DATA PREPROCESSING

DAYATA | YAP

CS 4103 | Intelligent Systems

September 6, 2023

OUTLINE

INTRODUCTION

Importance and significance

TECHNIQUES

Standardization

- Mean Removal
- Variance Scaling

Normalization

APPLICATION

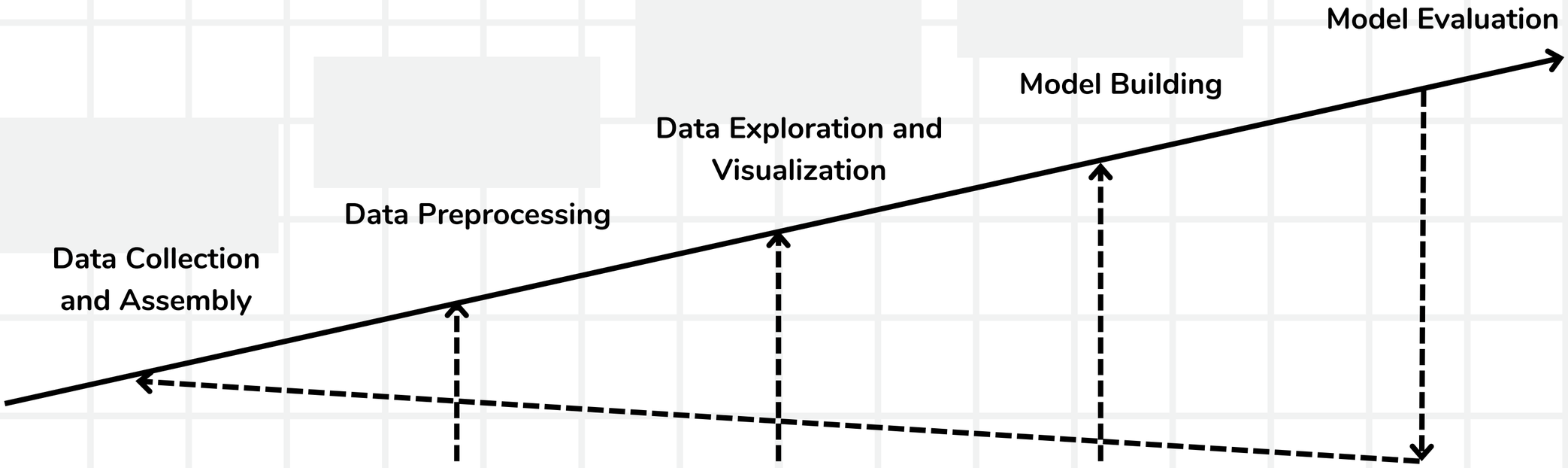
Examples and Exercises

What is data preprocessing?

INTRODUCTION

Data preprocessing is a fundamental step in the journey from raw data to actionable insights using machine learning models.

- **Techniques that transform and prepare the raw data into a suitable format for analysis and modeling.**
- **Ensures the accuracy and effectiveness of ML models.**
- **Addresses these challenges and enhance the quality of the data before feeding it into the models such as:**
  - **missing values, outliers, and differing scales.**

