4), the Integrity solution offers four critical capabilities for managing social media platforms: capture, control, compliance and communication.

For example, the SMC4 solution logs all social networking interaction for Security First, captures content, monitors incoming and outgoing messages and archives all communication for compliance reviews.

Because the solution uses open IBM Enterprise Content Management software, Security First can easily link it to critical company applications, databases and processes. For example, Content Collector for Email software automatically captures email content and attachments and sends an email back to the policyholder acknowledging receipt. In addition, Content Analytics with Enterprise Search software sifts through and analyzes the content of customers' posts and emails. The software then captures information gleaned from this analysis directly into claims documents to begin the claims process. Virtually all incoming communication from the company's web, the Internet and emails is pulled into a central FileNet Content Manager software repository to maintain control and route to the appropriate workflow.
"We can bring the customer conversation and any pictures and attachments into our policy and claims management system and use it to trigger our claims process and add to our documentation," says Kruck.

Prioritizing Communications with Access to Smarter Content

People whose homes have been damaged or destroyed by a hurricane are often displaced quickly, with little more than their clothes on their backs. Grabbing an insurance policy on the way out the door is often an afterthought. They're relying on their insurance companies to have the information they need to help them get their lives back in order as quickly as possible. When tens of thousands of policyholders require assistance within a short period of time, Security First must triage requests quickly. The Content Analytics with Enterprise Search software that anchors the SMC4 solution provides the information necessary to help the company identify and address the most urgent cases first.The software automatically sifts through data in email and social media posts, tweets and comments using text mining, text analytics, natural language processing and sentiment analytics to detect words and tones that identify significant property damage or that convey distress. Security First can then prioritize the messages and route them to the proper personnel to provide reassurance, handle complaints or process a claim. "With access to smarter content, we can respond to our customers in a more rapid, efficient and personalized way," says Kruck. "When customers are having a bad experience, it's really important to get to them quickly with the level of assistance appropriate to their particular situations."

RESULTS

Successfully Addressing Potential Compliance Issues

Companies in all industries must stay compliant with new and emerging regulatory requirements regarding social media. The text analysis capabilities provided in the IBM software help Security First filter inappropriate incoming communications and audit outbound communications, avoiding potential issues with message content.

The company can be confident that the responses its employees provide are compliant and controlled based on both Security First policies and industry regulations.

Security First can designate people or roles in the organization that are authorized to create and submit responses. The system automatically verifies these designations and analyzes outgoing message content, stopping any ineffective or questionable communications for further review. “Everything is recorded for compliance, so we can effectively track and maintain the process. We can control which employees respond, their level of authority and the content of their responses,” says Kruck.

These capabilities give Security First the confidence to expand its use of social media. Because compliance is covered, the company can focus on additional opportunities for direct dialog with customers. Before this solution, Security First filtered customer communications through agents. Now it can reach out to customers directly and proactively as a company.

“We’re one of the first insurance companies in Florida to make ourselves available to customers whenever, wherever and however they choose to communicate. We’re also managing internal processes more effectively and proactively, reaching out to customers in a controlled and compliant manner,” says Kruck.

Some of the prevailing business benefits of creative use of Web and social analytics include:

- Turns social media into an actionable communications channel during a major disaster.
- Speeds claims processes by initiating claims with information from email and social media posts.
- Facilitates prioritizing urgent cases by analyzing social media content for sentiments.
- Helps ensure compliance by automatically documenting social media communications.

2. Web Mining Overview

The Internet has forever changed the landscape of business as we know it. Because of the highly connected, flattened world and broadened competitive field, today’s companies are increasingly facing greater opportunities (being able to reach customers and markets that they may have never thought possible) and bigger challenges (a globalized and ever-changing competitive marketplace). Ones with the vision and capabilities to deal with such volatile environments are greatly benefiting from it, while others who resist are having a hard time surviving. Having an engaged presence on the Internet is not a choice anymore; it is a business requirement. Customers are expecting companies to offer their products and/or services over the Internet. They are not only buying products and services but also talking about companies and sharing their transactional and usage experiences with others over the Internet.

The growth of the Internet and its enabling technologies has made data creation, data collection, and data/information/opinion exchange easier. Delays in service, manufacturing, shipping, delivery, and customer inquiries are no longer private incidents, and are accepted as necessary evils.

Now, thanks to social media tools and technologies on the Internet, everybody knows everything. Successful companies are the ones who embrace these Internet technologies and use them for the betterment of their business processes so that they can better communicate with their customers, understand their needs and wants and serve them thoroughly and expeditiously. Being customer-focused and keeping customers happy have never been as important a concept for business as they are now, in this age of the Internet and social media.

The World Wide Web (or, for short, the Web) serves as an enormous repository of data and information on virtually everything one can conceive—business, personal, you name it; an abundant amount of it is there. The Web is perhaps the world's largest data and text repository, and the amount of information on the Web is growing rapidly. A lot of interesting information can be found online: whose homepage is linked to which other pages are linked to which other pages, how many people have links to a specific Web page, and how a particular site is organized. In addition, each visitor to a Web site, each search on a search engine, each click on a link, and each transaction on an e-commerce site creates additional data. Although unstructured textual data in the form of Web pages coded in HTML or XML is the dominant content of the Web, the Web infrastructure also contains hyperlink information (connections to other Web pages) and usage information (logs of visitors' interactions with Web sites), all of which provide rich data for knowledge discovery. Analysis of this information can help us make better use of Web sites and aid us in enhancing relationships and value for the visitors to our own Web sites.

Because of its sheer size and complexity, mining the Web is not an easy undertaking by any means. The Web also poses great challenges for effective and efficient knowledge discovery (Han and Kamber, 2006):

The Web is too big for effective data mining. The Web is so large and growing so rapidly that it is difficult to even quantify its size. Because of the sheer size of the Web, it is not feasible to set up a data warehouse to replicate, store, and integrate all the data on the Web, making data collection and integration a challenge.

The Web is too complex. The complexity of a Web page is far greater than a page in a traditional text document collection. Web pages lack a unified structure. They contain far more authoring style and content variation than any set of books, articles, or other traditional text-based document.

The Web is too dynamic. The Web is a highly dynamic information source. Not only does the Web grow rapidly, but its content is constantly being updated. Blogs, news stories, stock market results, weather reports, sports scores, prices, company advertisements, and numerous other types of information are updated regularly on the Web.

The Web is not specific to the domain. The Web serves a broad diversity of communities and connects billions of workstations. Web users have very different backgrounds, interests, and usage purposes. Most users may not have good knowledge of the structure of the information network and may not be aware of the heavy cost of a particular search that they perform.

The Web has everything. Only a small portion of the information on the Web is truly relevant or useful to someone (or some task). It is said that 99 percent of the information on the Web is useless to 99 percent of Web users. Although this may not seem obvious, it is true that a particular person is generally interested in only a tiny portion of the Web, whereas the rest of the Web contains information that is uninteresting to the user and may seemingly dissuade results. Finding the portion of the Web that is truly relevant to a person and the task being performed is a prominent issue in Web-related research.

These challenges have prompted many research efforts to enhance the effectiveness and efficiency of discovering and using data assets on the Web. Several index-based search engines constantly search the Web and index Web pages under certain keywords. Using these search engines, an experienced user may be able to locate documents by providing a set of tightly constrained keywords or phrases. However, a simple keyword-based search engine suffers from several deficiencies. First, a topic of any breadth can easily contain hundreds or thousands of documents. This can lead to many document entries being returned by the search engine, many of which are marginally relevant to the topic. Second, many documents that are highly relevant to a topic may not contain the exact keywords defining them. As we will cover in more detail later in this chapter, compared to keyword-based Web search , **Web mining** is a prominent (and more challenging) approach that can be used to substantially enhance the power of Web search engines because Web mining can identify authoritative Web pages, classify Web documents, and resolve many ambiguities and subtleties raised in keyword-based Web search engines.

Web mining (or Web data mining) is the process of discovering intrinsic relationships (i.e., interesting and useful information) from Web data, which are expressed in the form of textual, linkage, or usage information. The term Web mining was first used by Etzioni (1996); today, many conferences, journals, and books focus on Web data mining. It is a continually evolving area of technology and business practice. Web mining is essentially the same as data mining that uses data generated over the Web. The goal is to turn vast repositories of business transactions, customer interactions, and Web site usage data into actionable information (i.e., knowledge) to promote better decision making throughout the enterprise. Because of the increased popularity of term **analytics**, nowadays many have started to call Web mining **Web analytics**. However, these two terms are not the same. Although Web analytics is primarily focused on Web site usage data, Web mining is inclusive of all data generated via the Internet, including transaction, social, and usage data. While Web analytics aims to describe what has happened on the Web site (employing a predefined, metrics-driven descriptive analytics methodology), Web mining aims to discover previously unknown patterns and relationships (employing a novel predictive or prescriptive analytics methodology). From a big-picture perspective, Web analytics can be considered a part of Web mining.

As Figure 8.1 presents, a simple taxonomy of Web mining is divided into three main areas: Web content mining, Web structure mining, and Web usage mining. In the figure, the data sources used in these three main areas are also specified. Although these three areas are shown separately, as you will see in the following section, they are often used collectively and synergistically to address business problems and opportunities.

As Figure 8.1 indicates, Web mining relies heavily on data mining and text mining and their enabling tools and techniques. The figure also indicates that these three generic areas are further extended into several very well-known application areas.

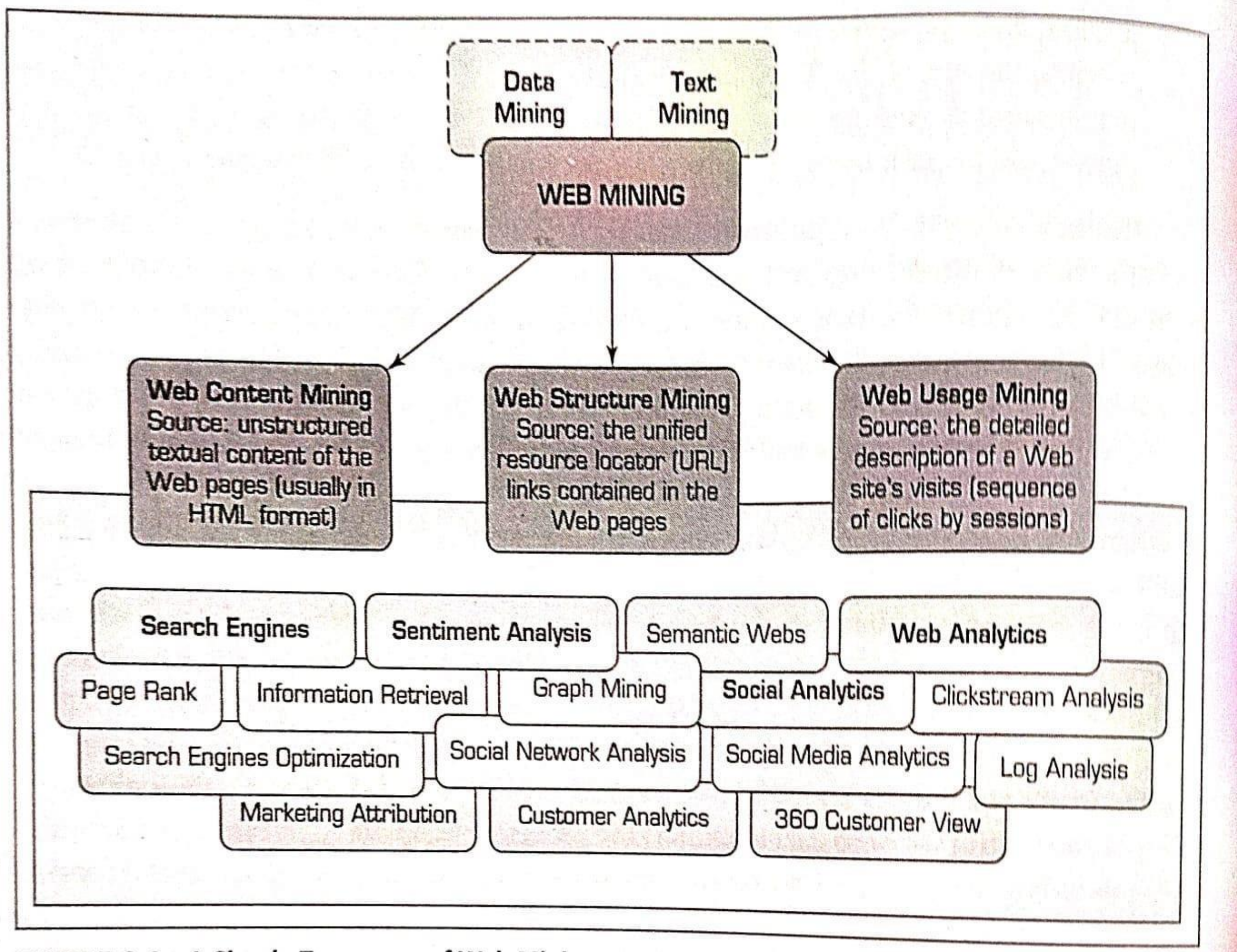


FIGURE 8.1 A Simple Taxonomy of Web Mining.

3. Web Content and Web Structure Mining

Web content mining refers to the extraction of useful information from Web pages. The documents may be extracted in some machine-readable format so that automated techniques can extract some information from these Web pages. Web crawlers (also called spiders) are used to read through the content of a Web site automatically. The information gathered may include document characteristics like what are used in text mining, but it may also include additional concepts, such as the document hierarchy. Such an automated (or semiautomated) process of collecting and mining Web content can be used for competitive intelligence (collecting intelligence about competitors' products, services, and customers). It can also be used for information/news/opinion collection and summarization, sentiment analysis, automated data collection, and structuring for predictive modeling. As an illustrative example to use Web content mining as an automated data collection tool, consider the following. For more than 10 years, two of the three authors of this book (Drs. Sharda and Delen) have been developing models to predict the financial success of Hollywood movies before their theatrical release. The data that they use for training the models come from several Web sites, each of which has a different hierarchical page structure. Collecting a large set of variables on thousands of movies (from the past several years) from these Web sites is a time-demanding, error-prone process. Therefore, they use Web content mining and spiders as an enabling technology to automatically collect, verify, validate (if the specific data item is available on more than one Web site, then the values are validated against each other and anomalies are captured and recorded), and store these values in a relational database. That way, they ensure the quality of the data while saving valuable time (days or weeks) in the process.

In addition to text, Web pages also contain hyperlinks pointing one page to another.Hyperlinks contain a significant amount of hidden human annotation that can potentially help to automatically infer the notion of centrality or authority. When a Web page developer includes a link pointing to another Web page, this may be regarded as the developer’s endorsement of the other page. The collective endorsement of a given page by different developers on the Web may indicate the importance of the page and may naturally lead to the discovery of authoritative Web pages (Miller, 2005).Therefore, the vast amount of Web linkage information provides a rich collection of information about the relevance, quality, and structure of the Web's contents, and thus is a rich source for Web mining.

Web content mining can also be used to enhance the results produced by search engines. In fact, search is perhaps the most prevailing application of Web content mining. Web structure mining. A search on the Web to obtain information on a specific topic (presented as a collection of keywords or a sentence) usually returns a few relevant, high-quality Web pages and a larger number of unusable Web pages. Use of a relevance index based on keywords and authoritative pages (or some measure of it) will improve the search results and ranking of relevant pages. The idea of authority (or authoritative pages) stems from earlier information retrieval work using citations among journal articles to evaluate the impact of research papers (Miller, 2005). Though that was the origin of the idea, there are significant differences between the citations in research articles and hyperlinks on Web pages. First, not every hyperlink represents an endorsement (some links are created for navigation purposes, and some are for paid advertisement). While this is true, if most of the hyperlinks are of the type of endorsement, then the collective opinion will still prevail. Second, for commercial and competitive interests, one authority will rarely have its Web page point to rival authorities in the same domain. For example, Microsoft may prefer not to include links on its web pages to Apple's Web sites, because this may be regarded as endorsement of its competitor's authority. Third, authoritative pages are seldom particularly descriptive; for example, the main Web page of Yahoo! may not contain the explicit self-description that it is in fact a Web search engine.

