Data Transformation

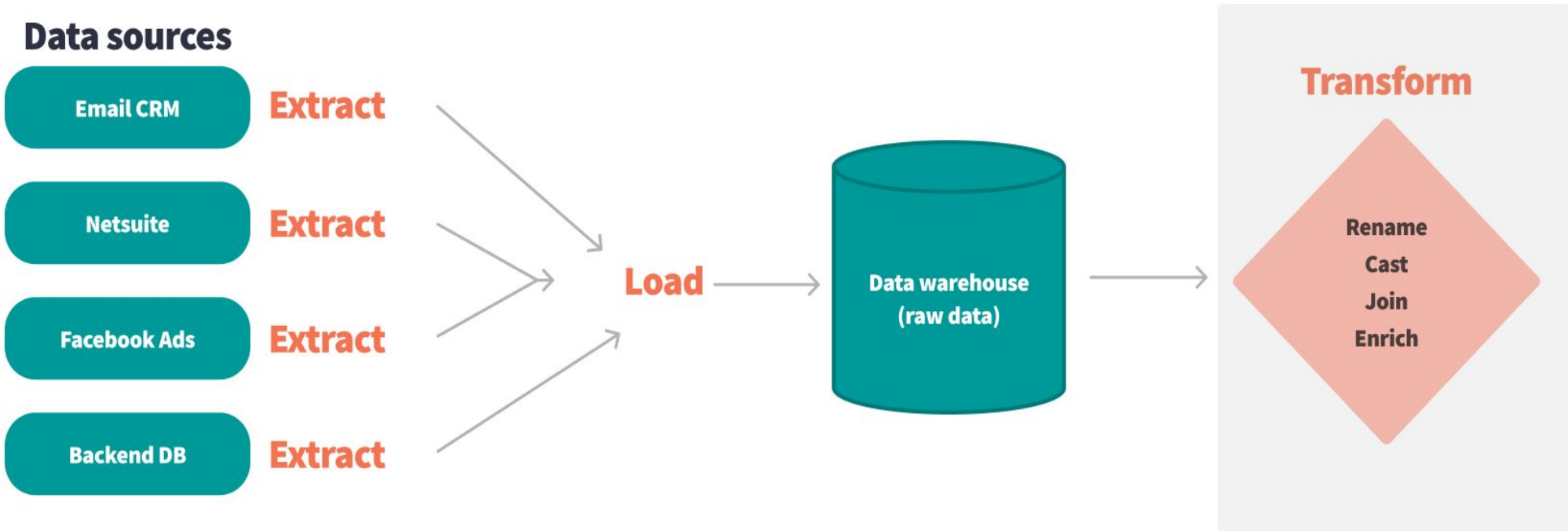
Data Transformation

- Data transformation is one of the techniques that we use in between data processing. This technique lets us convert the raw data into a required format so that the next procedures of data processing and data modelling can be performed efficiently.

- Technically, Data transformation changes the data structure, format and value and makes it clean and usable for the next processes. There can be two stages of data transformation processes because many organisations use data warehouses arranged in the ETL process, where data transformation is an in-between process.

- However, data transformation is an integral part of the entire data processing pipeline because it can be found right from data integration to the final stages of data wrangling.

