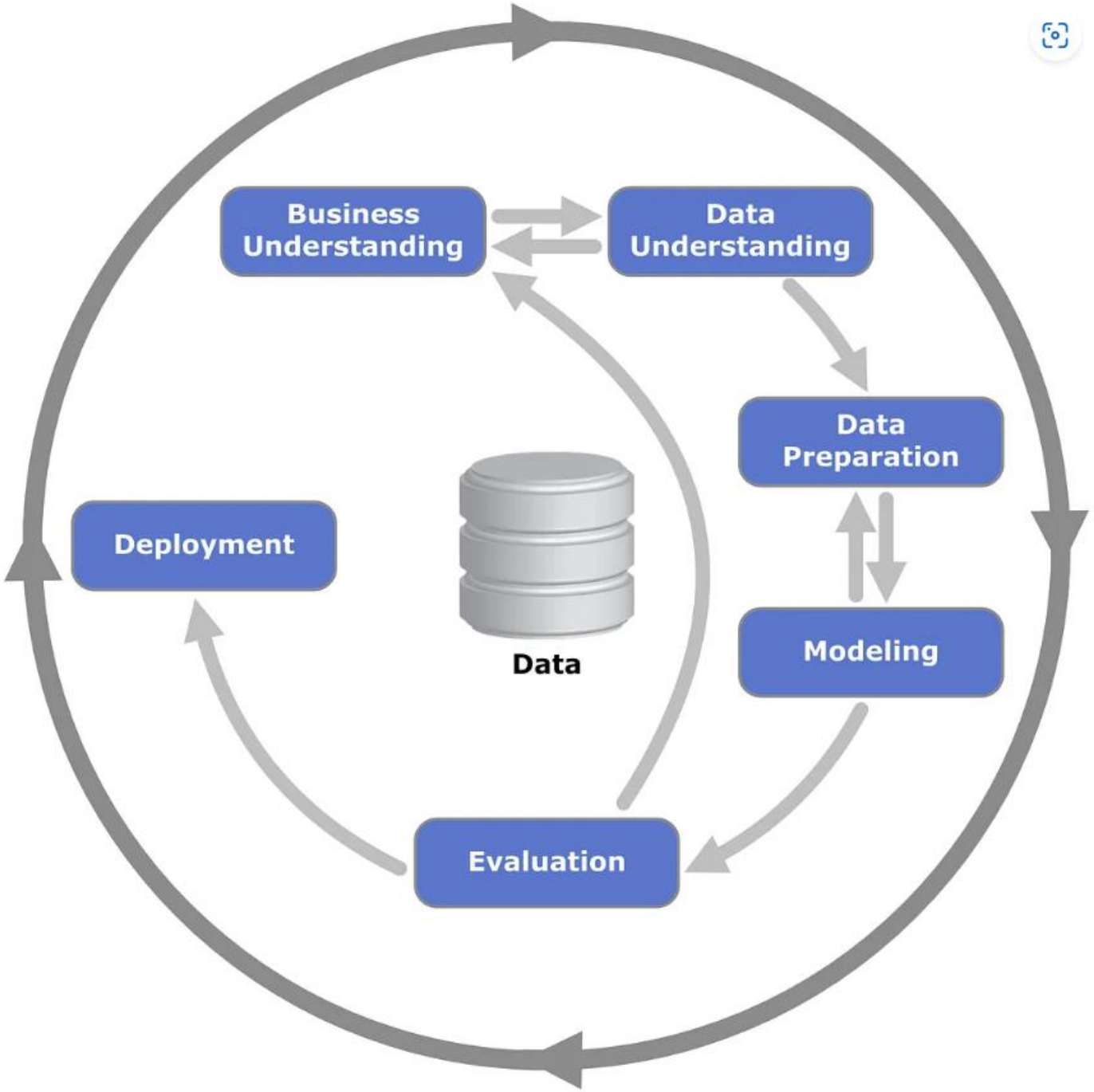
Data Science Workflow

It defines the phases (or steps) in a data science project.

It helps you plan, organize and implement your data science project

CRISP-DM (Cross Industry Standard Process for Data Mining)



CRISP-DM (Cross Industry Standard Process for Data Mining)

Business Understanding

- **Objective:** Define the problem and project goals from a business perspective.

Key Questions:

- What are we trying to achieve?
- What is the business impact of solving this problem?

Deliverables: A clear problem statement, success criteria, and project plan.

Example: A retail company wants to predict customer churn to improve retention rates. The goal might be to reduce churn by 10% in six months.

CRISP-DM (Cross Industry Standard Process for Data Mining)

Data Understanding

- **Objective:** Collect initial data, explore it, and identify issues.

Steps:

- Analyze data sources and quality.
- Perform exploratory data analysis (EDA).
- Document any gaps or anomalies.

Tools: Python (Pandas, Matplotlib, Seaborn), R, SQL.

Example: Analyze historical customer data to understand patterns like purchase frequency and churn behavior.

CRISP-DM (Cross Industry Standard Process for Data Mining)

Data Preparation

- **Objective:** Transform raw data into a usable format for modeling.

Steps:

- Clean and preprocess data (e.g., handling missing values, removing duplicates).
- Feature engineering.
- Split data into training, validation, and test sets.

Deliverables: A finalized dataset ready for modeling.

Example: Standardize numerical variables, encode categorical features, and create new features like “lifetime value.”

CRISP-DM (Cross Industry Standard Process for Data Mining)

Modeling

- **Objective:** Select and apply machine learning algorithms.

Steps:

- Choose algorithms based on the problem type (e.g., classification, regression).
- Train and tune models.
- Optimize hyperparameters.

Tools: Scikit-learn, TensorFlow, PyTorch, XGBoost.

Example: Train a random forest classifier to predict churn probability.

CRISP-DM (Cross Industry Standard Process for Data Mining)

Evaluation

- **Objective:** Assess the model's performance and ensure it meets business objectives.

Steps:

- Compare model metrics (e.g., accuracy, precision, recall) to success criteria.
- Validate against unseen data.
- Conduct a cost-benefit analysis.

Deliverables: A recommendation to proceed with or refine the model.

Example: The random forest model achieves 85% precision and recall, surpassing the 80% target.

CRISP-DM (Cross Industry Standard Process for Data Mining)

Deployment

- **Objective:** Deliver the results to stakeholders and integrate them into business processes.

Steps:

- Create dashboards or APIs for predictions.
- Monitor model performance in production.
- Gather feedback for improvement.

Deliverables: A deployed solution and documentation.

Example: Deploy the churn prediction model to alert customer success teams when high-risk customers are identified.