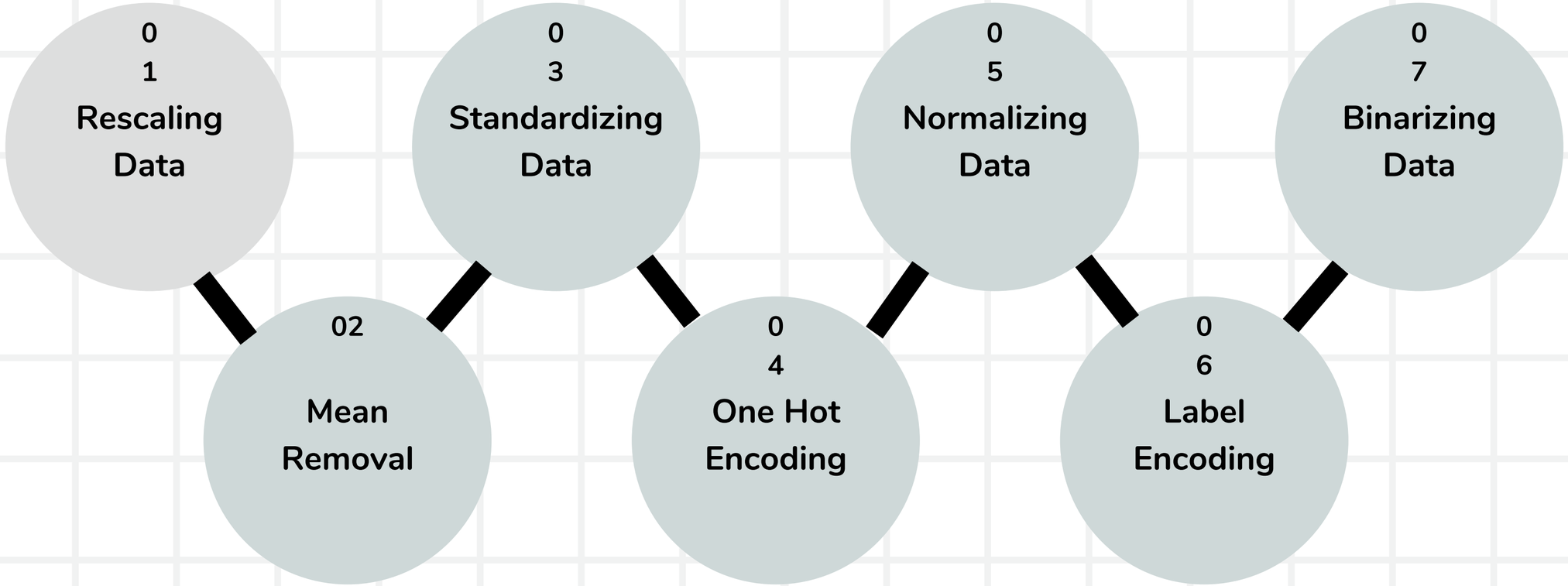

Model Evaluation
Model Building
Data Exploration and
Visualization
Data Preprocessing
Data Collection
and Assembly

NOTE:
Raw data collected from various sources is seldom ready to be used directly in machine learning models.

