

Introduction

With experience comes knowledge, and most Statistical programmers (SP) would agree, that experience unveils many truths behind data preparation and checks. Inexperienced programmers, tend to assume that data checks are limited to Data Management preparations. In reality, data checks are interconnected across DM and SP teams and a continuous workflow of checks and communication should be built between these two teams.

The growing volume and complexity of data, coupled with shorter turnover times and greater analytical opportunities are all contributors, that should light a fire, motivating SP to perform more extensive data checks. Accompanying these checks are an increased number of considerations required from the programming team, related to data preparations for statistical models, assumption confirmations, reduction in noise, and much more.

This poster considers the abovementioned elements, that contribute to the unspoken rules of data validation and preparation, which SP should be implementing in their daily activities.

Objectives

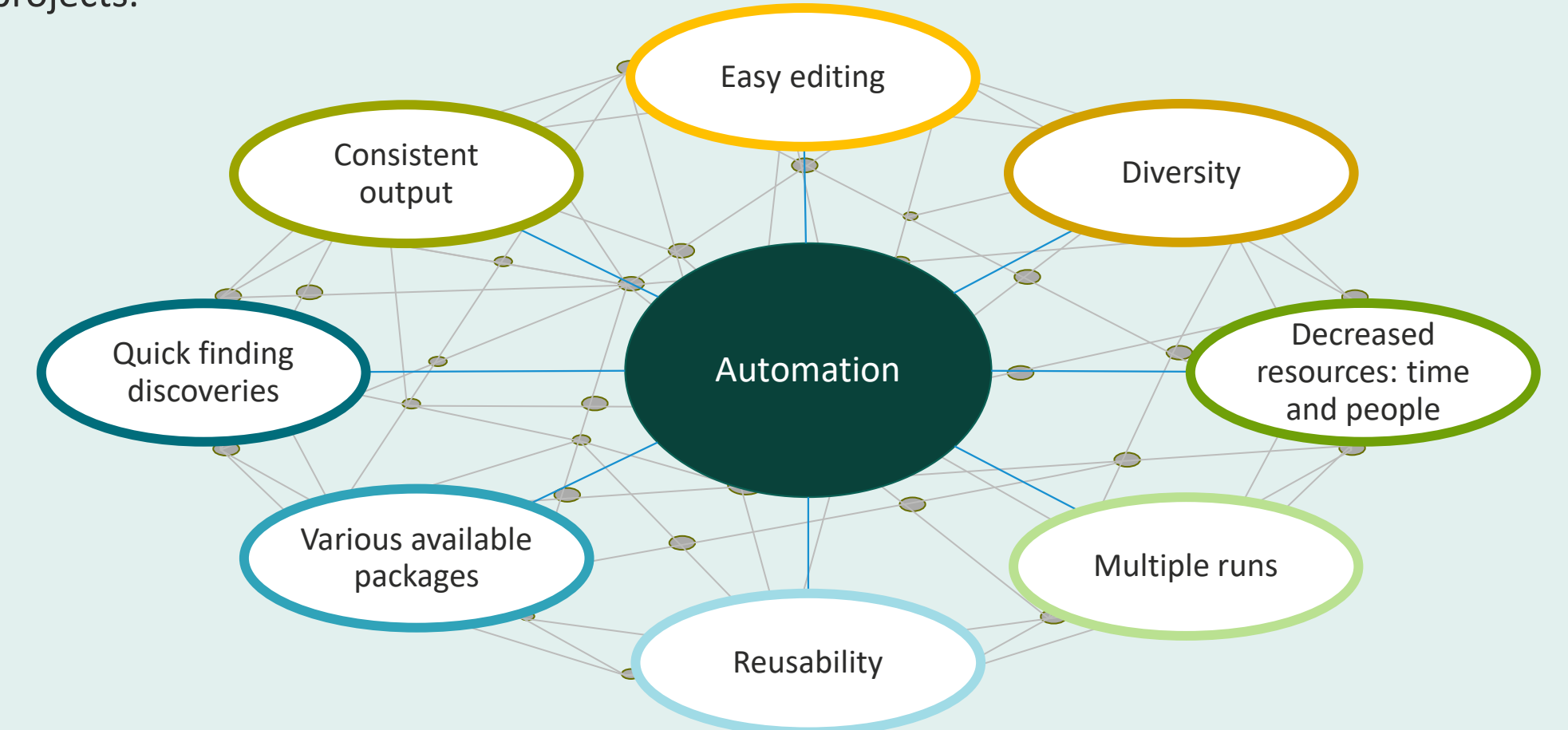
- [1] Evaluate data preparation expectations for Statistical Programmers.
- [2] Identify and motivate data preparation prospects.

Bias Considerations

For reliability purposes, it is crucial to meticulously identify and address any biases present in the dataset during the process of data preparation and review. Several types of biases should be taken into consideration, including sample bias, exclusion bias, measurement bias, observer bias, association bias, and others. For a more comprehensive understanding of these various biases, please refer to Njeri, N.R., (2022) ^[1]. Minimizing bias to the greatest extent possible holds paramount significance, and any concerns should always be brought forward.

