Introduction to Data Science

What is Data Science?

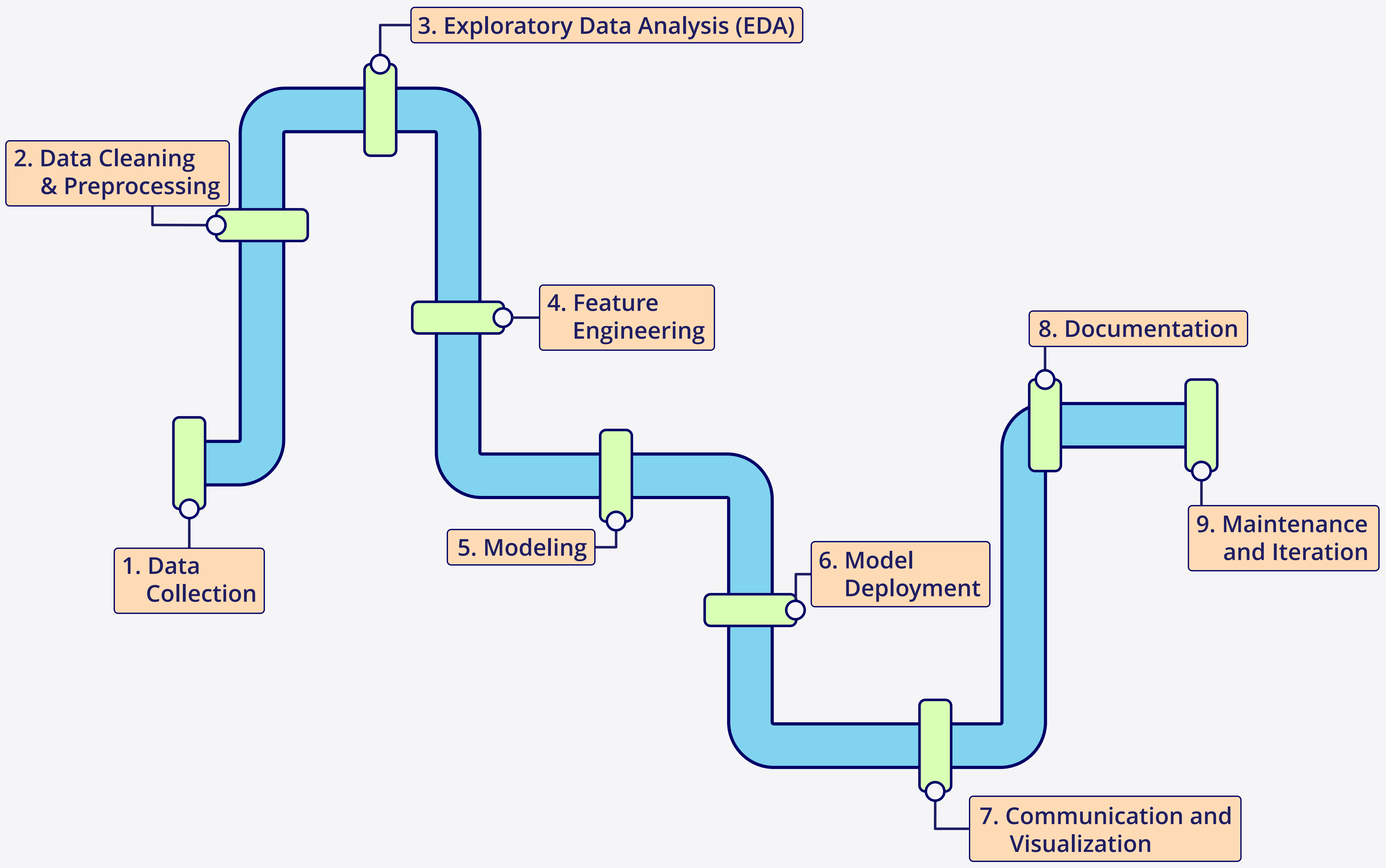
The multidisciplinary area of data science applies scientific procedures, systems, algorithms, and methodologies to organize the unstructured data to derive valuable knowledge and insights.

What is a Data Science Pipeline?

A data science pipeline is a collection of procedures and instruments for collecting, processing, analyzing, and visualizing data so that important insights can be drawn and choices can be made. Typically, a pipeline has multiple phases, each with a distinct set of tools and duties.