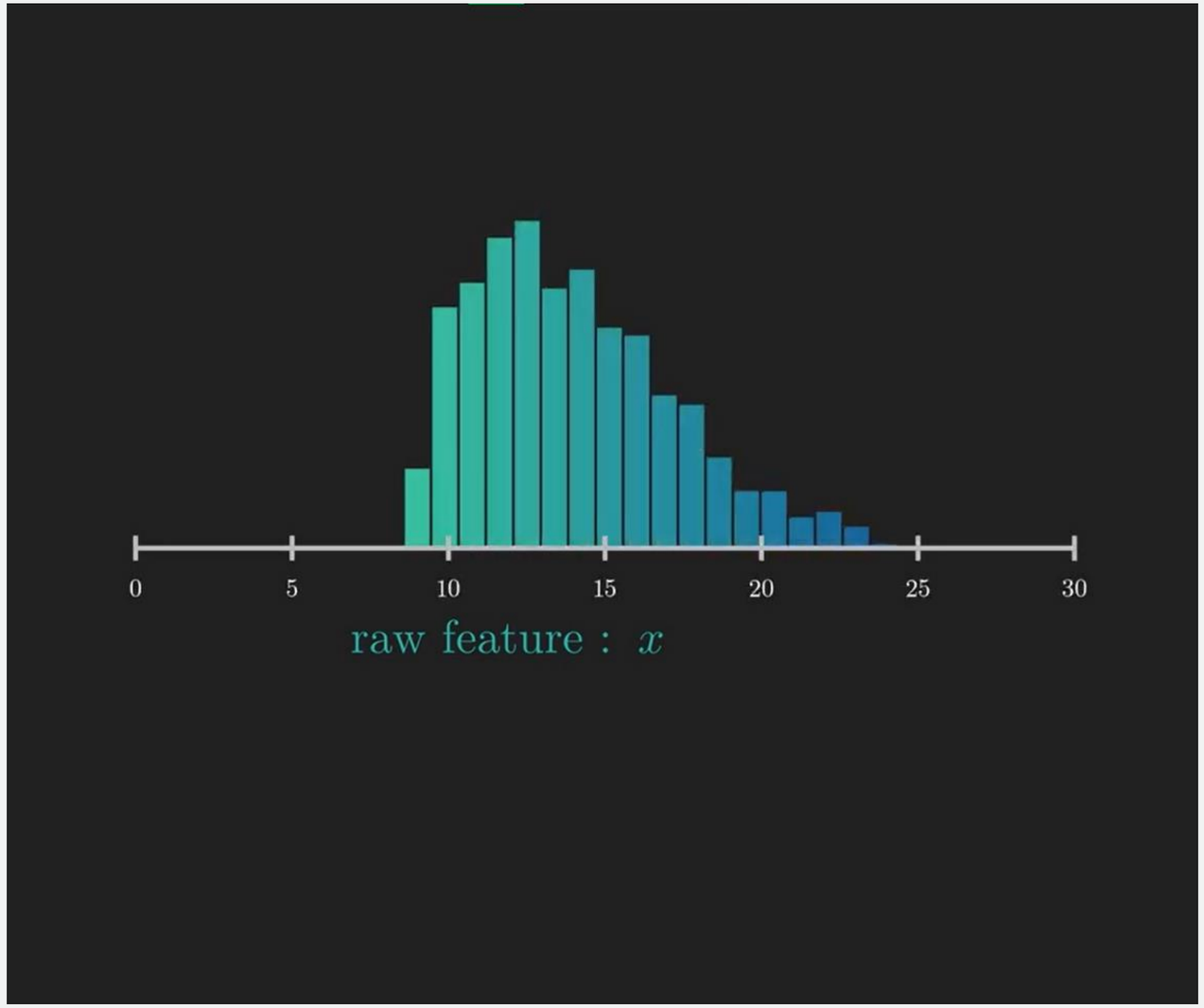

DATA BEFORE AND AFTER PREPROCESSING
DATA AFTER PREPROCESSING
DATA BEFORE PREPROCESSING

