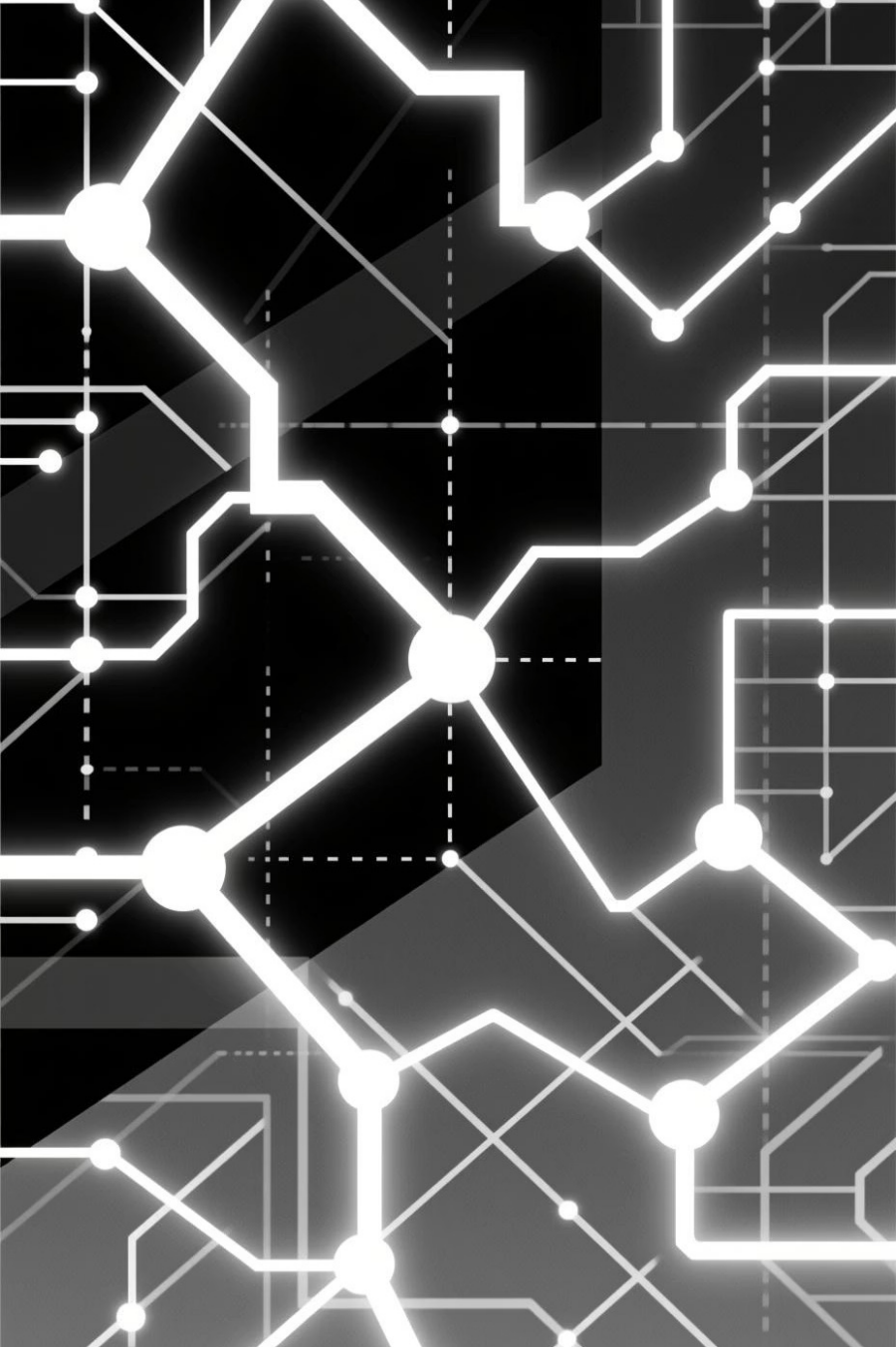




Data Collection and Cleansing: Foundations for Reliable Data Science

Why Data Collection and Cleansing Matter

Costs and Derailment

Poor data quality can lead to billions in losses and project failures, as seen with examples like Microsoft's Tay bot.

Insight Reliability

"Better data beats fancier algorithms"—clean data ensures that insights derived are trustworthy and actionable.

Raw Data Challenges

Raw data is inherently noisy, incomplete, and inconsistent, necessitating careful preparation before analysis.

Step 1: Data Collection Essentials



- Gather data directly relevant to your specific research question or problem scope.
- Utilise multiple sources carefully, ensuring you avoid duplication and the inclusion of irrelevant information.
- Thoroughly document metadata and collection methodologies to maintain crucial context and consistency for future use.

