AT82.02

DATA MODELING AND MANAGEMENT

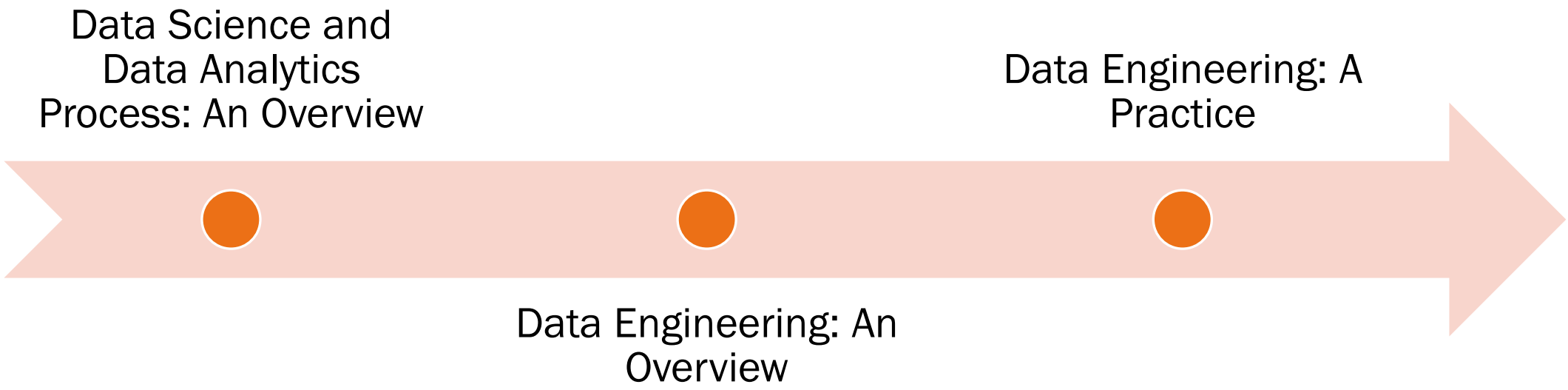
UNIT 5-1: DATA ENGINEERING

CHUTIPORN ANUTARIYA (CHUTI AT AIT DOT AC DOT TH)

Co-funded by the Erasmus+ Programme of the European Union

DS&AI

Outline

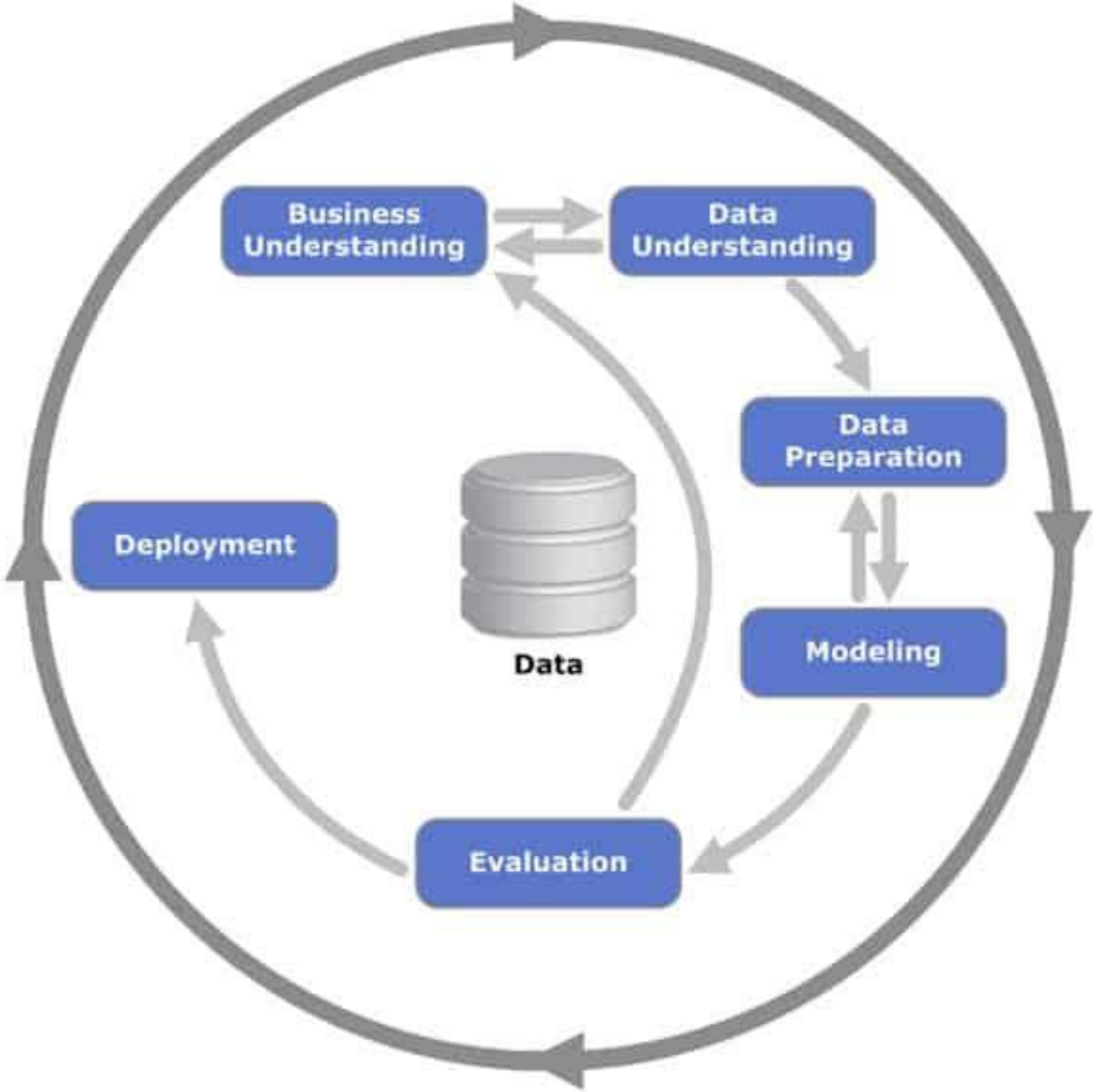

Data Science and Data Analytics Process: An Overview
Data Engineering: An Overview
Data Engineering: A Practice