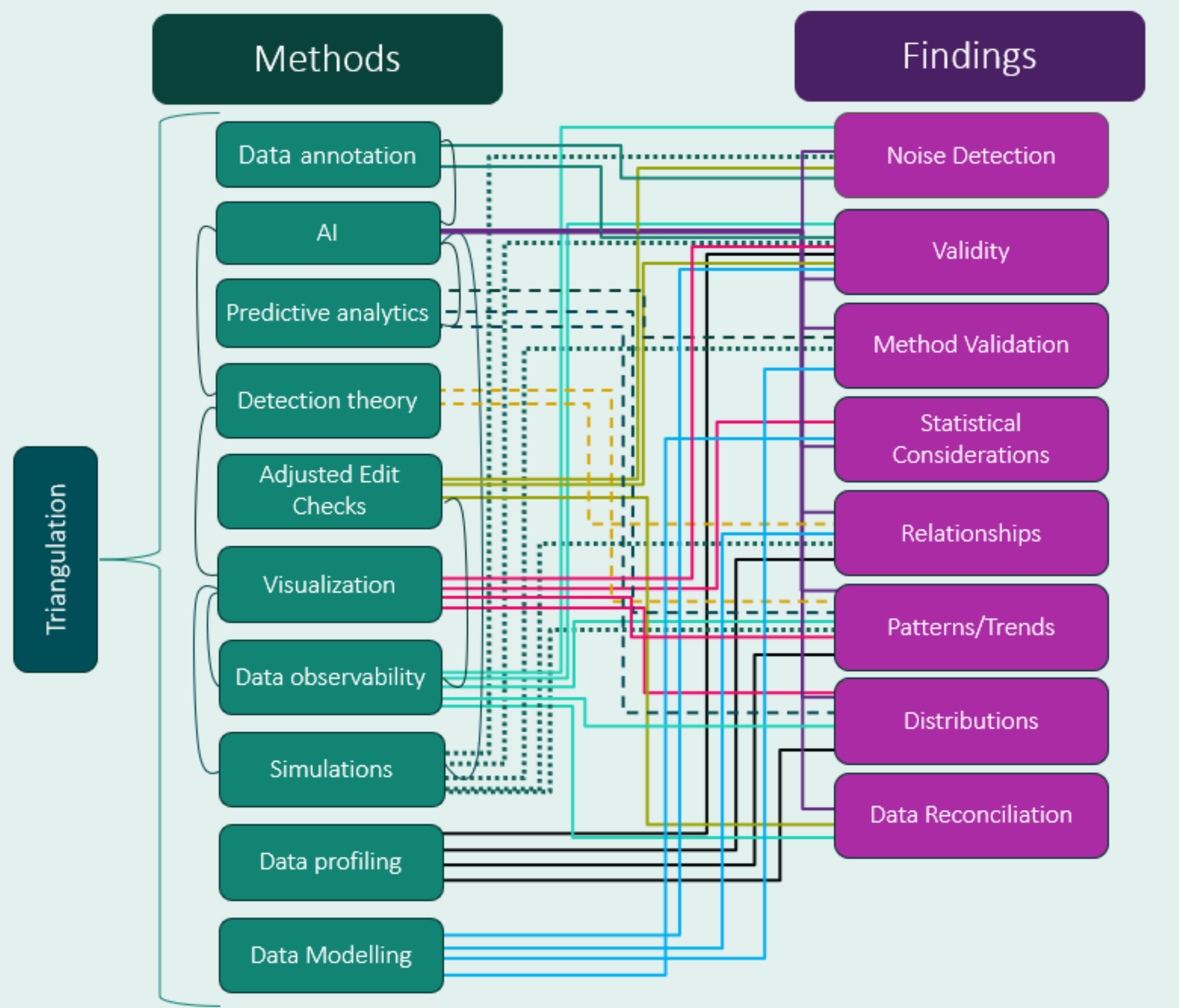
Automation

The process of data preparation and review can become quite lengthy and complicated. This is probably the primary reasons for SP's lack in motivation for following proper data and model validation checks. It is at this stage where macros are vital. Similar to analysis macros, these should be set up in such a way, that minimal effort is required to apply on new data. A once off set of good checks, could be run numerous times and for multiple projects.



Summary and Triangulation

Triangulation in data preparation refers to using multiple models or mixing methods to validate and cross-check data for reliability.