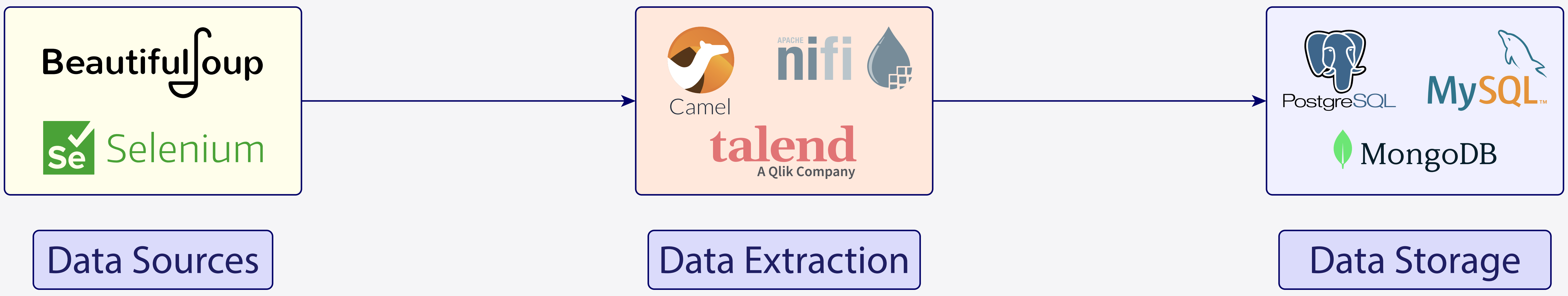
Let's explore the key stages involved in the data science pipeline.

Data Science Pipeline


3. Exploratory Data Analysis (EDA)
2. Data Cleaning & Preprocessing
4. Feature Engineering
8. Documentation
1. Data Collection
5. Modeling
6. Model Deployment
9. Maintenance and Iteration
7. Communication and Visualization

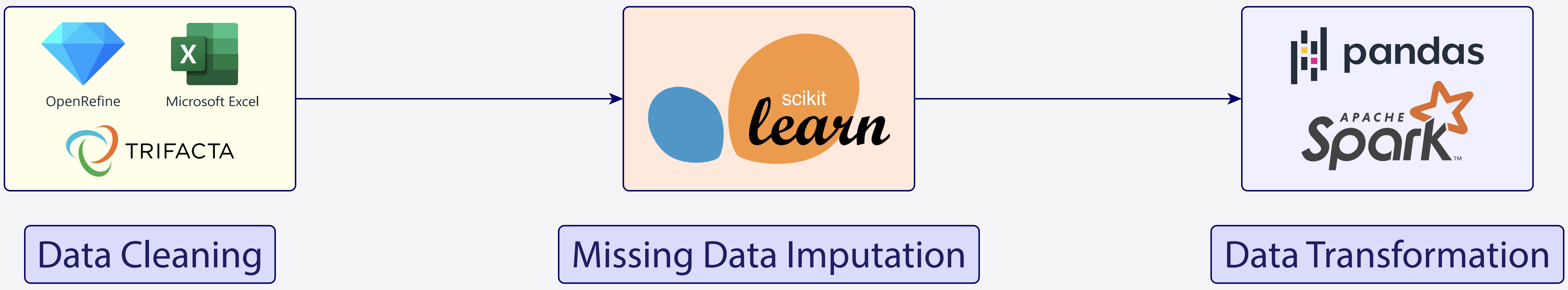
1 Data Collection

- Determine the sources of the data. Find the location of the pertinent data, whether in files, databases, APIs, or other sources.
- Gather information from multiple sources and put it in a format that can be examined.



2 Data Cleaning and Preprocessing

- Handle missing values.
- Clean the data by removing inconsistencies, outliers, or errors.
- Transform the data into a suitable format.