Data Science and Data Analytics Process

AN OVERVIEW

CRISP-DM

The CRoss Industry Standard Process for Data Mining (CRISP-DM) is a process model with six phases that naturally describes the data science life cycle.



CRISP-DM: Business Understanding

- Determine business objectives
- Assess situation
- Determine data mining goals
- Produce project plan

CRISP-DM: Data Understanding

Collect initial data

Describe data

Explore data

Verify data quality

CRISP-DM: Data Preparation

- Select data
- Clean data
- Construct data
- Integrate data
- Format data

CRISP-DM: Data Preparation

- Select data
- Clean data
- Construct data
- Integrate data
- Format data

CRISP-DM: Modeling

- Select modeling technique
- Generate test designs
- Build model
- Assess model

CRISP-DM: Evaluation

Evaluate results

Review process

Determine next steps

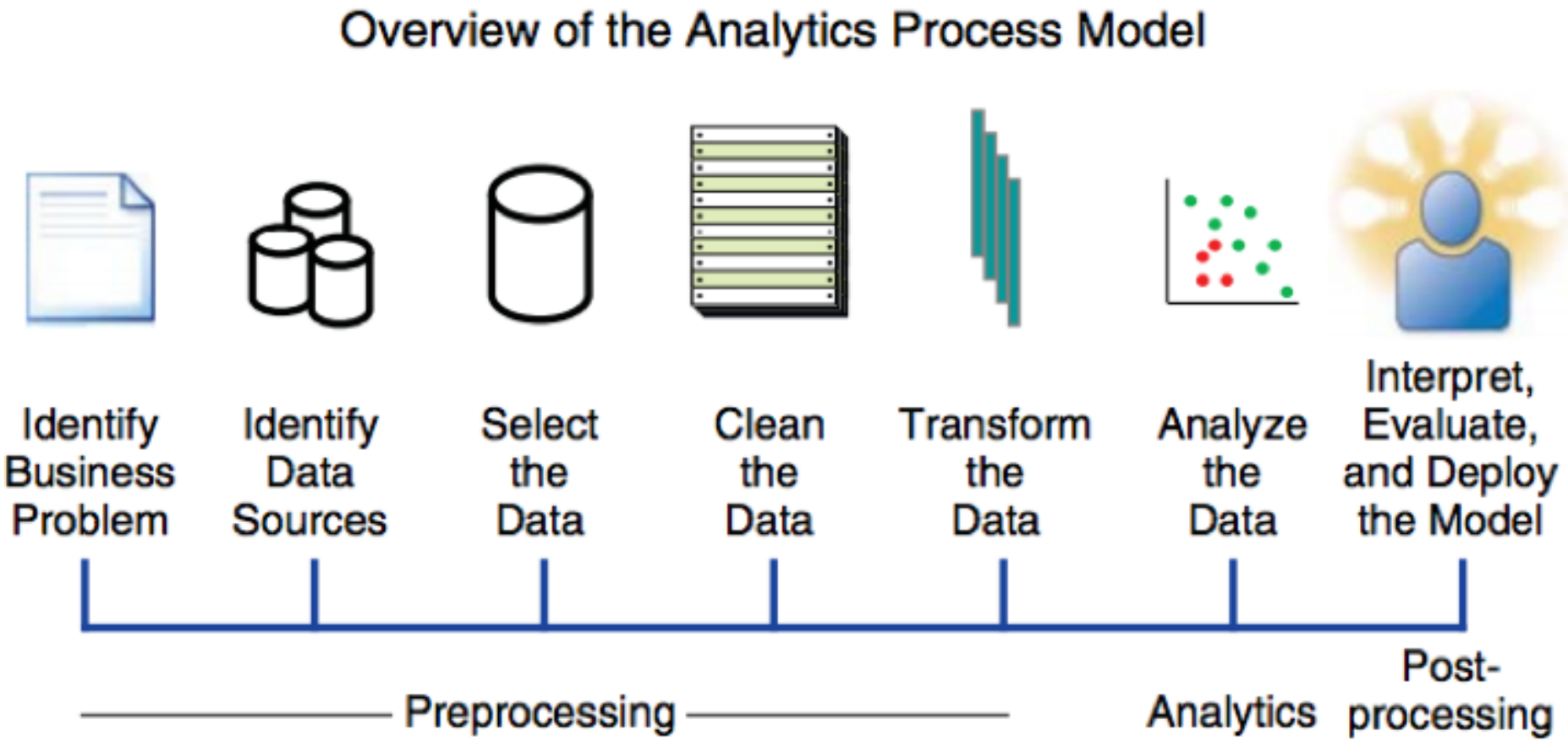
CRISP-DM: Deployment

Plan deployment

Plan monitoring and maintenance

Produce final report

Review project



Business Analytics Process Model

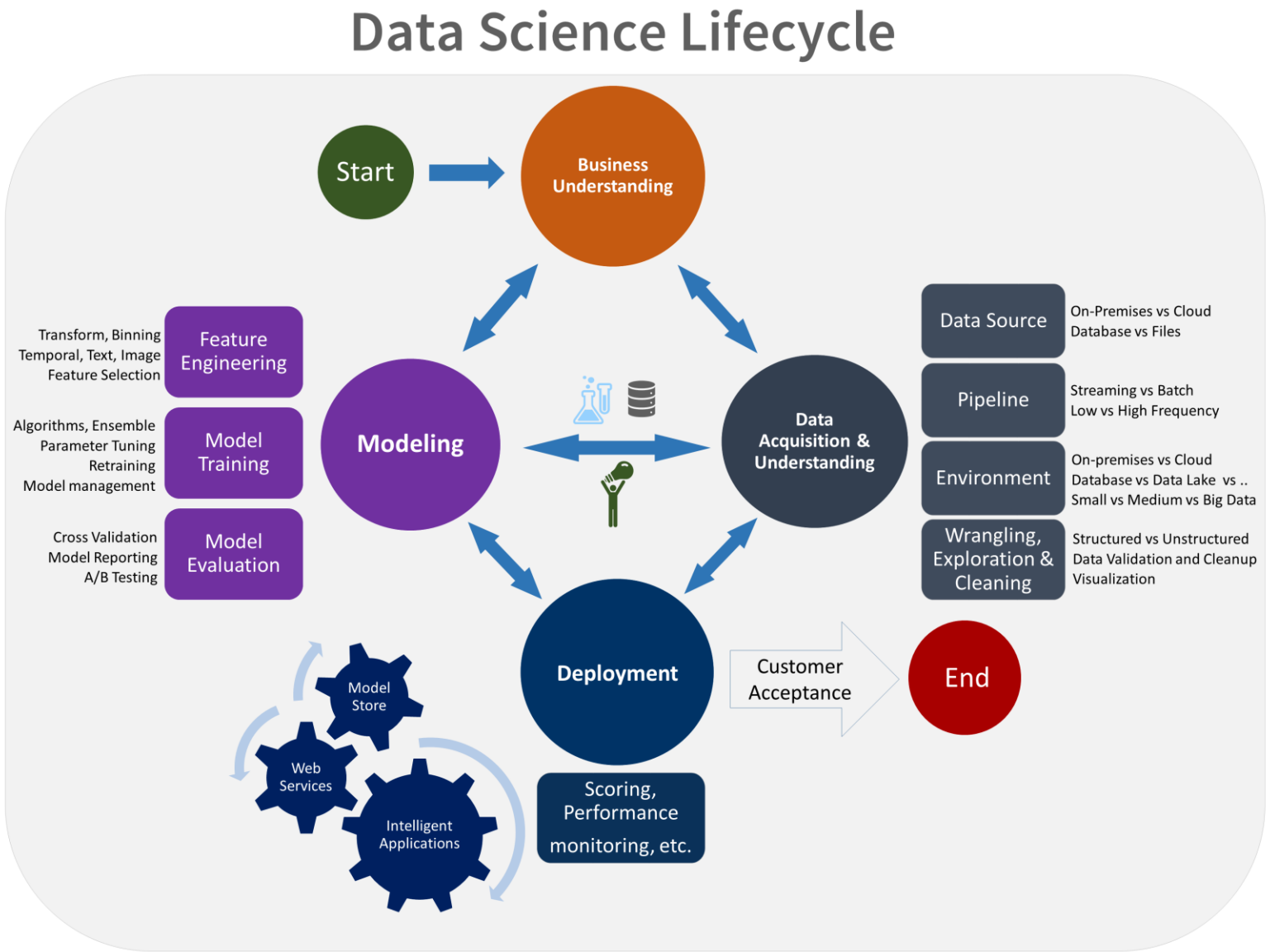
Src: https://blogs.sas.com/content/sgf/2019/05/14/big-data-in-business-analytics-talking-about-the-analytics-process-model/