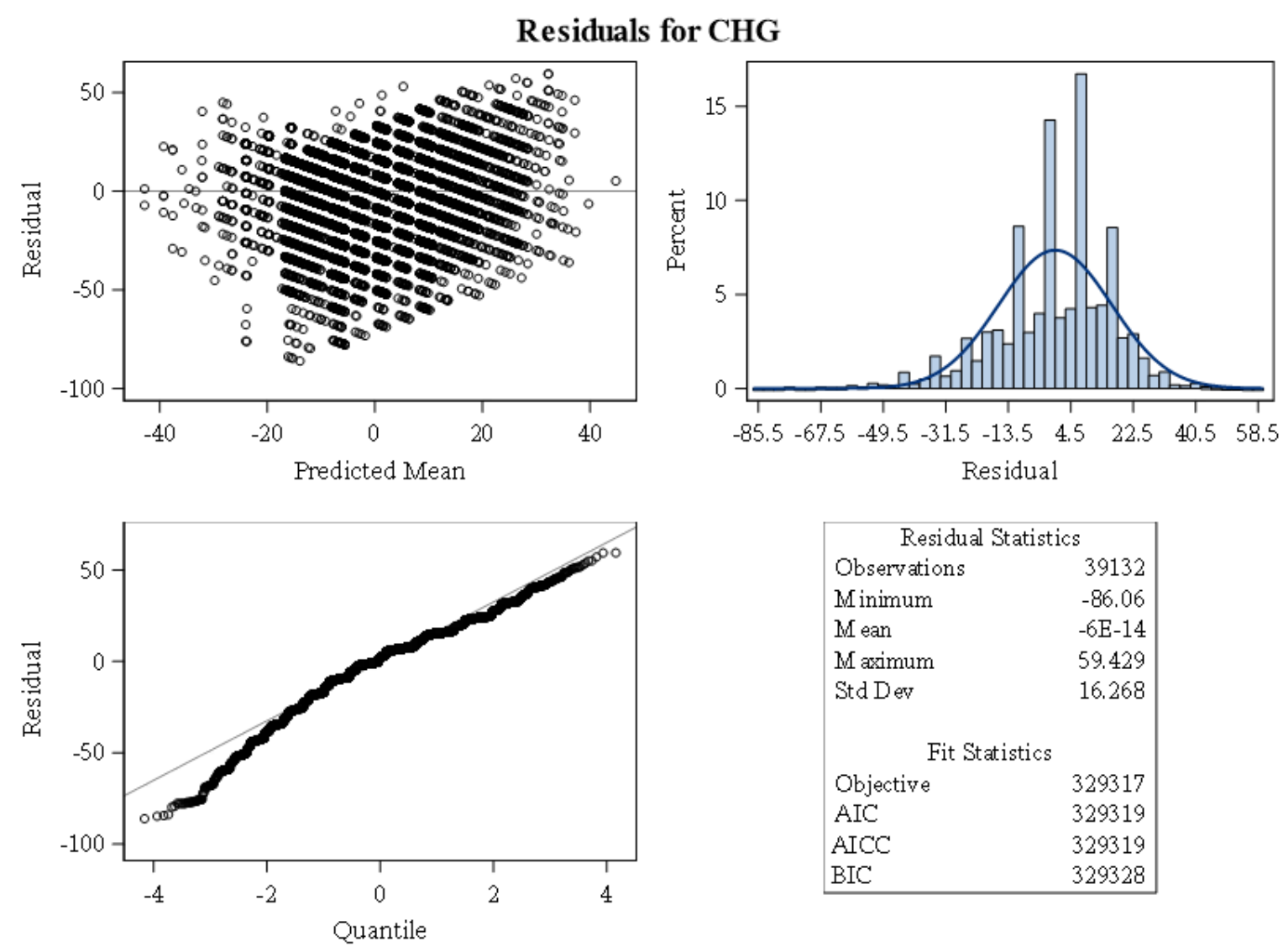


Findings: Data Preparation and Validation

For simplicity we categorized data checks into 4 groups, *i.e.*, Noisy data, Validity, Model Assumptions and Statistical Considerations. When any of these findings are present in the data, it can lead to incorrect conclusions and misinterpretations of relationships between variables.



- Non-contributing variables
- Single value variables
- Contributing rows
- Duplicate records
- Mostly missing
- Borderline results.
- Confirmed mislabeled instances.
- Missing values
- Data consistency
- Valid data values
- Skewness and normality
- Homogeneity of variances
- Linearity
- Independence
- Other model specific checks
- Outliers
- Near Zero variance
- Correlation

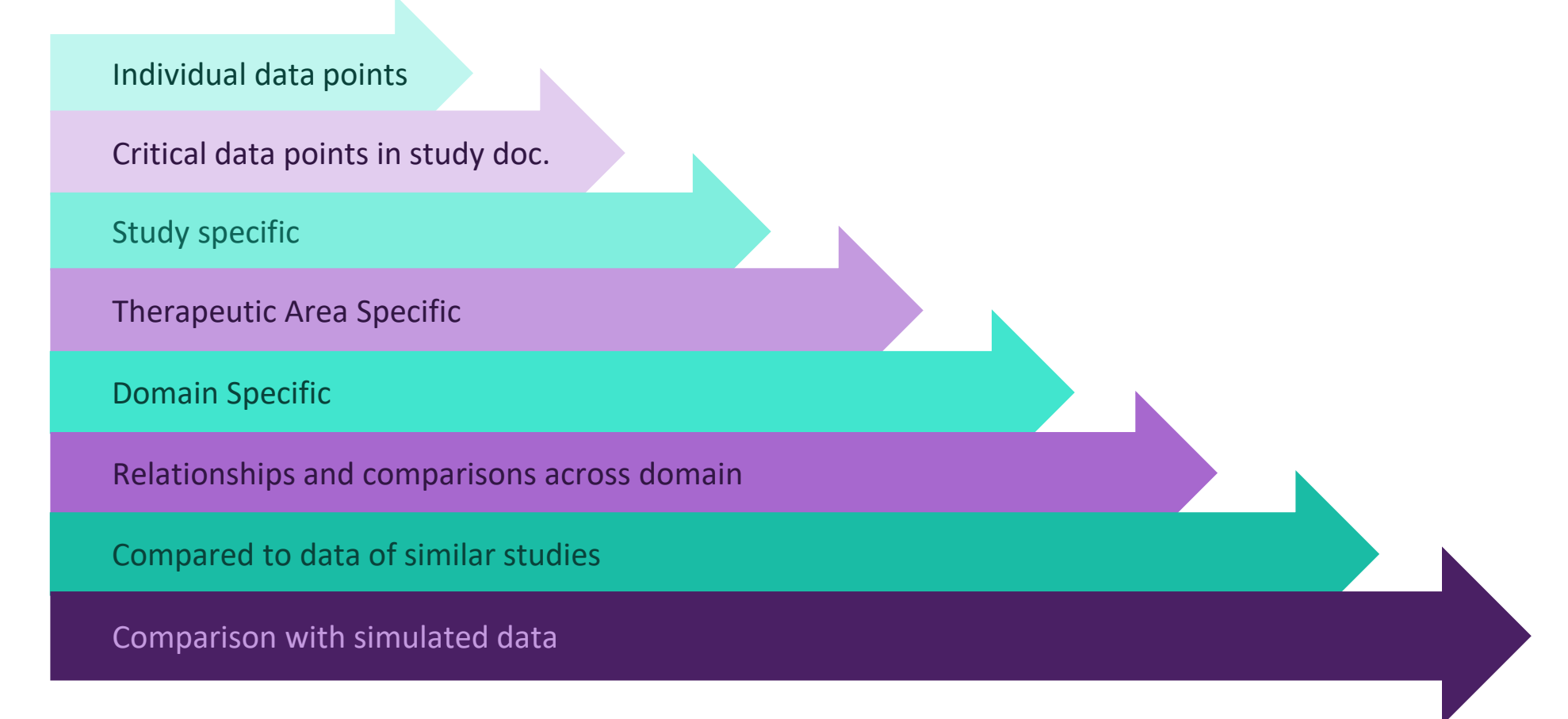


Validating assumptions enhances the model's interpretability and predictive accuracy and is a requirement prior to running the model. An example would be to check the underlining condition of normal residuals.