The structure of Web hyperlinks has led to another important category of Web pages called a hub. A hub is one or more Web pages that provide a collection of links to authoritative pages. Hub pages may not be prominent and only a few links may point to them; however, they provide links to a collection of prominent sites on a specific topic of interest. A hub could be a list of recommended links on an individual's homepage, recommended reference sites on a course Web page, or a professionally assembled resource list on a specific topic. Hub pages play the role of implicitly conferring the authorities on a narrow field. In essence, a close symbiotic relationship exists between good hubs and authoritative pages; a good hub is good because it points to many good authorities, and a good authority is good because it is being pointed to by many good hubs. Such relationships between hubs and authorities make it possible to automatically retrieve high-quality content from the Web.

The most popular publicly known and referenced algorithm used to calculate hubs and authorities is hyperlink-induced topic search (HITS). It was originally developed by Kleinberg (1999) and has since been improved by many researchers. HITS is a link-analysis algorithm that rates Web pages using the hyperlink information contained within them. In the context of Web search, the HITS algorithm collects a base document set for a specific query. It then recursively calculates the hub and authority values for each document. To gather the base document set, a root set that matches the query is fetched from a search engine.

For each document retrieved, a set of documents that point to the original document and another set of documents that is pointed to by the original document are added to the set as the original document’s neighborhood. A recursive process of document identification and link analysis continues until the hub and authority values converge. These values are then used to index and prioritize the document collection generated for a specific query.

Web structure mining is the process of extracting useful information from the links embedded in Web documents. It is used to identify authoritative pages and hubs, which are the cornerstones of the contemporary page-rank algorithms that are central to popular search engines such as Google and Yahoo!. Just as links going to a Web page may indicate a site’s popularity (or authority), links within the Web page (or the complete Web site) may indicate the depth of coverage of a specific topic. Analysis of links is very important in understanding the interrelationships among large numbers of Web pages, leading to a better understanding of a specific Web community, clan, or clique.

4. SEARCH ENGINES

Currently, there is no denying the importance of Internet search engines. As the size and complexity of the World Wide Web increases, finding what you want is becoming a complex and laborious process. People use search engines for a variety of reasons. We use them to learn about a product or a service before committing to buy (including who else is selling it, what the prices are at different locations/sellers, the common issues people are discussing about it, how satisfied previous buyers are, what other products or services might be better, etc.) and search for places to go, people to meet, and things to do.

In a sense, search engines have become the centerpiece of most Internet-based transactions and other activities. The incredible success and popularity of most Internet-based transactions and other activities. The incredible success and popularity of most Internet search engines, most notably Google, is a good testament to this claim. What is somewhat a mystery to many is how a search engine does what it is meant to do. In simplest terms, a search engine is a software program that searches for documents (Internet sites or files) based on the keywords (individual words, multi-word terms, or a complete sentence) that users have provided to do with the subject of their inquiry. Search engines are the workhorses of the Internet, responding to billions of queries.

Search engines are a popular term for information retrieval systems. Although Web search engines are the most popular, search engines are often used in a context other than the Web, such as desktop search engines or document retrieval systems. The overall goal of a search engine is to return one or more documents/pages (if more than one document/page applies, a rank-order list is often provided) that best matches the user's query. The two metrics that are often used to evaluate search engines are effectiveness (or quality—finding the right documents/pages) and efficiency (or speed—returning a response quickly). These two metrics tend to work in reverse direction; improving one tends to worsen the other. Often, based on user expectation, search engines focus on one at the expense of the other. Better search engines are the ones that excel in both at the same time. Because search engines not only search but, in fact, find and return the documents/pages, perhaps a more appropriate name for them would be "finding engines."

Anatomy of a Search Engine

Now let us dissect a search engine and look inside it. At the highest level, a search engine system is composed of two main cycles: a development cycle and a responding cycle. While one is interfacing with the World Wide Web, the other is interfacing with the user. One can think of the development cycle as a production process (manufacturing and inventorying documents/pages) and the responding cycle as a retailing process (providing customers/users with what they want).

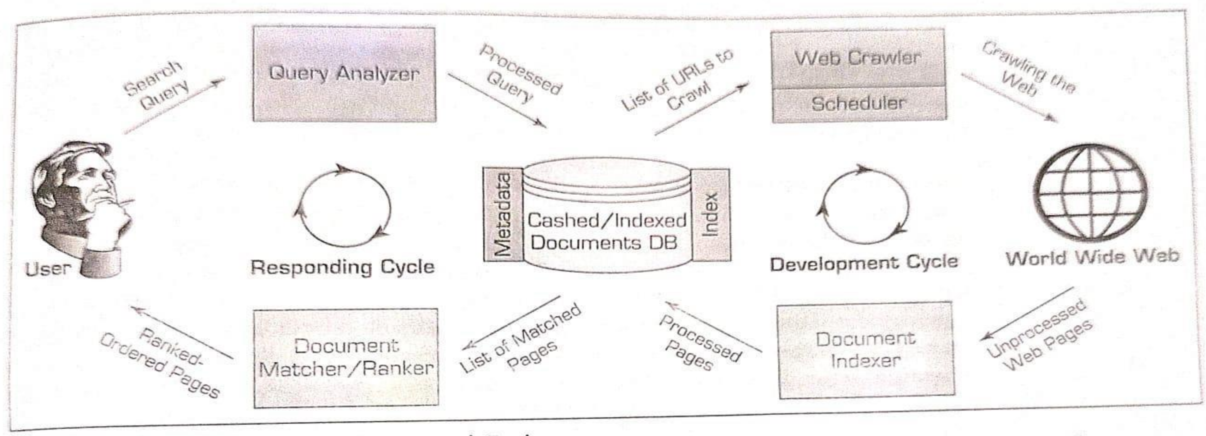


FIGURE 8.2 Structure of a Typical Internet Search Engine.

1. Development Cycle

The two main components of the development cycle are the Web crawler and document indexer. The purpose of this cycle is to create a huge database of documents/pages organized and indexed based on their content and information value. The reason for developing such a repository of documents/pages is quite obvious: Due to its sheer size and complexity, searching the Web to find pages in response to a user query is not practical (or feasible within a reasonable time frame); therefore, search engines "cache the Web" into their database, and uses the cashed version of the Web for searching and finding. Once created, this database allows search engines to rapidly and accurately respond to user queries.

Web Crawler

A Web crawler (also called a spider or a Web spider) is a piece of software that systematically browses (crawls through) the World Wide Web for the purpose of finding and fetching Web pages. Often Web crawlers copy all the pages they visit for later processing by other functions of a search engine.

A Web crawler starts with a list of URLs to visit, which are listed in the scheduler and often are called the seeds. These URLs may come from submissions made by Webmasters or, more often, they come from the internal hyperlinks of previously crawled documents/pages. As the crawler visits these URLs, it identifies all the hyperlinks in the page and adds them to the list of URLs to visit (i.e., the scheduler). URLs in the scheduler are recursively visited according to a set of policies determined by the specific search engine. Because there are large volumes of Web pages, the crawler can only download a limited number of them within a given time; therefore, it may need to prioritize its downloads.

Document Indexer

As the documents are found and fetched by the crawler, they are stored in a temporary staging area for the document indexer to grab and process. The document indexer is responsible for processing the documents (Web pages or document files) and placing them into the document database. In order to convert the documents/pages into the desired, easily searchable format, the document indexer performs the following tasks.

STEP 1: PREPROCESSING THE DOCUMENTS

Because the documents fetched by the crawler may all be in different formats, for the ease of processing them further, in this step they all are converted to some type of standard representation. For instance, different content types (text, hyperlink, image, etc.) may be separated from each other, formatted (if necessary), and stored in a place for further processing.

STEP 2: PARSING THE DOCUMENTS

This step is essentially the application of text mining (i.e., computational linguistic, natural language processing) tools and techniques to a collection of documents/pages. In this step, first the standardized documents are parsed into their components to identify index-worthy words/terms. Then, using a set of rules, the words/terms are indexed. More specifically, using tokenization rules, the words/terms/entities are extracted from the sentences in these documents. Using proper lexicons, the spelling errors and other anomalies in these words/terms are corrected. Not all the terms are discriminators. The nondiscrimination words/terms (also known as stop words) are eliminated from the list of index-worthy words/terms. Because the same word/term can be in many different forms, stemming is applied to reduce the words/terms to their root forms. Again, using lexicons and other language-specific resources (e.g., WordNet), synonyms and homonyms are identified, and the word/term collection is processed before moving into the indexing phase.

STEP 3: CREATING THE TERM-BY-DOCUMENT MATRIX

In this step, the relationships between the words/terms and documents/pages are identified. The weight can be as simple as assigning 1 for presence or 0 for absence of the word/term in the document/page. Usually more sophisticated weight schemas are used. For instance, as opposed to binary, one may choose to assign frequency of occurrence (number of times the same word/term is found in a document) as a weight. text mining research and practice have clearly indicated that the best weighting may come from the use of term-frequency divided by inverse-document-frequency (TF/IDF). This algorithm measures the frequency of occurrence of each word/term within a document and then compares that frequency against the frequency of occurrence in the document collection. As we all know, not all high-frequency words/term are good document discriminators; and a good document discriminator in a domain may not be one in another domain. Once the weighing schema is determined, the weights are calculated, and the term-by-document index file is created.

2. Response Cycle

The two main components of the responding cycle are the query analyzer and the document matcher/ranker.

Query Analyzer

The query analyzer is responsible for receiving a search request from the user (via the search engine's Web server interface) and converting it into a standardized data structure, so that it can be easily queried/matched against the entries in the document database. How the query analyzer does what it is supposed to do is quite like what the document indexer does (as we have just explained). The query analyzer parses the search string into individual words/terms using a series of tasks that include tokenization, removal of stop words, stemming, and word/term disambiguation (identification of spelling errors, synonyms, and homonyms).

The close similarity between the query analyzer and document indexer is not coincidental. In fact, it is quite logical, because both are working off of the document database; one is putting in documents/pages using a specific index structure, and the other is converting a query string into the same structure so that it can be used to quickly locate most relevant documents/pages.

Document Matcher/Ranker

This is where the structured query data is matched against the document database to find the most relevant documents/pages and rank them in the order of relevance/importance. The proficiency of this step is perhaps the most important component when different search engines are compared to one another. Every search engine has its own (often proprietary) algorithm that it uses to carry out this important step.

The early search engines used a simple keyword match against the document database and returned a list of ordered documents/pages, where the determinant of the order was a function that used the number of words/terms matched between the query and the document along with the weights of those words/terms. The quality and the usefulness of the search results were not all that good. Then, in 1997, the creators of Google came up with a new algorithm, called PageRank. As the name implies, PageRank is an algorithmic way to rank-order documents/pages based on their relevance and value/importance. Technology Insights 8.1 provides a high-level description of this patented algorithm

PageRank Algorithm

PageRank is a link analysis algorithm—named after Larry Page, one of the two inventors of Google, which started as a research project at Stanford University in 1996—used by the Google Web search engine. PageRank assigns a numerical weight to each element of a hyperlinked set of documents, such as the ones found on the World Wide Web, with the purpose of measuring its relative importance within a given collection.

It is believed that PageRank has been influenced by citation analysis, where citations in scholarly works are examined to discover relationships among researchers and their research topics. The applications of citation analysis range from identification of prominent experts in each field of study to providing invaluable information for a transparent review of academic achievements, which can be used for merit review, tenure, and promotion decisions.
The PageRank algorithm aims to do the same thing: identifying reputable/important/valuable documents/pages that are highly regarded by other documents/pages.

How Does PageRank Work?

Computationally speaking, PageRank extends the citation analysis idea by not counting links from all pages equally and by normalizing by the number of links on a page. PageRank is defined as follows:

Assume page A has pages P_1 through P_n pointing to it (with hyperlinks, which is like citations in citation analysis). The parameter d is a damping/smoothing factor that can assume values between 0 and 1. Also (A) is defined as the number of links going out of page A.
The simple formula for PageRank for page A can be written as follows:

$$PageRank(A)=(1-d)+d\sum_{i=1}^{n}\frac{PageRank(Pi)}{C(Pi)}$$

Note that the Page Ranks form a probability distribution over Web pages, so the sum of all web pages' Page Ranks will be 1. PageRank(A) can be calculated using a simple iterative algorithm and corresponds to the principal eigenvector of the normalized link matrix of the Web. The algorithm is so computationally efficient that a PageRank for 26 million Web pages can be computed in a few hours on a medium-size workstation (Brin and Page, 2012). Of course, there are more details to the actual calculation of PageRank in Google. Most of those details are either not publicly available or are beyond the scope of this simple explanation.

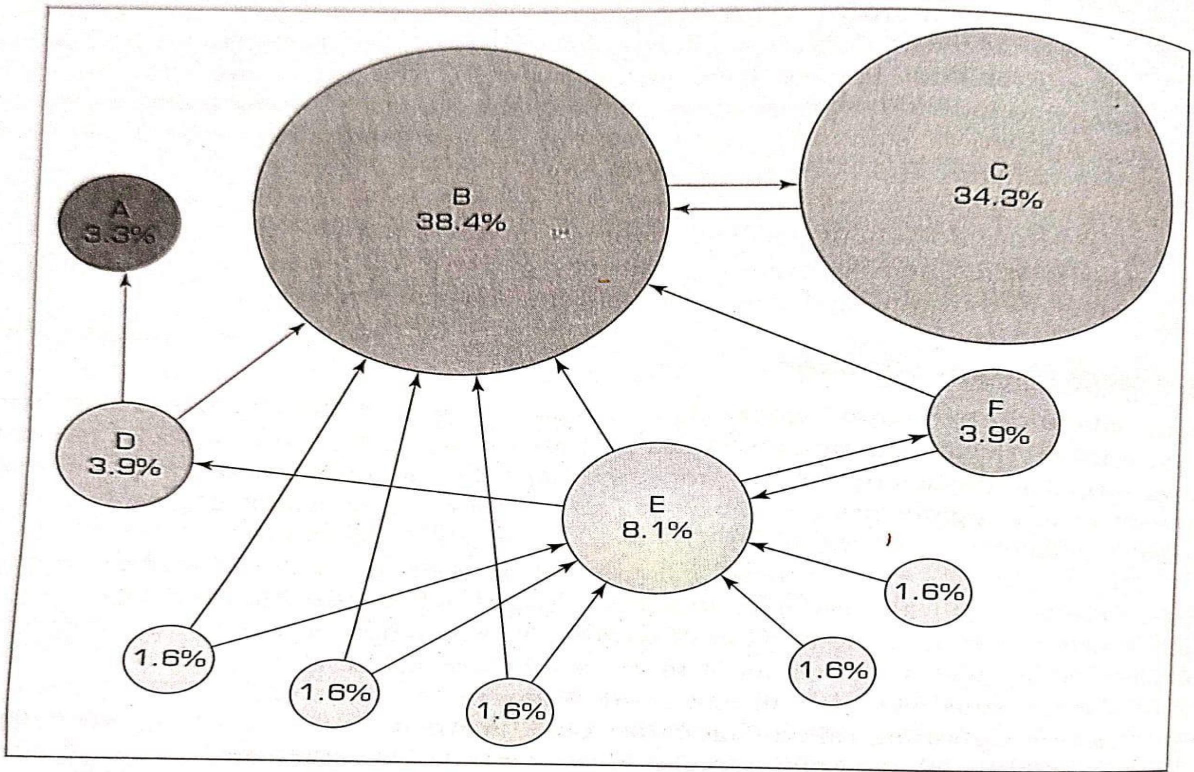


FIGURE 8.3 A Graphical Example for the PageRank Algorithm.

Justification of the Formulation

PageRank can be thought of as a model of user behavior. It assumes there is a random surfer who is given a Web page at random and keeps clicking on hyperlinks, never hitting back but eventually getting bored and starting on another random page. The probability that the random surfer visits a page is its PageRank. And the d damping factor is the probability at each page the random surfer will get bored and request another random page. One important variation is to only add the damping factor d to a single page, or a group of pages. This allows for personalization and can make it nearly impossible to deliberately mislead the system to get a higher ranking.

