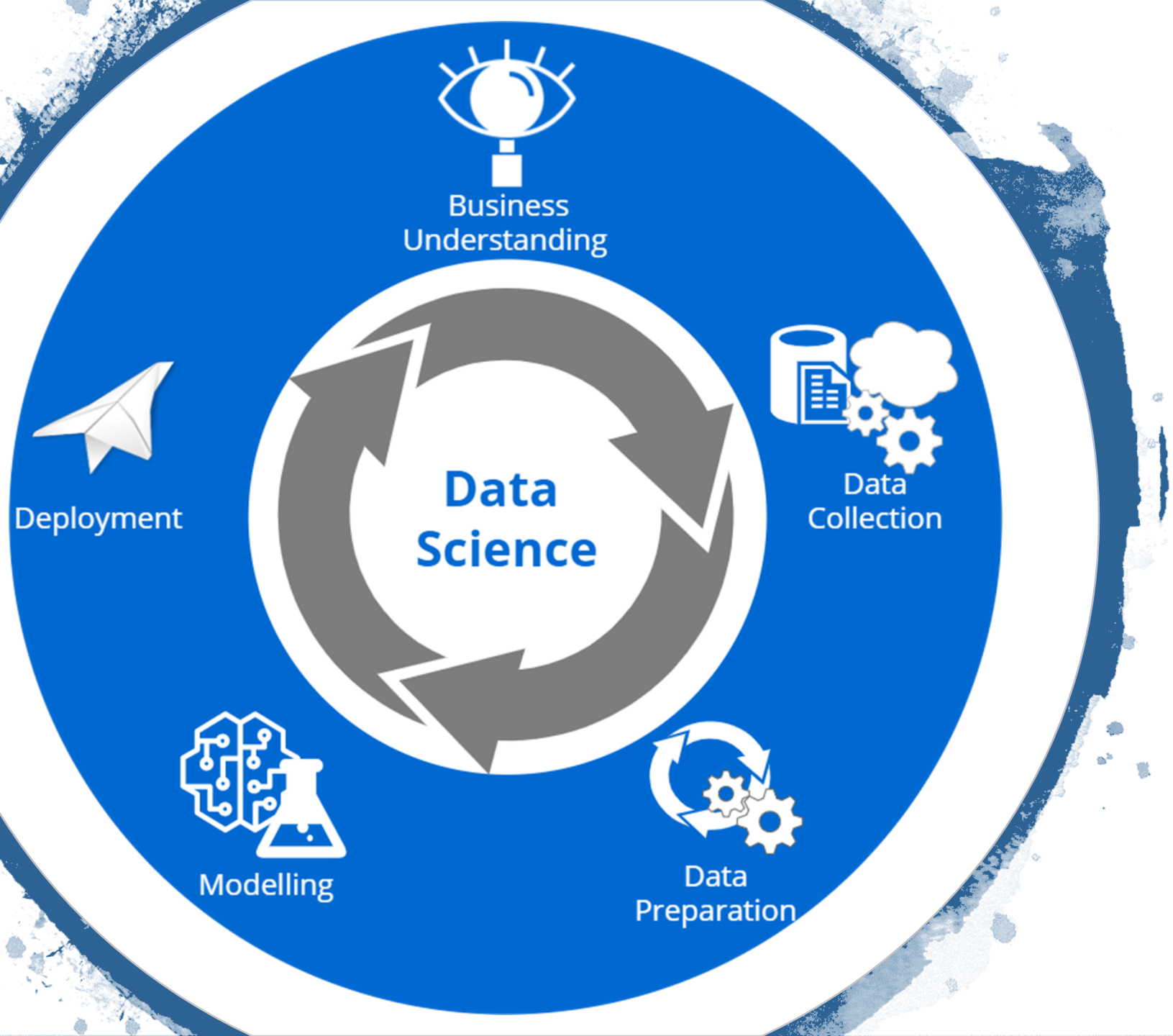