Methodologies: Data preparations and Validation

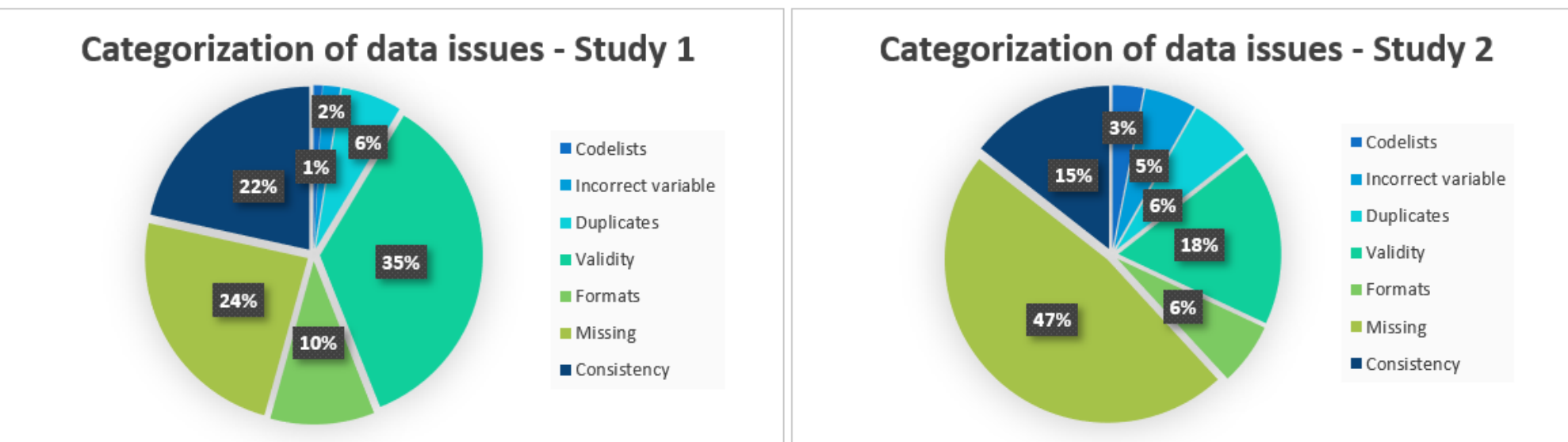
Adjusted Edit Checks

In accordance with study requirements, adjusted edit checks can be formulated and executed whenever new data becomes available, to identify potential data issues. These checks may be derived from protocol, specifications, study objectives and critical data points. It is crucial that appropriate checks are applied to critical analysis data points. The adjusted checks include the creation of code and macros and can be accomplished through a collaborative effort involving study stakeholders (especially DM), with reference to the CRF and data expectations.



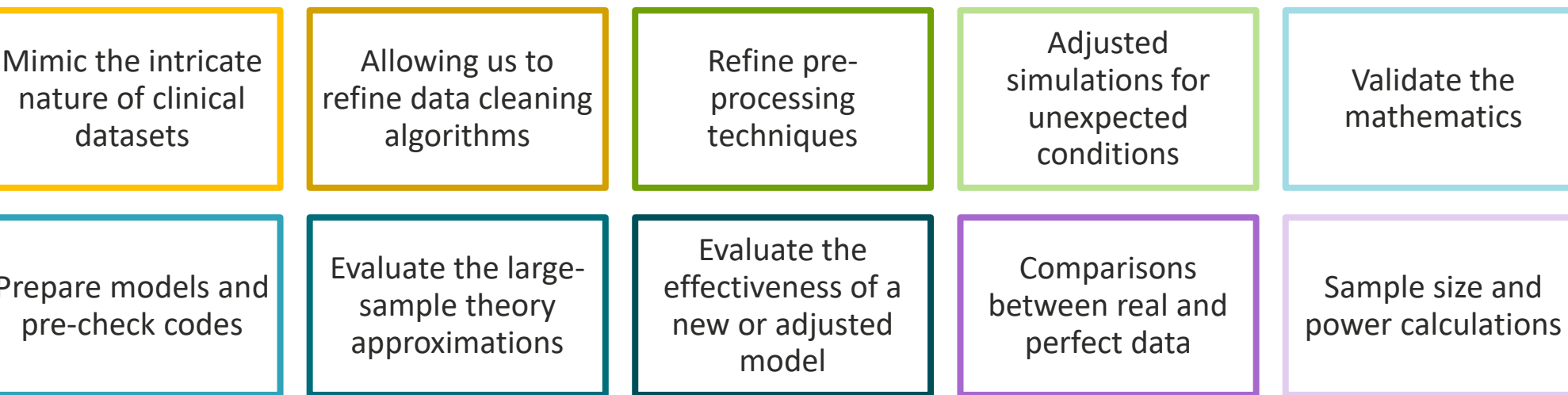
Data Observability

Data observability refers to the ability to monitor and understand the data throughout its lifecycle. It involves visualizing both real-time and historical data, to detect any anomalies, errors, or issues that may affect the integrity and reliability of the clinical trial results. Data observability is crucial for maintaining data quality and ensuring compliance with regulatory requirements ^[C].