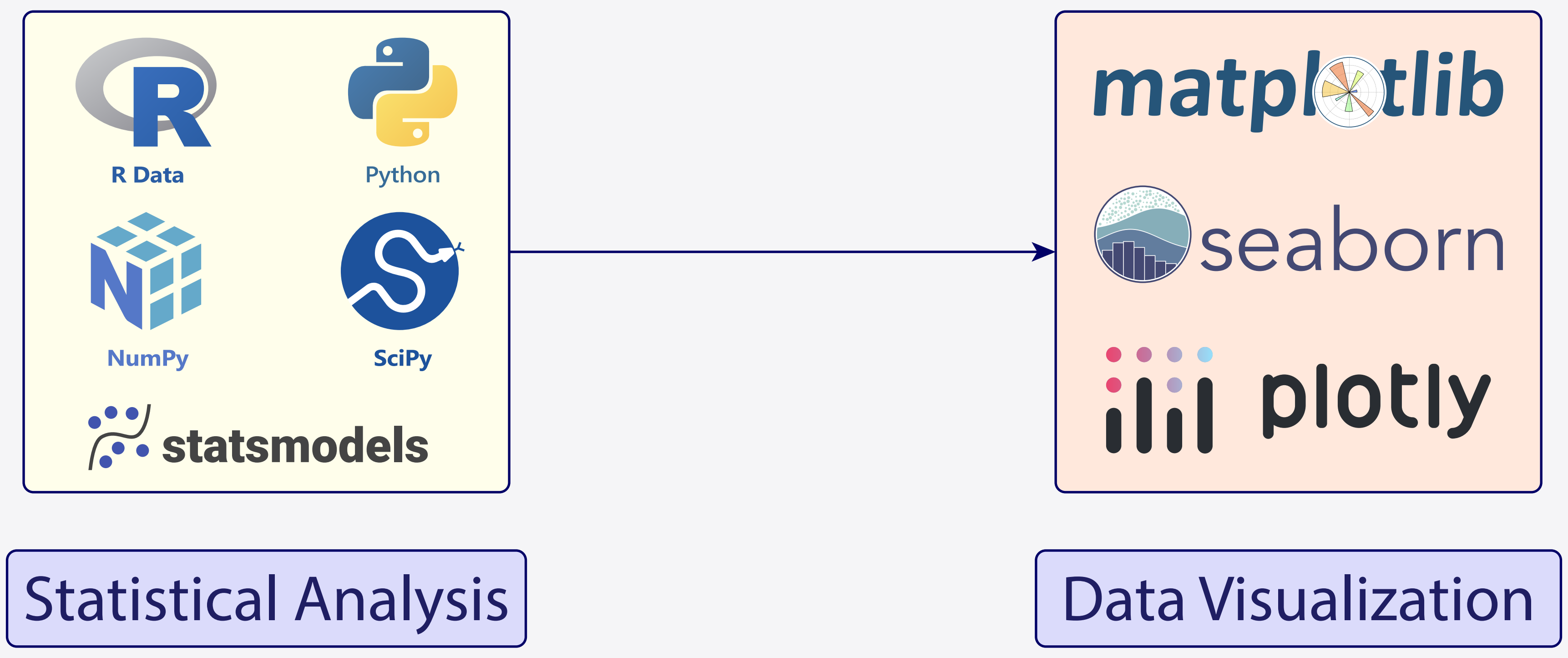
Data Cleaning

Missing Data Imputation

Data Transformation

3 Exploratory Data Analysis (EDA)

- Explore the dataset to get insights into the structure, distribution, and relationship between variables by statistical analysis.
- Visualize the data by creating multiple plots to understand the patterns and trends in the data.