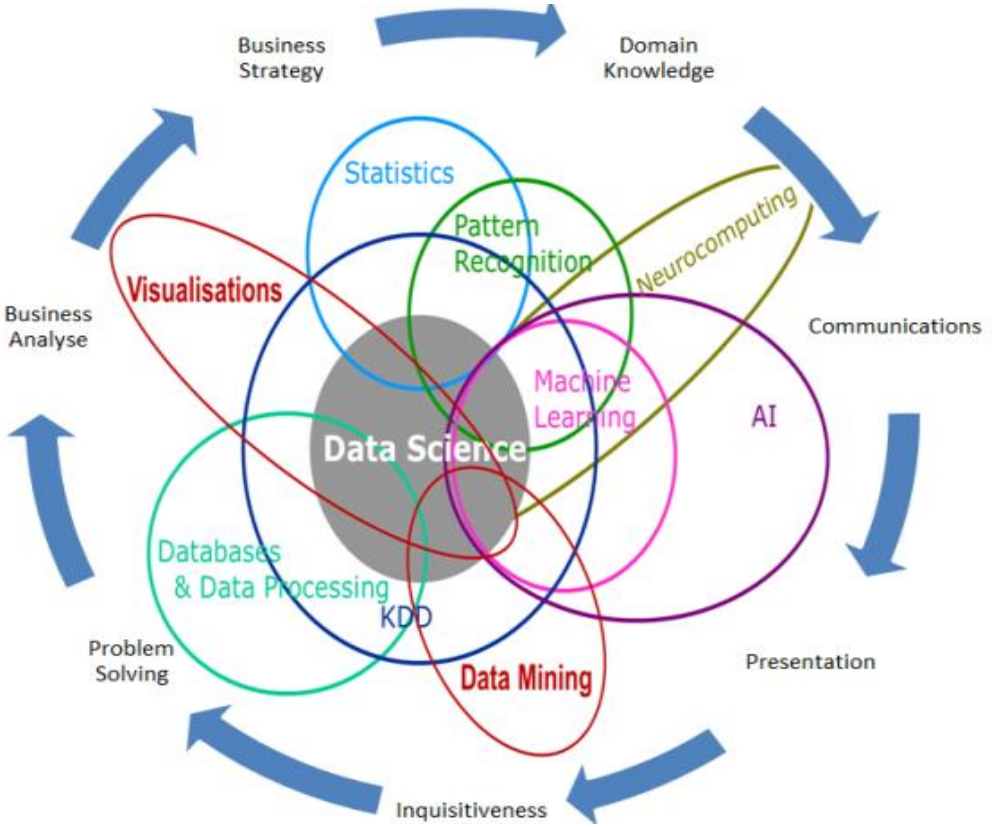

