Data Analytics Process Model

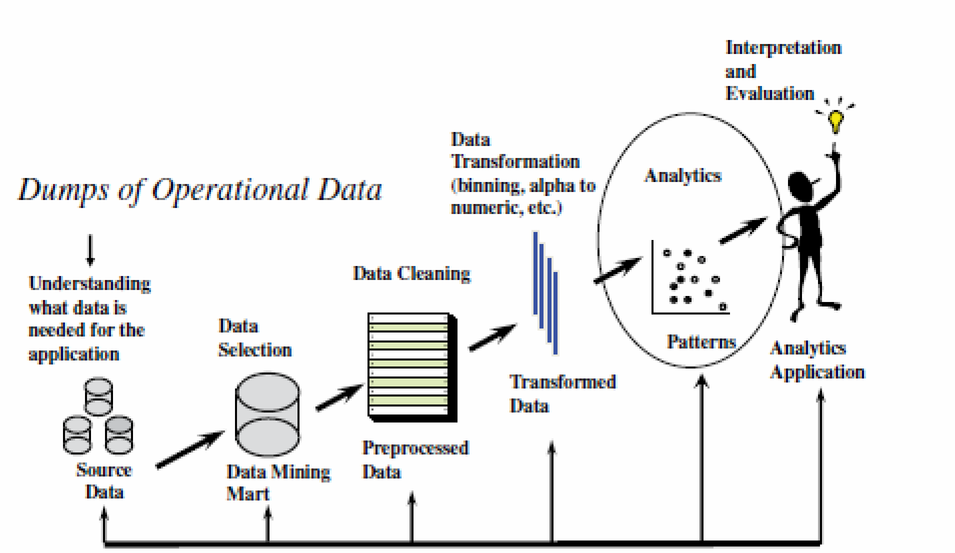


Figure 1.2 The Analytics Process Model

The **Data Analytics Process Model** provides a structured, iterative framework for converting raw operational data into useful, actionable knowledge that supports business decision-making. The major stages of the model are explained below:

1. **Business Problem Definition**
   The first and most crucial step is to clearly define the business problem or objective to be solved using analytics. A well-defined problem ensures that the analysis remains focused and relevant. Examples include churn prediction, fraud detection, or customer segmentation.

2. **Data Identification and Selection**
   Once the problem is defined, all potential source data relevant to the problem are identified(eg: population of people). Specific variables regarding a population (e.g., Age and Income) may be specified and obtained. Data is the key ingredient in analytics, and the choice of data has a deterministic impact on the quality of analytical models. Data may come from multiple operational systems and external sources. Data may be numerical or categorical.

3. **Data Collection and Staging**.

   Data Collection is the process of gathering information on targeted variables identified as data requirements. The emphasis is on ensuring accurate and honest collection of data. Data Collection ensures that data gathered is accurate such that the related decisions are valid. Data Collection provides both a baseline to measure and a target to improve. The selected data are gathered into a staging area, such as a data mart or data warehouse. This consolidated storage enables efficient access and further processing of data.

4. **Exploratory Data Analysis (EDA)**.

   The data that is collected must be processed or organized for analysis. This includes structuring the data as required for the relevant Analysis Tools. For example, the data might have to be placed into rows and columns in a table within a Spreadsheet or Statistical Application. A Data Model might have to be created. Preliminary analysis is carried out to understand data characteristics using techniques such as **OLAP** (roll-up, drill-down, slicing, and dicing). This step helps identify patterns, trends, and potential data quality issues

5. **Data Cleaning**
   Data cleaning is performed to remove inconsistencies such as missing values, outliers, noise, and duplicate records. This step is critical for improving data quality and reliability of the analytical results. There are several types of Data Cleaning that depend on the type of data. For example, while cleaning the financial data, certain totals might be compared against reliable published numbers or defined thresholds

6. **Data Transformation**
   The cleaned data are transformed into a suitable format for analysis. Common transformations include binning, alphanumeric-to-numeric coding, normalization, and geographical aggregation. The output of this stage is preprocessed and transformed data.