Techniques for Data Preprocessing

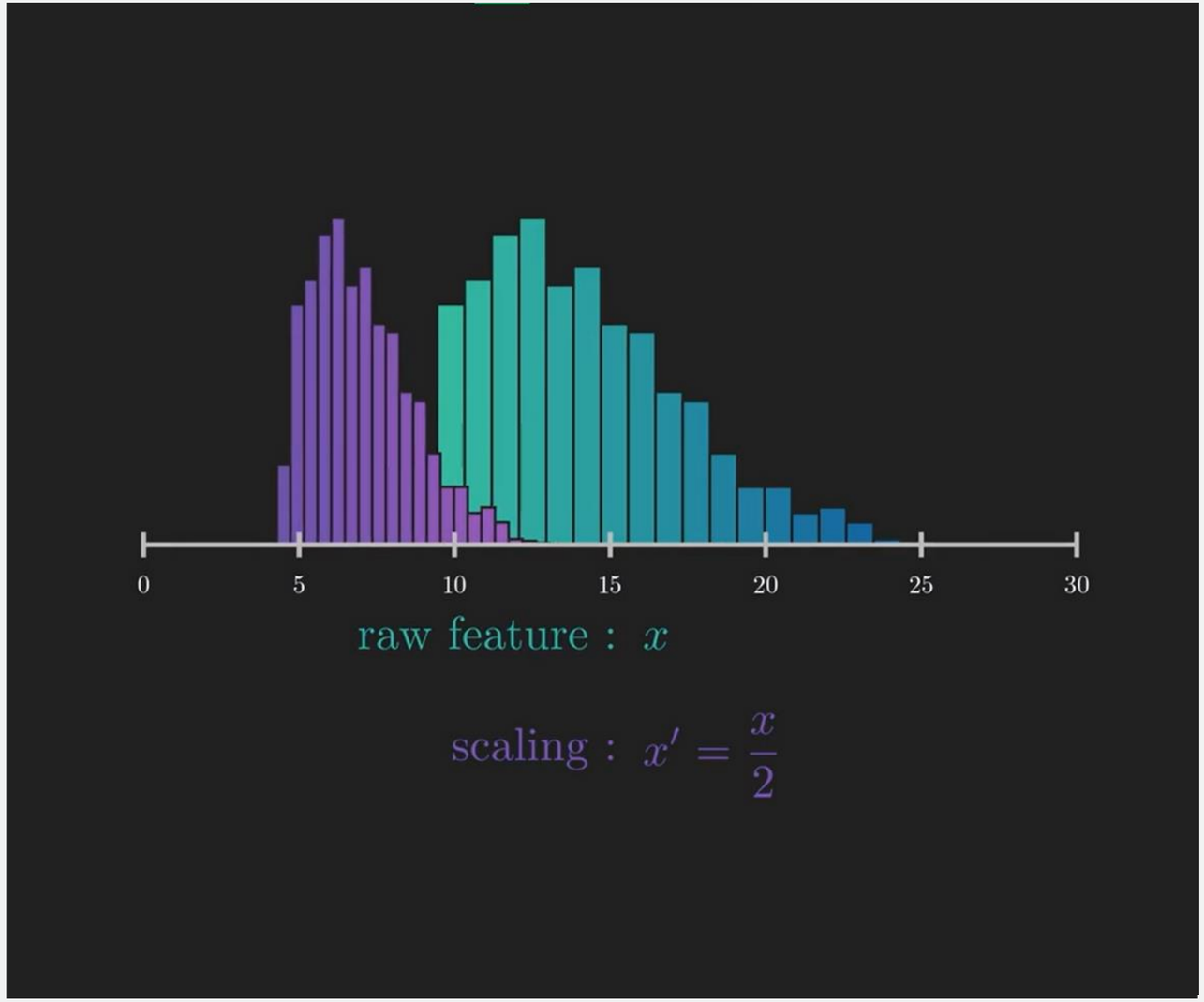
Data Preprocessing for Machine Learning