Business
Understanding
Data
Collection
Data
Preparation
Modelling
Deployment
Data
Science

# What is Data Science?

Data science is a multidisciplinary blend of **data inference, algorithmm development, and technology** in order to solve analytically complex problems.



# What is not Data Science?

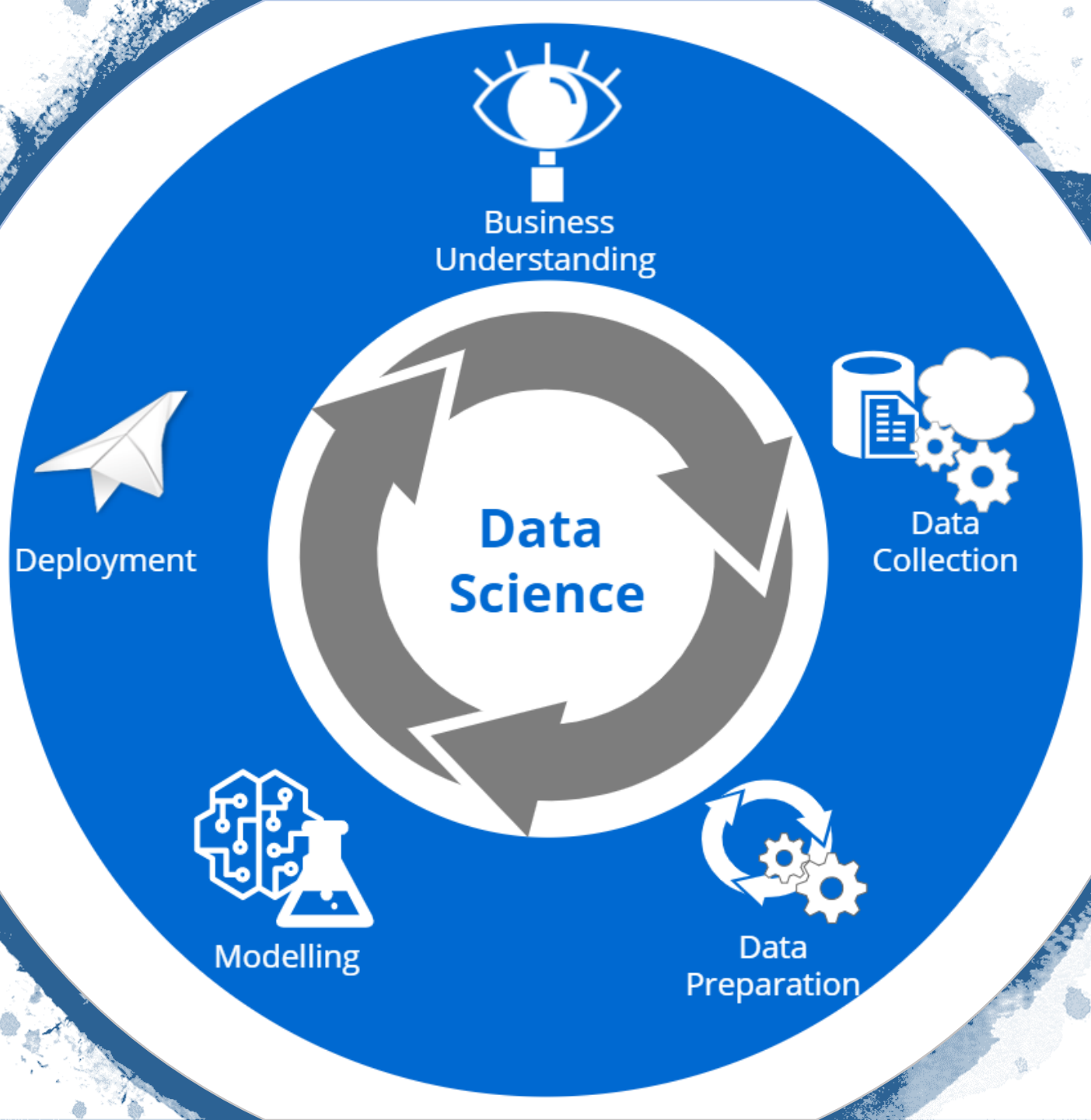
**Data Science ≠ Business Intelligence**
**Data Science ≠ Big Data**
**Data Science ≠ Machine Learning**

**Machine Learning is not a branch of Data science.** Machine Learning originated from Artificial Intelligence. Data science is only using ML as a tool. The reason is that it produces amazing and autonomous results for specific tasks

**It's not the salvation of companies that never measured anything and now want to get insights from their data**. "Garbage in, garbage out" Data science will be as good as the data generated on the following years.

**Just present data using some Excel charts** without any insight about the data.

# Lifecycle


Business Understanding
Data Collection
Data Preparation
Modelling
Deployment
Data Science

# Business Understanding

**Goals**

Specify key variables

Identify data sources

**How to do it**

- Identify the ***key business variables*** that the analysis needs to predict
- Define the ***project goals*** by asking and refining "sharp" questions that are relevant, specific, and unambiguous
- Define:
  - project team
  - roles & responsibilities
- Define the success ***metrics*** -SMART-
- Find the relevant data

**Artefact**

Charter document

Data dictionaries

List data sources

Data Collection

**Goals**

Clean and high quality data

Architecture of the data pipeline

**How to do it**

- ***Ingest the data*** into the target analytic environment
- ***Explore the data*** to determine if the data quality is adequate to answer the question
- ***Set up a data pipeline*** to score new or regularly refreshed data

**Artefact**

Data quality report

Solution architecture

Data Preparation

**Goals**

Data Analysis

Optimal data features

**How to do it**

- ***Handling missing data***
- ***Correcting invalid values***
- ***Removing duplicates***
- ***Structuring the data to be fed into an algorithm***
- ***Feature engineering***: Create data features from the raw data to facilitate model training

**Artefact**

Analysis report

Feature sets

Checkpoint decision

Modelling

**Goals**

Machine learning model

**How to do it**

- ***Model training:*** Find the model that answers the question most accurately by comparing their success metrics
  - *Split the input data*
  - *Build the models by using the training data set*
  - *Evaluate the training and the test data set*
  - *Determine the "best" solution*
- Determine if your model is ***suitable for production***

**Artefact**

Machine learning model

Model report

Checkpoint decision

Deployment

**Goals**

Deploy model

**How to do it**

- ***Operationalize the model:*** Deploy the model and pipeline to a production or production-like environment for application consumption