7. **Analytics / Model Building**
   In this step, analytical models are developed using appropriate techniques such as classification, clustering, regression, or association rule mining. The choice of technique depends on the business problem (e.g., churn prediction, fraud detection, market basket analysis).

8. **Interpretation and Evaluation**
   The results of the analytical model are interpreted and evaluated by business experts. While many patterns may be trivial, the focus is on discovering unexpected, interesting, and actionable patterns, often called knowledge diamonds, that add real business value.

9. **Deployment and Monitoring**
   Once validated, the model is deployed as an analytics application such as a decision support system or scoring engine. The model output must be presented in a user-friendly manner and integrated with other systems. Continuous monitoring and backtesting are required to ensure long-term effectiveness.

Iterative Nature of the Process
The analytics process is **iterative**, meaning analysts may revisit earlier steps as new data requirements or issues arise. Notably, data selection and preprocessing are the most time-consuming stages, often accounting for nearly 80% of the total effort.

Thus, the Data Analytics Process Model systematically transforms raw data into valuable insights through well-defined, interconnected, and iterative steps, enabling informed and effective business decision-making.

Analytical Model Requirements

A good analytical model must satisfy several important requirements to ensure that it delivers value, is reliable, and can be successfully used in real-world applications. These requirements are explained below:

1. **Business Relevance**
   The most critical requirement of an analytical model is that it should be directly aligned with the business problem it is intended to solve. The problem must be clearly defined, qualified, and agreed upon by all stakeholders at the beginning of the project. A technically sound model is useless if it does not address the original business objective.

2. **Statistical Performance**
   An analytical model should demonstrate strong statistical significance and predictive power. The way performance is measured depends on the type of analytics used. For example, in classification problems such as churn or fraud detection, the model should have good discrimination ability, while in clustering, the formed clusters should be highly homogeneous.

3. **Interpretability**
   Interpretability refers to how easily the model's results and patterns can be understood by business users. In domains such as credit risk analysis and medical diagnosis, high interpretability is essential to explain decisions and gain trust. Less interpretability may be acceptable in areas like fraud detection or marketing response modeling.

4. **Justifiability**
   Justifiability measures how well a model's outcomes align with prior business knowledge and intuition. A model may be interpretable but still unacceptable if it contradicts well-established domain logic. Therefore, model results must make business sense in addition to being statistically valid.

5. **Balance between Performance and Transparency**
   There is often a trade-off between statistical performance and interpretability. High-performing models such as neural networks act as "black boxes," while simpler models like linear regression are transparent but have limited modeling power. Choosing the right balance depends on the application context.

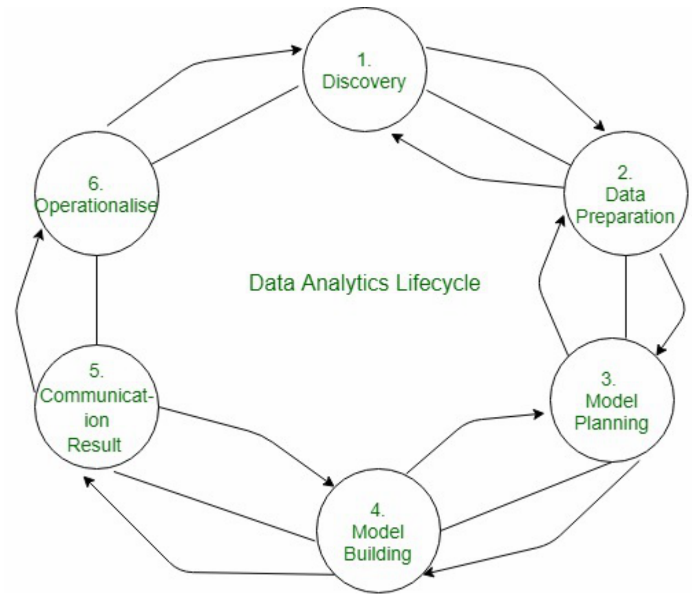
6. **Operational Efficiency**
   The model should be efficient to implement and run in practice. This includes the effort required for data collection, preprocessing, model evaluation, and integration with business applications. Operational efficiency is especially important in real-time environments such as online fraud detection.