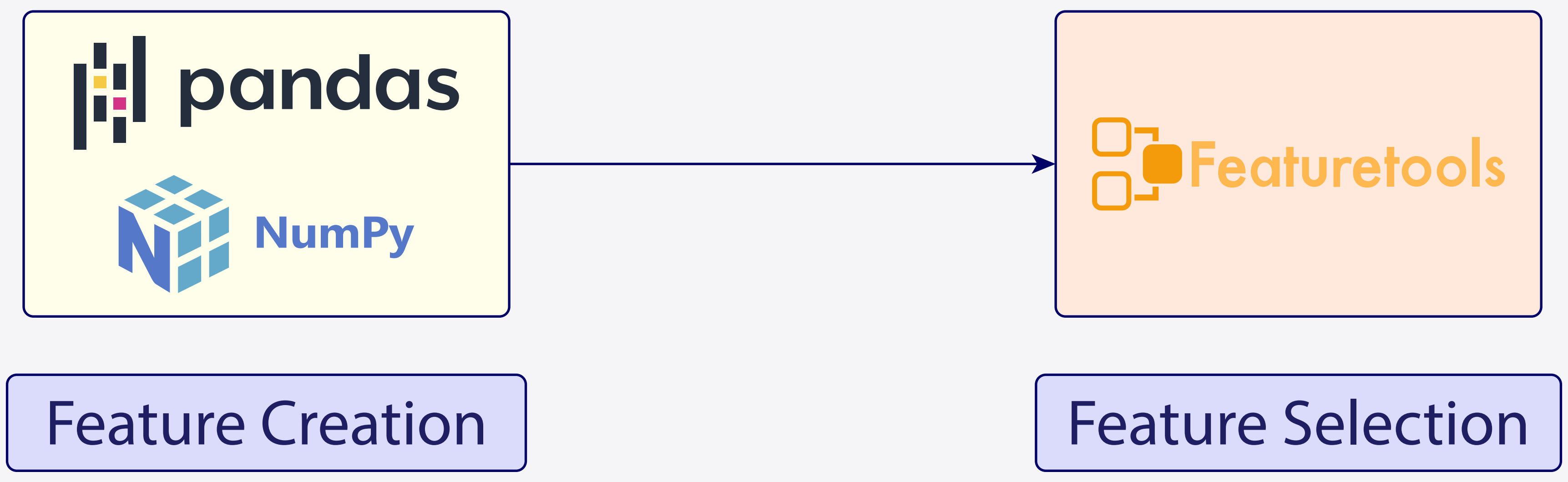


Statistical Analysis

Data Visualization

4 Feature Engineering

- Generate new features to enhance the predictive power of the model.
- Choose the important features to enhance the model's accuracy.

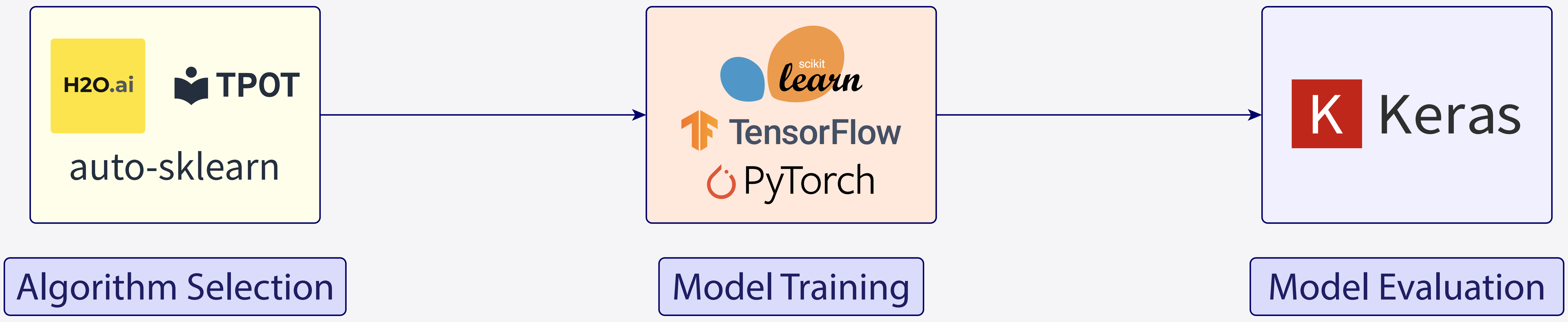


Feature Creation

Feature Selection

5 Modeling

- Select a machine learning model based on the nature of the problem, e.g., classification, regression, clustering, etc.
- Use the historical data to train the model.
- Assess the model's performance using accuracy, precision, recall, or F1 score metrics.