Data (Analytics) Workflow

Src: DataCamp

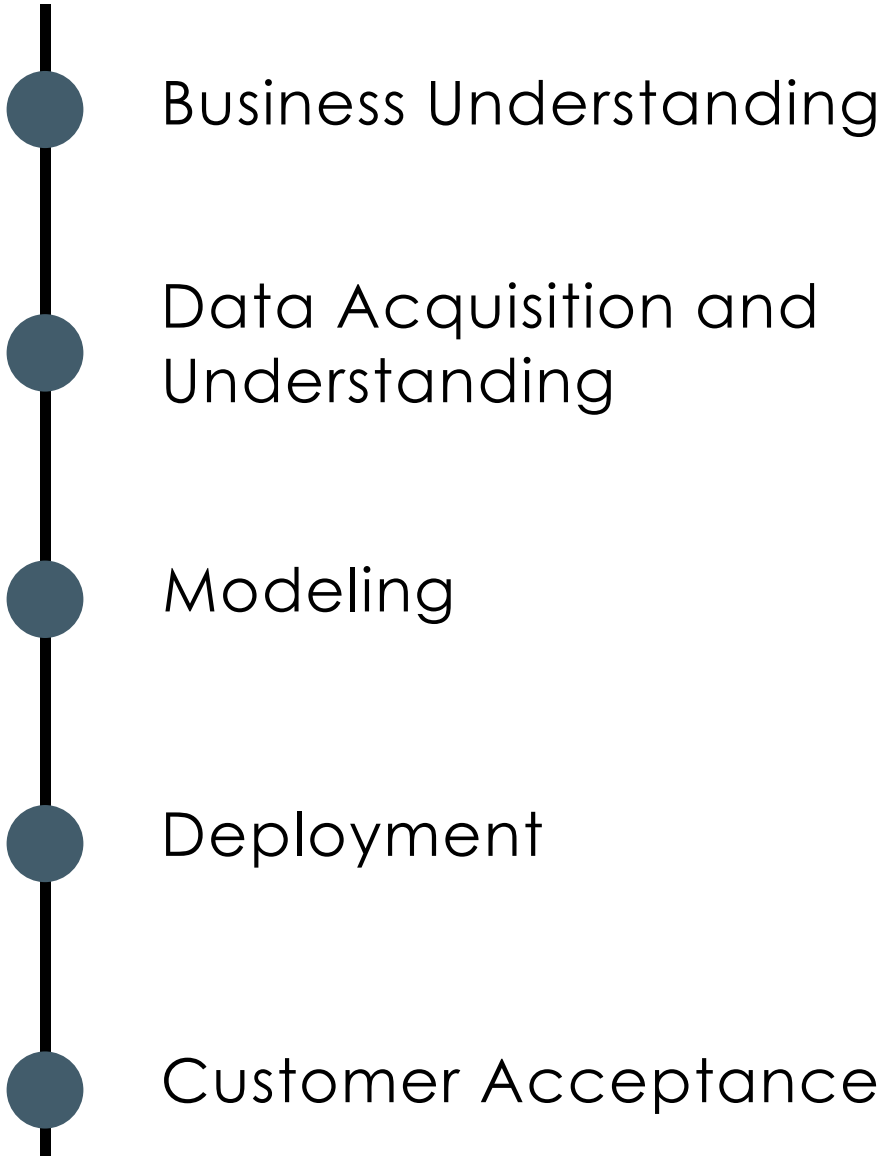
TDSP: Team Data Science Process

- An agile, iterative data science methodology to deliver predictive analytics solutions and intelligent applications efficiently
- Combination of Scrum and CRISP-DM



Ref: Microsoft Team Data Science Lifecycle (Microsoft, 2020)

Data Science Process Lifecycle



TDSP: BUSINESS UNDERSTANDING

TASK 1.1

Define objectives: discuss with the client as well as with the various stakeholders in order to understand and identify the business challenges. Formulating questions that define the business objectives that data science techniques can achieve.

- Step1: Identify the main business variables that the analysis needs to predict. These variables should be interpreted as the model destinations and associated metrics used to determine the success of the project.
- Step2: Define project goals by asking "precise" questions that are relevant, specific and unambiguous.

TDSP: BUSINESS UNDERSTANDING

Usually five types of questions are used:

- How much or how many? (regression)
- What category? (classification)
- What group? (group)
- Is that weird? (anomaly detection)
- Which option should be adopted? (recommendation)

It must be determined, however, what questions to ask and how the answer will achieve your business objectives.

TDSP: BUSINESS UNDERSTANDING

Step3: Define the project team specifying the roles and responsibilities of the most varied members. Develop a plan in order to find out more information or even their sources.

Step4: Define the metrics of success.
The metrics should be SMART :

- Specific
- Measurable
- Achievable
- Relevant
- Time-bound

TDSP: BUSINESS UNDERSTANDING

TASK 1.2

Identify data sources: Find the relevant data that helps to answer the questions that define the project objectives.

Step1: Look up the following data:

- Data relevant to the question. Do you have target measures and resources related to the target?
- Data that is an accurate measure of your model destination and resources of interest.

For example, you may consider that existing systems may need to collect and record additional types of data to solve the problem. In this situation, you may need to look for external data sources or update your systems to collect new data.

TDSP: DATA ACQUISITION AND UNDERSTANDING

TASK 2.1

Ingest the data into the target analytic environment.

Step1: Process configuration to move data from the source to the destination where the analytical operations are performed.

TDSP: DATA ACQUISITION AND UNDERSTANDING

TASK 2.2

Explore the data to determine if the data quality is adequate to answer the question.

Step1: Before training the models, it is necessary to develop a good understanding of the data. The data sets collected from the most diverse sources are often noisy, have no value or have a number of other discrepancies.

Step2: Compress and analyze the data to audit its quality and provide the necessary information to process the data before they are ready for modeling. This process is often iterative.

TDSP: DATA ACQUISITION AND UNDERSTANDING

TASK 2.3

Set up a data pipeline to score new or regularly refreshed data.

Step1: In addition to initial data ingestion and cleaning, it is necessary to set up a process to obtain new data or update it regularly as part of an ongoing learning process. At this stage, a data pipeline solution architecture must be developed.

Step2: The pipeline should be situated in parallel with the next step of the data science project. Depending on the business needs and constraints of the existing systems into which this solution is being integrated, the pipeline can be one of the following options:

- Batch-based
- Streaming or in real time
- A hybrid

TDSP: MODELING

TASK 3.1

Feature engineering: Create data features from the raw data to facilitate model training.

Step1: Feature engineering involves the inclusion, aggregation and transformation of raw variables to create the resources used in the analysis.

Step2: It is essential to understand how resources relate to each other and how machine learning algorithms should use these resources.

Step3: Feature engineering is an act of finding and including informational variables, and trying to avoid unrelated variables.

TDSP: MODELING

TASK 3.2

Model training: Find the model that answers the question most accurately by comparing their success metrics.

Step1: Divide the input data randomly for modeling into a training data set and a test data set.

Step2: Build the models using the training data set.

Step3: Evaluate the training and test data set. Choose some competing learning algorithms as well as the various associated adjustment parameters that are oriented to answer the question of interest with the current data.