Another intuitive justification is that a page can have a high PageRank if there are many pages that point to it, or if there are some pages that point to it and have a high PageRank. Intuitively, pages that are well cited from many places around the Web are worth looking at. Also, pages that have perhaps only one citation from something like the Yahoo! homepage are also generally worth looking at. If a page was not high quality, or was a broken link, it is quite likely that Yahoo!'s homepage would not link to it. The formulation of PageRank handles both cases and everything in between by recursively propagating weights through the link structure of the Web.

PageRank is an innovative way to rank documents/pages, it is an augmentation to the process of retrieving relevant documents from the database and ranking them based on the weights of the words/terms. Google does all of these collectively and more to come up with the most relevant list of documents/pages for a given search request. Once an ordered list of documents/pages is created, it is pushed back to the user in an easily digestible format. At this point, users may choose to click on any of the documents in the list, and it may not be the one at the top.

If they click on a document/page link that is not at the top of the list, then can we assume that the search engine did not do a good job ranking them? Perhaps, yes. Leading search engines like Google monitor the performance of their search results by capturing, recording, and analyzing postdelivery user actions and experiences. These analyses often lead to more rules to further refine the ranking of the documents/pages so that the links at the top are more preferable to the end users.

How Does Google Do It?

Even though complex low-level computational details are trade secrets and are not known to the public, the high-level structure of the Google search system is well-known and quite simple. From the infrastructure standpoint, the Google search system runs on a distributed network of tens of thousands of computers/servers and can, therefore, carry out its heavy workload effectively and efficiently using sophisticated parallel processing algorithms (a method of computation in which many calculations can be distributed to many servers and performed simultaneously, significantly speeding up data processing).At the highest level, the Google search system has three distinct parts (googleguide.com):

1. Googlebot, a Web crawler that roams the Internet to find and fetch Web pages
2. The indexer, which sorts every word on every page and stores the resulting index of words in a huge database
3. The query processor, which compares your search query to the index and recommends the documents that it considers most relevant.

1. Googlebot

Googlebot is Google's Web crawling robot, which finds and retrieves pages on the Web and hands them off to Google indexer. It's easy to imagine Googlebot as a little spider scurrying across the strands of cyberspace, but Googlebot doesn't traverse the Web at all. It functions, much like your Web browser, by sending a request to a Web server for a Web page, downloading the entire page, and then handing it off to Google's indexer. Googlebot consists of many computers requesting and fetching pages much more quickly than you can with your Web browser. In fact, Googlebot can request thousands of different pages simultaneously. To avoid overwhelming Web servers, or crowding out requests from human users, Googlebot deliberately makes requests of each individual Web server more slowly than it's capable of doing.

When Googlebot fetches a page, it removes all the links appearing on the page and adds them to a queue for subsequent crawling. Googlebot tends to encounter little spam because most Web authors link only to what they believe are high-quality pages. By harvesting links from every page it encounters, Googlebot can quickly build a list of links that can cover broad reaches of the Web. This technique, known as deep crawling, also allows Googlebot to probe deep within individual sites. Because of their massive scale, deep crawls can reach almost every page on the Web. To keep the index current, Google continuously recrawls popular frequently changing Web pages at a rate roughly proportional to how often the pages change. Such crawls keep an index current and are known as fresh crawls. Newspaper pages are downloaded daily; pages with stock quotes are downloaded much more frequently. Of course, fresh crawls return fewer pages than the deep crawl. The combination of the two types of crawls allows Google to both make efficient use of its resources and keep its index reasonably current.

2. Google Indexer

Google Indexer Googlebot gives the indexer the full text of the pages it finds. These pages are stored in Google's index database. This index is sorted alphabetically by search term, with each index entry storing a list of documents in which the term appears and the location within the text where it occurs. This data structure allows rapid access to documents that contain user query terms. To improve search performance, Google ignores common words, called stop words (such as the, is, on, or, of, a, an, as well as certain single digits and single letters). Stop words are so common that they do little to narrow a search, and therefore they can safely be discarded. The indexer also ignores some punctuation and multiple spaces, as well as converting all letters to lowercase, to improve Google's performance.

3. Google Query Processor

The query processor has several parts, including the user interface (search box), the "engine" that evaluates queries and matches them to relevant documents, and the results formatter**.**

Google uses a proprietary algorithm, called PageRank, to calculate the relative rank order of a given collection of Web pages. PageRank is Google's system for ranking Web pages. A page with a higher PageRank is deemed more important and is more likely to be listed above a page with a lower PageRank. Google considers over a hundred factors in computing a PageRank and determining which documents are most relevant to a query, including the popularity of the page, the position and size of the search terms within the page, and the proximity of the search terms to one another on the page.

Google also applies machine-learning techniques to improve its performance automatically by learning relationships and associations within the stored data. For example, the spelling-correcting system uses such techniques to figure out likely alternative spellings. Google closely guards the formulas it uses to calculate relevance; they're tweaked to improve quality and performance, and to outwit the latest devious techniques used by spammers.

Indexing the full text of the Web allows Google to go beyond simply matching single search terms. Google gives more priority to pages that have search terms near each other and in the same order as the query. Google can also match multi-word phrases and sentences. Because Google indexes HTML code in addition to the text on the page, users can restrict searches based on where query words appear (e.g., in the title, in the URL, in the body, and in links to the page, options offered by Google's Advanced Search Form and Using Search Operators).

Understanding the internals of popular search engines helps companies, who rely on search engine traffic, better design their e-commerce sites to improve their chances of getting indexed and highly ranked by search providers

5. SEARCH ENGINE OPTIMIZATION

Search engine optimization (SEO) is the intentional activity of affecting the visibility of an e-commerce site or a Web site in a search engine's natural (unpaid or organic) search results. In general, the higher ranked on the search results page, and more frequently a site appears in the search results list, the more visitors it will receive from the search engine's users. As an Internet marketing strategy, SEO considers how search engines work, what people search for, the actual search terms or keywords typed into search engines, and which search engines are preferred by their target audience. Optimizing a Web site may involve editing its content, HTML, and associated coding to both increase its relevance to specific keywords and to remove barriers to the indexing activities of search engines. Promoting a site to increase the number of backlinks, or inbound links, is another SEO tactic.

In the early days, in order to be indexed, all Webmasters needed to do was to submit the address of a page, or URL, to the various engines, which would then send a “spider” to “crawl” that page, extract links to other pages from it, and return information found on the page to the server for indexing. The process, as explained before, involves a search engine spider downloading a page and storing it on the search engine’s own server, where a second program, known as an indexer, extracts various information about the page, such as the words it contains and where these are located, as well as any weight for specific words, and all links the page contains, which are then placed into a scheduler for crawling at a later date. Nowadays search engines are no longer relying on Webmasters submitting URLs (even though they still can); instead, they are proactively and continuously crawling the Web, and finding, fetching, and indexing everything about it.

Being indexed by search engines like Google, Bing, and Yahoo! is not good enough for businesses. Getting ranked on the most widely used search engines (see Technology Insights 8.2 for a list of most widely used search engines) and getting ranked higher than your competitors are what make the difference. A variety of methods can increase the ranking of a Web page within the search results. Cross-linking between pages of the same Web site to provide more links to the most important pages may improve its visibility. Writing content that includes frequently searched keyword phrases, to be relevant to a wide variety of search queries, will tend to increase traffic. Updating content to keep search engines crawling back frequently can give additional weight to a site. Adding relevant keywords to a Web page’s metadata, including the title tag and meta description, will tend to improve the relevancy of a site’s search listings, thus increasing traffic. URL normalization of Web pages so that they are accessible via multiple URLs and using canonical link element and redirects can help make sure links to different versions of the URL all count toward the page’s link popularity score.

Methods for Search Engine Optimization

In general, SEO techniques can be classified into two broad categories: techniques that search engines recommend as part of good site design, and those techniques of which search engines do not approve. The search engines attempt to minimize the effect of the latter, which is often called spamdexing (also known as search spam, search engine spam, or search engine poisoning). Industry commentators have classified these methods, and the practitioners who employ them, as either white-hat SEO or black-hat SEO (Goodman, 2005). White hats tend to produce results that last a long time, whereas black hats anticipate that their sites may eventually be banned either temporarily or permanently once the search engines discover what they are doing.

Top 15 Most Popular Search Engines (March 2013)

Rank	Name	Estimated Unique Monthly Visitors
1	Google	900,000,000
2	Bing	165,000,000
3	Yahoo! Search	160,000,000
4	Ask	125,000,000
5	AOL Search	33,000,000
6	MyWebSearch	19,000,000
7	blekko	9,000,000
8	Lycos	4,300,000
9	Dogpile	2,900,000
10	WebCrawler	2,700,000
11	Info	2,600,000
12	Infospace	2,000,000
13	Search	1,450,000
14	Excite	1,150,000
15	GoodSearch	1,000,000

An SEO technique is considered white hat if it conforms to the search engines' guidelines and involves no deception. Because search engine guidelines are not written as a series of rules or commandments, this is an important distinction to note. White-hat SEO is not just about following guidelines but about ensuring that the content of search engine indexes and subsequent ranks is the same content a user will see. White-hat advice is generally summed up as creating content for users, not for search engines, and then making that content easily accessible to the spiders, rather than attempting to trick the algorithm from its intended purpose.

White-hat SEO is in many ways like Web development that promotes accessibility, although the two are not identical.
Black-hat SEO attempts to improve rankings in ways that are disapproved by the search engines or involve deception. One black-hat technique uses text that is hidden, either as text colored like the background, in an invisible div, or positioned off-screen. Another method gives a different page depending on whether the page is being requested by a human visitor or a search engine, a technique known as cloaking. Search engines may penalize sites they discover using black-hat methods, either by reducing their rankings or eliminating their listings from their databases altogether. Such penalties can be applied either automatically by the search engines' algorithms, or by a manual site review. One example was the February 2006 Google removal of both BMW Germany and Ricoh Germany for use of unapproved practices (Cutts, 2006). Both companies, however, quickly apologized, fixed their practices, and were restored to Google’s list.

For some businesses SEO may generate significant return on investment. However, one should keep in mind that search engines are not paid for organic search traffic, their algorithms change constantly, and there are no guarantees of continued referrals. Due to this lack of certainty and stability, a business that relies heavily on search engine traffic can suffer major losses if the search engine decides to change its algorithms and stop sending visitors. According to Google's CEO, Eric Schmidt, in 2010, Google made over 500 algorithm changes—almost 1.5 per day. Because of the difficulty in keeping up with changing search engine rules, companies that rely on search traffic practice one or more of the following: (1) Hire a company that specializes in search engine optimization (there seem to be an abundant number of those nowadays) to continuously improve your site's appeal to changing practices of the search engines;(2) pay the search engine providers to be listed on the paid sponsors sections; and (3) consider liberating yourself from dependence on search engine traffic.

Either originating from a search engine (organically or otherwise) or coming from other sites and places, what is most important for an e-commerce site is to maximize the likelihood of customer transactions.

6. WEB USAGE MINING (WEB ANALYTICS)

Web usage mining (also called **Web analytics**) is the extraction of useful information from data generated through Web page visits and transactions. Masand et al. (2002) states that at least three types of data are generated through Web page visits:

1. Automatically generated data stored in server access logs, referrer logs, agent logs, and client-side cookies
2. User profiles
3. Metadata, such as page attributes, content attributes, and usage data.

Analysis of the information collected by Web servers can help us better understand user behavior. Analysis of this data is often called clickstream analysis. By using data and text mining techniques, a company might be able to discern interesting patterns from the clickstreams. For example, it might learn that 60 percent of visitors who searched for “hotels in Maui” had searched earlier for “airfares to Maui.” Such information could be useful in determining where to place online advertisements. Clickstream analysis might also be useful for knowing when visitors access a site. For example, if a company knew that 70 percent of software downloads from its Web site occurred between 7 and 11 p.m., it could plan for better customer support and network bandwidth during those hours.

The process of extracting knowledge from clickstream data and how the generated knowledge is used to improve the process, improve the Web site, and, most important, increase customer value.

Web mining has a wide range of business applications. For instance, Nasraoui (2006) listed the following six most common applications:

1. Determine the lifetime value of clients.
2. Design cross-marketing strategies across products.
3. Evaluate promotional campaigns.
4. Target electronic ads and coupons at user groups based on user access patterns.
5. Predict user behavior based on previously learned rules and users' profiles.
6. Present dynamic information to users based on their interests and profiles.

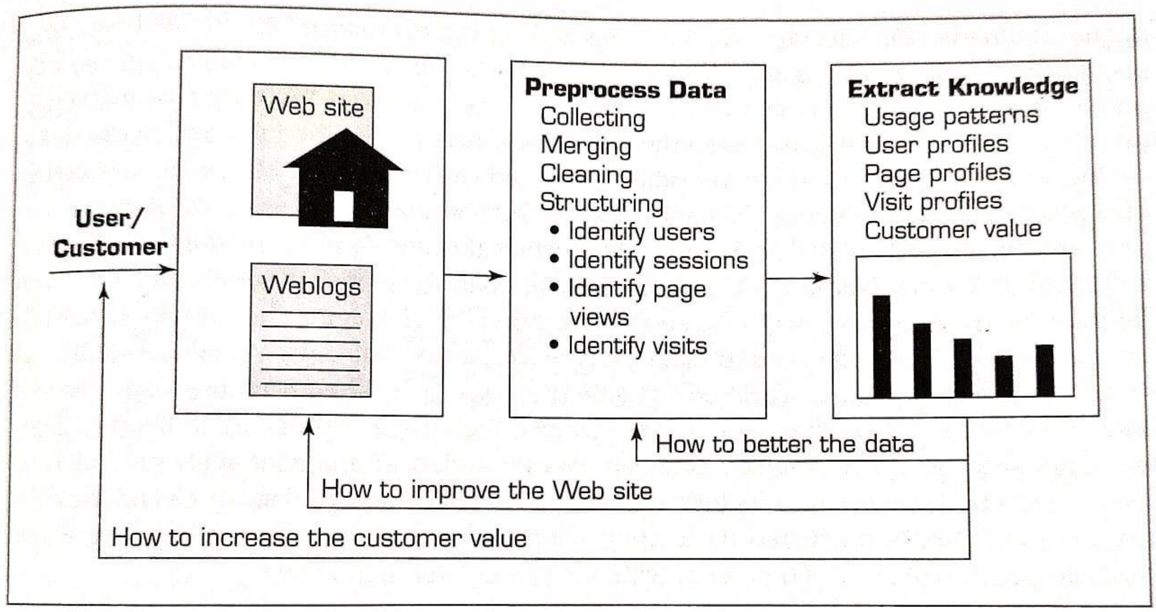


FIGURE 8.4 Extraction of Knowledge from Web Usage Data.

Amazon.com provides an excellent example of how Web usage history can be leveraged dynamically. A registered user who revisits Amazon.com is greeted by name. This is a simple task that involves recognizing the user by reading a cookie (i.e., a small text written by a Web site on the visitor's computer). Amazon.com also presents the user with a choice of products in a personalized store, based on previous purchases and an association analysis of similar users. It also makes special “Gold Box” offers that are good for a short amount of time. All these recommendations involve a detailed analysis of the visitor as well as the user's peer group developed through the use of clustering, sequence pattern discovery, association, and other data and text mining techniques.

Web Analytics Technologies

There are numerous tools and technologies for Web analytics in the marketplace. Because of their power to measure, collect, and analyze Internet data to better understand and optimize Web usage, the popularity of Web analytics tools is increasing. Web analytics hold the promise to revolutionize how business is done on the Web. Web analytics is not just a tool for measuring Web traffic; it can also be used as a tool for e-business and market research, and to assess and improve the effectiveness of an e-commerce Web site. Web analytics applications can also help companies measure the results of traditional print or broadcast advertising campaigns. It can help estimate how traffic to a Web site changes after the launch of a new advertising campaign. Web analytics provides information about the number of visitors to a Web site and the number of page views. It helps gauge traffic and popularity trends, which can be used for market research.

There are two main categories of web analytics: off-site and on-site. Off-site Web analytics refers to Web measurement and analysis about you and your products that takes place outside your Web site. It includes the measurement of a Web site's potential audience (prospect or opportunity), share of voice (visibility or word-of-mouth), and buzz (comments or opinions) that is happening on the Internet.