7. **Maintainability, Monitoring, and Backtesting**
   Analytical models should be easy to monitor, backtest, and re-estimate when data or business conditions change. This ensures long-term reliability and consistent performance of the model in production.

8. **Economic Cost and Cost–Benefit Consideration**
The total cost of developing and deploying the model must be justified by its expected benefits. Costs include data acquisition, preprocessing, analysis, software, computational resources, and skilled personnel. A thorough cost–benefit analysis should be conducted before model development.

9. **Regulatory and Legal Compliance**
Analytical models must comply with local and international regulations and legislation. Examples include Basel II/III in credit risk and Solvency II in insurance. Privacy regulations and data usage laws, such as restrictions on cookies in web analytics, must also be strictly followed.
In summary, a good analytical model should be business-relevant, statistically sound, interpretable and justifiable, operationally efficient, cost-effective, and compliant with regulations. Meeting these requirements ensures that the model is both effective and sustainable in practice.

Various phases of data analytics lifecycle.



The Data Analytics Lifecycle describes a structured, iterative approach for transforming data into meaningful insights and deploying them for business use. It typically consists of six interconnected phases, as explained below:

1. Discovery

This is the initial phase of the lifecycle where the business problem is clearly defined.

- Understand the business objectives, goals, and success criteria.
- Identify key stakeholders and their expectations.
- Assess available data sources, tools, skills, timelines, and constraints.

- Formulate initial hypotheses and analytical approaches.

This phase ensures that the analytics effort is aligned with business needs.

2. Data Preparation

In this phase, data required for analysis is collected and prepared.

- Data is gathered from multiple sources (databases, files, external data).
- Data cleaning is performed to handle missing values, outliers, inconsistencies, and duplicates.
- Data is transformed through normalization, aggregation, binning, encoding, etc.
- The final output is a clean, structured dataset ready for modeling.

This phase is often the most time-consuming, taking a major portion of the project effort.

3. Model Planning

Here, the analytical strategy is designed.

- Select appropriate analytical techniques (e.g., classification, regression, clustering).
- Decide which variables (features) will be used in the model.
- Choose evaluation metrics (accuracy, precision, recall, RMSE, etc.).
- Split data into training and testing datasets.

The goal is to plan how models will be built and assessed before actual implementation.

4. Model Building

This phase involves developing and training the analytical models.

- Apply selected algorithms using prepared data.
- Train, test, and validate models.
- Tune model parameters to improve performance.
- Compare multiple models and select the best-performing one.

The output is a validated model that effectively addresses the business problem.

5. Communicate Results

In this phase, insights derived from the model are interpreted and communicated to stakeholders.

- Translate technical results into business-friendly language.
- Use visualizations, dashboards, and reports to present findings.
- Highlight actionable insights and recommendations.
- Validate results with domain experts to ensure business relevance.

Effective communication ensures stakeholder understanding and acceptance.

6. Operationalize

This is the final phase where analytics is put into real-world use.

- Deploy the model into production systems (e.g., decision support systems, scoring engines).
- Integrate with existing applications and workflows.
- Monitor model performance and conduct backtesting.
- Update or retrain models as data and business conditions change.

This phase ensures sustained value from the analytics solution.

The data analytics lifecycle is iterative, not linear. Insights gained at any phase may require revisiting earlier steps, such as acquiring more data or refining the model.

The data analytics lifecycle provides a systematic framework consisting of discovery, data preparation, model planning, model building, result communication, and operationalization. Following these phases ensures that analytics solutions are accurate, actionable, and aligned with business objectives.