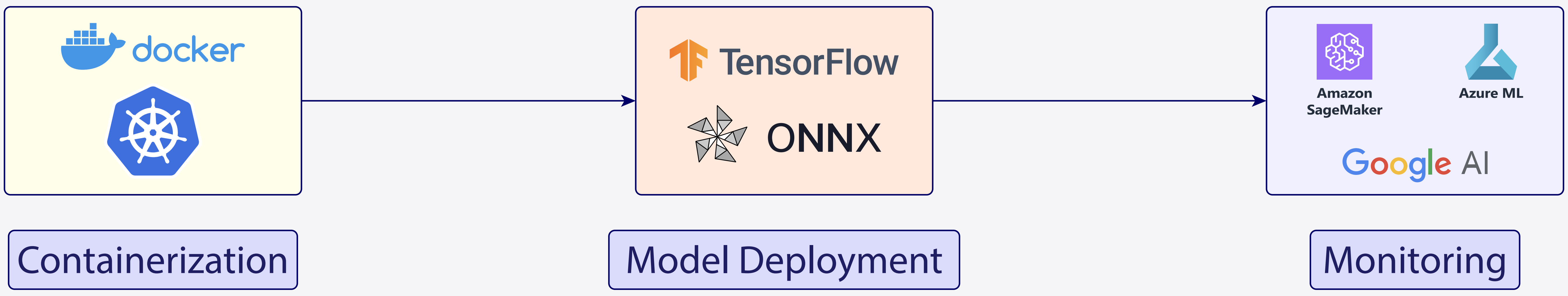
Algorithm Selection

Model Training

Model Evaluation

6 Model Deployment

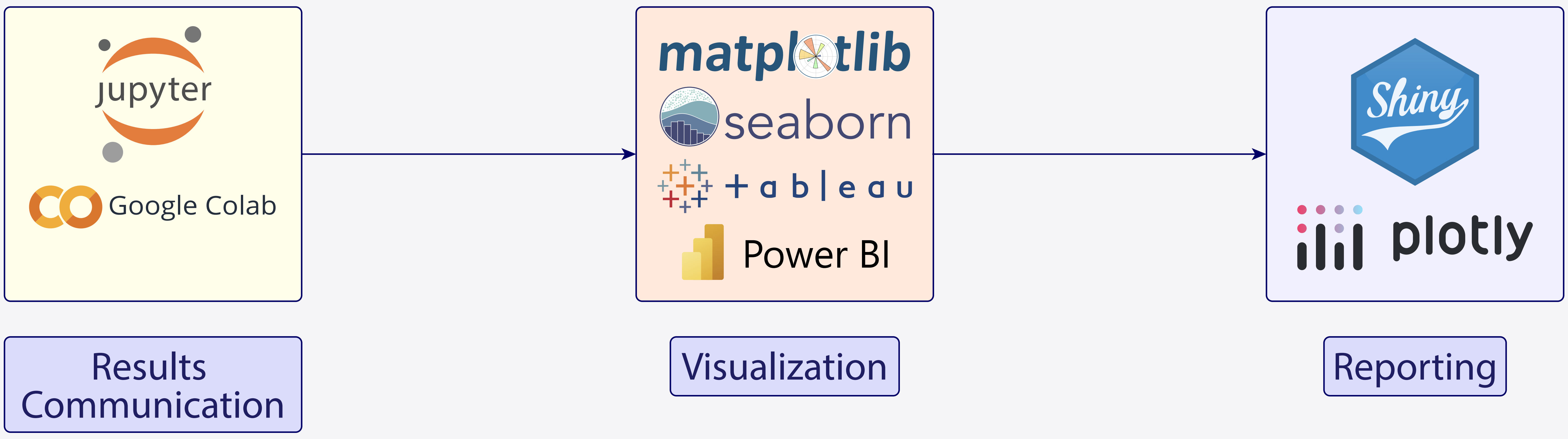
- Deploy the model to make real-time predictions.
- Continuously monitor the model’s performance and retain it if necessary.



Containerization → Model Deployment → Monitoring

7 Communication and Visualization

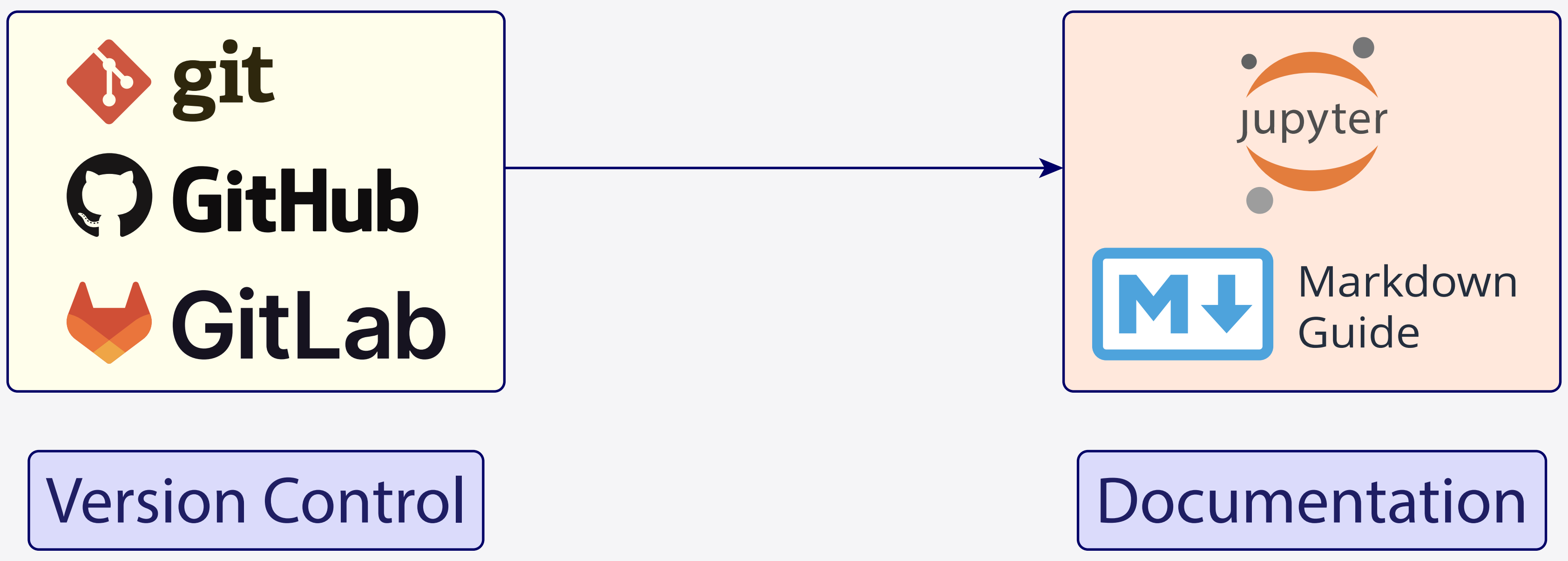
- Communicate the findings, results, and insights to the stakeholders.
- Visualize the results using charts, graphs, and dashboards to convey complex information.