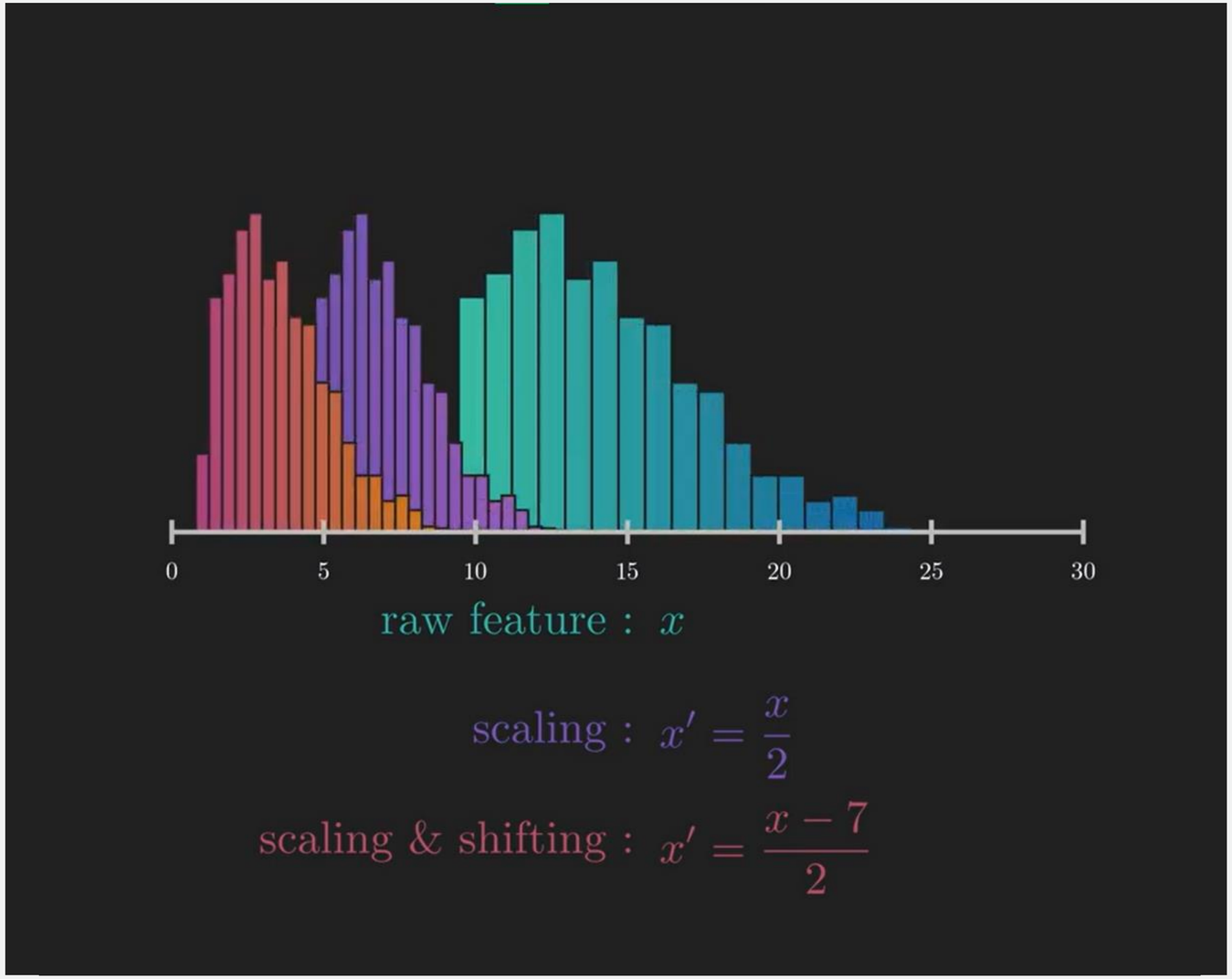
SCALING AND SHIFTING DATA



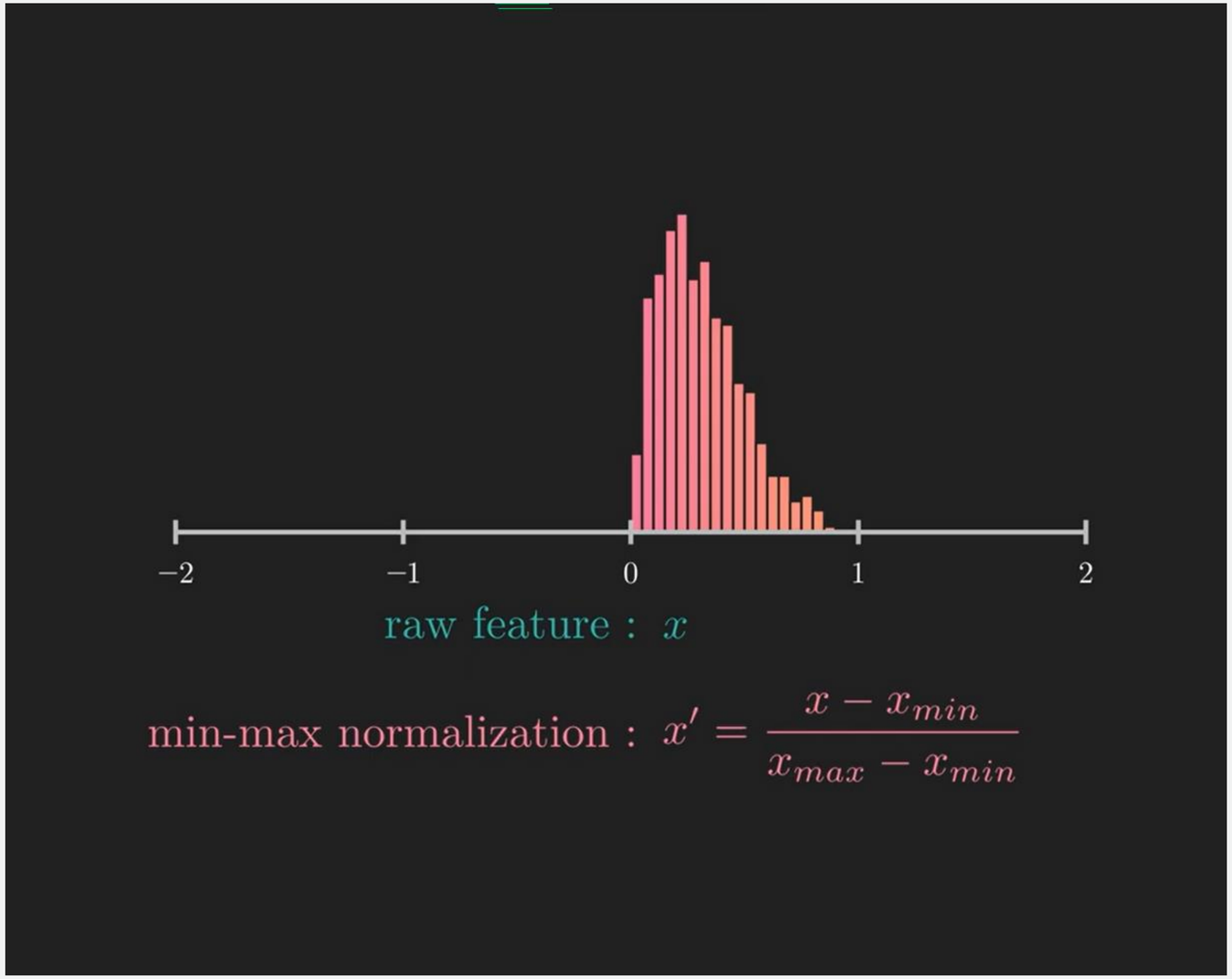
