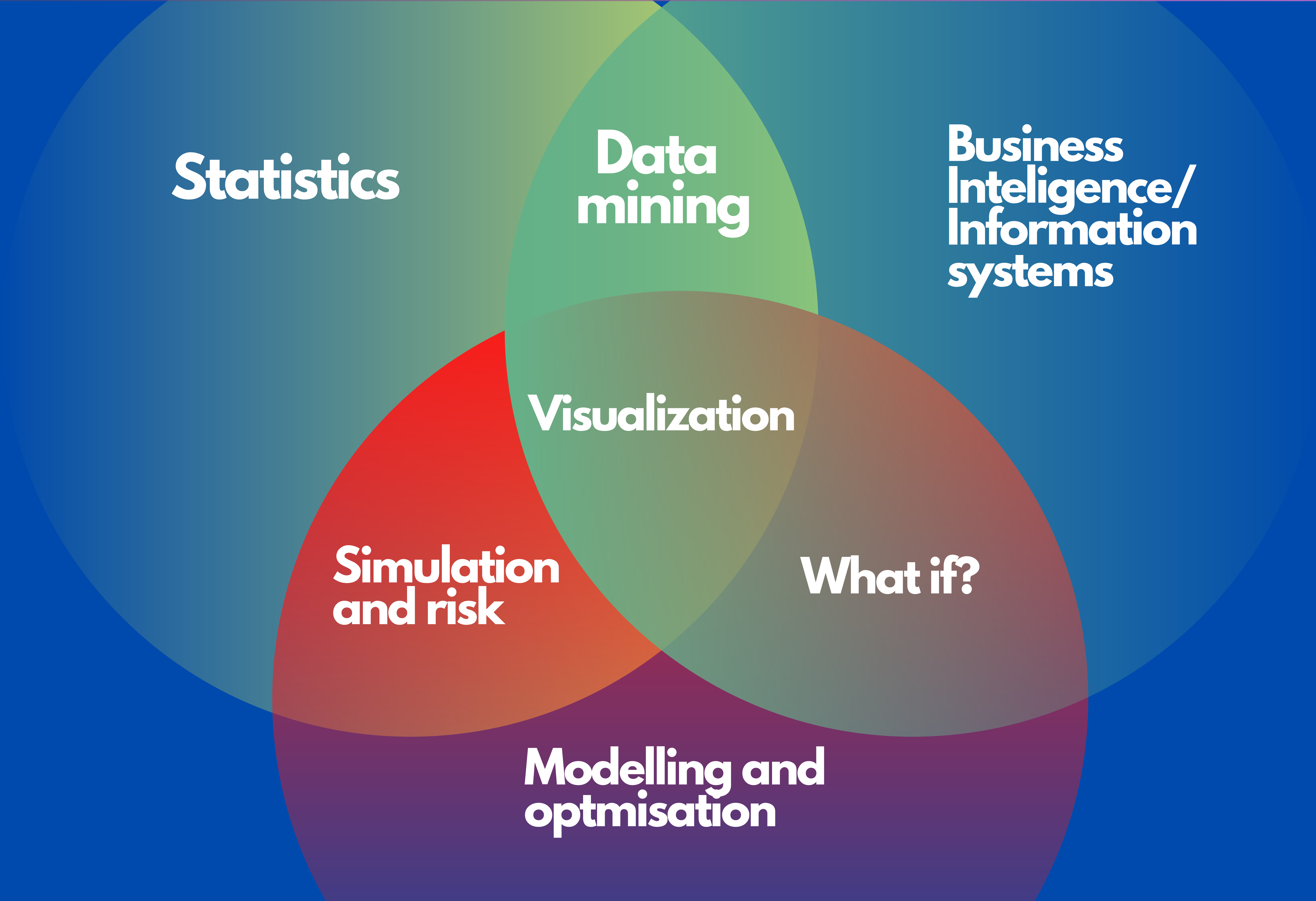
BUSINESS ANALYTICS UNIT II

A VISUAL PERSPECTIVE OF BUSINESS ANALYTICS



Statistics

Data mining

Business Inteligence/ Information systems

Visualization

Simulation and risk

What if?

Modelling and optmisation

DATA STRUCTURES

A data structure is a specialized format for organizing, processing, retrieving and storing data. There are several basic and advanced types of data structures, all designed to arrange data to suit a specific purpose. Data structures make it easy for users to access and work with the data they need in appropriate ways. Most importantly, data structures frame the organization of information so that machines and humans can better understand it.