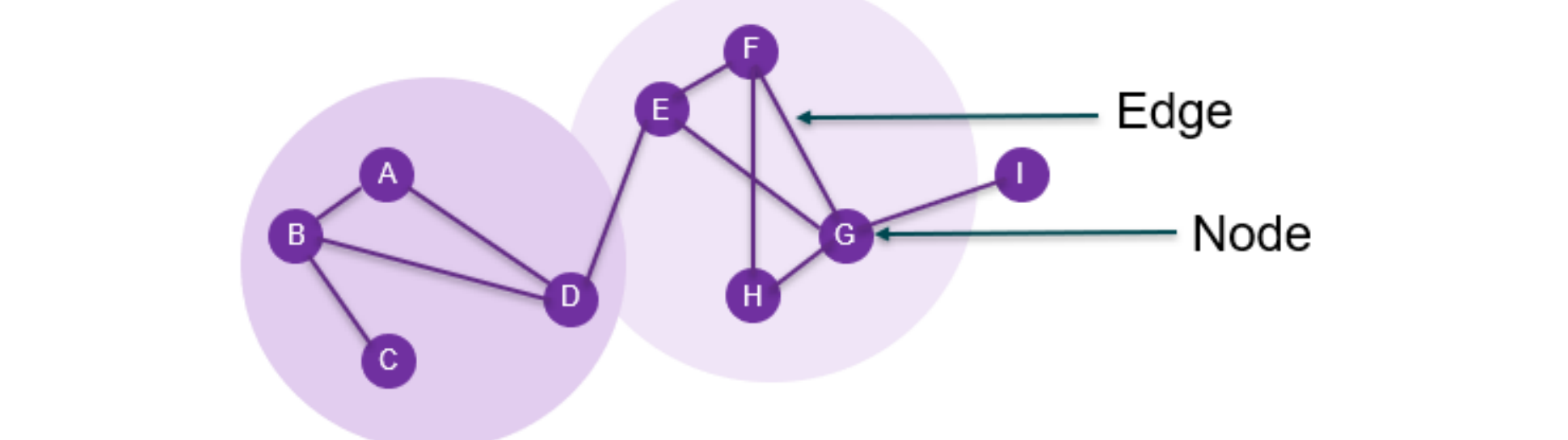
Simulations

When we refer to simulations, we in fact refer to computerized experiments, which uses simulated random sampling, to mimic data according to some predefined characteristics. SP programmers can utilize simulations in a variety of ways, starting with easily imitating real-world data scenarios. Being able to mimic real-world data, gives us the opportunity to evaluate and create model analysis without having to look at the actual study data, allowing us to stay blinded and start our investigations prior to data availability.