Step4: Determine the "best" solution to answer the question by comparing the success metrics between alternative methods.

TDSP: MODELING

TASK 3.3

Determine if your model is suitable for production.

TDSP: DEPLOYMENT

TASK 4.1

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

Step1: Once you have a set of models that work well, they should be operated for other applications to consume. Depending on the business needs, forecasts are made in real time or in batch. To implement models, the API interface must be exposed.
The interface allows the model to be easily consumed from various applications, for example:

- Online sites
- Spreadsheets
- Instrument panels
- Line-of-business applications
- Back-end applications

TDSP: CUSTOMER ACCEPTANCE

TASK 5.1

System validation: Confirm that the deployed model and pipeline meet the customer's needs.

TASK 5.2

Project hand-off: Hand the project off to the entity that's going to run the system in production.

Outline

Data Science and Data Analytics Process: An Overview

Data Engineering: An Overview

Data Engineering

AN OVERVIEW

Data Engineer

Data engineers deliver the **correct data** in the **right form** to the **right people** as **efficiently** as possible.

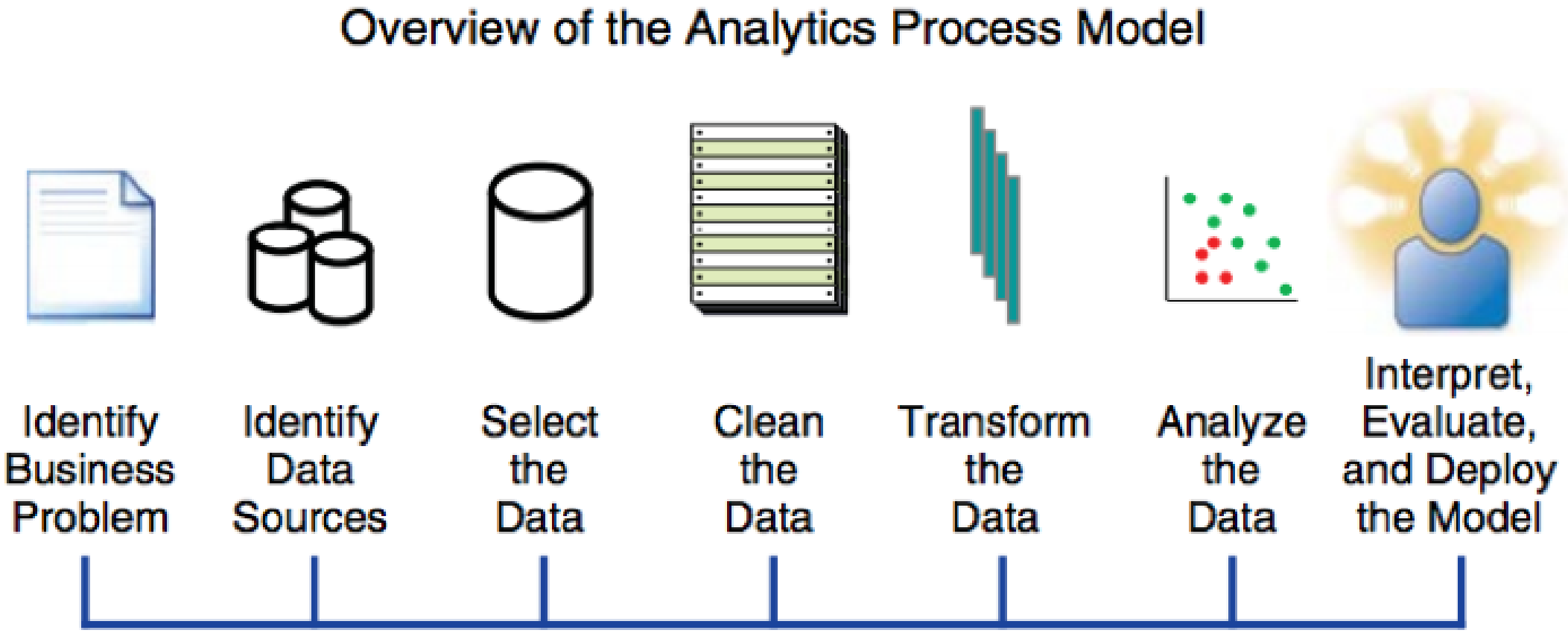
Data engineers primarily focus on the following areas:

Build and maintain the organization’s data pipeline systems.