DATA ORGANIZATION

Data organization is the practice of categorizing and classifying data to make it more usable. Similar to a file folder, where we keep important documents, you'll need to arrange your data in the most logical and orderly fashion, so you — and anyone else who accesses it — can easily find what they're looking for.

IMPORTANCE

Good data organization strategies are important because your data contains the keys to managing your company's most valuable assets. Getting insights out of this data could help you obtain better business intelligence and play a major role in decision making.

DATA ORGANIZATION IN AN OPTIMAL WAY

- Establish consistent and clear naming practices. Name your files in a descriptive and clear way. If you need to rename multiple files, you can use a file renaming application to do it automatically.
- Keep file titles short. Avoid symbols. If you use dates, keep a consistent format.
- Use consistent file version management. This means that you create a new file using an updated name, instead of saving over the old file. This is also known as “file versioning.”
- Create and use a data dictionary to standardize categories and provide a definition around the role of each. This will allow all your company’s stakeholders to get the most out of the datasets you’ve collected.

INCOMPLETE DATA

This is by far the most common issue when dealing with DQ. Key columns are missing information, failing ETL jobs or causing downstream analytics impact. The best way to fix this is to put in place a reconciliation framework control. The control would check the number of records passing through your analytical layers and alert when records have gone missing.

DEFAULT VALUES

Ever analysed your data and found 01/01/1891 as a date for a transaction? Unless your customer base comprises 130-year-old individuals, this is likely a case of using default values. This is especially a problem if there is a lack of documentation. The best way to fix this is to profile the data and understand the pattern of why default values were used. Usually, engineers use this data when a real-life alternative date is unavailable.

DATA FORMATS INCONSISTENCIES

String columns predominantly suffer from this problem, where data can be stored in many formats. For example, a customer’s first and last name is stored in different cases or an email address without the correct formatting. It occurs when multiple systems store information without an agreed data format. To fix this, data needs to be homogenised (standardised) across the source system or at least in the data pipeline when fed to the data lake or warehouse.

DUPLICATE DATA

Reasonably straightforward to spot, quite tricky to fix. If the critical attribute is populated with dirty data duplicates, it will break all the key downstream processes. It can also cause other DQ issues. To fix this, a master data management control needs to be implemented, even as basic as a uniqueness check. This control will check for exact duplicates of records and purge one record. It can also send a notification for the other record to the data engineer or steward for investigation.

CROSS-SYSTEM INCONSISTENCIES

Very common in large organisations that have grown by acquisitions and mergers. Multiple source legacy systems all have a slightly different view of the world. Customer name, address or DOB all have inconsistent or incorrect information. Like 4, a master data management solution must be implemented to ensure all the different information is matched into a single record. This matching doesn't need to be exact; it could be fuzzy based on a threshold of match percentage.

DRASTIC DATA CHANGES

Imagine you receive a data file full of customer addresses every day. You generally receive 5000 records daily, but today you only see 30 records. The file still passes other checks like uniqueness, validity and accuracy; however, it is still missing data. A reconciliation framework, as proposed in 1. would help resolve this issue within your analytical layers. If this data comes from a different source, you will have to implement a control file confirming the records sent. An automated process would then reconcile this file with the records received after the data transfer. This control would help catch and fix this issue.

ORPHANED DATA

This DQ issue relates to data inconsistency problems where data exists in one system and not the other. A customer exists in table A, but their account doesn't exist in table B. It would be classed as an orphan customer. On the other hand, if an account exists in table B but has no associated customer, it would be classed as an orphan account. A data quality rule that checks for consistency each time data is ingested in tables A and B will help spot the issue. To remediate this, the source system would need to check the underlying cause of this inconsistency.

DYSFUNCTIONAL HISTORY MAINTENANCE

History maintenance is critical for any data warehousing implementation. Now data is being received chronologically, and history is being maintained using SCD (slowly changing dimensions) Type 2. However, the incorrect rows are being opened and closed, leading to a false representation of the latest valid record. In turn, breaking the history maintenance method and downstream processes. To fix this, ensure the correct date column is used to determine the history maintenance.