What is more mainstream is on-site Web analytics. Historically, Web analytics referred to on-site visitor measurement. However, in recent years this has blurred, mainly because vendors are producing tools that span both categories-site Web analytics measure a visitors' behavior once they are on your Web site. This includes its drivers and conversions, for example, the degree to which different landing pages are associated with online purchases. On-site Web analytics measure the performance of your Web site in a commercial context. response.

This data collected on the Web site is then compared against key performance indicators for performance and used to improve a Web site's or marketing campaign's audience response. Even though Google Analytics is the most widely used on-site Web analytics service, there are others provided by Yahoo! and Microsoft, and newer and better tools are emerging constantly that provide additional layers of information.

For on-site Web analytics, there are two technical ways of collecting data. The first and more traditional method is the server log file analysis, where the Web server records file requests made by browsers. The second method is page tagging, which uses JavaScript embedded in the site page code to make image requests to a third-party analytics-dedicated server whenever a page is rendered by a

Web browser (or when a mouse click occurs). Both collect data that can be processed to produce Web traffic reports. In addition to these two main streams, other data sources may also be added to augment Web site behavior data. These other sources may include e-mail, direct-mail campaign data, sales and lead history, or social media-originated data.

Web Analytics Metrics

Using a variety of data sources, Web analytics programs provide access to a lot of valuable marketing data, which can be leveraged for better insights to grow your business and better document your ROI. The insight and intelligence gained from Web analytics can be used to effectively manage the marketing efforts of an organization and its various products or services. Web analytics programs provide nearly real-time data, which can document your marketing campaign successes or empower you to make timely adjustments to your current marketing strategies.

While Web analytics provides a broad range of metrics, there are four categories of metrics that are generally actionable and can directly impact your business objectives (TWG, 2013). These categories include:

1. Web site usability: How were they using my Web site?

2. Traffic sources: Where did they come from?

3. Visitor profiles: What do my visitors look like?

4. Conversion statistics: What does all this mean for the business?

1. Web Site Usability

Beginning with your Web site, let's look at how well it works for your visitors. This is where you can learn how “user-friendly” it really is or whether you are providing the right content.

1. **Page views**. The most basic of measurements, this metric is usually presented as the “average page views per visitor.” If people come to your Web site and don’t view many pages, then your Web site may have issues with its design or structure. Another explanation for low page views is a disconnect in the marketing messages that brought them to the site and the content that is available.

2. **Time on site**. Like page views, it’s a fundamental measurement of a visitor’s interaction with your Web site. Generally, the longer a person spends on your Web site, the better it is. That could mean they’re carefully reviewing your content, utilizing interactive components you have available, and building toward an informed decision to buy, respond, or take the next step you’ve provided. On the contrary, the time on site also needs to be examined against the number of pages viewed to make sure the visitor isn’t spending his or her time trying to locate content that should be more readily accessible.

3. **Downloads.** This includes PDFs, videos, and other resources you make available to your visitors. Consider how accessible these items are as well as how well they're promoted. If your Web statistics, for example, reveal that 60 percent of the individuals who watch a demo video also make a purchase, then you'll want to strategize to increase viewership of that video.

4. **Click map.** Most analytics programs can show you the percentage of clicks each item on your Web page received. This includes clickable photos, text links in your copy, downloads, and, of course, any navigation you may have on the page. Are they clicking the most important items?

5. **Click paths.** Although an assessment of click paths is more involved, it can quickly reveal where you might be losing visitors in a specific process. A well-designed Web site uses a combination of graphics and information architecture to encourage visitors to follow "predefined" paths through your Web site. These are not rigid pathways but rather intuitive steps that align with the various processes you've built into the Web site. One process might be that of "educating" a visitor who has minimum understanding of your product or service. Another might be a process of "motivating" a returning visitor to consider an upgrade or repurchase. A third process might be structured around items you market online. You'll have as many process pathways in your Web site as you have target audiences, products, and services. Each can be measured through Web analytics to determine how effective they are.

2. Traffic Sources

Your Web analytics program is an incredible tool for identifying where your Web traffic originates. Basic categories such as search engines, referral Web sites, and visits from bookmarked pages (i.e., direct) are compiled with little involvement by the marketer. With a little effort, however, you can also identify Web traffic that was generated by your various offline or online advertising campaigns.

1. **Referral Web sites.** Other Web sites that contain links that send visitors directly to your Web site are considered referral Web sites. Your analytics program will identify each referral site your traffic comes from, and a deeper analysis will help you determine which referrals produce the greatest volume, the highest conversions, the newest visitors, etc.
2. **Search engines**. Data in the search engine category is divided between paid search and organic (or natural) search. You can review the top keywords that generated Web traffic to your site and see if they are representative of your products and services. Depending upon your business, you might want to have hundreds (or thousands) of keywords that draw potential customers. Even the simplest product search can have multiple variations based on how the individual phrases the search query.

3. **Direct.**Direct searches are attributed to two sources. An individual who bookmarks one of your Web pages in their favorites and clicks that link will be recorded as a direct search. Another source occurs when someone types your URL directly into their browser. This happens when someone retrieves your URL from a business card, brochure, print ad, radio commercial, etc. That's why it's good strategy to use coded URLs.

4. **Offline campaigns.** If you utilize advertising options other than Web-based campaigns, your Web

analytics program can capture performance data if you'll include a mechanism for sending them to your Web site. Typically, this is a dedicated URL that you include in your advertisement (i.e., "www.mycompany.com/offer50") that delivers those visitors to a specific landing page. You now have data on how many responded to that ad by visiting your Web site.

5. Online campaigns. If you are running a banner ad campaign, search engine advertising campaign, or even e-mail campaigns, you can measure individual campaign effectiveness by simply using a dedicated URL like the offline campaign strategy.

3. Visitor Profiles

One of the ways you can leverage your Web analytics into a powerful marketing tool is through segmentation. By blending data from different analytics reports, you'll begin to see a variety of user profiles emerge.

1. **Keywords.** Within your analytics report, you can see what keywords visitors used in search engines to locate your Web site. If you aggregate your keywords by similar attributes, you'll begin to see distinct visitor groups that are using your Web site. For example, the search phrase that was used can indicate how well they understand your product or its benefits. If they use words that mirror your own product or service descriptions, then they probably are already aware of your offerings from effective advertisements, brochures, etc. If the terms are more general in nature, then your visitor is seeking a solution for a problem and has happened upon your Web site. If this second group of searchers is sizable, then you'll want to ensure that your site has a strong education component to convince them they've found their answer and then move them into your sales channel.

2. **Content groupings.** Depending upon how you group your content, you may be able to analyze sections of your Web site that correspond with specific products, services, campaigns, and other marketing tactics. If you conduct a lot of trade shows and drive traffic to your Web site for specific product literature, then your Web analytics will highlight the activity in that section.

3. **Geography.**Analytics permits you to see where your traffic geographically originates, including country, state, and city locations. This can be especially useful if you see geo-targeted campaigns or want to measure your visibility across a region.

4. **Time of day.** Web traffic generally peaks at the beginning of the workday, during lunch, and toward the end of the workday. It's not unusual, however, to find strong Web traffic entering your Web site up until late evening. You can analyze this data to determine when people browse versus buy and make decisions on what hours you should offer customer service.

5. **Landing page profiles.** If you structure your various advertising campaigns properly, you can drive each of your targeted groups to a different landing page, which your Web analytics will capture and measure. By combining these numbers with the demographics of your campaign media, you can know what percentage of your visitors fit each demographic.

4. Conversion Statistics

Each organization will define a "conversion" according to its specific marketing objectives. Some Web analytics programs use the term "goal" to benchmark certain Web site objectives, whether that be a certain number of visitors to a page, a completed registration form, or an online purchase.

1. **New visitors.** If you're working to increase visibility, you'll want to study the trends in your new visitor’s data. Analytics identifies all visitors as either new or returning.

2. **Returning visitors.** If you're involved in loyalty programs or offer a product that has a long purchase cycle, then your returning visitor’s data will help you measure progress in this area.

3. **Leads.** Once a form is submitted and a thank-you page is generated, you have created a lead. Web analytics will permit you to calculate a completion rate (or abandonment rate) by dividing the number of completed forms by the number of Web visitors that came to your page. A low completion percentage would indicate a page that needs attention.

4. **Sales/conversions.** Depending upon the intent of your Web site, you can define a “sale” by an online purchase, a completed registration, an online submission, or any number of other Web activities. Monitoring these figures will alert you to any changes (or successes!) that occur further upstream.

5. **Abandonment/exit rates.** Just as important as those moving through your Web site are those who began a process and quit or came to your Web site and left after a page or two. In the first case, you’ll want to analyze where the visitor terminated the process and whether there are a few visitors quitting at the same place. Then investigate the situation for resolution. In the latter case, a high exit rate on a Web site or a specific page generally indicates an issue with expectations. Visitors click to your Web site based on some message contained in an advertisement, a presentation, etc., and expect some continuity in that message. Make sure you’re advertising a message that your Web site can reinforce and deliver.

Within each of these items there are metrics that can be established for your specific organization. You can create a weekly dashboard that includes specific numbers or percentages that will indicate where you’re succeeding—or highlight a marketing challenge that should be addressed. When these metrics are evaluated consistently and used in conjunction with other available marketing data, they can lead you to a highly quantified marketing program. Figure 8.6 shows a Web analytics dashboard created with freely available Google Analytics tools.

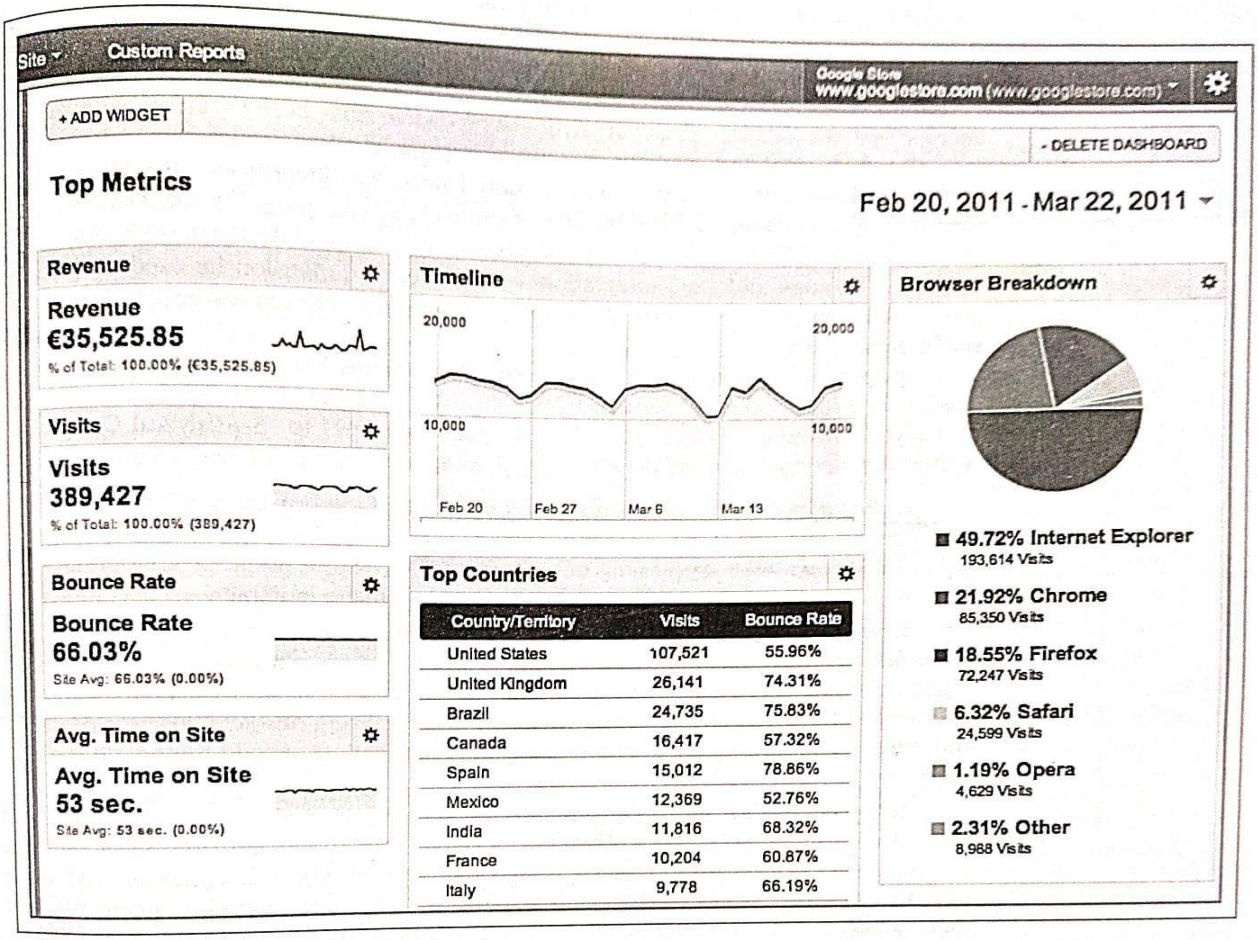


FIGURE 8.6 A Sample Web Analytics Dashboard.

7. WEB ANALYTICS MATURITY MODEL AND WEB ANALYTICS TOOLS

The term "maturity" relates to the degree of proficiency, formality, and optimization of business models, moving "ad hoc" practices to formally defined steps and optimal business processes. A maturity model is a formal depiction of critical dimensions and their competency levels of business practice. Collectively, these dimensions and levels define the maturity level of an organization in that area of practice. It often describes an evolutionary improvement path from ad hoc, immature practices to disciplined, mature processes with improved quality and efficiency.

A good example of maturity models is the BI Maturity Model developed by The Data Warehouse Institute (TDWI). In the TDWI BI Maturity Model the main purpose was to gauge where organization data warehousing initiatives are at a point in time and where it should go next. It was represented in a six-stage framework (Management Reporting → Spreadsheets → Data Marts → Data Warehouse → Enterprise Data Warehouse → BI Services).

Another related example is the simple business analytics maturity model, moving from simple descriptive measures to predicting future outcomes, to obtaining sophisticated decision systems (i.e., Descriptive Analytics → Predictive Analytics → Prescriptive Analytics).

For Web analytics perhaps the most comprehensive model was proposed by Stéphane Hamel (2009).
In this model, Hamel uses six dimensions—
(1) Management, Governance and Adoption,
(2) Objectives Definition,
(3) Scoping,
(4) The Analytics Team and Expertise,
(5) Continuous Improvement Process and Analysis Methodology,
(6) Tools, Technology and Data Integration—
and for each dimension he used six levels of proficiency/competence. Figure shows Hamel’s six dimensions and the respective proficiency levels.

The proficiency/competence levels have different terms/labels for each of the six dimensions, describing specifically what each level means. Essentially, the six levels are indications of analytical maturity ranging from "0–Analytically Impaired" to "5–Analytical Competitor."

A short description of each of the six levels of competencies is given here (Hamel, 2009)

1. **Impaired:** Characterized by the use of out-of-the-box tools and reports; limited resources lacking formal training (hands-on skills) and education (knowledge). Web analytics is used on an ad hoc basis and is of limited value and scope. Some tactical objectives are defined, but results are not well communicated and there are multiple versions of the truth.

2. **Initiated:** Works with metrics to optimize specific areas of the business (such as marketing or the e-commerce catalogue). Resources are still limited, but the process is getting streamlined. Results are communicated to various business stakeholders (often director level). However, Web analytics might be supporting obsolete business processes and, thus, be limited in the ability to push for optimization beyond the online channel. Success is mostly anecdotal.

3. **Operational:** Key performance indicators and dashboards are defined and aligned with strategic business objectives. A multidisciplinary team is in place and uses various sources of information such as competitive data, voice of customer, and social media or mobile analysis. Metrics are exploited and explored through segmentation and multivariate testing. The Internet channel is being optimized; personas are being defined. Results start to appear and be considered at the executive level. Results are centrally driven but broadly distributed.

4. **Integrated:** Analysts can now correlate online and offline data from various sources to provide a near 360-degree view of the whole value chain. Optimization encompasses complete processes, including back-end and front-end. Online activities are defined from the user perspective and persuasion scenarios are defined. A continuous improvement process and problem-solving methodologies are prevalent. Insight and recommendations reach the CXO level.