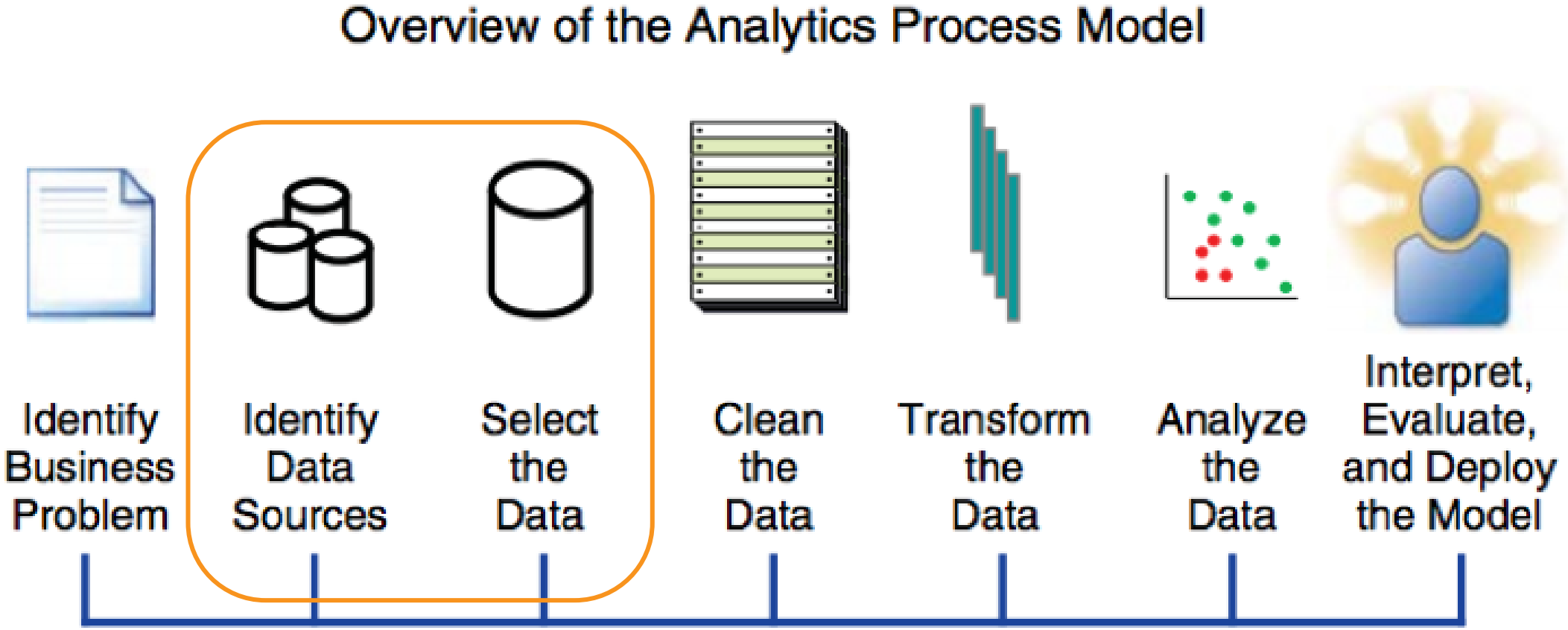
Clean and wrangle data into a usable state.

ETL: Extract > Transform > Load

ELT: Extract > Load > Transform



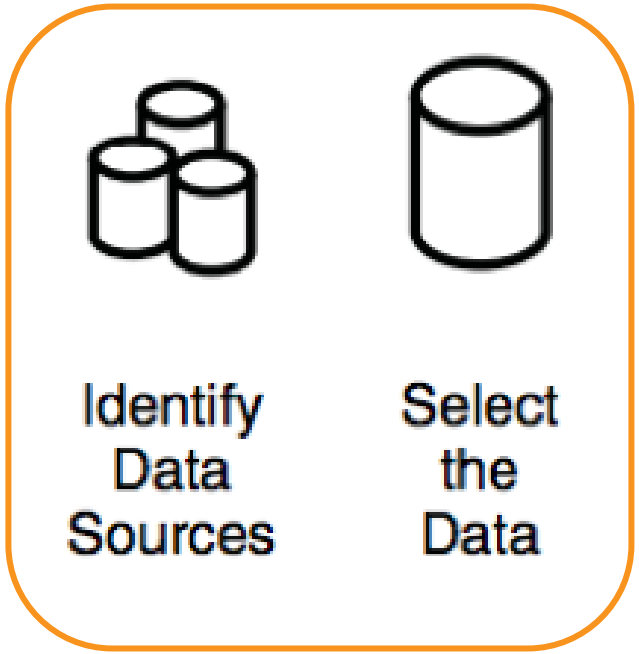
Which steps are part of Data Engineering process?



Construct your dataset

Dataset: Size

"Garbage in, garbage out"



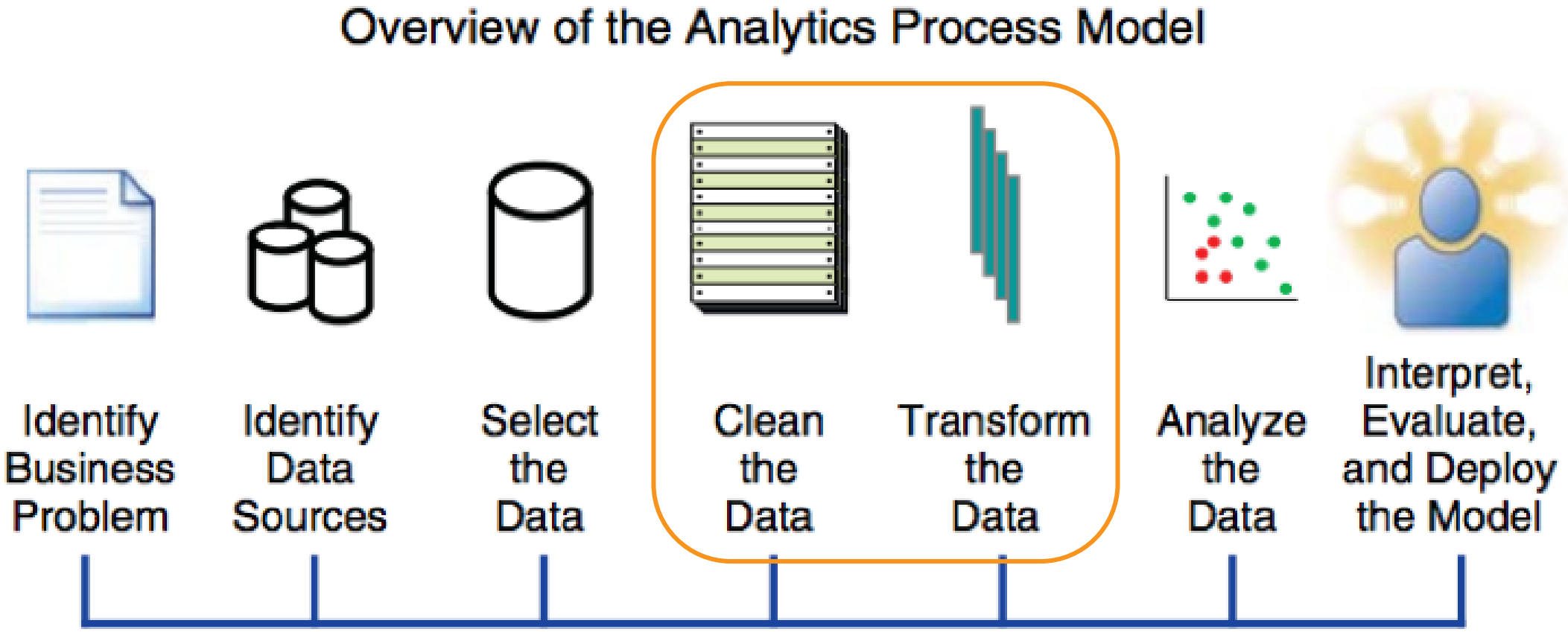
How much data? It depends on the project!