IRRELEVANT DATA

Nothing is more frustrating than capturing ALL the available information. Besides the regulatory restrictions of data minimisation, capturing all the available data is more expensive and less sustainable. To fix this, data capturing principles need to be agreed upon; each data attribute should have an end goal; otherwise, it should not be captured.

UNCLEAR DATA DEFINITIONS

Speak to Sam in Finance and Jess in Customer Services, both interpreting the same data point differently; sounds familiar? Clarity is a DQ dimension that is not discussed much, as in the modern data stack, it is part of the business glossary or data catalogue. Fundamentally, this is a DQ issue. Fixing this requires aligning data definitions each time a new metric/data point is created.

REDUNDANT DATA

Multiple teams across the organisation, capturing the same data repeatedly. In an organisation with an online and high street presence, capturing the same information numerous times will lead to data being available in various systems leading to data redundancy. Not only is this poor for the company's bottom line, but it is also a poor customer experience. To fix this, a singular base system should be utilised where all the organisations' agents receive their data, yet again a master data implementation.

OLD & STALE DATA

Storing data beyond a certain period adds no value to your data stack. It costs more money, confuses the engineer, and it impacts your ability to conduct analytics. It also makes the data irrelevant (. To fix this, apply the GDPR principle on retention, and store it for "no longer than is necessary".

INCONSISTENT KEYS

Imagine building a new data warehouse with Primary & Surrogate keys for your core data model. Once the data warehouse matures and receives new data daily, including seasonal peaks, you realise that the natural keys are not unique. This finding breaks the model's design, leading to a breach of referential integrity. To fix this, comprehensive profiling of the data has to be carried out, including seasonal data, to ensure the key on which the surrogate key is dependent is always unique.

POOR DATA ACCESSIBILITY

Good quality data should be accessible to the people that require it to make informed decisions. Having data locked away in a data warehouse with no integration and access to the data analysts, stewards or scientists is useless. To fix this, you should implement an operating model including permissions on how teams would access the data.

DATA RECEIVED TOO LATE

Data needs to be timely enough to make the critical decision in that period. If your marketing campaigns are running weekly, you must receive the required data by the set day of the week to trigger them. Otherwise, too late data could lead to poor responses on your campaigns. You must agree on an appropriate time window with the engineering team to fix this. And work backwards to ensure your source systems can adhere to those SLAs (Service Level Agreements).

WHAT IS DATA QUALITY ?

Data quality is how we describe the state of any given dataset. It measures objective elements such as completeness, accuracy, and consistency. But it also measures more subjective factors, such as how well-suited a dataset is to a particular task. This subjective aspect makes determining data quality challenging at times. If data quality is high, you can use a dataset for its intended purpose. This might be to make key spending decisions, improve operations, or inform future growth. Yet if data quality is low, all these areas are negatively affected.

HOW DO YOU MEASURE DATA QUALITY?

a good baseline is to look at the six characteristics of quality data.

- Validity
- Accuracy
- Completeness
- Consistency
- Uniformity
- Relevance

WHAT IS DATA QUALITY ?

- Data quality is essential for one main reason: You give customers the best experience when you make decisions using accurate data. A great customer experience leads to happy customers, brand loyalty, and higher revenue for your business. If you're using poor-quality data, you're mostly guessing at what your customers want. Worse still, you might be actively doing things your customers dislike.

- Collecting trustworthy data and updating existing records gives you a better understanding of your customers. It also lets you keep in contact with them using verified email addresses, mailing information, and phone numbers. This information helps you market effectively and use resources efficiently.

- Maintaining data quality can help you stay ahead of your competitors, too. Reliable data keeps your business agile. You can spot trends and industry changes sooner so you can take advantage of new opportunities or tackle challenges before your competitors.