5. Competitor: This level is characterized by several attributes of companies with a strong analytical culture (Davenport and Harris, 2007):

a. One or more senior executives strongly advocate fact-based decision making and analytics
b. Widespread use of not just descriptive statistics, but predictive modelling and complex optimization techniques
c. Substantial use of analytics across multiple business functions or processes
d. Movement toward an enterprise-level approach to managing analytical tools, data, and organizational skills and capabilities.

6. Addicted: This level matches Davenport's "Analytical Competitor" characteristics: deep strategic insight, continuous improvement, integrated, skilled resources, top management commitment, fact-based culture, continuous testing, learning, and most important: far beyond the boundaries of the online channel.

In figure 8.7, one can mark the level of proficiency in each of the six dimensions to create their organization's maturity model (which would look like a spider diagram) such an assessment can help organizations better understand what dimensions they are lagging behind and take corrective actions to mitigate it.

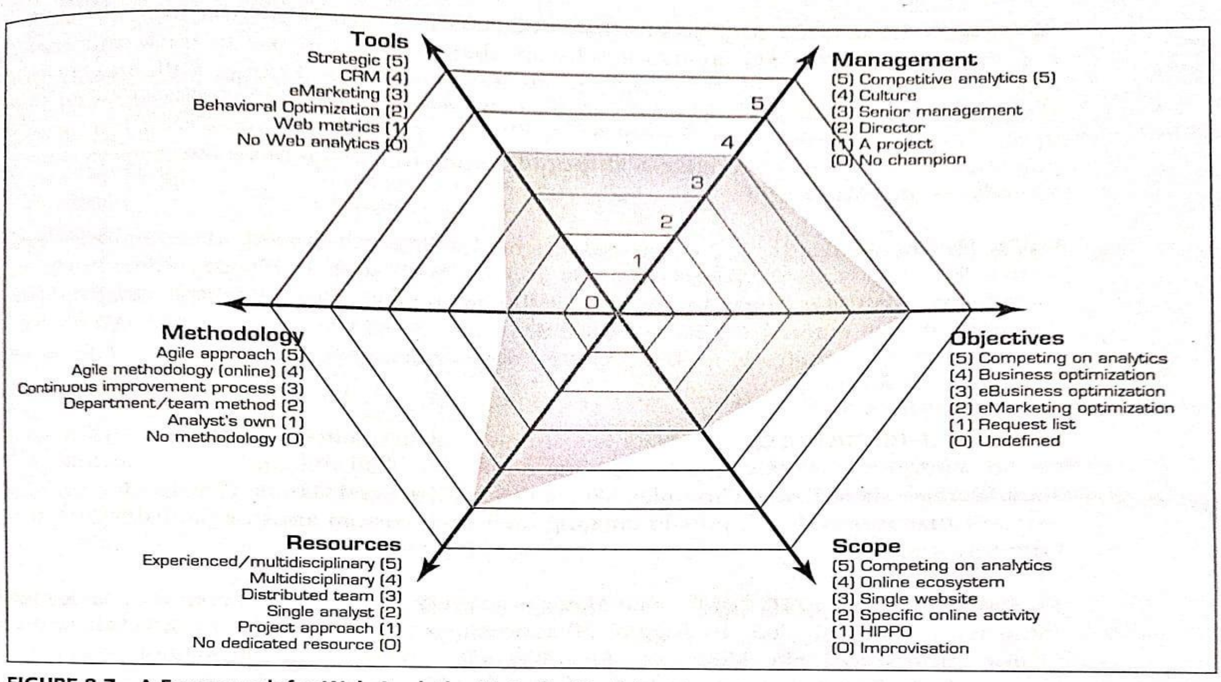


FIGURE 8.7 A Framework for Web Analytics Maturity Model.

Web Analytics Tools

There are plenty of Web analytics applications (downloadable software tools and Web-based/on-demand service platforms) in the market. Companies (large, medium, or small) are creating products and services to grab their fair share from the emerging Web analytics marketplace.What is most interesting is that many of the most popular Web analytics tools are free—yes, free to download and use for whatever reasons, commercial or nonprofit.
The following are among the most popular free (or almost free) Web analytics tools:

GOOGLE WEB ANALYTICS (GOOGLE.COM/ANALYTICS) This is a service offered by Google that generates detailed statistics about a Web site's traffic and traffic sources and measures conversions and sales. The product is aimed at marketers as opposed to the Webmasters and technologists from which the industry of Web analytics originally grew. It is the most widely used Web analytics service. Even though the basic service is free of charge, the premium version is available for a fee.

YAHOO! WEB ANALYTICS (WEB.ANALYTICS.YAHOO.COM) Yahoo! Web analytics is Yahoo!'s alternative to the dominant Google Analytics. It's an enterprise-level, robust Web-based third-party solution that makes accessing data easy, especially for multiple user groups. It's got all the things you'd expect from a comprehensive Web analytics tool, such as pretty graphs, custom-designed (and printable) reports, and real-time data tracking.

OPEN WEB ANALYTICS (OPENWEBANALYTICS.COM) Open Web Analytics (OWA) is a popular open-source Web analytics software that anyone can use to track and analyze how people use Web sites and applications. OWA is licensed under GPL and provides Web site owners and developers with easy ways to add Web analytics to their sites using simple JavaScript, PHP, or REST-based APIs. OWA also comes with built-in support for tracking Web sites made with popular content management frameworks such as WordPress and Media Wiki.

PIWIK (PIWIK.ORG) Piwik is one of the leading self-hosted, decentralized, open-source Web analytics platforms, used by 460,000 Web sites in 150 countries. Piwik was founded by Matthieu Aubry in 2007. Over the last 6 years, more talented and passionate members of the community have joined the team. As is the case in many open-source initiatives, they are actively looking for new developers, designers, datavis architects, and sponsors to join them.

FIRESTAT (FIRESTATS.CC) FireStats is a simple and straightforward Web analytics application written in PHP/MySQL. It supports numerous platforms and set-ups including C# sites, Django sites, Drupal, Joomla! WordPress, and several others. FireStats has an intuitive API that assists developers in creating their own custom apps or publishing platform components.

SITE METER (SITEMETER.COM) Site Meter is a service that provides counter and tracking information for Web sites. By logging IP addresses and using JavaScript or HTML to track visitor information, Site Meter provides Web site owners with information about their visitors, including how they reached the site, the date and time of their visit, and more.

WOOPRA (WOOPRA.COM) Woopra is a real-time customer analytics service that provides solutions for sales, service, marketing, and product teams. The platform is designed to help organizations optimize the customer life cycle by delivering live, granular behavioral data for individual Web site visitors and customers. It ties this individual-level data to aggregate analytics reports for a full life-cycle view that bridges departmental gaps.

AWSTATS (AWSTATS.ORG) AWStats is open-source Web analytics reporting tool, suitable for analyzing data from Internet services such as Web, streaming media, mail, and FTP servers. AWStats parses and analyzes server log files, producing HTML reports. Data is visually presented within reports by tables and bar graphs. Static reports can be created through a command line interface, and on-demand reporting is supported through a Web browser CGI program.

SNOOP (REINVIGORATE.NET) Snoop is a desktop-based application that runs on the Mac OS X and Windows XP/Vista platforms. It sits nicely on your system status bar/system tray, notifying you with audible sounds whenever something happens. Another outstanding Snoop feature is the Name Tags option, which allows you to "tag" visitors for easier identification. So, when Joe over at the accounting department visits your site, you'll instantly know.

MOCHIBOT (MOCHIBOT.COM) MochiBot is a free Web analytics/tracking tool especially designed for Flash assets. With MochiBot, you can see who's sharing your Flash content, how many times people view your content, as well as help you track where your Flash content is to prevent piracy and content theft. Installing MochiBot is a breeze; you simply copy a few lines of ActionScript code in the .FLA files you want to monitor.

In addition to these free Web analytics tools, Table provides a list of commercially available Web analytics tools.
Commercial Web Analytics Software Tools

s.no	Product Name	Description	URL
1	Angoss Knowledge WeberMin	Combines ANGOSS Knowledge STUDIO and clickstream analysis	angoss.com
2	ClickTracks	Visitor patterns can be shown on Web site	clicktracks.com, now at Lyris.com
3	LiveStats from DeepMetrix	Real-time log analysis, live demo on site	deepmetrix.com
4	Megaputer WebAnalyst	Data and text mining capabilities	megaputer.com/site/textanalyst.php
5	MicroStrategy Web Traffic Analysis Module	Traffic highlights, content analysis, and Web visitor analysis reports	microstrategy.com/Solutions/Applications/WTAM
6	SAS Web Analytics	Analyzes Web site traffic	sas.com/solutions/webanalytics
7	SPSS Web Mining for Clementine	Extraction of Web events	www-01.ibm.com/software/analytics/spss/
8	WebTrends	Data mining of Web traffic information	webtrends.com
9	XML Miner	A system and class library for mining data and text expressed in XML, using fuzzy logic expert system rules	scientio.com

Putting It All Together-A Web Site Optimization Ecosystem

It seems that just about everything on the Web can be measured-every click can be recorded, every view can be captured, and every visit can be analyzed all in an effort notions of "infinite measurability" and "automatic optimization" in the online channel are to continually and automatically optimize the online experience. Unfortunately, the far more complex than most realize. The assumption that any single application of Web mining techniques will provide the necessary range of insights required to understand a holistic view to Web site visitor behavior is deceptive and potentially risky. Ideally, customer experience is needed that can only be captured using both quantitative and qualitative data. Forward-thinking companies have already taken steps toward capturing and analyzing a holistic view of the customer experience, which has led to significant gains, both in terms of incremental financial growth and increasing customer loyalty and satisfaction.
According to Peterson (2008), the inputs for Web site optimization efforts can be classified along two axes describing the nature of the data and how that data can be used. On one axis are data and information-data being primarily quantitative and information being primarily qualitative. On the other axis are measures and action measures being reports, analysis, and recommendations all designed to drive actions, the actual changes being made in the ongoing process of site and marketing optimization. Each quadrant created by these dimensions leverages different technologies and creates different out-puts, but much like a biological ecosystem, each technological niche interacts with the others to support the entire online environment (see Figure 8.8).

Most believe that the Web site optimization ecosystem is defined by the ability to log, parse, and report on the clickstream behavior of site visitors. The underlying tech-nology of this ability is generally referred to as Web analytics. Although Web analytics tools provide invaluable insights, understanding visitor behavior is as much a function of qualitatively determining interests and intent as it is quantifying clicks from page to page. Fortunately, there are two other classes of applications designed to provide a more qualitative view of online visitor behavior designed to report on the overall user experi-ence and report direct feedback given by visitors and customers: customer experience management (CEM) and voice of customer (VOC):

- Web analytics applications focus on "where and when" questions by aggregating, mining, and visualizing large volumes of data, by reporting on online marketing and visitor acquisition efforts, by summarizing page-level visitor interaction data, and by summarizing visitor flow through defined multistep processes.

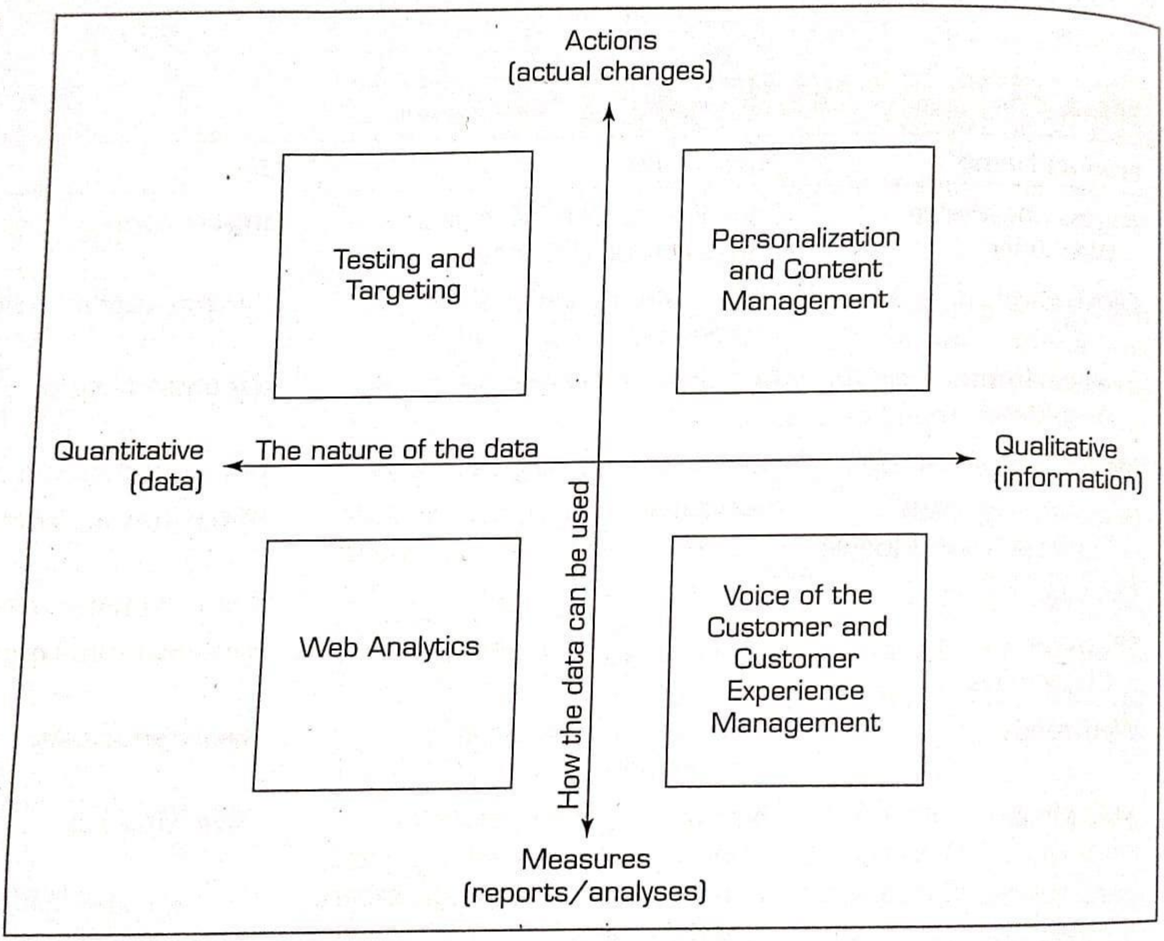


FIGURE 8.8 **Two-Dimensional View of the Inputs for Web Site Optimization.**

- Voice of customer applications focus on "who and how" questions by gathering and reporting direct feedback from site visitors, by benchmarking against other sites and offline channels, and by supporting predictive modelling of future visitor behaviour.

- Customer experience management applications focus on "what and why" questions by detecting Web application issues and problems, by tracking and resolving business process and usability obstacles, by reporting on site performance and availability, by enabling real-time alerting and monitoring, and by supporting deep diagnosis of observed visitor behavior.