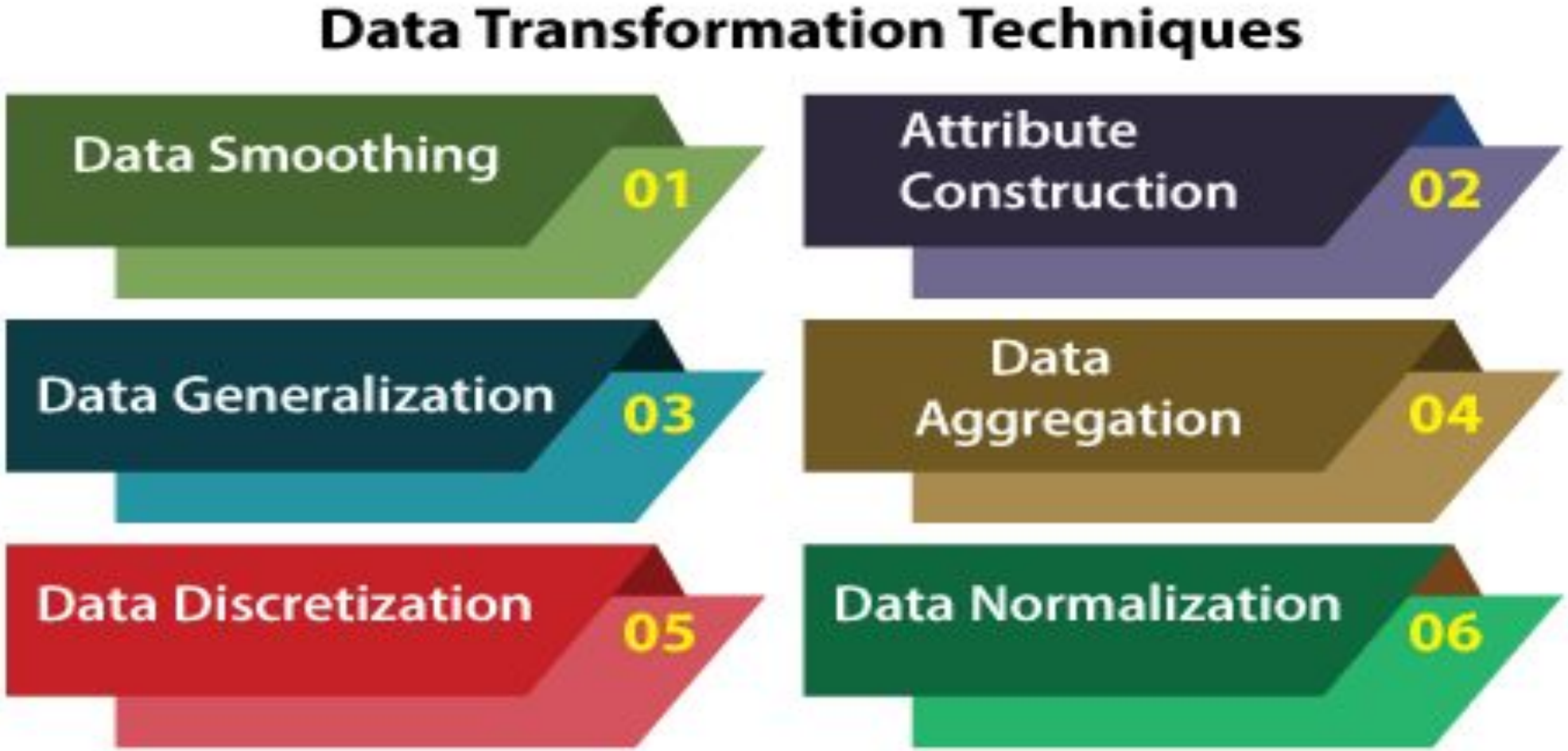
ELT


Data sources
Email CRM
Netsuite
Facebook Ads
Backend DB
Extract
Extract
Extract
Extract
Load
Data warehouse
(raw data)
Transform
Rename
Cast
Join
Enrich

Type of data transformation

- **Constructive:** The process of adding, copying or replicating data.
- **Destructive:** deleting the records or field.
- **Aesthetic:** standardising the data to make it valuable.
- **Structural**: reorganising the data by moving, merging and renaming the columns.

Data Transformation Techniques



Various general data transformation techniques:

1. Data smoothing

- We can think of data smoothing as the process of removing noise from the data that can also involve some algorithms. This transformation makes the important data feature more visible and helps predict the pattern.

2. Attribute construction

- In this data transformation technique, we add new attributes to the data based on the already existing attributes. This new attribute smoothens the other processes by simplifying the way to get accurate data.