Step 2: Identifying Data Issues

Detecting Errors

Look for duplicates, missing values, and structural errors such as typos or inconsistent formatting.

Spotting Outliers

Identify unusual data points; these can be genuine errors or potentially valuable insights.

Using Visualisation

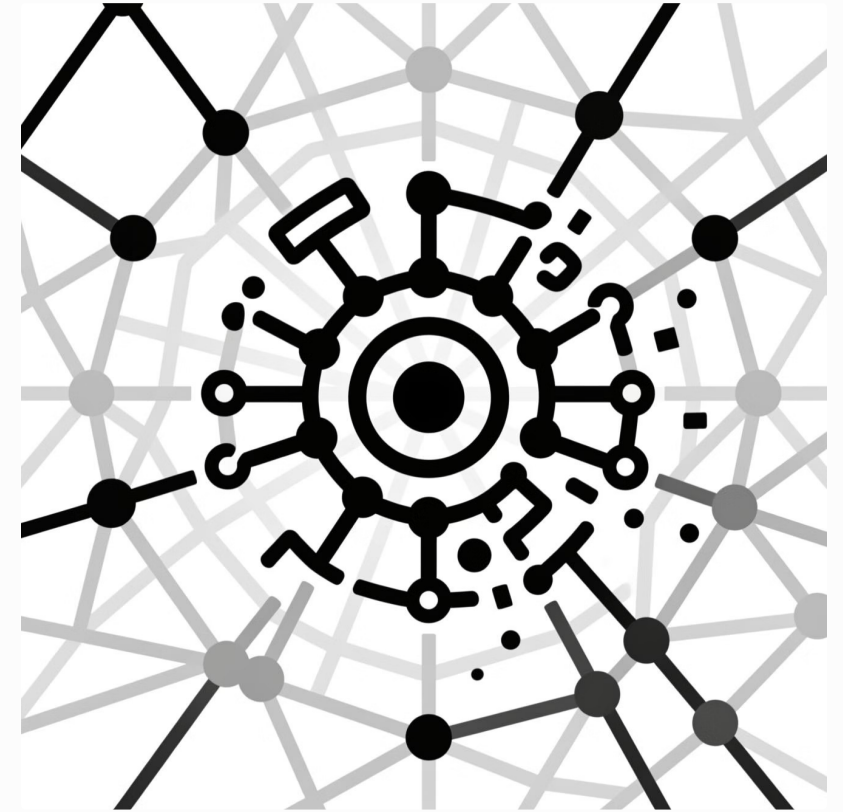
Employ tools like box plots and histograms, alongside statistical methods (Z-scores, IQR), to uncover issues.

Step 3: Removing Duplicates and Irrelevant Data

Duplicate records can significantly inflate dataset size and introduce bias into your analytical results, leading to inaccurate conclusions.

Irrelevant observations, such as data from non-target demographics, can distract from the core analysis and hinder the discovery of meaningful patterns.

Implementing robust de-duplication and filtering techniques is crucial for improving overall dataset quality and enhancing analytical performance, ensuring a cleaner, more focused approach.



Step 4: Fixing Structural Errors

1

Standardise Naming

Ensure consistency across all entries (e.g., convert "N/A" to "Not Applicable").

2

Correct Typos

Eliminate spelling mistakes and unify capitalisation to prevent misclassification.

3

Consistent Data Types

Verify that all data types and formats are uniform across every column.