SCALING AND SHIFTING DATA



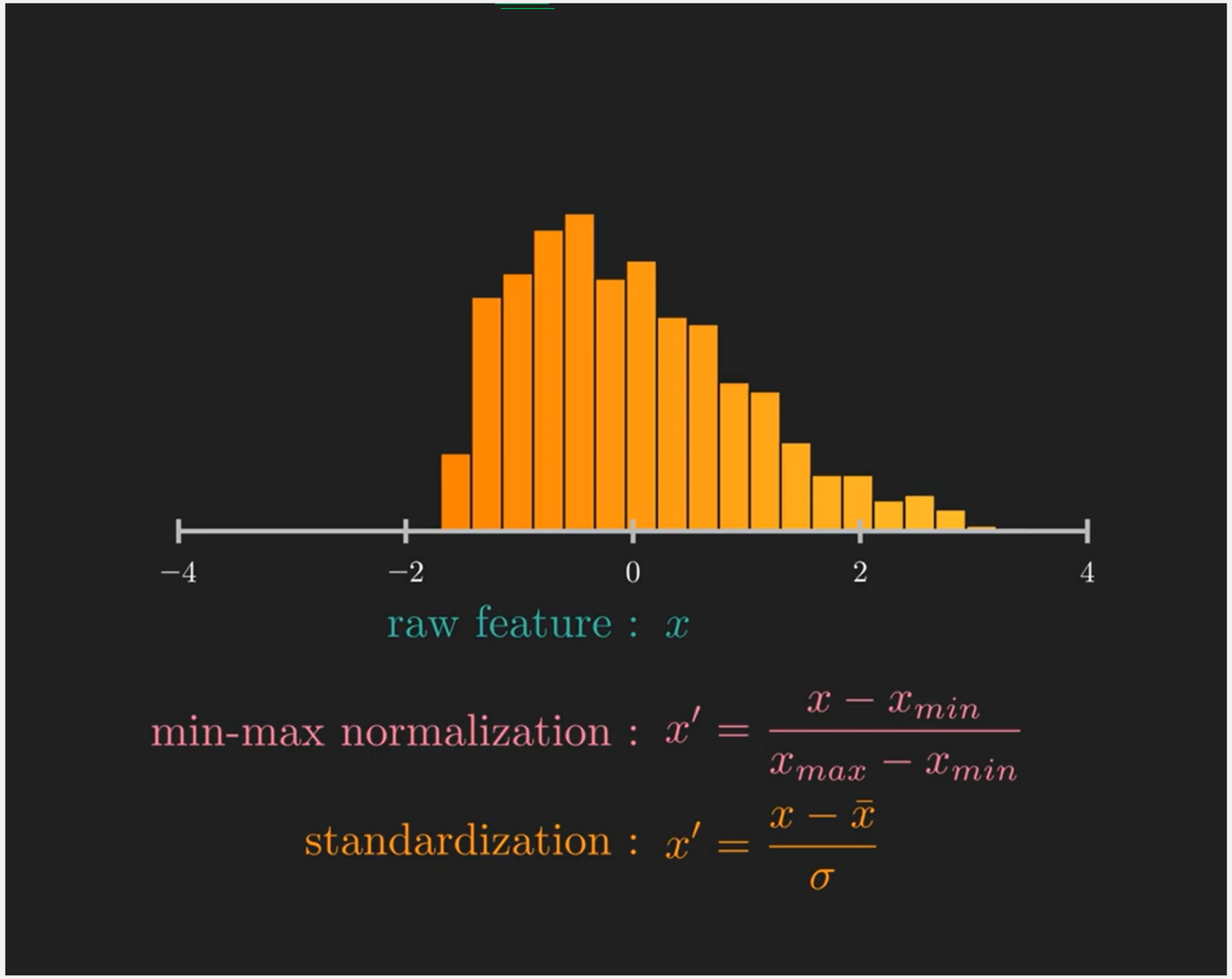
SCALING AND SHIFTING DATA