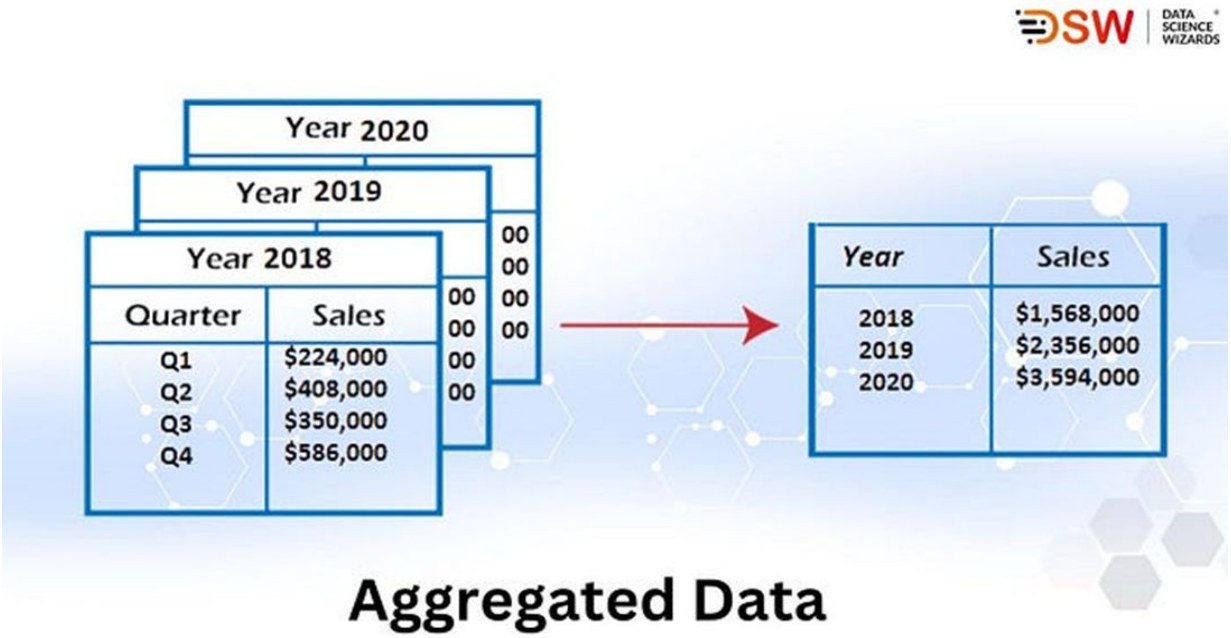
3. Data Generalization

- This data transformation technique depends on the concept of hierarchy to transform low-level data attributes into high-level data attributes.
- Through this transformation, we get a clear picture of the data, and this process can be divided into two different approaches:
  1. Data cube process (OLAP: Online analytical processing) approach.
  2. Attribute-oriented induction (AOI) approach.
- A simple example of data generalization is to convert age data from numerical to categories.

	ORIGINAL DATA	GENERALIZED DATA
AGES	16	10-19 (2)
	18	20-29 (3)
	21	30-39 (5)
	23	40-49 (5)
	27	
	32	
	32	
	36	
	38	
	39	
	44	
	47	
	47	

4. Data aggregation

- Data aggregation or data collection is a technique of presenting data in a summary form. There is a huge chance of data coming from different sources, and integrating all the coming data into a description is data aggregation. This part of data processing is crucial because it depends on the quality and quantity of the data we use.
- An example of this process is making an annual sales report based on the given quarterly or monthly data.



5. Data Discretization

- We can think of data discretization as the process of converting continuous data into a set of intervals so that small interval labels can substitute them. Just like normalization, this technique makes the data more interpretable. If a data analysis work uses continuous values, then discrete forms of values can be replaced by constant quality attributes.
- We can also call this technique the data reduction technique, as it transforms discrete data into a set of categorical data.
- We can classify this technique into supervised and unsupervised discretization. Where supervised data discretization uses class information, on the other hand, unsupervised data discretization uses the processing direction of the involved data process.
- For example, we can classify people in numeric age groups such as(0–5,5–10…) or in classes like kid, youth or adult.