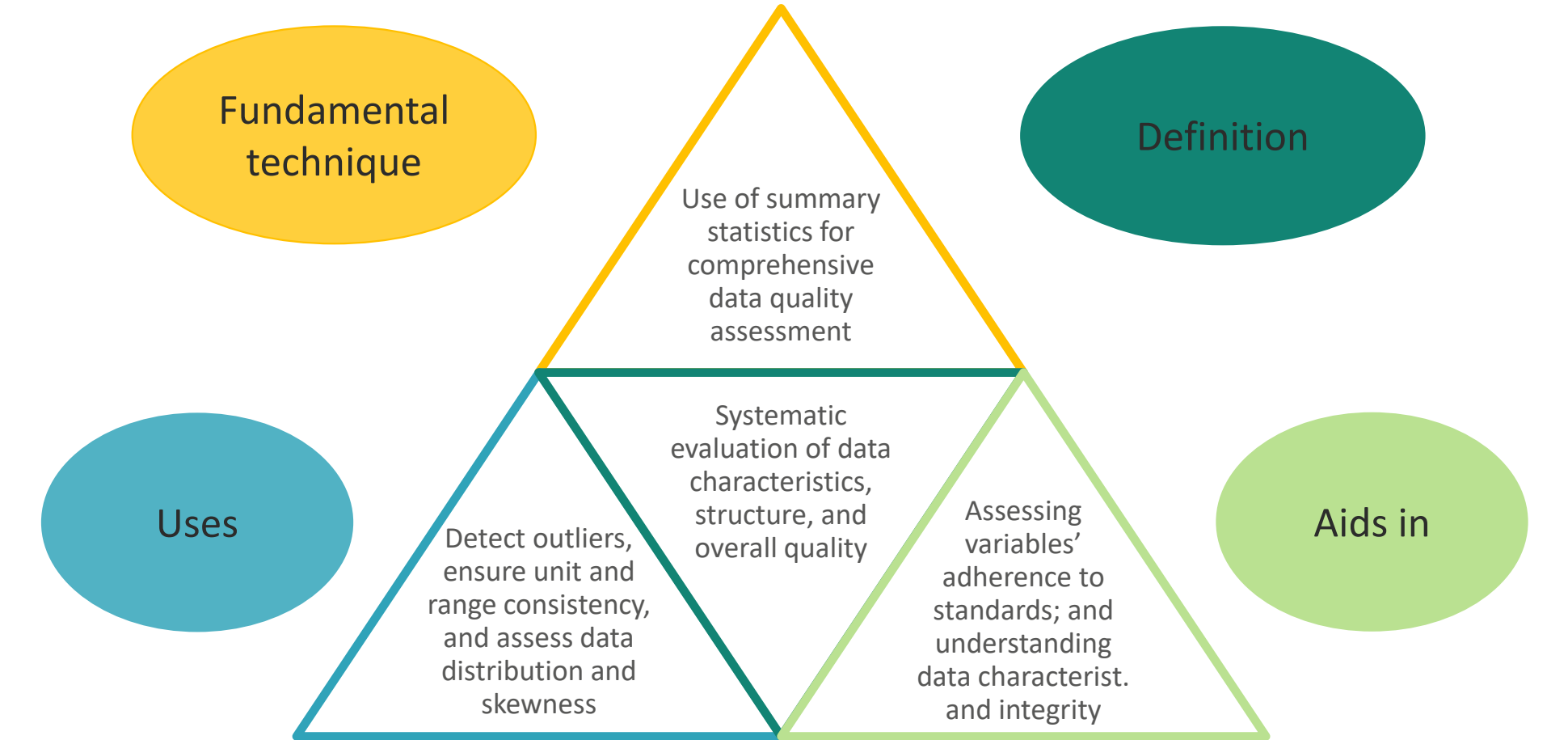


Detection Theory

The focus of detection theory is to determine relationships and patterns in data; and is optimally used along with other visualization techniques. Detection theory defines events/states as nodes, and edges as connections between them ^[B]. This method could be applied to adverse event data to identify unexpected events and assist in identifying noncompliance.