SCALING AND SHIFTING DATA



SCALING AND SHIFTING DATA



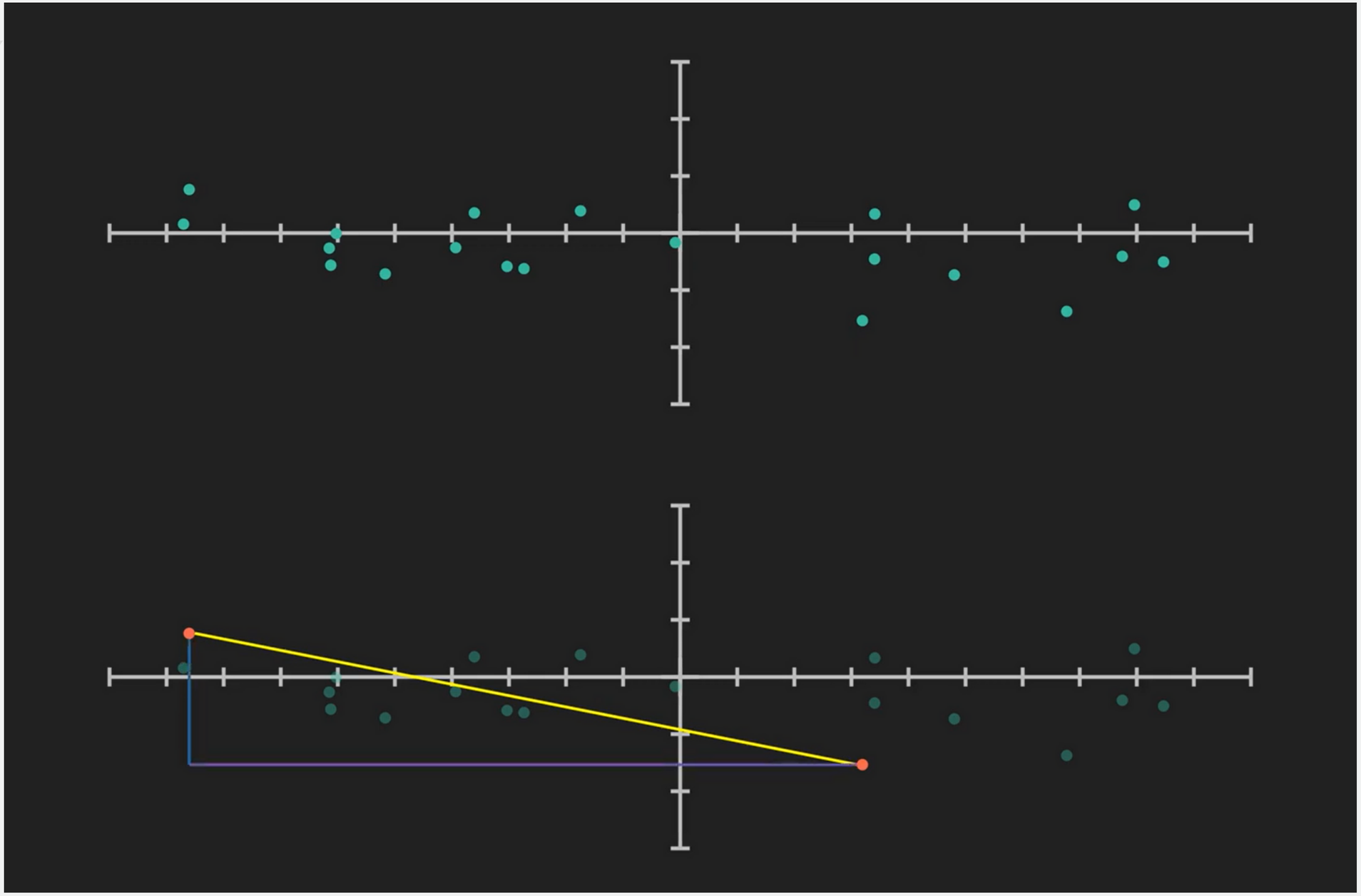
Why perform Feature Scaling?