6. Data Normalization

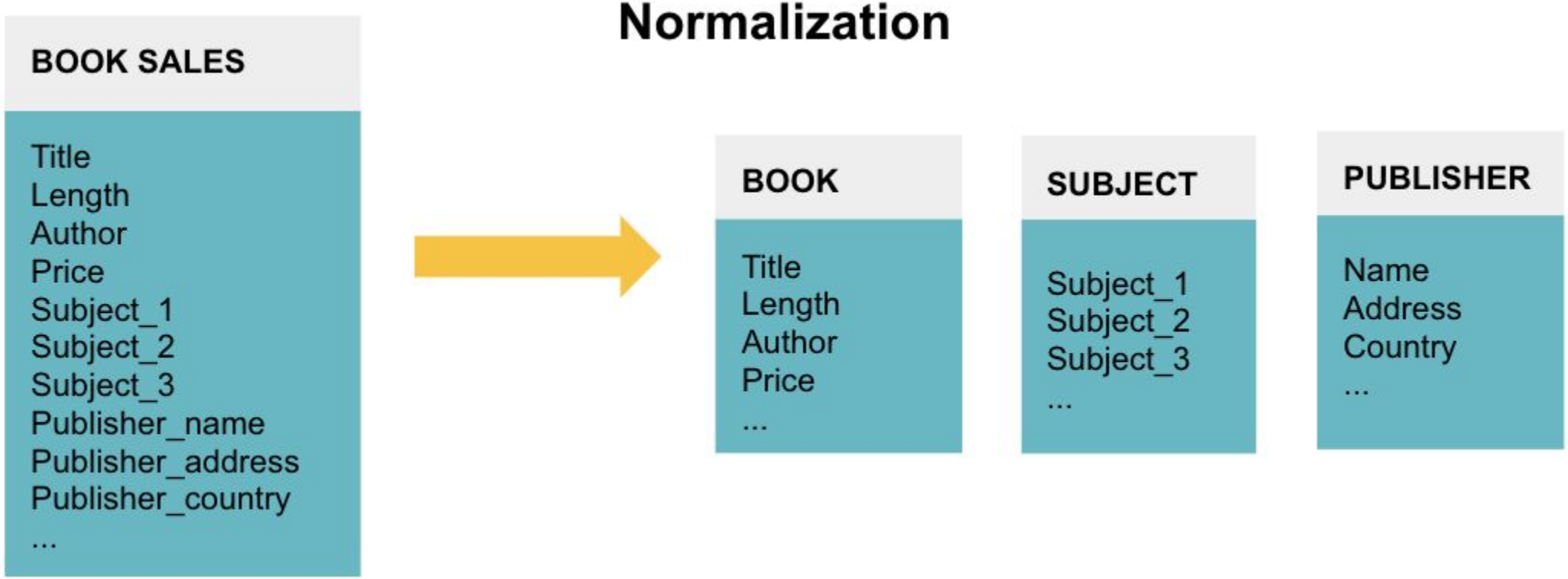
- Data normalization is another technique of data transformation that involves scaling the data in a smaller range, such as scaling data between 0 to 1 or -1 to 1.

- The main objective of data normalization is to eliminate data redundancy and improve data consistency and accuracy. By organizing data into smaller, more manageable ranges, it becomes easier to update and maintain the data.