All three applications are needed to have a complete view of the visitor behavior where each application plays a distinct and valuable role. Web analytics, CEM, and VOC applications form the foundation of the Web site optimization ecosystem that supports the online business's ability to positively influence desired outcomes (a pictorial representation of this process view of the Web site optimization ecosystem is given in Figure 8.9). These similar-yet-distinct applications each contribute to a site operator's ability to recognize, react, and respond to the ongoing challenges faced by every Web site owner. Fundamental to the optimization process is measurement, gathering data and information that can then be transformed into tangible analysis, and recommendations for improvement using Web mining tools and techniques. When used properly, these applications allow for convergent validation-combining different sets of data collected for the same audience to provide a richer and deeper understanding of audience behavior. The convergent validation model-one

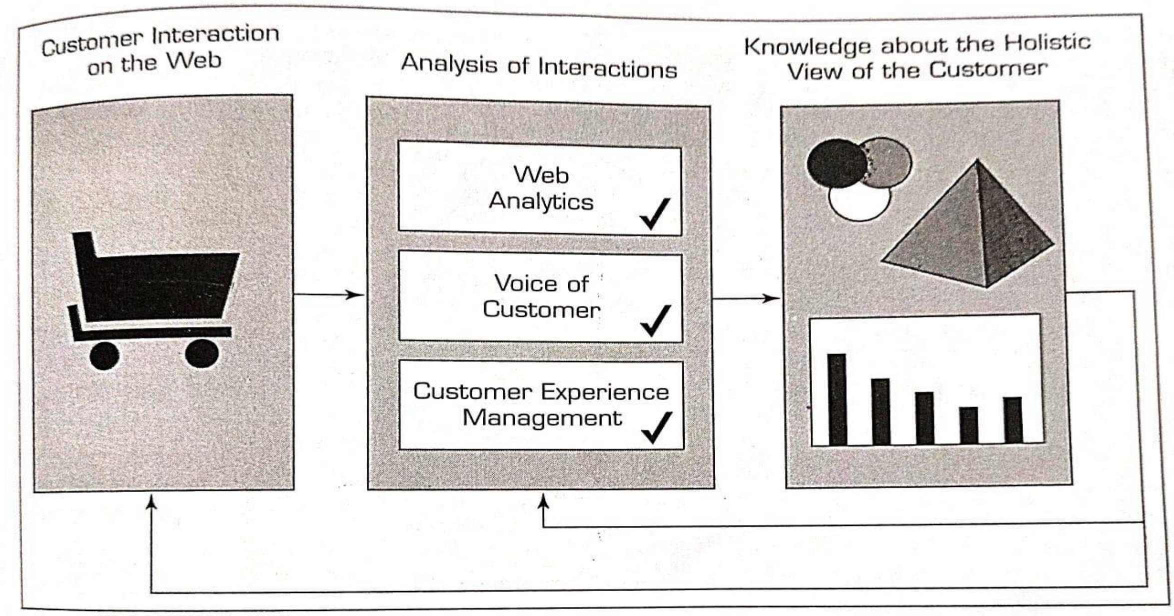


FIGURE 8.9 **A Process View of the Web Site Optimization Ecosystem.**

where multiple sources of data describing the same population are integrated to increase the depth and richness of the resulting analysis—forms the framework of the Web site optimization ecosystem. On one side of the spectrum are the primarily qualitative inputs from VOC applications; on the other side are the primarily quantitative inputs from CEM bridging the gap by supporting key elements of data discovery. When properly implemented, all three-systems sample data from the same audience. The combination of these data, either through data integration projects or simply via the process of conducting good analysis, supports far more actionable insights than any of the ecosystem members individually.

A Framework for Voice of the Customer Strategy

Voice of the customer (VOC) is a term usually used to describe the analytic process of capturing a customer's expectations, preferences, and aversions. It essentially is a market research technique that produces a detailed set of customer wants and needs, organized into a hierarchical structure, and then prioritized in terms of relative importance and satisfaction with current alternatives. Attensity, one of the innovative service providers in the analytics marketplace, developed an intuitive framework for VOC strategy that they called LARA, which stands for Listen, Analyze, Relate, and Act. It is a methodology that outlines a process by which organizations can take user-generated content (UGC), whether generated by consumers talking in Web forums, on micro-blogging sites like Twitter and social networks like Facebook, or in feedback surveys, e-mails, documents, research, etc., and using it as a business asset in a business process. Figure 8.10 shows a pictorial depiction of this framework.

LISTEN To “listen” is a process in itself that encompasses both the capability to listen to the open Web (forums, blogs, tweets, you name it) and the capability to seamlessly access enterprise information (CRM notes, documents, e-mails, etc.). It takes a listening post, deep federated search capabilities, scraping and enterprise class data integration, and a strategy to determine who and what you want to listen to.

ANALYZE This is the hard part. How can you take all this mass of unstructured data and make sense of it? This is where the “secret sauce” of text analytics comes into play. Look for solutions that include keyword, statistical, and natural language approaches that will allow you to essentially tag or barcode every word and the relationships among words, making it data that can be accessed, searched, routed, counted, analysed, charted, reported on, and even reused. Keep in mind that, in addition to technical capabilities, it has to be easy to use, so that your business users can focus on the insights, not the technology. It should have an engine that doesn't require the user to define keywords or terms that they want their system to look for or include in a rule base. Rather, it should automatically identify terms ("facts," people, places, things, etc.) and their relationships with other terms or combinations of terms making it easy to use, maintain, and be more accurate, so you can rely on the insights as actionable.

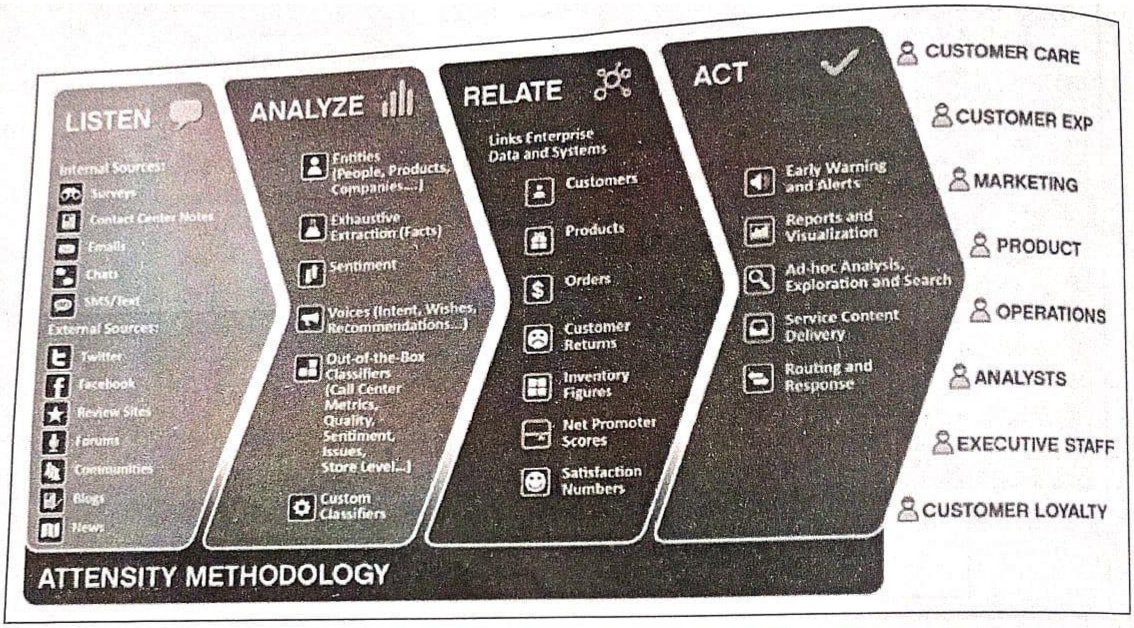


FIGURE 8.10 Voice of the Customer Strategy Framework. *Source:* **Attensity.com.** Used with permission.

RELATE Now that you have found the insights and can analyze the unstructured data, the real value comes when you can connect those insights to your "structured" data: your customers (which customer segment is complaining about your product most?); your products (which product is having the issue?); your parts (is there a problem with a specific part manufactured by a specific partner?); your locations (is the customer who is tweeting about wanting a sandwich near your nearest restaurant?); and so on. Now you can ask questions of your data and get deep, actionable insights.

ACT Here is where it gets exciting, and your business strategy and rules are critical. What do you do with the new customer insight you've obtained? How do you leverage the problem resolution content created by a customer that you just identified? How do you connect with a customer who is uncovering issues that are important to your business or who is asking for help? How do you route the insights to the right people? And how do you engage with customers, partners, and influencers once you understand what they are saying? You understand it; now you've got to act.

Short Questions:

1. Define Web Analytics.
2. List of two key benefits of Web Analytics for businesses.
3. What was the primary goal of Security First Insurance in applying Web Mining?
4. Define Web Mining.
5. List the three categories of Web Mining.
6. What is Web Content Mining?
7. Define a search engine.
8. What is the difference between a crawler-based and human-powered search engine?
9. What is Search Engine Optimization (SEO)
10. Define Web Usage Mining.
11. What is the Web Analytics Maturity Model (WAMM)?
12. What is the main function of Google Analytics?

Essay Questions:

1. Explain the Web Analytics process with suitable examples.
2. Discuss how Security First Insurance used Web Mining to improve customer connection.
3. Compare Web Content Mining, Web Structure Mining, and Web Usage Mining with examples.
4. Evaluate the importance of Web Mining in digital marketing strategies.
5. Analyze the role of Web Content Mining in search engines.
6. Explain the architecture of a search engine with a neat diagram
7. Differentiate between Black Hat and White Hat SEO techniques with examples.
8. Discuss the applications of Web Usage Mining in recommendation systems.
9. Evaluate the ethical issues related to Web Usage Mining.
10. Explain the stages of the Web Analytics Maturity Model with examples.

…………………………………. END……………………………………

M. PRAVARSHA REDDY

WEB AND SOCIAL MEDIA ANALYTICS

UNIT-V

Topics:

a) Social Analytics and Social Network Analysis:

1. **Social Analytics**
2. **Social Network Analysis**
3. **Social Media Definitions and Concepts**
4. **Social Media Analytics**

b) Prescriptive Analytics - Optimization and Multi-Criteria Systems:

1. **Multiple Goals**
2. **Sensitivity Analysis**
3. **What-If Analysis**
4. **Goal Seeking.**

a) Social Analytics and Social Network Analysis:

1. Social Analytics

Social analytics may mean different things to different people, based on their worldview and field of study. For instance, the dictionary definition of social analytics refers to a philosophical perspective developed by the Danish historian and philosopher Lars-Henrik Schmidt in the 1980s. The theoretical object of the perspective is *socius*, a kind of "commonness" that is neither a universal account nor a communality shared by every member of a body (Schmidt, 1996). Thus, social analytics differs from traditional philosophy as well as sociology. It might be viewed as a perspective that attempts to articulate the contentions between philosophy and sociology.
Our definition of social analytics is somewhat different; as opposed to focusing on the "social" part (as is the case in its philosophical definition), we are more interested in the "analytics" part of the term. Gartner defined social analytics as "monitoring, analysing, measuring and interpreting digital interactions and relationships of people, topics, ideas and content." Social analytics include mining the textual content created in social media (e.g., sentiment analysis, natural language processing) and analysing socially established networks (e.g., influencer identification, profiling, prediction) for the purpose of gaining insight about existing and potential customers' current and future behaviours, and about the likes and dislikes toward a firm's products and services. Based on this definition and the current practices. Social analytics can be classified into two different, but not necessary mutually exclusive, branches: social network analysis and social media analytics.

2. Social Network Analysis

A **social network** is a social structure composed of individuals/people (or groups of individuals or organizations) linked to one another with some type of connections/relationships. The social network perspective provides a holistic approach to analysing structure and dynamics of social entities. The study of these structures uses social network analysis to identify local and global patterns, locate influential entities, and examine network dynamics. Social networks and the analysis of them is essentially an interdisciplinary field that emerged from social psychology, sociology, statistics, and graph theory. Development and formalization of the mathematical extent of social network analysis dates to the 1950s; the development of foundational theories and methods of social networks dates back to the 1980s (Scott and Davis, 2003). Social network analysis is now one of the major paradigms in business analytics, consumer intelligence, and contemporary sociology, and is also employed in several other social and formal sciences.
A social network is a theoretical construct useful in the social sciences to study relationships between individuals, groups, organizations, or even entire societies (social units). The term is used to describe a social structure determined by such interactions. The ties through which any given social unit connects represent the convergence of the various social contacts of that unit. In general, social networks are self-organizing, emergent, and complex, such that a globally coherent pattern appears from the local interaction of the elements (individuals and groups of individuals) that make up the system.

Following are a few typical social network types that are relevant to business activities.

COMMUNICATION NETWORKS

Communication studies are often considered a part of both the social sciences and the humanities, drawing heavily on fields such as sociology, psychology, anthropology, information science, biology, political science, and economics.

Many communications concepts describe the transfer of information from one source to another and thus can be represented as a social network. Telecommunication companies are tapping into this rich information source to optimize their business practices and to improve customer relationships.

COMMUNITY NETWORKS

Traditionally, community referred to a specific geographic location, and studies of community ties had to do with who talked, associated, traded, and attended social activities with whom. Today, however, there are extended “online” communities developed through social networking tools and telecommunications devices. Such tools and devices continuously generate large amounts of data, which can be used by companies to discover invaluable, actionable information.

CRIMINAL NETWORKS

In criminology and urban sociology, much attention has been paid to the social networks among criminal actors. For example, studying gang murders and other illegal activities as a series of exchanges between gangs can lead to better understanding and prevention of such criminal activities. Now that we live in a highly connected world (thanks to the Internet), many of the criminal networks’ formations and their activities are being watched/pursued by security agencies using state-of-the-art Internet tools and tactics. Even though the Internet has changed the landscape for criminal networks and law enforcement agencies, the traditional social and philosophical theories still mostly still apply.

INNOVATION NETWORKS

Business studies on diffusion of ideas and innovations in a network environment focus on the spread and use of ideas among the members of the social network. The idea is to understand why some networks are more innovative, and why some communities are early adopters of ideas and innovations.

Social Network Analysis Metrics

Social network analysis (SNA) is the systematic examination of social networks. Social network analysis views social relationships in terms of network theory, consisting of nodes (representing individuals or organizations within the network) and ties/connections (which represent relationships between the individuals or organizations, such as friendship, kinship, organizational position, etc.). These networks are often represented using social network diagrams, where nodes are represented as points and ties are represented as lines.

Connections

- Homophily: This refers to the extent to which actors form ties with others who are similar versus dissimilar. Similarity can be based on gender, race, age, occupation, educational achievement, status, values, or other important characteristics.
- Multiplexity: This is the number of different content-forms in a connection. For example, two people who are friends and work together have a multiplexity of 2. Higher multiplexity links are associated with stronger relationships.
- Mutuality/reciprocity: This measures how much two actors reciprocate each other's friendship or interactions.
- Network closure: This is about the completeness of relational triads. It describes the assumption that one's friends are also friends with each other, a trait known as transitivity. This helps fulfil individuals' need for cognitive closure.
- Propinquity: This is the tendency for actors to form more ties with people who are geographically close to them.

Distributions

- Bridge: A bridge is an individual whose weak ties connect two individuals or clusters, filling a "structural hole." Bridges provide the only link between these groups, and often offer the shortest route between them when longer routes are unfeasible due to message distortion or delivery failure.
- Centrality: Refers to metrics quantifying the importance or influence of a node or group in a network. Examples include betweenness centrality, closeness centrality, eigenvector centrality, alpha centrality, and degree centrality.
- Density: The proportion of direct ties in a network relative to the total number possible.
- Distance: The minimum number of ties required to connect two particular actors.
- Structural holes: The absence of ties between two parts of a network. Exploiting structural holes provides competitive advantage by connecting otherwise disconnected groups, also known as an alternate form of social capital.
- Tie strength: Defined by a combination of time, emotional intensity, intimacy, and reciprocity (or mutuality). Strong ties are associated with homophily, propinquity, and transitivity, while weak ties are often bridges.

Segmentation Metrics

- Cliques and social circles: Cliques are groups where every individual is directly tied to every other; social circles involve less direct cohesion or less stringent direct contact.
- Clustering coefficient: Measures the likelihood that two associates of a node are also associates; a higher coefficient implies greater cliquishness or network transitivity.
- Cohesion: The degree to which actors are connected by cohesive bonds; structural cohesion refers to the minimum number of members who, if removed, would disconnect the group.

3. SOCIAL MEDIA DEFINITIONS AND CONCEPTS

Social media refers to the enabling technologies of social interactions among people in which they create, share, and exchange information, ideas, and opinions in virtual communities and networks. It is a group of Internet-based software applications that build on the ideological and technological foundations of Web 2.0, and that allow the creation and exchange of user-generated content (Kaplan and Haenlein, 2010). Social media depends on mobile and other Web-based technologies to create highly interactive platforms for individuals and communities to share, co-create, discuss, and modify user-generated content. It introduces substantial changes to communication between organizations, communities, and individuals.