Results Communication → Visualization → Reporting

8 Documentation

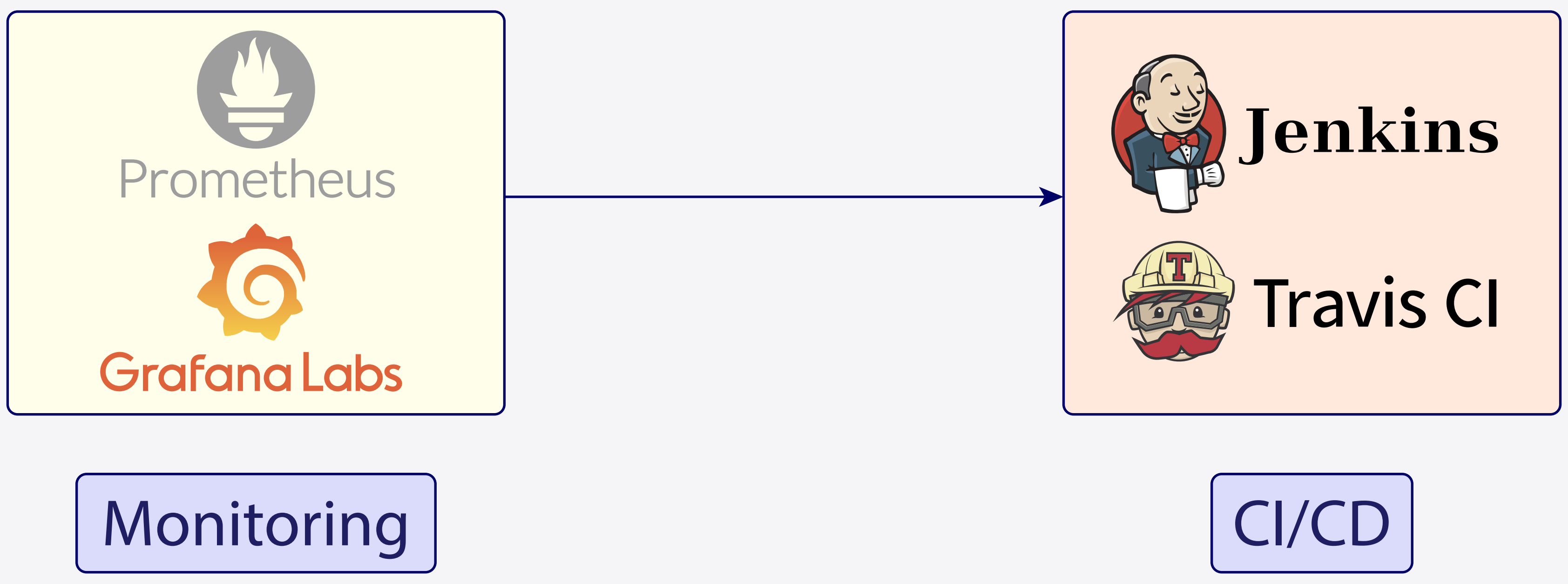
- Document the complete process and maintain documents for all stages of data science.



Version Control → Documentation

9 Maintenance and Iteration

- Retrain the model with new data to ensure the accuracy of the model.
- Continuously improve the pipeline based on feedback and changing requirements.



Monitoring → CI/CD