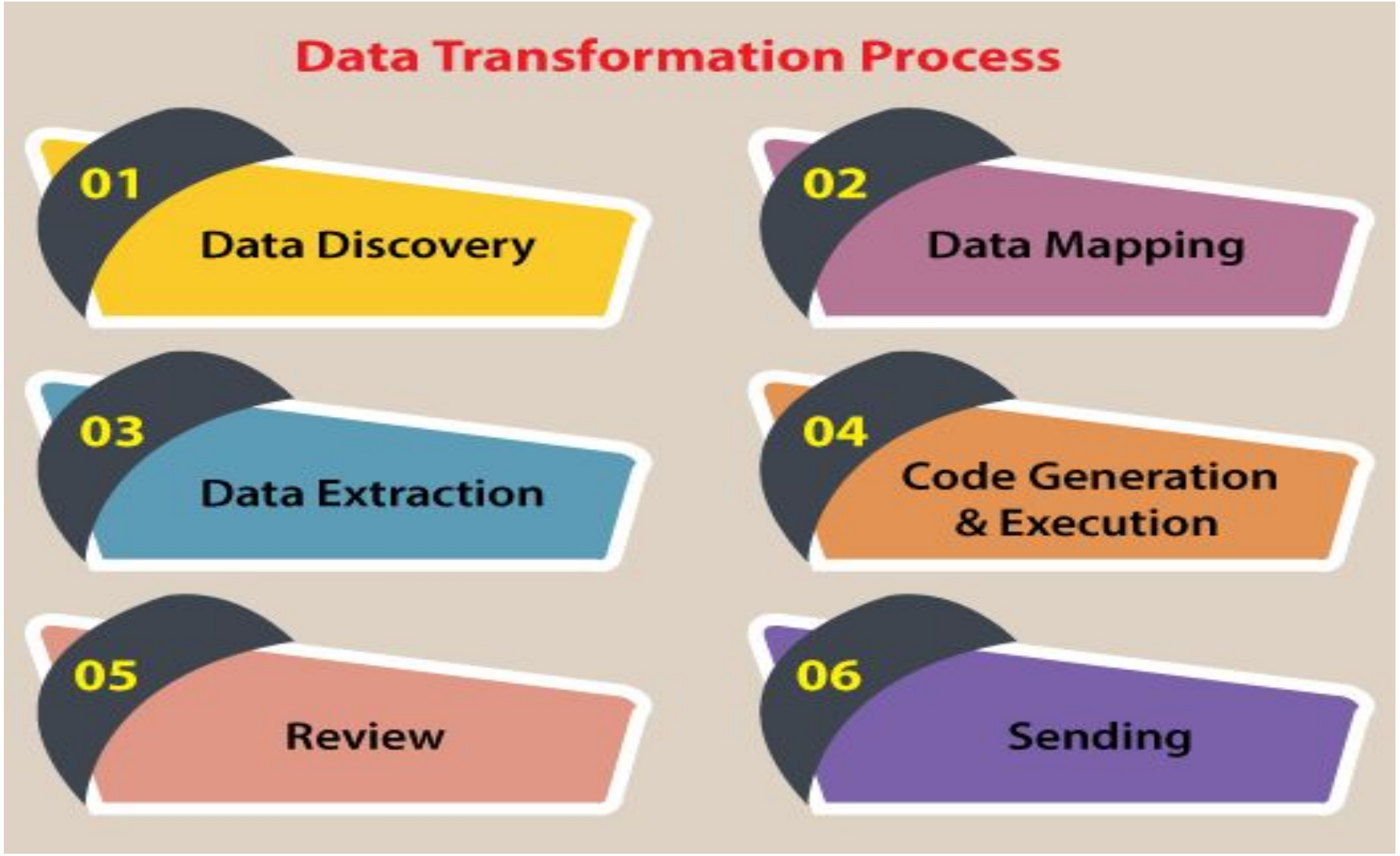
- Normalization also helps to reduce the likelihood of data inconsistencies, which can arise when the same information is stored in multiple locations.

- There are various types of data normalization techniques we use, such as
  - Min-Max normalization,
  - Z-score normalization,
  - Decimal Scaling etc.,

Normalization



Data transformation Process



Data transformation process

1.Data Discovery: Before applying ETL in any organisation, it is important to know and understand the data source. Usually, data profiling tools we use to do so. With the help of this step, one gets to know what ahead is required to get data in the required format.

2. Data Mapping: This step of the process is performed to determine how any field is mapped, modified, filtered, joined or aggregated.