Data set	Size (number of examples)
Iris flower data set	150 (total set)
MovieLens (the 20M data set)	20,000,263 (total set)
Google Gmail SmartReply	238,000,000 (training set)
Google Books Ngram	468,000,000,000 (total set)
Google Translate	trillions

Characteristics of Dataset Quality

Dataset: Quality

Characteristic	How to measure
Accuracy	Is the information correct in every detail?
Completeness	How comprehensive is the information?
Reliability	Does the information contradict other trusted resources?
Relevance	Do you really need this information?
Timeliness	How up- to-date is information? Can it be used for real-time reporting?



Clean and Transform Data

Data Cleansing

Data Cleansing (or Data Cleaning) is the process of identifying incorrect, irrelevant incomplete or the “dirty” parts of a dataset and then cleaning them.

The process of data cleansing may involve the removal of typographical errors, data validation, and data enhancement.

Data Cleansing Methods:

- Removal of unwanted data
- Removal of duplicate data
- Removal of irrelevant data
- Correct mistaken/erroneous data
- Filter-out outliers
- Resolving inconsistencies / missing data:
  - Dropping vs Imputing a value

Data Transformation

The process of converting **data** from one format to another, typically from the format of a source system into the required format of a destination system.

Cleansing – (sometimes cleansing is defined as part of data transformation) inconsistencies and missing values in the data are resolved.

Standardization – formatting rule are applied to the data set.

Deduplication – redundant data is excluded or discarded.

Verification – unusable data is removed and anomolies are flagged.

Sorting – data is organized/grouped.

Other tasks – any additional/optional rules can be applied to improve data quality.

- Filtering (e.g. Selecting only certain data fields/records).
- Enriching (e.g. Full name to First Name , Middle Name , Last Name).
- Splitting a column into multiple columns and vice versa.
- Joining together data from multiple sources.
- Labeling.

Data Engineering

A PRACTICE

Movie Rating Prediction

- Movie rating is an important element to decide movie quality.
- People prefer to use rating as reference to decide whether to watch a movie or not.
- We plan to use historical values of the movie as features (e.g. genres) and the user profiles (e.g., age, gender) to predict movie rating before the movie released.

Movie Rating Prediction

Datasets: Movies, User Profiles, User Rating

The variables used for **input**:

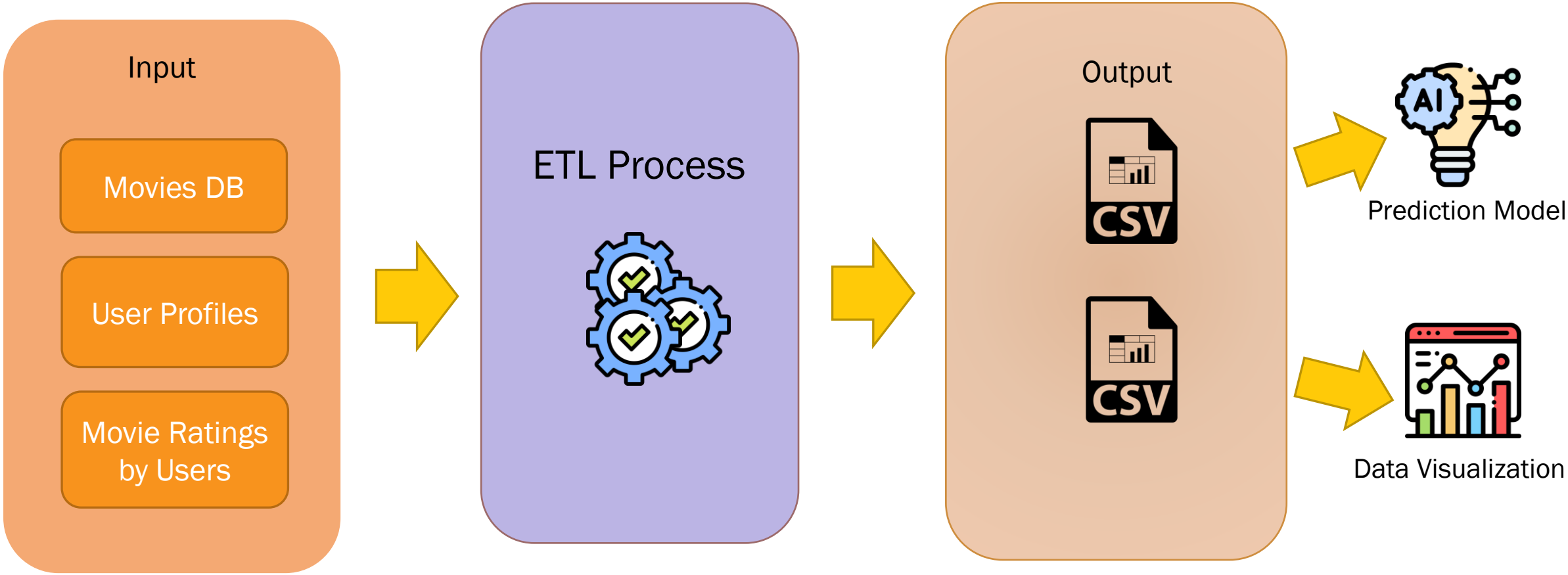
- Age, Gender, Occupation, Movie Category