Since their emergence in the early 1990s, Web-based social media technologies have seen a significant improvement in both quality and quantity. These technologies take on many different forms, including online magazines, Internet forums, Web logs, social blogs, microblogging, wikis, social networks, podcasts, pictures, video, and product/service evaluations/ratings. By applying a set of theories in the field of media research (social presence, media richness) and social processes (self-presentation, self-disclosure), Kaplan and Haenlein (2010) created a classification scheme with six different types of social media: collaborative projects (e.g., Wikipedia), blogs and microblogs (e.g., Twitter), content communities (e.g., YouTube), social networking sites (e.g., Facebook), virtual game worlds (e.g., World of Warcraft), and virtual social worlds (e.g., Second Life).

Web-based social media are different from traditional/industrial media, such as newspapers, television, and film, as they are comparatively inexpensive and accessible to enable anyone (even private individuals) to publish or access/consume information. Industrial media generally require significant resources to publish information, as in most cases the articles (or books) go through many revisions before being published (as was the case in the publication of this very book). Here are some of the most prevailing characteristics that help differentiate between social and industrial media (Morgan et al., 2010).

Quality: In industrial publishing—mediated by a publisher—the typical range of quality is substantially narrower than in niche, unmediated markets. The main challenge posed by content in social media sites is the fact that the distribution of quality has high variance: from very high-quality items to low-quality, sometimes abusive, content.

Reach: Both industrial and social media technologies provide scale and are capable of reaching a global audience. Industrial media, however, typically use a centralized framework for organization, production, and dissemination, whereas social media are by their very nature more decentralized, less hierarchical, and distinguished by multiple points of production and utility.

Frequency: Compared to industrial media, updating and reposting on social media platforms is easier, faster, and cheaper, and therefore practiced more frequently, resulting in fresher content.

Accessibility: The means of production for industrial media are typically government and/or corporate (privately owned), and are costly, whereas social media tools are generally available to the public at little or no cost.

Usability: Industrial media production typically requires specialized skills and training. Conversely, most social media production requires only modest reinterpretation of existing skills; in theory, anyone with access can operate the means of social media production.

Immediacy: The time lag between communications produced by industrial media can be long (weeks, months, or even years) compared to social media (which can be capable of virtually instantaneous responses).

Updatability: Industrial media, once created, cannot be altered (once a magazine article is printed and distributed, changes cannot be made to that same article), whereas social media can be altered almost instantaneously by comments or editing.

How Do People Use Social Media?

Not only are the numbers on social networking sites growing, but so is the degree to which they are engaged with the channel. Brogan and Bastone (2011) presented research results that stratify users according to how actively they use social media and tracked evolution of these user segments over time. They listed six different engagement levels (Figure 8.11). According to the research results, the online user community has been steadily migrating upwards on this engagement hierarchy. The most notable change is among Inactive. Forty-four percent of the online population fell into this category.

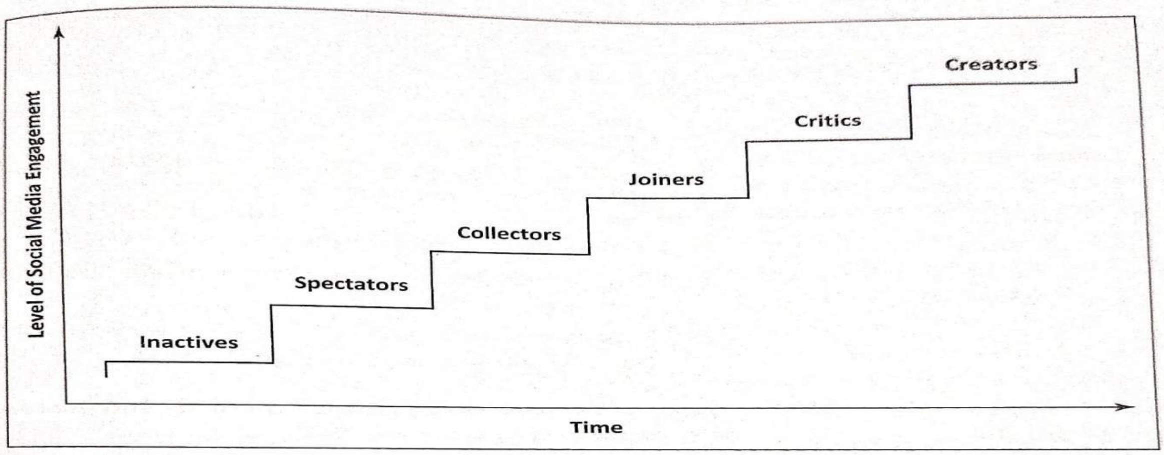


FIGURE 8.11 Evolution of Social Media User Engagement.

Two years later, more than half of those Inactive had jumped into social media in some form or another. “Now roughly 82 percent of the adult population online is in one of the upper categories,” said Bastone. “Social media has truly reached a state of mass adoption.”

4. SOCIAL MEDIA ANALYTICS

Social media analytics refers to the systematic and scientific ways to consume the vast amount of content created by Web-based social media outlets, tools, and techniques for the betterment of an organization's competitiveness. Social media analytics is rapidly becoming a new force in organizations around the world, allowing them to reach out to and understand consumers as never before. In many companies, it is becoming the tool for integrated marketing and communications strategies.

The exponential growth of social media outlets, from blogs, Facebook, and Twitter to LinkedIn and YouTube, and analytics tools that tap into these rich data sources offer organizations the chance to join a conversation with millions of customers around the globe every day. This aptitude is why nearly two-thirds of the 2,100 companies who participated in a recent survey by Harvard Business Review (HBR) Analytic Services said they are either currently using social media channels or have social media plans in the works (HBR, 2010). But many still say social media is an experiment, as they try to understand how to best use the different channels, gauge their effectiveness, and integrate social media into their strategy.

Despite the vast potential social media analytics brings, many companies seem focused on social media activity primarily as a one-way promotional channel and have yet to capitalize on the ability to not only listen to, but also analyse, consumer conversations and turn the information into insights that impact the bottom line. Here are some of the results from the HBR Analytic Services survey (HBR, 2010):

- Three-quarters (75%) of the companies in the survey said they did not know where their most valuable customers were talking about them.
- Nearly one-third (31%) do not measure effectiveness of social media.
- Less than one-quarter (23%) are using social media analytic tools.
- A fraction (7%) of participating companies can integrate social media into their marketing activities.

While still searching for best practice and measurements, two-thirds of the companies surveyed are convinced their use of social media will grow, and many anticipate investing more in it next year, even as spending in traditional media declines. So, what is it specifically that companies are interested in measuring in social media?

Measuring the Social Media Impact

For organizations, small or large, there is valuable insight hidden in all the user-generated content on social media sites. But how do you dig it out of dozens of review sites, thousands of blogs, millions of Facebook posts, and billions of tweets? Once you do that, how do you measure the impact of your efforts? These questions can be addressed by the analytics extension of the social media technologies. Once you decide on your goal for social media (what it is that you want to accomplish), there is a multitude of tools to help you get there. These analysis tools usually fall into three broad categories:

- Descriptive analytics: Uses simple statistics to identify activity characteristics and trends, such as how many followers you have, how many reviews were generated on Facebook, and which channels are being used most often.
- Social network analysis: Follows the links between friends, fans, and followers to identify connections of influence as well as the biggest sources of influence.
- Advanced analytics: Includes predictive analytics and text analytics that examine the content in online conversations to identify themes, sentiments, and connections that would not be revealed by casual surveillance.

Sophisticated tools and solutions to social media analytics use all three categories of analytics (i.e., descriptive, predictive, and prescriptive) in a somewhat progressive fashion.

Best Practices in Social Media Analytics

As an emerging tool, social media analytics is practiced by companies in a somewhat haphazard fashion. Because there are not well-established methodologies, everybody is trying to create their own by trial and error. What follows are some of the field-tested best practices for social media analytics proposed by Paine and Chaves (2012).

THINK OF MEASUREMENT AS A GUIDANCE SYSTEM, NOT A RATING SYSTEM Measurements are often used for punishment or rewards; they should not be. They should be about figuring out what the most effective tools and practices are, what needs to be discontinued because it doesn't work, and what needs to be done more because it does work very well. A good analytics system should tell you where you need to focus. Maybe all that emphasis on Facebook doesn't really matter, because that is not where your audience is. Maybe they are all on Twitter, or vice versa. According to Paine and Chaves, channel preference won't necessarily be intuitive, "We just worked with a hotel that had virtually no activity on Twitter for one brand but lots of Twitter activity for one of their higher brands." Without an accurate measurement tool, you would not know.

TRACK THE ELUSIVE SENTIMENT Customers want to take what they are hearing and learning from online conversations and act on it. The key is to be precise in extracting and tagging their intentions by measuring their sentiments. As we have seen in Chapter 7, text analytic tools can categorize online content, uncover linked concepts, and reveal the sentiment in a conversation as "positive," "negative," or "neutral," based on the words people use. Ideally, you would like to be able to attribute sentiment to a specific product, service, and business unit. The more precise you can get in understanding the tone and perception that people express, the more actionable the information becomes, because you are mitigating concerns about mixed

polarity. A mixed-polarity phrase, such as "hotel in great location but bathroom was smelly" should not be tagged as "neutral" because you have positives and negatives offsetting each other. To be actionable, these types of phrases are to be treated separately; "bathroom was smelly" is something someone can own and improve upon. One can classify and categorize these sentiments, look at trends over time, and see significant differences in the way people speak either positively or negatively about you. Furthermore, you can compare sentiment about your brand to your competitors.

CONTINUOUSLY IMPROVE THE ACCURACY OF TEXT ANALYSIS An industry-specific text analytics package will already know the vocabulary of your business. The system will have linguistic rules built into it, but it learns over time and gets better. Much as you would tune a statistical model as you get more data, better parameters, or new techniques to deliver better results, you would do the same thing with the natural language processing that goes into sentiment analysis. You set up rules, taxonomies, categorization, and meaning of words; watch what the results look like; and then go back and do it again.

LOOK AT THE RIPPLE EFFECT It is one thing to get a great hit on a high-profile site, but that's only the start. There's a difference between a great hit that just sits there and goes away versus a great hit that is tweeted, retweeted, and picked up by influential bloggers. Analysis should show you which social media activities go "viral" and which quickly go dormant—and why.

LOOK BEYOND THE BRAND One of the biggest mistakes people make is to be concerned only about their brand. To successfully analyse an act on social media, you need to understand not just what is being said about your brand, but the broader conversation about the spectrum of issues surrounding your product or service, as well. Customers don't usually care about a firm's message or its brand; they care about themselves. Therefore, you should pay attention to what they are talking about, where they are talking, and where their interests are.

IDENTIFY YOUR MOST POWERFUL INFLUENCERS Organizations struggle to identify who has the most power in shaping public opinion. It turns out, your most important influencers are not necessarily the ones who advocate specifically for your brand; they are the ones who influence the whole realm of conversation about your topic. You need to understand whether they are saying nice things, expressing support, or simply making observations or critiquing. What is the nature of their conversations? How is my brand being positioned relative to the competition in that space?

LOOK CLOSELY AT THE ACCURACY OF YOUR ANALYTIC TOOL Until recently, computer-based automated tools were not as accurate as humans for sifting through online content. Even now, accuracy varies depending on the media. For product review sites, hotel review sites, and Twitter, it can reach anywhere between 80 to 90 percent accuracy, because the context is more boxed in. When you start looking at blogs and discussion forums, where the conversation is more wide-ranging, the software can deliver 60 to 70 percent accuracy (Paine and Chaves, 2012). These figures will increase over time, because the analytics tools are continually upgraded with new rules and improved algorithms to reflect field experience, new products, changing market conditions, and emerging patterns of speech.

INCORPORATE SOCIAL MEDIA INTELLIGENCE INTO PLANNING Once you have big-picture perspective and detailed insight, you can begin to incorporate this information into your planning cycle. But that is easier said than done. A quick audience poll revealed that very few people currently incorporate learning from online conversations into their planning cycles (Paine and Chaves, 2012). One way to achieve this is to find time-linked associations between social media metrics and other business activities or market events. Social media is typically either organically invoked or invoked by something your organization does; therefore, if you see a spike in activity at some point in time, you want to know what was behind that.

Social Media Analytics Tools and Vendors

Monitoring social media, identifying interesting conversations among potential customers, and inferring what they are saying about your company, your products, and services is an essential yet challenging task for many organizations.

There are two main paths that an organization can take to attain social media analytics (SMA) capabilities: in-house development or outsourcing.

Building an effective SMA system requires extensive knowledge in several related fields (e.g., Web, text mining, predictive analytics, reporting, visualization, performance management, etc.), so most organizations choose outsourcing, except for very large enterprises.
Due to the exponential growth in SMA, many start-up companies have emerged, claiming to provide practical and cost-effective solutions to organizations of all sizes.

However, not all succeeded, but many have survived and evolved to offer integrated approaches for product development, customer support, public outreach, lead generation, market research, and campaign management.
Below are brief definitions of 10 SMA tools/vendors. This is not a complete list, but 10 with which the authors have some familiarity:

- **ATTENSITY360:** Attensity360 operates on four key principles: listen, analyse, relate, and act. It helps monitor trending topics, influencers, and the reach of your brand, and recommends ways to join the conversation. Attensity Analyse applies text analytics to unstructured text to extract meaning and uncover trends, while Attensity Respond automates routing of incoming social media mentions. Clients include Whirlpool, Vodafone, Versatel, TMobile, Oracle, and Wiley.
- **RADIAN6/SALESFORCE CLOUD:** Purchased by Salesforce in 2011, Radian 6 works with brands to help them listen more intelligently to consumers, competitors, and influencers. Their dashboard tracks mention on more than 100 million social media sites and offers an engagement console to coordinate responses across blogs, Twitter, and Facebook accounts. Clients include Red Cross, Adobe, AAA, Cirque du Soleil, H&R Block, March of Dimes, Microsoft, Pepsi, and Southwest Airlines.
- **SYSOMOS:** Sysomos manages real-time conversations with the Heartbeat tool, providing constantly updated snapshots of social media conversations, workflow management, collaboration, and influencer engagement. Clients include IBM, HSBC, Roche, Ketchum, Sony Ericsson, Philips, ConAgra, Edelman, Shell Oil, Nokia, Sapient, Citi, and Interbrand. Owner: Marketwire.
- **COLLECTIVE INTELLECT:** Boulder, Colorado–based, started out by providing monitoring to financial firms, and has evolved into a top-tier player in social media intelligence gathering.

WEBTRENDS Webtrends offers services geared toward monitoring, measuring, analyzing, profiling, and targeting audiences for a brand. The partner-based platform allows for crowd-sourced improvements and problem solving, creating transparency for their products and services. Their clients include CBS, NBC Universal, 20th Century Fox, AOL, Electronic Arts, Lifetime, and Nestle.
CRIMSON HEXAGON Cambridge, Massachusetts-based Crimson Hexagon taps into billions of conversations taking place in online media and turns them into actionable data for better brand understanding and improvement. Based on a technology licensed from Harvard, its VoxTrot Opinion can analyze vast amounts of qualitative information and determine quantitative proportion of opinion. Their clients include CNN, Hanes, AT&T, HP, Johnson & Johnson, Mashable, Microsoft, Monster, Thomson Reuters, Rubbermaid, Sybase, and The Wall Street Journal
CONVERSEON, New York-based social media consulting firm Converseon, named a leader in the social media monitoring sector by Forrester Research, builds tailored dashboards for its enterprise installations and offers professional services around every step of the social business intelligence process. Converseon starts with technology and adds human analysis, resulting in high-quality data

and impressive functionality. Their clients include Dow, Amway, Graco, and other major brands.
SPIRAL16 Spiral16 takes an in-depth look at who is saying what a brand and compares results with those of top competitors. The goal is to help you monitor the effectiveness of your social media strategy, understand the sentiment behind conversations online, and mine large amounts of data. It uses impressive 3D displays and a standard dashboard. Their clients include Toyota, Lee, and Cadbury.
BUZZLOGIC BuzzLogic uses its technology platform to identify and organize the conversation universe, combining both conversation topics and audience to help brands reach audiences who are passionate on everything from the latest tech craze and cloud computing to parenthood and politics. Their clients include Starbucks, American Express, HBO, and HP.
SPROUTSOCIAL Founded in 2010 in Chicago, Illinois, SproutSocial is an innovative social media analytics company that provides social analytics services to many well-known firms and organizations. Their clients include Yahoo!, Nokia, Pepsi, St. Jude Children's Research Center, Hyatt Regency, McDonalds, and AMD. A sample screen shot of their social media solution dashboard is shown in Figure 8.12.