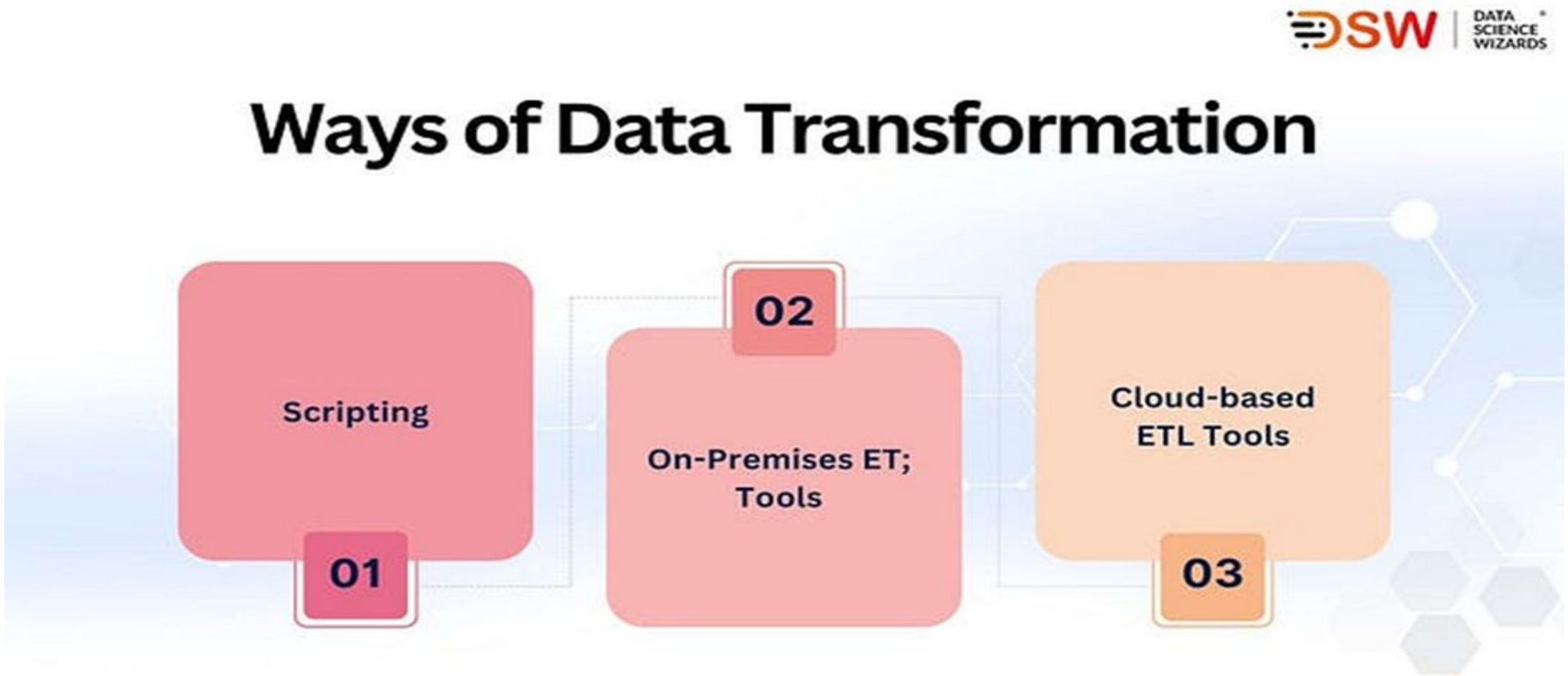
3. Data Extraction: This step involves the extraction of data from its original source. Examples of sources can be databases or streaming sources such as sales log files from web applications.

4. Code Execution: The foremost step that involves actual data transformation is code generation and execution. Here code is generated to transform data in the required format.

5. Review: This step ensures that the code executed on data is transforming the data accurately.

6. Sending: Till here, the data is transformed, but the last thing which remains to complete is to send the data to its target destination. Here the destination can be a relational database or warehouse that handles both unstructured and structured data.

Ways of data transformation



Three general ways of data transformation are as follows.

- **By scripting:** this way, we write codes in python, SQL or other languages to query and transform the data. However, using python and SQL has its own benefits because they help automate various tasks, as well as they require less coding to perform than a traditional programming language requires.
- **Using ETL tools:** these tools are designed to make the data extraction and transformation work easier than scripting on-premises. These tools can be hosted on the organisation's server and can save a lot of time. But these tools often require expertise in using tools and significant infrastructure costs.
- **Cloud-based ETL tools:** these tools are based on a relevantly new technology where it can be hosted in the cloud. These tools are designed to provide the easiest way to extract and transform data even a non-technical person can perform operations on it. Using these tools, we can store the data in any cloud source, and when required easily, we can load the data in a data warehouse. With the help of such ETL tools, we can schedule data pulling while monitoring the usage.

Advantages of Data Transformation

- **Improved data quality:** we all know of the risks and costs associated with low-quality data, so data transformation helps organisations eliminate data quality issues.
- **Faster Queries:** it becomes easy and quick to extract transformed data from its storage or location.
- **Efficient Data Management:** as data is non-stop generating in different sources, it becomes challenging to build and understand metadata. Data transformation helps in refining the metadata, and this is how data management becomes more efficient.
- **Better organisation:** Data transformation makes the process easy because it is more interpretable for both humans and computers.
- **More use of data:** In the process of generating and collecting data continuously, a lot of data gets unanalysed. Data transformation lets the get the most out of data by standardising it and making it more usable.

What is Normalization?

- Normalization in machine learning is a data preprocessing technique used to change the value of the numerical column in the dataset to a common scale without distorting the differences in the range of values or losing information.
- In simple terms, Normalization refers to the process of transforming features in a dataset to a specific range. This range can be different depending on the chosen normalization technique.
- The two most common normalization techniques are Min-Max Scaling and Z-Score Normalization, which is also called Standardization.