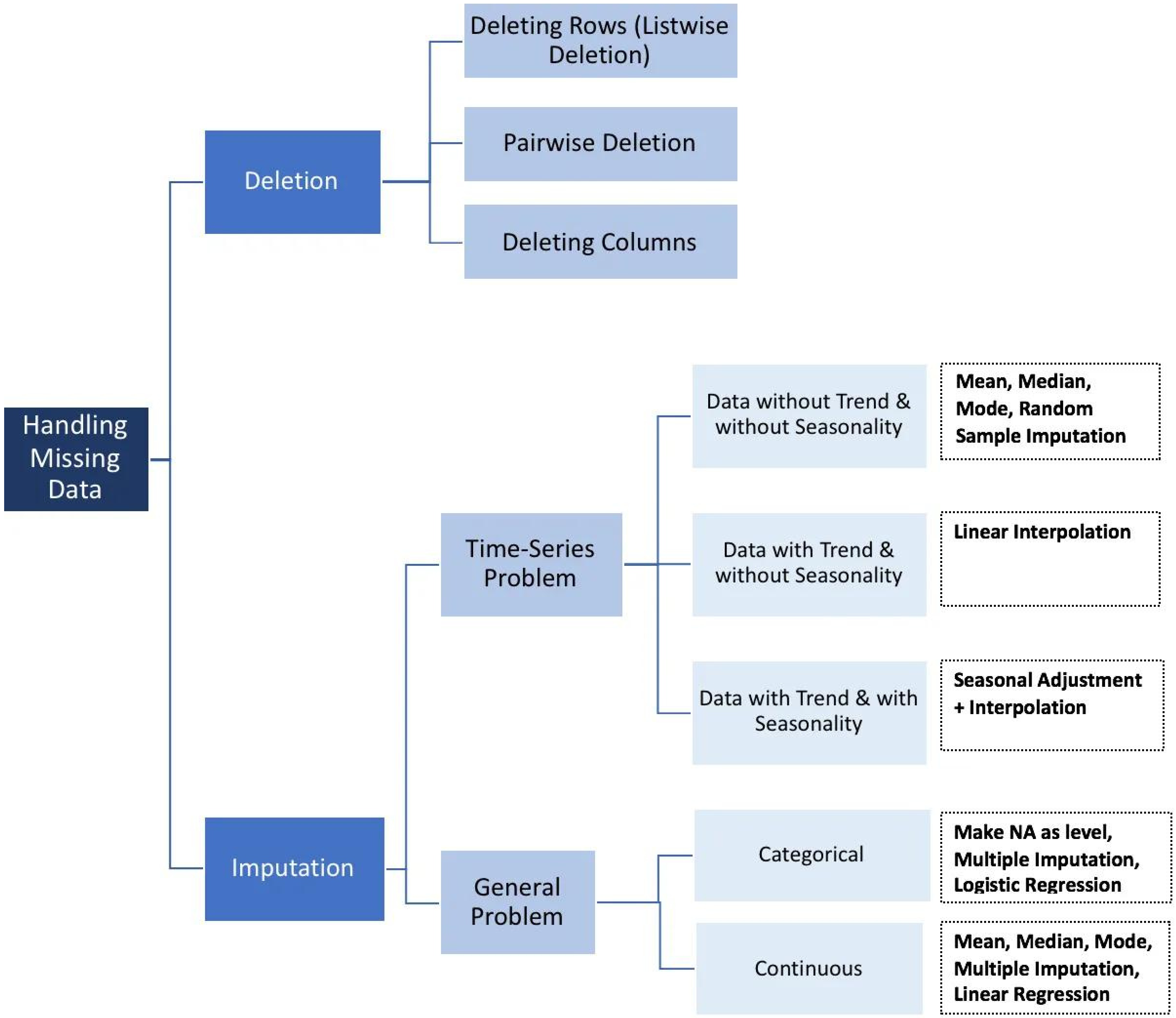
- If you want to preserve good data quality, you must constantly manage it to get the best results. Luckily, modern data tools and platforms from companies like Experian help automate and streamline your day-to-day data validation and management.

Dealing with Missing or incomplete data

Missing Completely at Random (MCAR): The fact that a certain value is missing has nothing to do with its hypothetical value and with the values of other variables.

Missing at Random (MAR): Missing at random means that the propensity for a data point to be missing is not related to the missing data, but it is related to some of the observed data.

Missing not at Random (MNAR): Two possible reasons are that the missing value depends on the hypothetical value (e.g. People with high salaries generally do not want to reveal their incomes in surveys) or missing value is dependent on some other variable’s value (e.g. Let’s assume that females generally don’t want to reveal their ages! Here the missing value in age variable is impacted by gender variable).