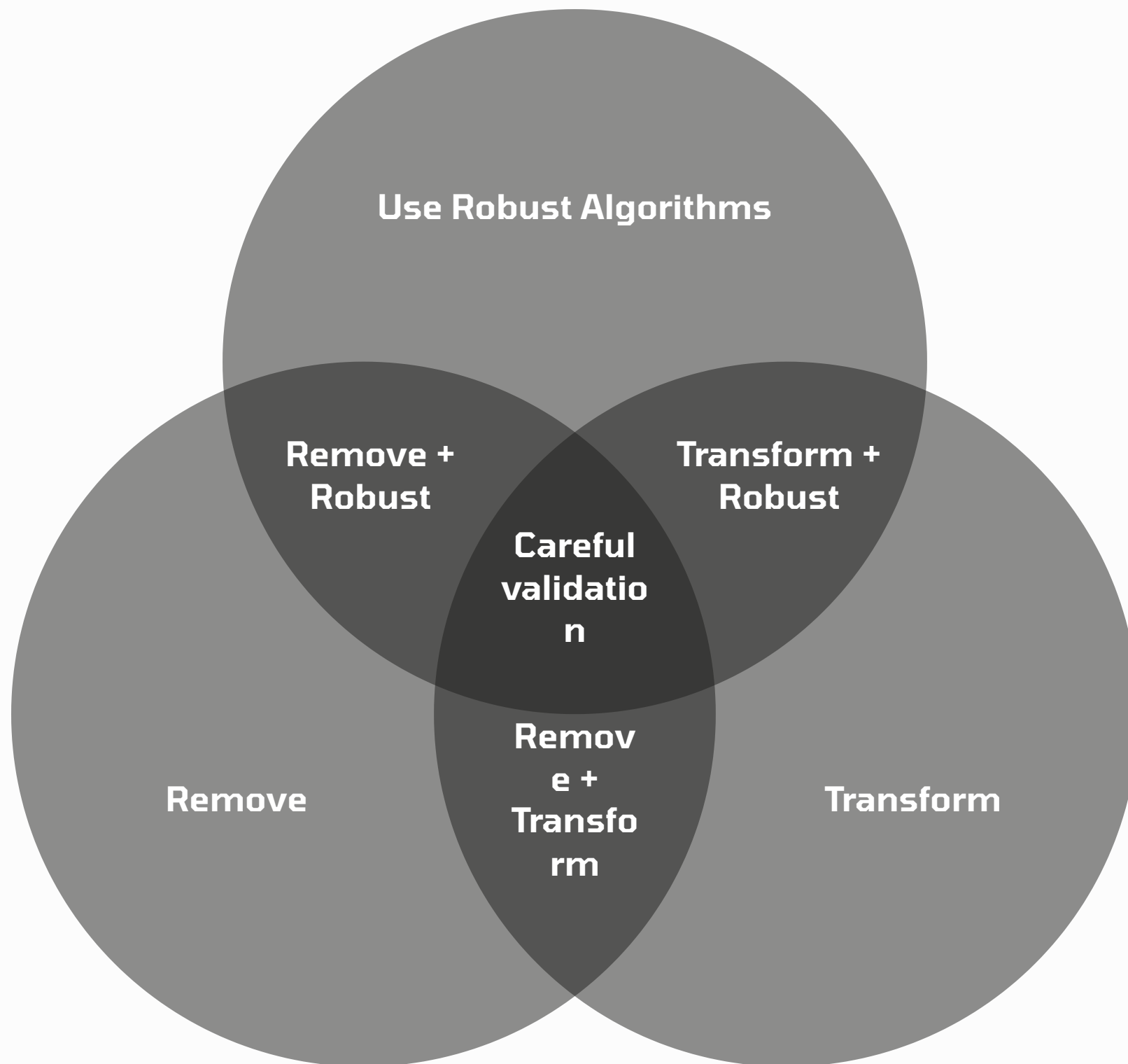
Step 5: Handling Missing Data

Addressing missing data is a nuanced process, requiring careful consideration of available options:

- **Removal:** Delete incomplete records, but be mindful that this can reduce your sample size, potentially impacting the statistical power of your analysis.
- **Imputation:** Fill in missing values using methods such as the mean, median, or interpolation. However, this can introduce assumptions that may bias your results.
- **Adjust Analysis:** Adapt your analytical methods to account for missing data without altering the dataset itself.

⊗ **Beware:** Dropping too much data reduces representativeness; imputing can distort actual distributions. Choose a strategy based on data context and analysis goals.

Step 6: Managing Outliers



The primary decision when managing outliers is determining if they are genuine errors or meaningful extreme values. An example is a test score of 2 in a group averaging 30: this could be a data entry error or a true low score. The chosen technique—removal, transformation (e.g., log, winsorisation), or using robust algorithms—must align with the outlier's nature and your analytical objectives.

Tools and Techniques for Data Cleaning

Python Libraries

- **Pandas:** Utilise functions like `isnull()`, `dropna()`, and `fillna()` for efficient missing data handling.
- **NumPy:** Essential for numerical operations and array manipulation.
- **Scikit-learn preprocessing:** For more advanced data transformation and scaling techniques.

Best Practices

- **Automated Scripts:** Implement scripts for consistency checks and validation, streamlining repetitive tasks.
- **Maintain Raw Data:** Always keep copies of your original raw data before any cleaning, ensuring traceability and reversibility.
- **Document Steps:** Meticulously document all cleaning steps for transparency and reproducibility of your analysis.

Conclusion: Clean Data, Confident Decisions

Data cleaning is not merely a task but a critical, iterative process fundamentally shaping the quality of your analysis.

Investing time upfront in thorough data preparation prevents costly errors and misleading conclusions down the line.

By mastering the principles of data collection and cleansing, you empower yourself to unlock the true potential of data science, leading to more confident and reliable decisions.