Now, let's discuss Min-Max Scaling.

- **Min-Max Scaling**
- This method rescales the features the features to a fixed range, usually 0 to 1. The formula for calculating the scaled value of a feature is:
- Normalized Value = Value - Min/ Max - Min

where,

- Value: Original Value of the feature
- Min: Minimum value of the feature across all the data points.
- Max: Maximum value of the feature across all the data points.

When to use Normalization?

- When using algorithms that assume the input features are on a similar scale or bounded range, such as neural networks. These algorithms often assume input values are in the range [0,1].

- When you want to speed up the convergence of gradient descent by ensuring all features contribute equally to the cost function.

- If the data doesn't follow a Gaussian distribution.

- For models where the magnitude of variables is important, such as k-nearest neighbours.

Advantages of Normalization

- **Improves Algorithm Performance:** Normalization can lead to faster convergence and improve the performance of machine learning algorithms, especially those that are sensitive to the scale of input features.
- **Consistent Scale:** It brings all the variables to the same scale, making it easier to compare the importance of features directly.
- **Reduces the Impact of Outliers:** Methods like Min-Max scaling can reduce the impact of outliers, although this can also be a disadvantage in cases where outliers are important.
- **Necessary for Certain Algorithms:** Some algorithms, like k-nearest Neighbors (k-NN) and neural networks, require data to be normalized for effective performance.
- **Easier to Learn:** When features are on a similar scale, gradient descent (used in training many machine learning models) can converge more quickly.

Disadvantages

- **Data Dependency:** The normalization process makes the training data dependent on the specific scale, which might not be appropriate for all kinds of data distributions.
- **Loss of Information:** In some cases, normalization can lead to a loss of information, especially if the data is sparse and the normalization compresses different values into a small range.
- **Sensitivity to New Data:** The parameters used for normalization (min, max, mean, standard deviation) can change with the introduction of new data, requiring re-normalization with updated parameters.
- **Time and Resources:** The normalization process adds extra steps to data preprocessing, which requires additional computation time and resources.

What is Standardization?

- Standardization is a data preprocessing technique used in statistics and machine learning to transform the features of your dataset so that they have a mean of 0 and a standard deviation of 1. This process involves rescaling the distribution of values so that the mean of observed values is aligned to 0 and the standard deviation to 1.
- Standardization aims to adjust the scale of data without distorting differences in the ranges of values or losing information.
- Unlike other scaling techniques, standardization maintains all original data points' information (except for cases of constant columns).
- It ensures that no single feature dominates the model's output due to its scale, leading to more balanced and interpretable models.