**Artefact**

Deployment details

solution architecture document

Business

**Goals**

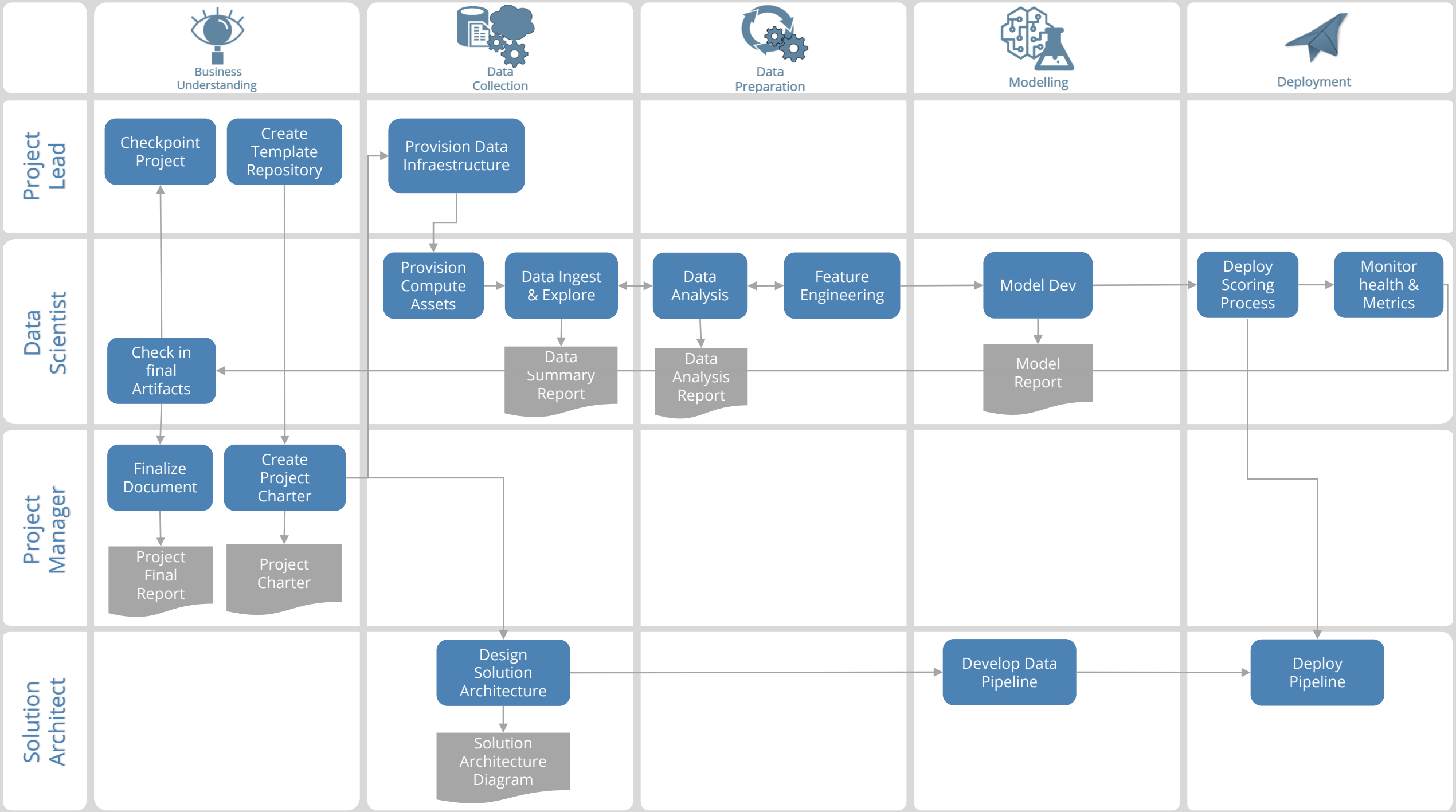
Finalize the project deliverables

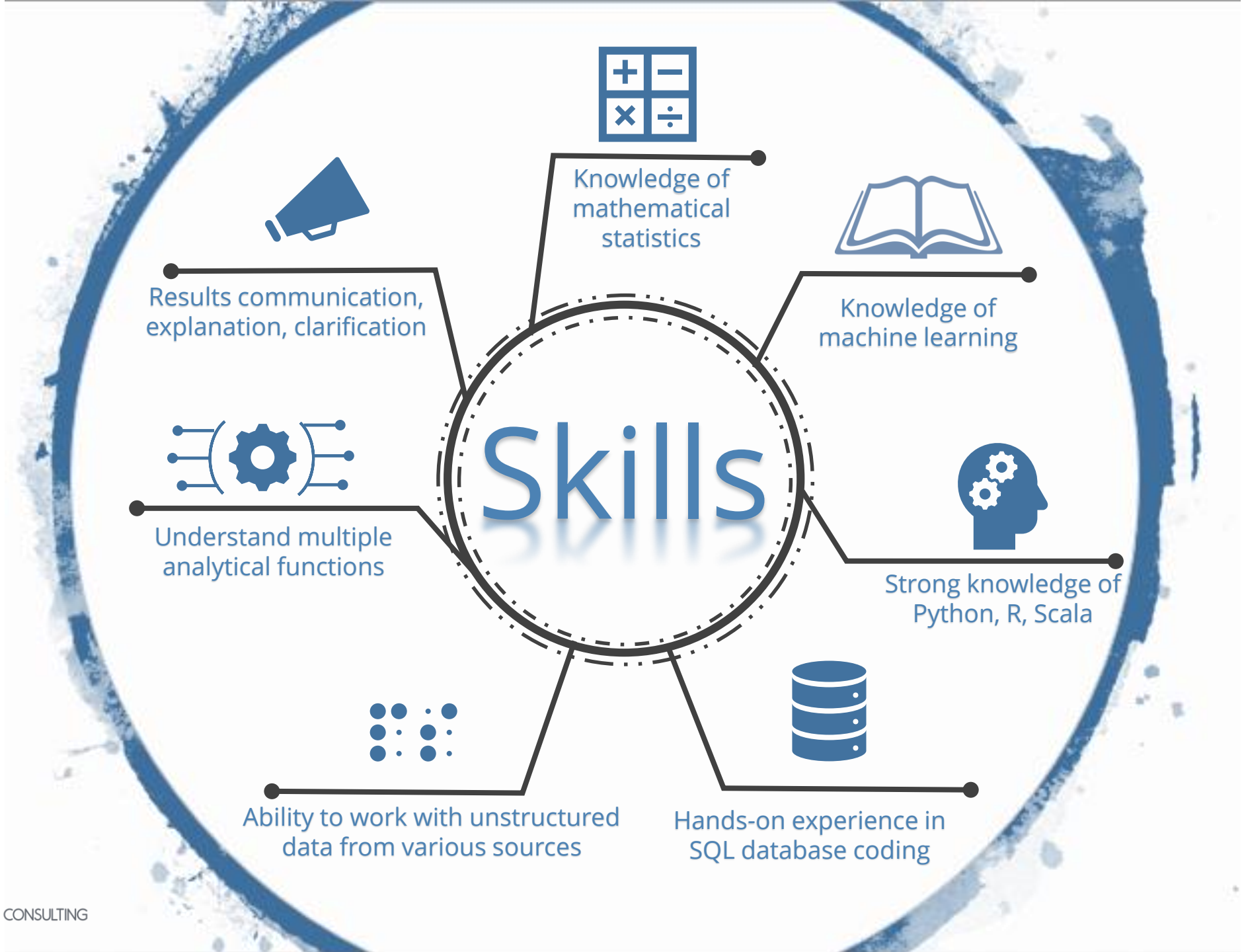
**How to do it**

- ***System validation:*** Confirm that the deployed model and pipeline meet the customer's needs
- ***Project hand-off:*** Hand the project off to the entity that's going to run the system in production

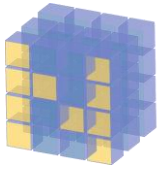
**Artefact**

Exit report of the project for the customer


Business Understanding
Data Collection
Data Preparation
Modelling
Deployment
Project Lead
Data Scientist
Project Manager
Solution Architect
Checkpoint Project
Create Template Repository
Provision Data Infraestructure
Provision Compute Assets
Data Ingest & Explore
Data Analysis
Feature Engineering
Model Dev
Deploy Scoring Process
Monitor health & Metrics
Check in final Artifacts
Data Summary Report
Data Analysis Report
Model Report
Finalize Document
Create Project Charter
Project Final Report
Project Charter
Design Solution Architecture
Develop Data Pipeline
Deploy Pipeline
Solution Architecture Diagram


Skills
Knowledge of mathematical statistics
Results communication, explanation, clarification
Knowledge of machine learning
Understand multiple analytical functions
Strong knowledge of Python, R, Scala
Ability to work with unstructured data from various sources
Hands-on experience in SQL database coding

