CRISP-DM Process – Retail Case
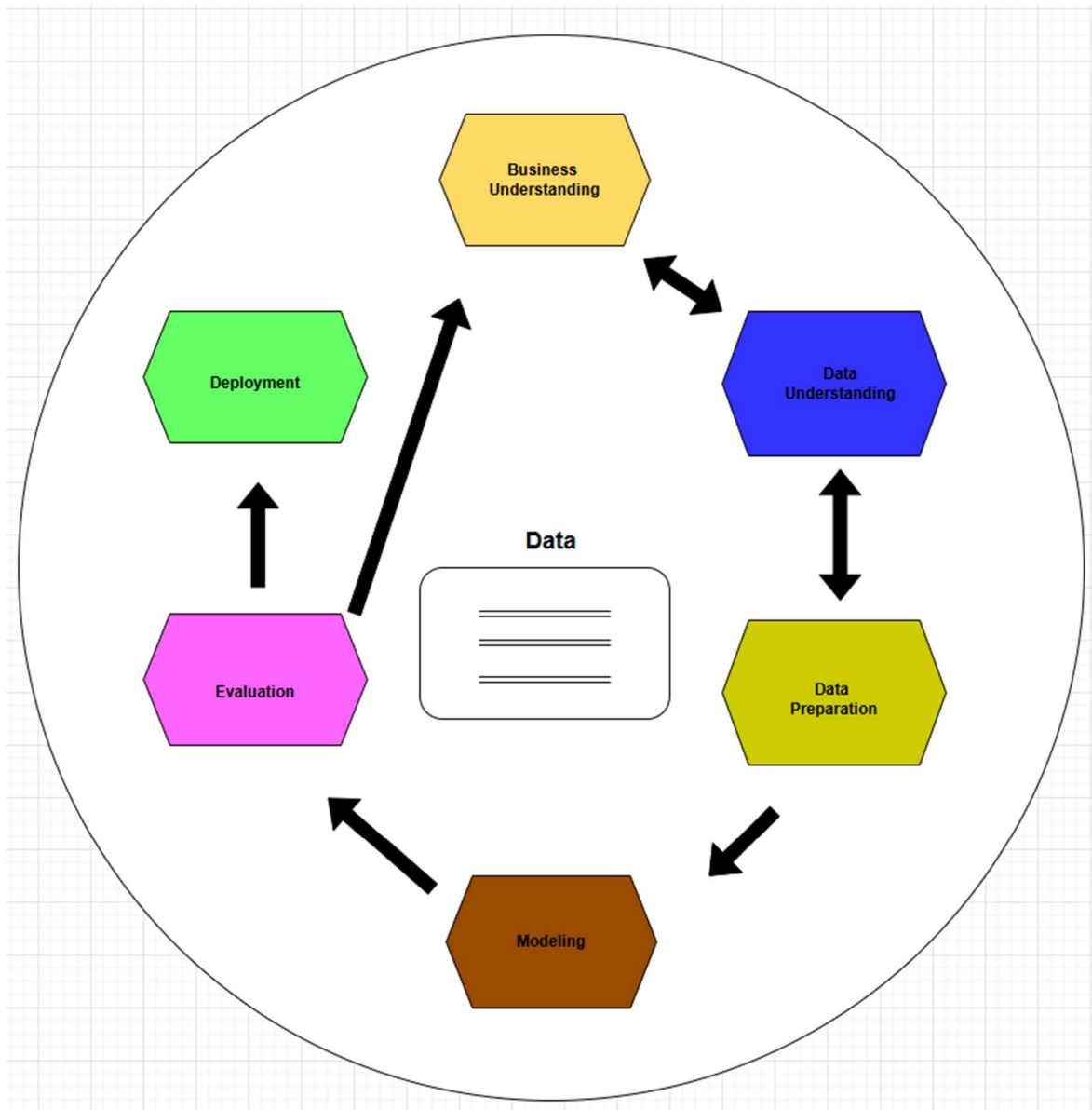

Business Understanding:

1. Determine business objectives.
    a. Increase sales during peak season by understanding customer buying behavior.
2. Assess situation.

    a. Review past seasonal sales and current marketing strategies to identify challenges.
3. Determine data mining goals.
    a. Use historical purchase data to predict which products to promote.
4. Produce project plan.
    a. Create a timeline and assign team roles for data collection, modelling, and deployment.

Data Understanding:

1. Collect initial data
    a. Gather sales transactions, customer loyalty data, store locations, and online orders.
2. Describe the data
    a. Summarize key fields like purchase date, product ID, customer ID, and store region.
3. Explore the data
    a. Look for missing values, outliers, or unexpected trends (e.g., sudden drops in sales).
4. Verify data quality
    a. Ensure data is complete, accurate, and consistent across different store branches.

Data Preparation:

1. Select data
    a. Choose data from the last 12 months and focus on top-selling product categories.
2. Clean data
    a. Remove duplicate transactions and correct errors in pricing or product IDs.
3. Construct data
    a. Create new fields such as "holiday season flag" or "total basket value per customer."
4. Integrate data
    a. Merge sales, inventory, and customer profile data into a unified dataset.

Modelling:

1. Select modeling techniques

    a. Choose decision trees and clustering to identify purchase behavior and customer segments.

2. Generate test design
   a. Split data into training and tests sets to evaluate model performance reliably.
3. Build model
   a. Train model to forecast product demand and segment customers by purchase patterns.
4. Assess model
   a. Check metrics like accuracy and precision to determine how well the model performs

Evaluation:

1. Evaluate results
   a. Test if the model correctly predicts demand for top-selling items during holidays.
2. Review process
   a. Double-check if all data preparation and modeling steps follow business requirements.
3. Determine next steps
   a. Decide whether to refine the model or move forward with deployment based on accuracy and business relevance.

Deployment

1. Plan deployment
   a. Outline how model insights will be used in inventory restocking and marketing campaigns.
2. Plan monitoring and maintenance
   a. Set up alerts and regular checks to ensure the model still performs well as new sales data arrives.
3. Produce final report
   a. Deliver a report detailing goals, methods, findings, and ROI projections to leadership.
4. Review project
   a. Hold a post-mortem to identify what went well, what didn't, and how future projects can improve.

Data Detective – Mini Retail Dataset

Customer ID	Purchase Amount	Payment Method	Product Category	Store Location	Loyalty Member
1001	54.99	Credit Card	Clothing	New York	Yes
1002	23.5	Cash	Accessories	Miami	No
1003	120	Credit Card	Electronics	New York	Yes
1004	35.25	Mobile Pay	Clothing	Chicago	No
1005	88.1	Cash	Electronics	Miami	Yes

- What are the instances? The instances are individual customer transactions (rows).
- What are the features? The features (columns) that describe each transaction are:
    1. Customer ID
    2. Purchase Amount
    3. Payment Method
    4. Product Category
    5. Store Location
    6. Loyalty Member
- Which features are categorical? Which are continuous?
    1. Categorical
        I. Payment Method (e.g., Credit Card, Cash)
        II. Product Category (e.g., Clothing, Electronics)
        III. Store Location (e.g., New York, Miami)
        IV. Loyalty Member (Yes/No)
    2. Continuous
        a. Purchase Amount (numeric currency value)
    3. ID-type (not used in modeling):
        a. Customer ID (an identifier, typically not a predictive feature)
- Is this data labeled or unlabeled? If the goal is to predict something like Loyalty member, Product Category, or Payment Method using the other features. These can serve as labels in a supervised learning task. If the goal is to group similar customers without a specific label using Clustering, then the data is unlabeled data.