Formula of Standardization

- Z = (x-mean)/standard deviation

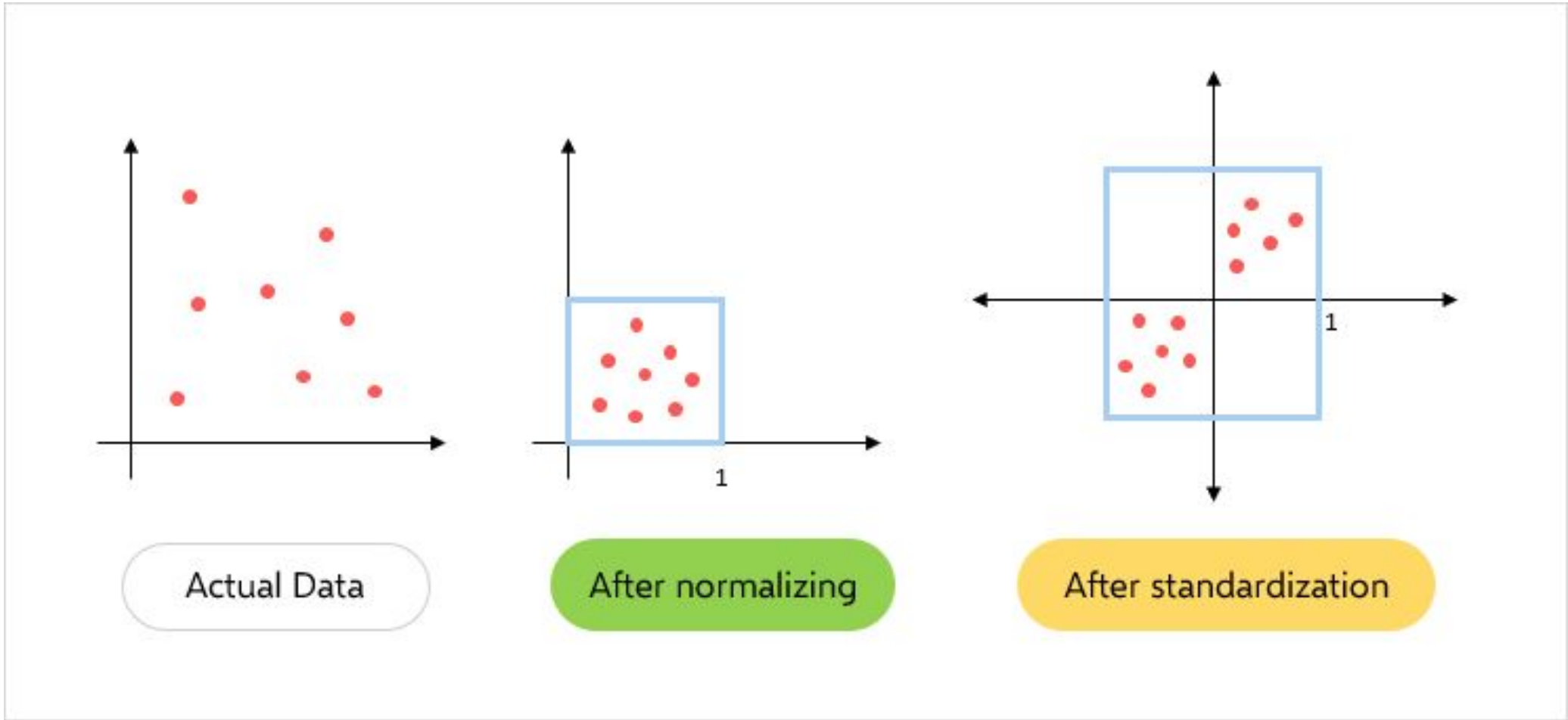
$$x^{(i)}_{std} = \frac{x^{(i)} - \mu_x}{\sigma_x}$$

$$x^{(i)}_{norm} = \frac{x^{(i)} - \mathbf{x}_{min}}{\mathbf{x}_{max} - \mathbf{x}_{min}}$$

standardization

min-max scaling ("normalization")

	input	standardized	normalized
0	0	-1.46385	0.0
1	1	-0.87831	0.2
2	2	-0.29277	0.4
3	3	0.29277	0.6
4	4	0.87831	0.8
5	5	1.46385	1.0


1
1
Actual Data
After normalizing
After standardization

When to use Standardization?

- Algorithms that assume the input features are normally distributed with zero mean and unit variance, such as Support Vector Machines, Logistic Regression, etc.
- Standardization can be a better choice if your data contains many outliers as it scales the data based on the standard deviation.
- It is often used before applying Principal Component Analysis (PCA) to ensure that each feature contributes equally to the analysis.
- "If the data features exhibit a Gaussian distribution, meaning that the data is normally distributed."

Advantages of Standardization

- **Improves Convergence Speed:** Standardization can speed up the convergence of many machine learning algorithms by ensuring features have the same scale.

- **Handles Outliers Better:** It is less sensitive to outliers compared to Min-Max scaling because it scales data based on the distribution's standard deviation.

- **Useful for Algorithms Assuming Normal Distribution:** Many machine learning algorithms assume that the input features are normally distributed. Standardization makes this assumption more valid.

- **Necessary for Certain Algorithms:** Algorithms like Support Vector Machines (SVM), k-nearest Neighbors (k-NN), and Principal Component Analysis (PCA) often perform better with standardized data.

Disadvantages of Standardization

- **Not Bound to a Specific Range:** Unlike Min-Max scaling, standardization does not bound features to a specific range, which might be a requirement for certain algorithms.

- **May Hide Useful Information:** In some cases, the process of standardizing can hide useful information about outliers that could be beneficial for the model.

- **Requirement for Recalculation:** Whenever new data is added to the dataset, the standardization process may need to be recalculated and applied again to maintain consistency.

- **Computational Resources:** The process requires additional computations, which can be a concern for very large datasets or limited computational resources.