The variables used for **model prediction**:

- User Rating

source: https://www.kaggle.com/sherinclaudia/movie-rating-prediction

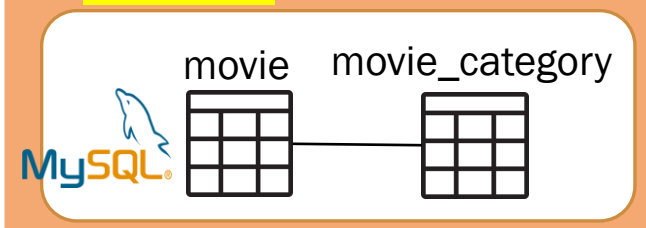
Movie Rating Prediction



Movie Rating Prediction

Input

Movies DB



Relational Database

User Profiles

CSV File

Movie Rating by Users

MongoDB

ETL Process

Extract

Step1: get movie data from RDB and convert data to CSV file
Step2: get user profile in CSV format
Step3: get movie rating by users from MongoDB
3.1 export data JSON file
3.2 convert Json to CSV file

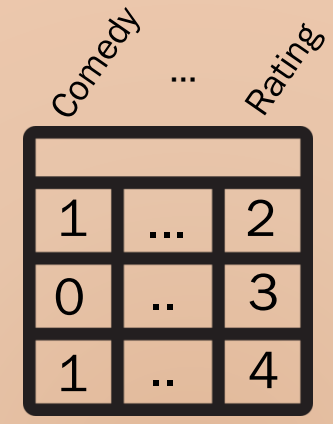
Transform

Step.4 Cleans, Transform and Integrate
4.1 Remove Duplicate Records
4.2 Dealing with Missing Values
4.3 Integrate 3 datasets
4.4 Convert birthdate to age
4.5 Convert gender to numeric
4.6 Filter age >= 7

Load

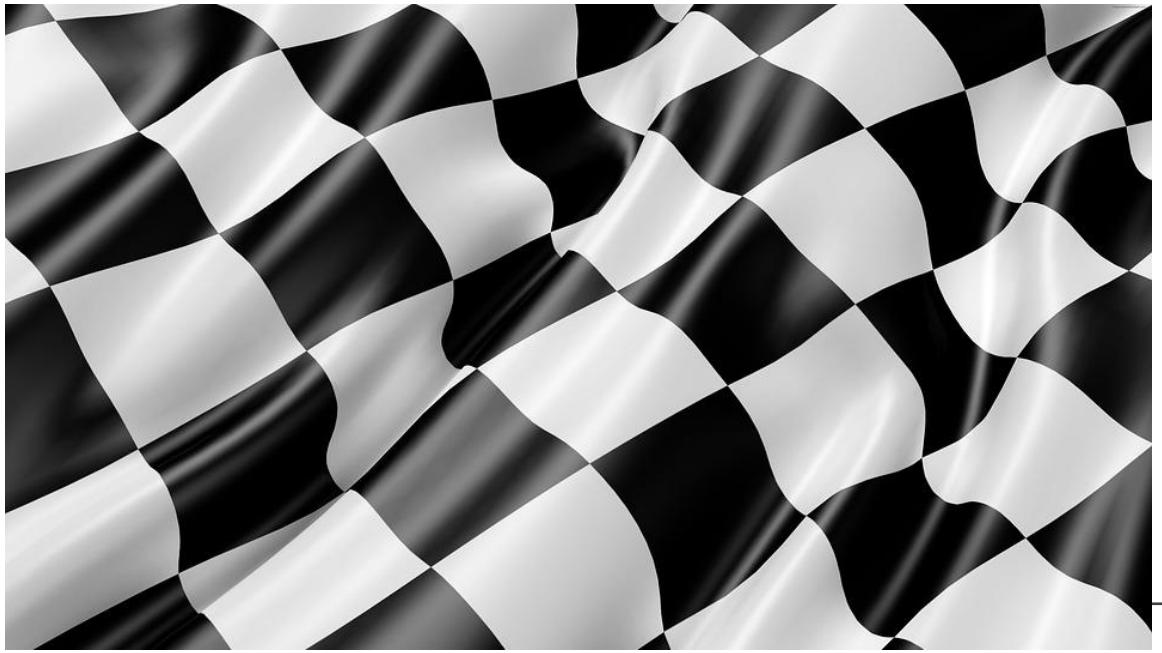
Step.5 Write output as CSV files

Output



CSV for Prediction

CSV for Visualization



Thank you.

Let’s Summarize!