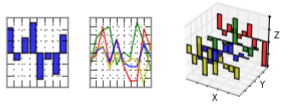
# Tools



NumPy

pandas

$y_{it} = \beta' x_{it} + \mu_i + \epsilon_{it}$



matplotlib

Seaborn


spaCy

PYTORCH

django

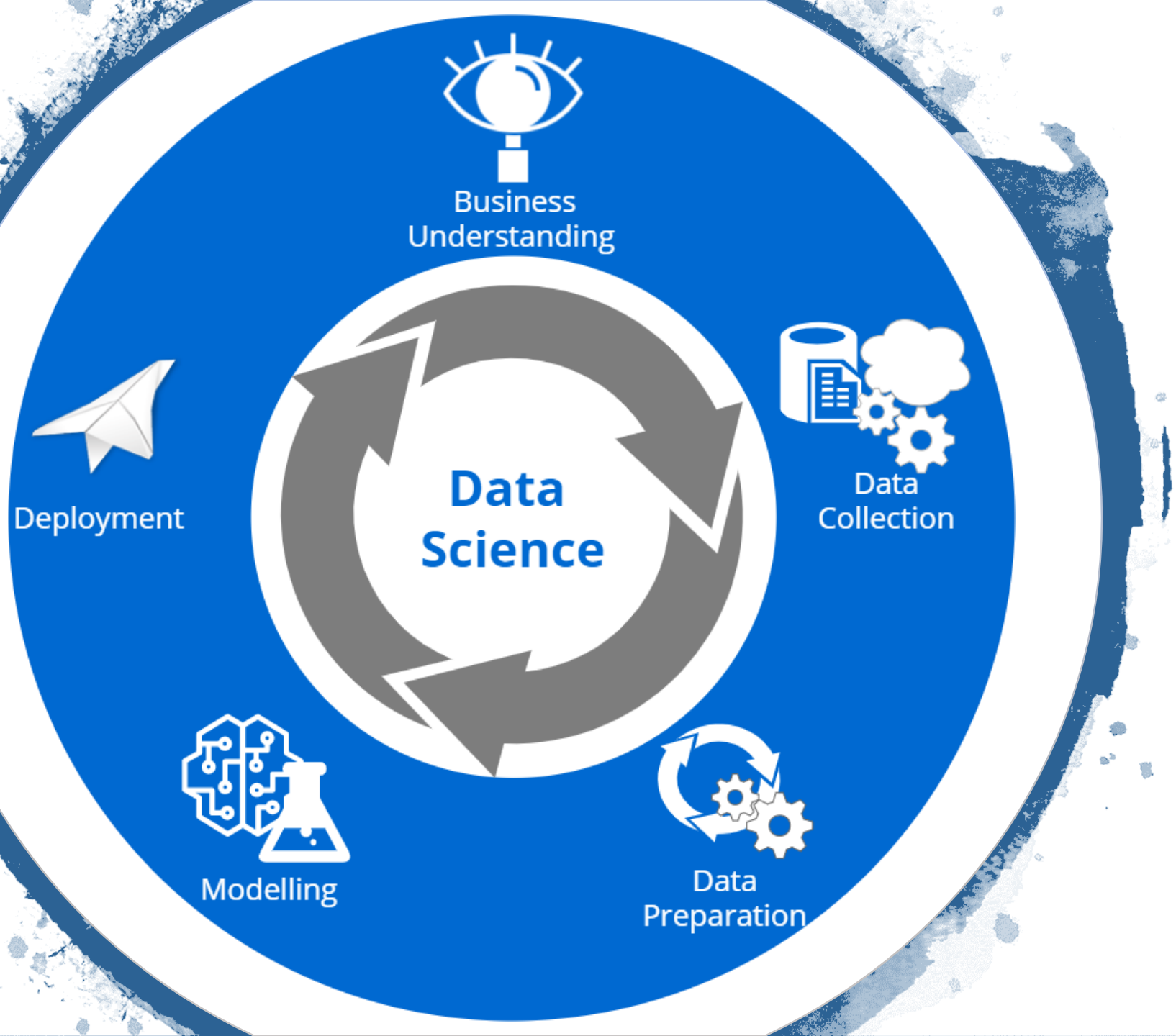


Spark

TensorFlow

mlflow


Keras


Business
Understanding
Data
Collection
Data
Preparation
Modelling
Deployment
Data
Science