Handling Missing Data

Deletion

Deleting Rows (Listwise Deletion)

Pairwise Deletion

Deleting Columns

Imputation

Time-Series Problem

Data without Trend & without Seasonality

Mean, Median, Mode, Random Sample Imputation

Data with Trend & without Seasonality

Linear Interpolation

Data with Trend & with Seasonality

Seasonal Adjustment + Interpolation

General Problem

Categorical

Make NA as level, Multiple Imputation, Logistic Regression

Continuous

Mean, Median, Mode, Multiple Imputation, Linear Regression

DELETION METHODS

The deletion methods only work for certain datasets where participants have missing fields. There are several deleting methods – two common ones include Listwise Deletion and Pairwise Deletion. It means deleting any participants or data entries with missing values. This method is particularly advantageous to samples where there is a large volume of data because values can be deleted without significantly distorting readings. Alternatively, data scientists can fill out the missing values by contacting the participants in question. The problem with this method is that it may not be practical for large datasets. Furthermore, some corporations obtain their information from third-party sources, which only makes it unlikely that organisations can fill out the gaps manually. Pairwise deletion is the process of eliminating information when a particular data point, vital for testing, is missing. Pairwise deletion saves more data compared to likewise deletion because the former only deletes entries where variables were necessary for testing, while the latter deletes entire entries if any data is missing, regardless of its importance.

IMPUTATION TECHNIQUES

Data scientists use two data imputation techniques to handle missing data: Average imputation and common-point imputation. Average imputation uses the average value of the responses from other data entries to fill out missing values. However, a word of caution when using this method – it can artificially reduce the variability of the dataset. Common-point imputation, on the other hand, is when the data scientists utilise the middle point or the most commonly chosen value. For example, on a five-point scale, the substitute value will be 3. Something to keep in mind when utilising this method is the three types of middle values: mean, median and mode, which is valid for numerical data (it should be noted that for non-numerical data only the median and mean are relevant).

REGRESSION ANALYSIS

Regression is useful for handling missing data because it can be used to predict the null value using other information from the dataset. There are several methods of regression analysis, like Stochastic regression. Regression methods can be successful in finding the missing data, but this largely depends on how well connected the remaining data is. Of course, the one drawback with regression analysis is that it requires significant computing power, which could be a problem if data scientists are dealing with a large dataset.

WHAT IS DATA CLASSIFICATION?

The process of analyzing unstructured or structured data and categorizing it based on contents, file type, and other metadata is referred to as data classification.

Purpose of Data Classification

Risk Mitigation

- Access to personally identifiable information is limited (PII)
- Control the location of intellectual property and its access (IP)
- Reduce the attack surface area on data that is sensitive.
- The classification should be integrated into DLP and other policy-enforcing applications.

Governance/Compliance

- Determine which data is governed by GDPR, HIPAA, CCPA, PCI, SOX, and other regulations.
- To enable additional tracking and controls, apply metadata tags to protected data.
- Legal holds, quarantining, archiving, and other needed actions can all be enabled.
- Facilitate Data Subject Access Requests and the "Right to be Forgotten" (DSARs)

Efficiency and Optimization

- Allow efficient access to content based on type, usage, and other factors.
- Finds and removes stale or redundant data.
- Move data that is frequently accessed to faster devices or cloud-based infrastructure.

Analytics

- To improve business operations, enable metadata tagging.
- Inform the organization about where the data is stored and used.

DATA SENSITIVITY LEVELS

The data sensitivity classification levels are high, medium, or low.

- **High Sensitivity Data**

If compromised or destroyed in an unauthorized transaction, the organization or individuals would suffer catastrophic consequences. Financial records, intellectual property, and authentication data are just a few data classification examples.

- **Medium Sensitivity Data**

Intended for internal use only but would not have a catastrophic impact on the organization or individuals if compromised or destroyed. e.g., Documents and Emails with zero confidential information.

- **Low Sensitivity Data**

They are intended to be used by the general public. E.g., content of a public website.

TYPES OF DATA CLASSIFICATION

Data bracket substantially entails multiple markers that define types of data and their integrity and confidentiality. In data classification processes, availability may also be taken into account. Data sensitivity is frequently classified based on various levels of importance or privacy, linked to the security measures implemented to defend each classification level.

There are three types of data classification that are widely used in the industry:

- Content-based classification examines and interprets files in search of sensitive data.
- Context-based classification considers characteristics such as creator, application, and location as indirect markers.
- User-based: The classification of each document is based on a manual selection by the end-user. To sensitive flag documents, user-based classification depends on user knowledge and discretion during creation, edit, or review.