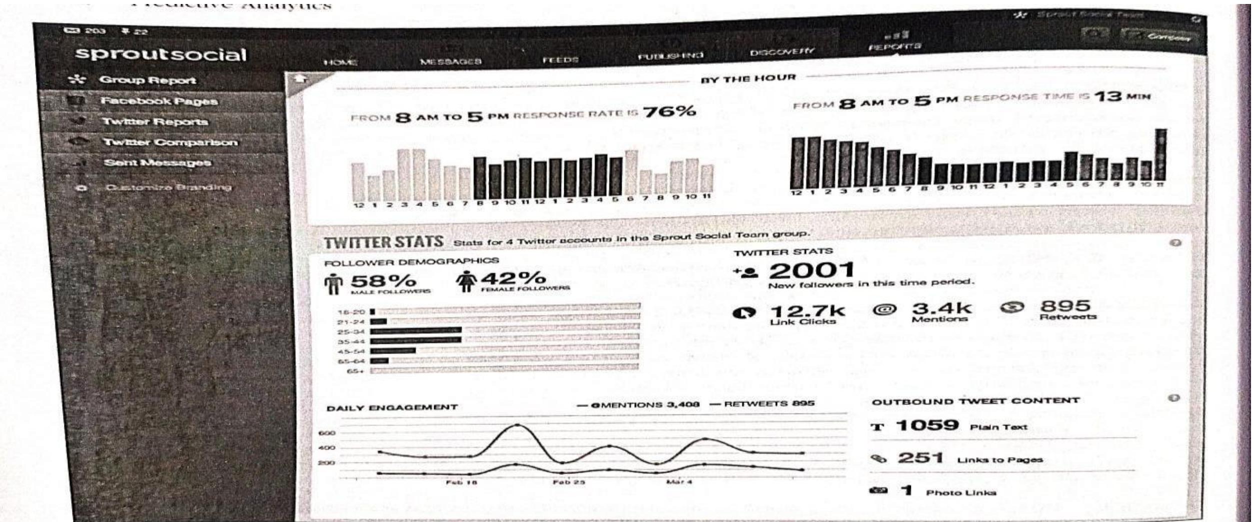


FIGURE 8.12 A Social Media Analytics Screenshot. *Source:* Courtesy of sproutsocial.com.

b) Prescriptive Analytics - Optimization and Multi-Criteria Systems:

1. Multiple Goals

The analysis of management decisions aims at evaluating, to the greatest possible extent, how far each alternative advances managers toward their goals. Unfortunately, managerial problems are seldom evaluated with a single simple goal, such as profit maximization. Today's management systems are much more complex, and one with a single goal is rare. Instead, managers want to attain *simultaneous goals*, some of which may conflict. Different stakeholders have different goals. Therefore, it is often necessary to analyze each alternative in light of its determination of each of several goals (see Koksalan and Zionts, 2001).

For example, consider a profit-making firm. In addition to earning money, the company wants to grow, develop its products and employees, provide job security to its workers, and serve the community. Managers want to satisfy the shareholders and at the same time enjoy high salaries and expense accounts, and employees want to increase their take-home pay and benefits. When a decision is to be made—say, about an investment project—some of these goals complement each other, whereas others conflict. Kearns (2004) described how the analytic hierarchy process (AHP), which we will introduce in Section 9.9, combined with integer programming, addressed multiple goals in evaluating IT investments.

Many quantitative models of decision theory are based on comparing a single measure of effectiveness, generally some form of utility to the decision maker. Therefore, it is usually necessary to transform a multiple-goal problem into a single-measure-of-effectiveness problem before comparing the effects of the solutions. This is a common method for handling multiple goals in an LP model.

Certain difficulties may arise when analyzing multiple goals:

- It is usually difficult to obtain an explicit statement of the organization's goals.
- The decision maker may change the importance assigned to specific goals over time or for different decision scenarios.
- Goals and sub-goals are viewed differently at various levels of the organization and within different departments.
- Goals change in response to changes in the organization and its environment.
- The relationship between alternatives and their role in determining goals may be difficult to quantify.
- Complex problems are solved by groups of decision makers, each of whom has a personal agenda.
- Participants assess the importance (priorities) of the various goals differently.

Several methods of handling multiple goals can be used when working with MSS. The most common ones are:

- Utility theory
- Goal programming
- Expression of goals as constraints, using LP
- A points system

2. Sensitivity Analysis

A model builder makes predictions and assumptions regarding input data, many of which deal with the assessment of uncertain futures. When the model is solved, the results depend on these data. Sensitivity analysis attempts to assess the impact of a change in the input data or parameters on the proposed solution (i.e., the result variable).

Sensitivity analysis is extremely important in MSS because it allows flexibility and adaptation to changing conditions and to the requirements of different decision-making situations, provides a better understanding of the model and the decision-making situation it attempts to describe, and permits the manager to input data in order to increase the confidence in the model. Sensitivity analysis tests relationships such as the following:

- The impact of changes in external (uncontrollable) variables and parameters on the outcome variable(s)
- The impact of changes in decision variables on the outcome variable(s)
- The effect of uncertainty in estimating external variables
- The effects of different dependent interactions among variables
- The robustness of decisions under changing conditions

Sensitivity analyses are used for:

- Revising models to eliminate too-large sensitivities
- Adding details about sensitive variables or scenarios
- Obtaining better estimates of sensitive external variables
- Altering a real-world system to reduce actual sensitivities
- Accepting and using the sensitive (and hence vulnerable) real world, leading to the continuous and close monitoring of actual results

The two types of sensitivity analyses are automatic and trial-and-error.

AUTOMATIC SENSITIVITY ANALYSIS Automatic sensitivity analysis is performed in standard quantitative model implementations such as LP. For example, it reports the range within which a certain input variable or parameter value (e.g., unit cost) can vary without having any significant impact on the proposed solution. Automatic sensitivity analysis is usually limited to one change at a time, and only for certain variables. However, it is very powerful because of its ability to establish ranges and limits very fast (and with little or no additional computational effort). For example, automatic sensitivity analysis is part of the LP solution report for the MBI Corporation product-mix problem described earlier. Sensitivity analysis is provided by both Solver and Lindo. Sensitivity analysis could be used to determine that if the right-hand side of the marketing constraint on CC-8 could be decreased by one unit, then the net profit would increase by $1,333.33. This is valid for the right-hand side decreasing to zero. For details, see Hillier and Lieberman (2005) and Taha (2006) or later editions of these textbooks.

TRIAL-AND-ERROR SENSITIVITY ANALYSIS The impact of changes in any variable, or in several variables, can be determined through a simple trial-and-error approach. You change some input data and solve the problem again. When the changes are repeated several times, better solutions may be discovered. Such experimentation, which is easy to conduct when using appropriate modelling software, such as Excel, has two approaches: what-if analysis and goal seeking.

3. What-If Analysis

What-if analysis is structured as *What will happen to the solution if an input variable, an assumption, or a parameter value is changed?* Here are some examples:

- What will happen to the total inventory cost if the cost of carrying inventories increases by 10 percent?
- What will be the market share if the advertising budget increases by 5 percent?

With the appropriate user interface, it is easy for managers to ask a computer model these types of questions and get immediate answers. Furthermore, they can perform multiple cases and thereby change the percentage, or any other data in the question, as desired. The decision maker does all this directly, without a computer programmer.

Figure 9.10 shows a spreadsheet example of a what-if query for a cash flow problem. When the user changes the cells containing the initial sales (from 100 to 120) and the sales growth rate (from 3% to 4% per quarter), the program immediately recomputes the value of the annual net profit cell (from $127 to $182). At first, initial sales were 100, growing at 3 percent per quarter, yielding an annual net profit of $127. Changing the initial sales cell to 120 and the sales growth rate to 4 percent causes the annual net profit to rise to $182. What-if analysis is common in expert systems. Users are given the opportunity to change their answers to some of the system's questions, and a revised recommendation is found.

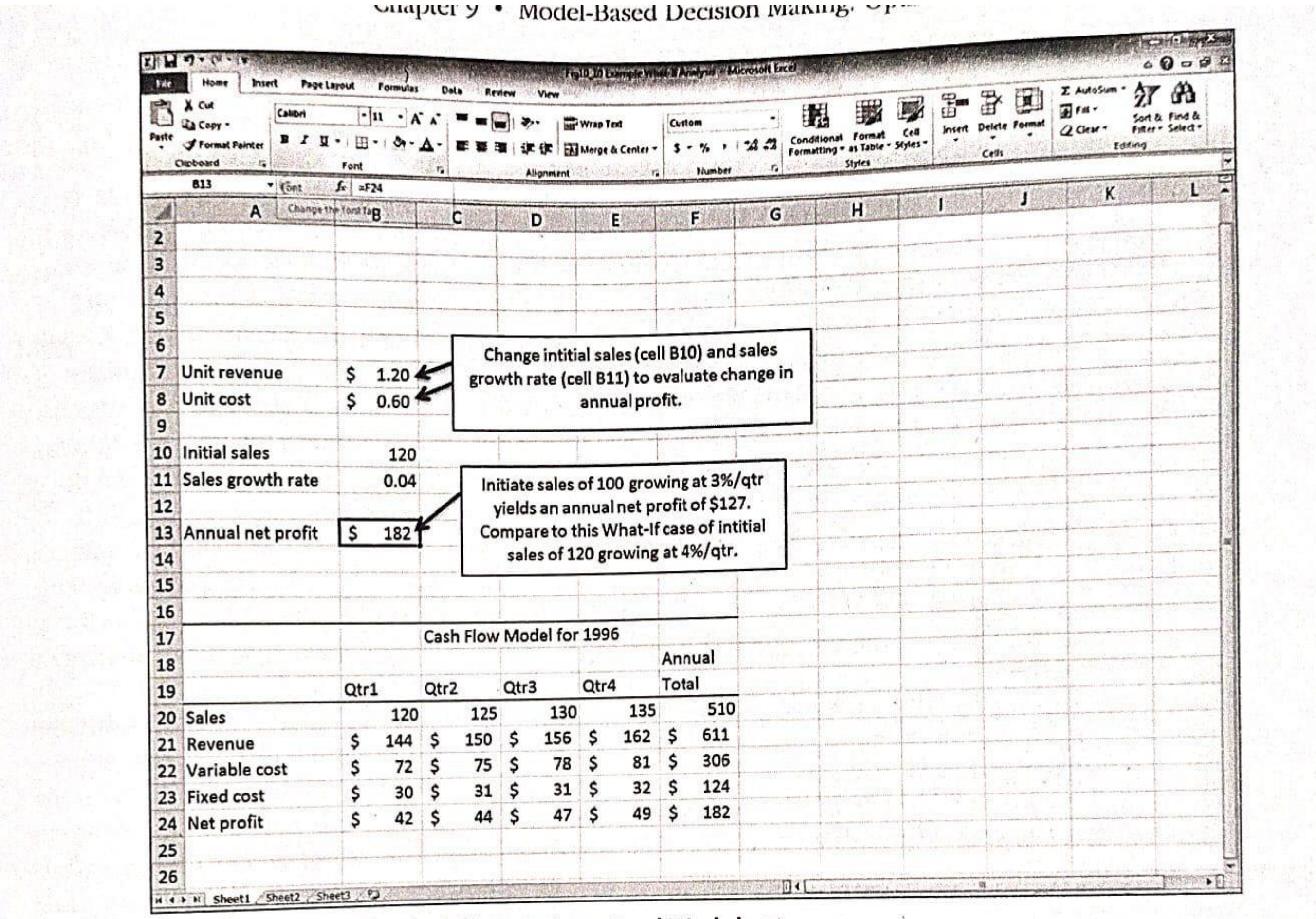


FIGURE 9.10 Example of a What-If Analysis Done in an Excel Worksheet.

4. Goal Seeking

Goal seeking calculates the values of the inputs necessary to achieve a desired level of an output (goal). It represents a backward solution approach. The following are some examples of goal seeking:

- What annual R&D budget is needed for an annual growth rate of 15 percent by 2018?
- How many nurses are needed to reduce the average waiting time of a patient in the emergency room to less than 10 minutes?

An example of goal seeking is shown in Figure 9.11. For example, in a financial planning model in Excel, the internal rate of return is the interest rate that produces a net present value (NPV) of zero. Given a stream of annual returns in Column E, we can compute the net present value of planned investment. By applying goal seeking, we can determine the internal rate of return where the NPV is zero. The goal to be achieved is NPV equal to zero, which determines the internal rate of return (IRR) of this cash flow, including the investment. We set the NPV cell to the value 0 by changing the interest rate cell. The answer is 38.77059 percent.

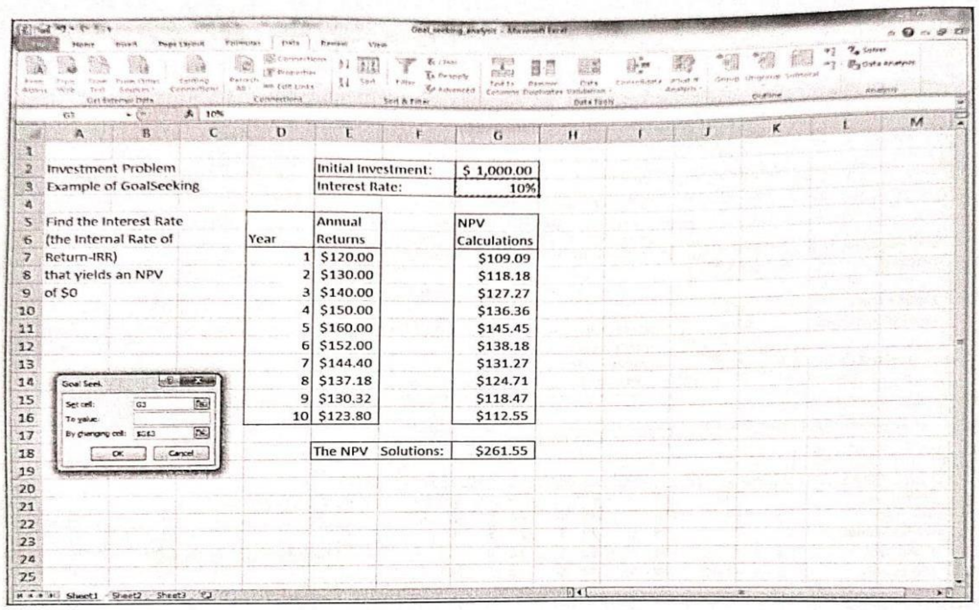


FIGURE 9.11 Goal-Seeking Analysis.

COMPUTING A BREAK-EVEN POINT BY USING GOAL SEEKING Some modelling software packages can directly compute break-even points, which is an important application of goal seeking. This involves determining the value of the decision variables (e.g., quantity to produce) that generate zero profit.

In many general applications programs, it can be difficult to conduct sensitivity analysis because the prewritten routines usually present only a limited opportunity for asking what-if questions. In a DSS, the what-if and the goal-seeking options must be easy to perform.

WEB AND SOCIAL MEDIA ANALYTICS

SHORT QUESTIONS:

1. Define Social Analytics.
2. What is the importance of Social Network Analysis in modern business?
3. List of two key metrics used in Social Network Analysis.
4. Define Social Media.
5. Differentiate between traditional media and social media.
6. What are Social Media Analytics?
7. List any two tools commonly used for Social Media Analytics.
8. Define Prescriptive Analytics.
9. Differentiate between Descriptive, Predictive, and Prescriptive Analytics.
10. What is a Multiple Goal Decision System?
11. Define Sensitivity Analysis in decision-making.
12. Differentiate between What-If Analysis and Goal Seeking.

ESSAY QUESTIONS:

1. Evaluate the challenges and limitations of Social Analytics in organizational decision-making.
2. Discuss how social media has transformed communication and marketing strategies.
3. Analyze the role of Social Media Analytics in measuring brand reputation.
4. Explain the importance of Optimization in Prescriptive Analytics with an example.
5. Evaluate the effectiveness of Goal Seeking in achieving business objectives.

…………………………………………………………….END ……………………………………………………………….

M. PRAVARSHA REDDY