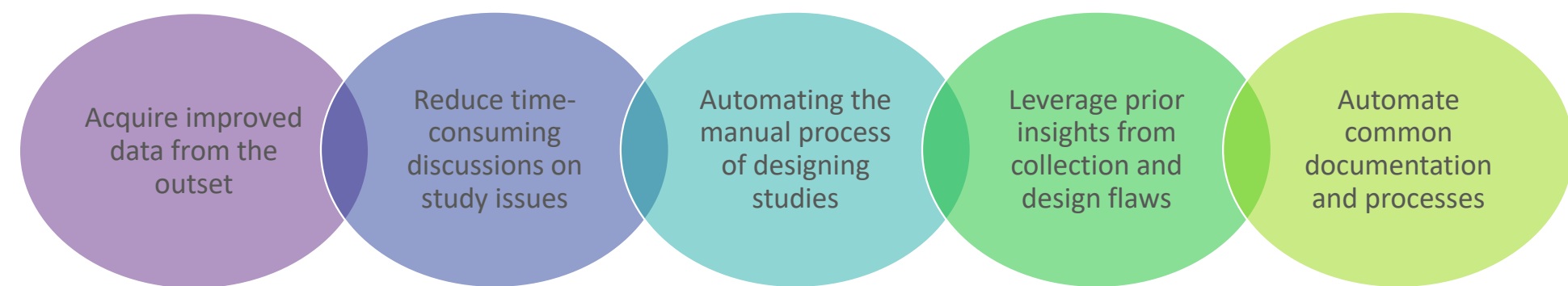


Data Profiling

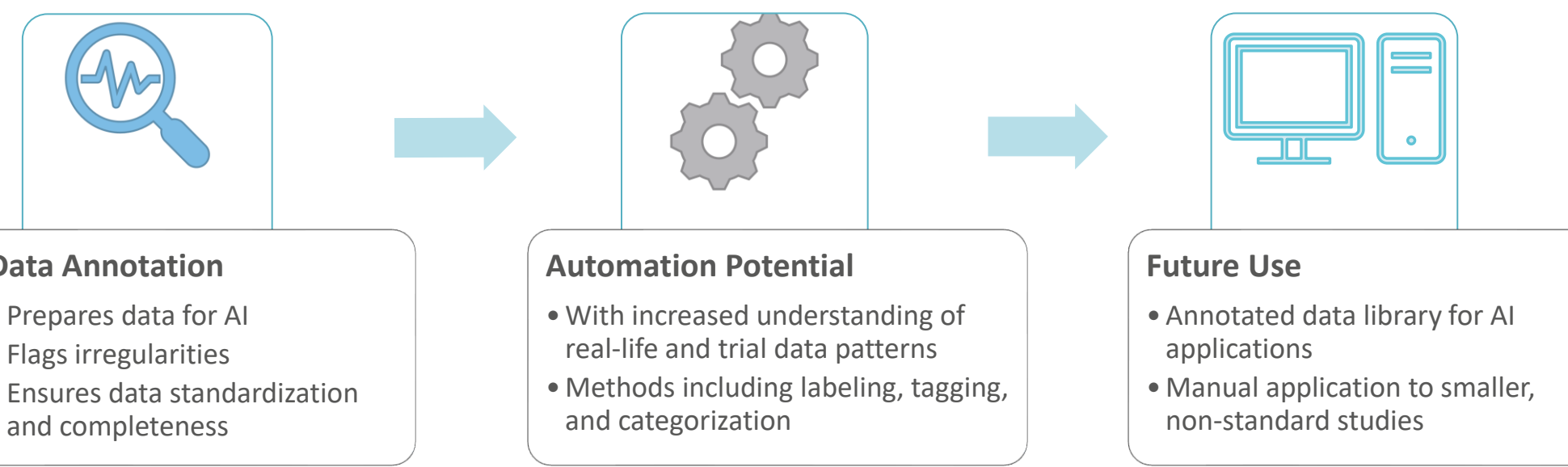


Data Modelling

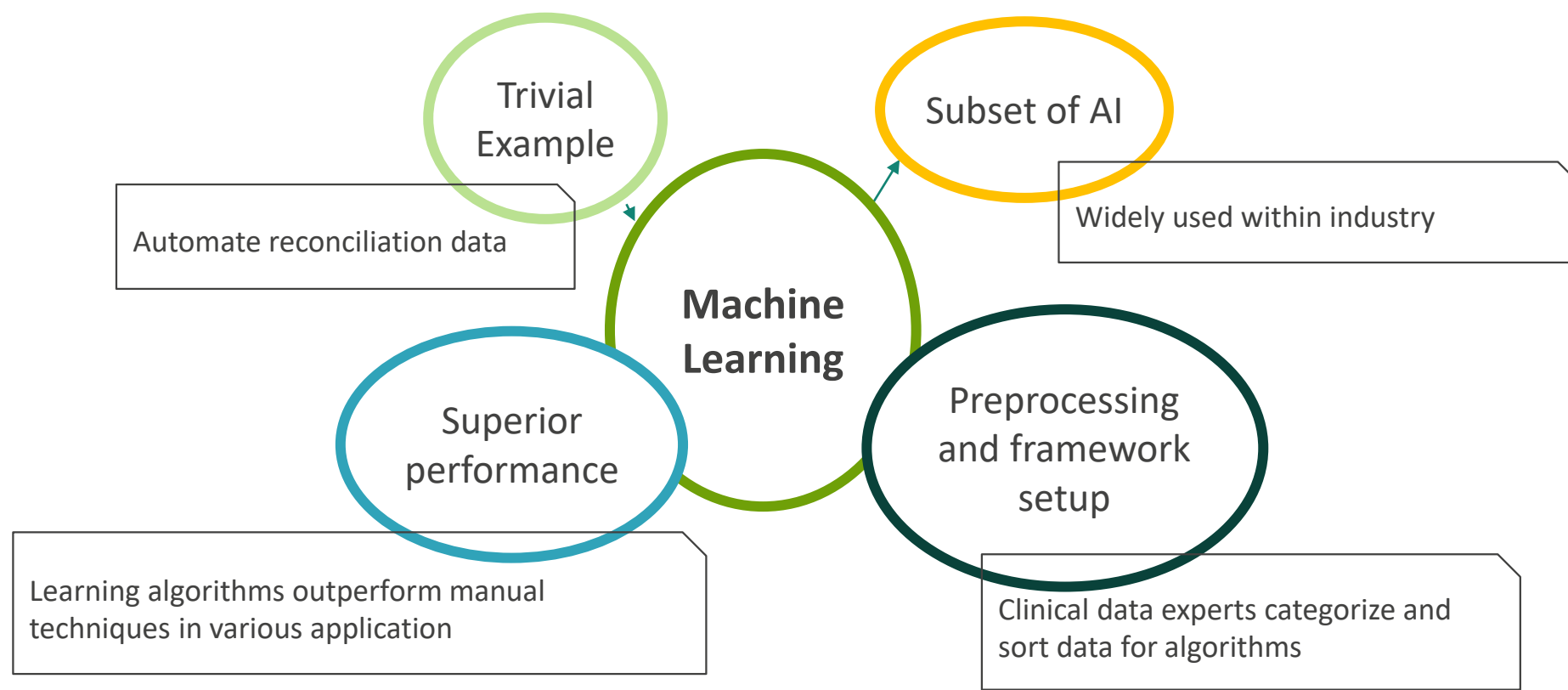
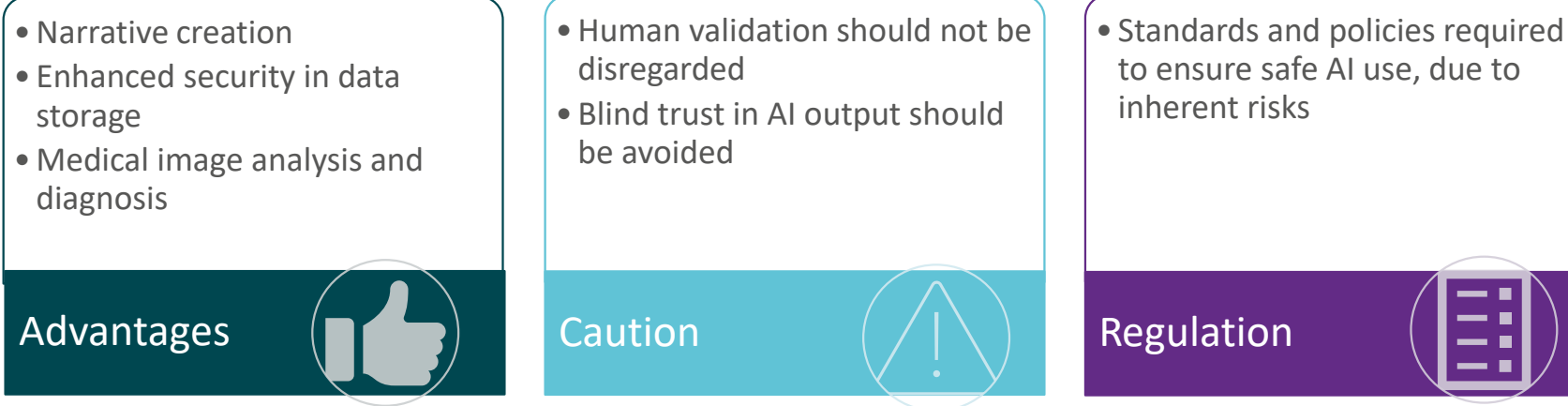
Data modelling is the process of evaluating and understanding data on a granular level before analysis. Benefits include but are not limited to:



Data Annotation



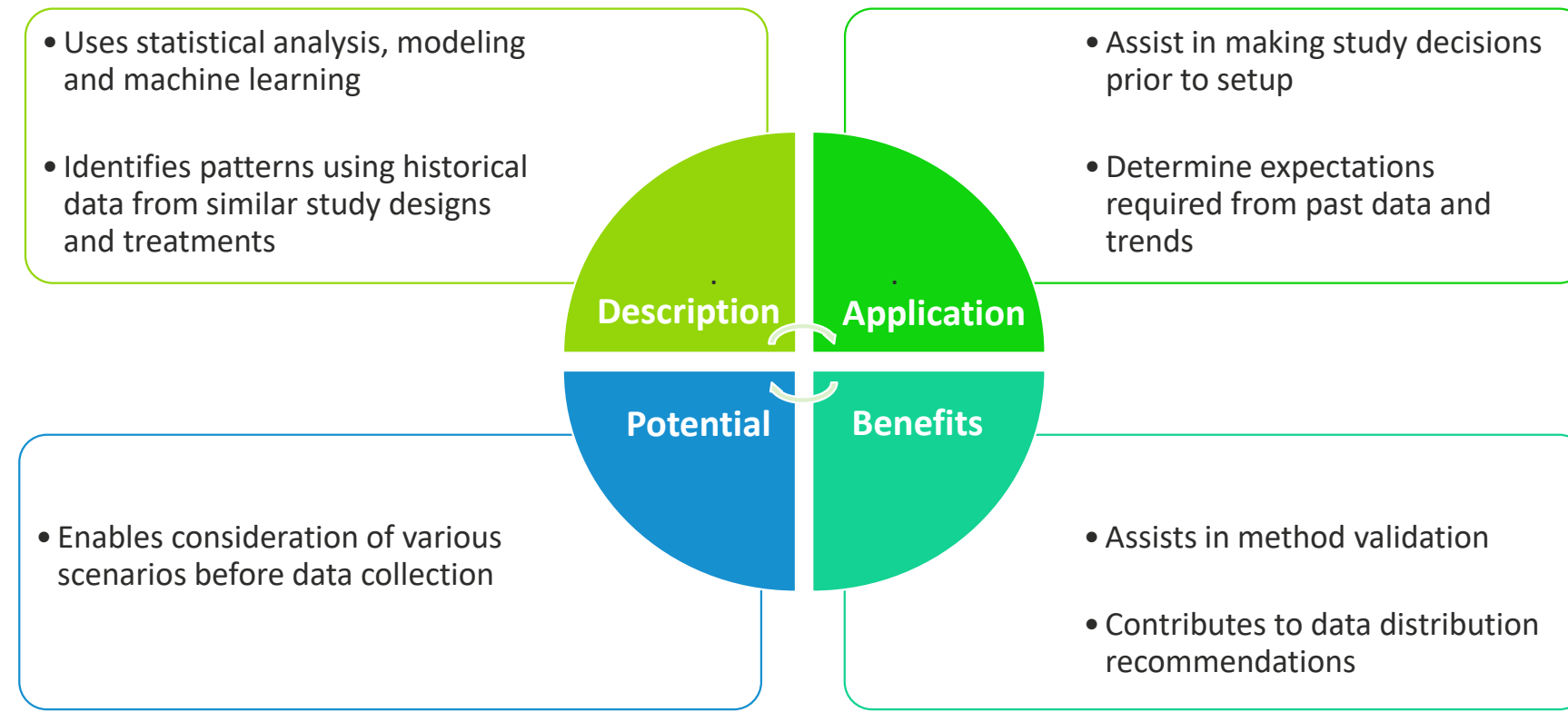
Artificial Intelligence (AI)



Visualization

Charts and figures are an easy way of detecting anomalies. We can visualize patterns and trends in data with time series plots; the distribution of data with density plots; PK concentration figures can assist in detection of concentrations for placebo drug or a non-zero concentration at baseline. While charts are very helpful for cleaning data, combining visualization with simulation can take data cleaning to the next level. In a study described in Knepper, D., Lindblad, A.S., *et al.* (2016), studies were simulated to detect potential fabricated data and noncompliance from sites ^[A].

Predictive Analytics



Conclusion



Our main roadblock is not the availability of resources but rather the unawareness and incentive to use them.

Acknowledgements

We would like to express our special thanks of gratitude to ICON Strategic Solutions management for their generous support and financial contribution to the poster and opportunity to present it at the PHUSE UK Connect Conference. In addition, we would like to thank our direct managers for their unconditional support and guidance.

References

[A] Knepper D, Lindblad A.S., *et al.*, 2016. Statistical Monitoring in Clinical Trials: Best Practices for Detecting Data Anomalies Suggestive of Fabrication or Misconduct. Therapeutic Innovation & Regulatory Science, Vol. 50, pp. 144-154.
[B] Miao, X., Wu, J., *et al.*, 2021. A comprehensive survey on graph anomaly detection with deep learning. IEEE Transactions on Knowledge and Data Engineering.
[C] Musa, A.E., 2022. What Is Data Observability, and Why Do You Need It? Towards Data Science. Available from: <https://towardsdatascience.com/what-is-data-observability-and-why-do-you-need-it-87463f054bea>
[D] Njeri, N.R., 2022. Data Preparation for Machine Learning Modelling. International Journal of Computer Applications Technology and Research, Vol. 11, pp. 231-235.
[E] Chantala, K., 2021. Cutting-Edge Techniques and SAS Macros to Clean, Prepare and Manage Data. SAS Global Forum Proceedings 2021, pp. 1177 - 2021.
[F] Manral, A., 2019. The Importance of Predictive Models in Clinical Trials. Nalashaa healthcare solutions. Available from: <https://blog.nalashaahealth.com/importance-predictive-models-clinical-trials/>
[G] Morris, T.P., White, I.R. & Crowther, M.J., 2019. Using simulation studies to evaluate statistical methods. Statistics in Medicine, Vol. 38, pp. 2074-2102.
[H] Noble H. & Heale, R., 2019. Triangulation in research, with examples Evidence-Based Nursing, Vol. 22, pp. 67-68.
[I] Naumann, F., 2014. Data Profiling Revisited. ACM SIGMOD, Vol. 42, pp. 40-49.
[J] Verma, J.P., & Abdel-Salam, G., 2019. Testing Statistical Assumptions in Research. John Wiley & Sons, Inc.