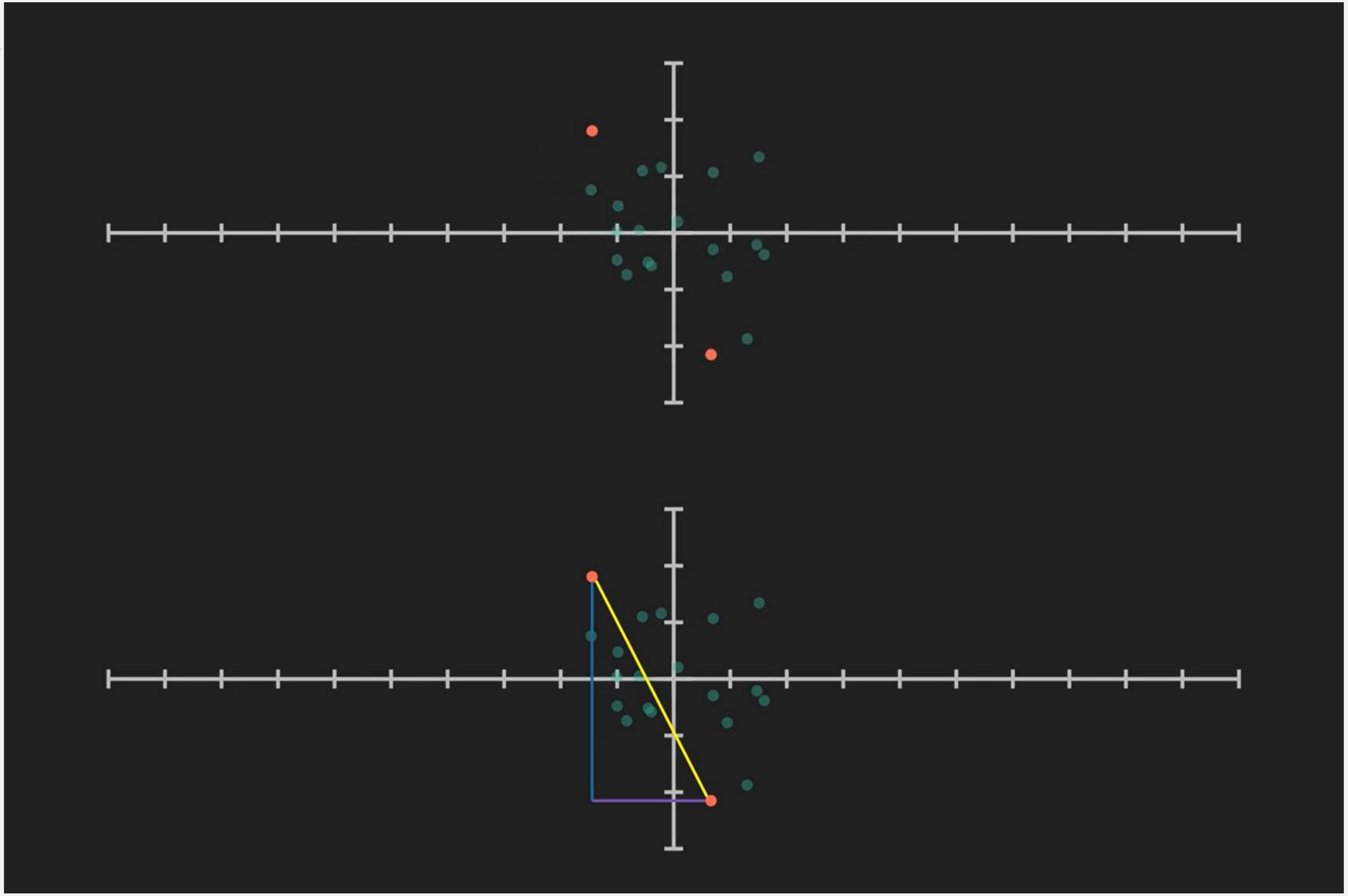
REASONS FOR SCALING DATA

Faster Convergence

Improved efficiency and interpretability

Computing distance appropriately

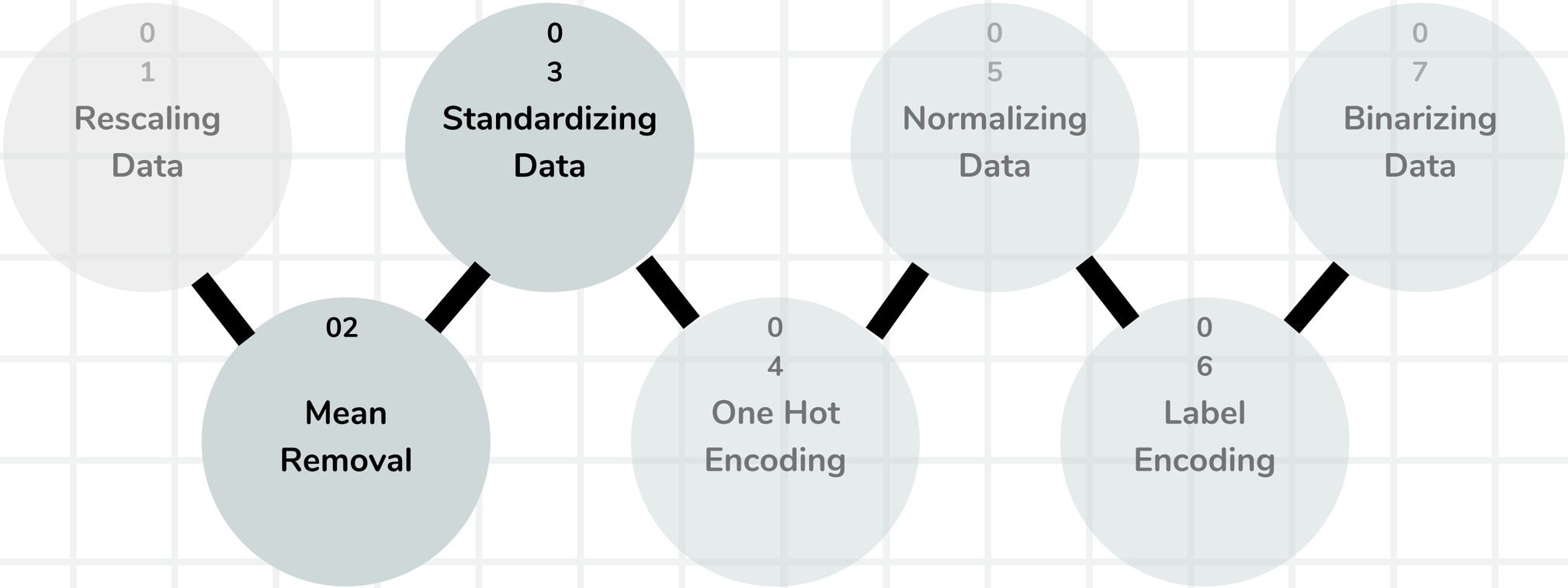




FEATURE SCALING

- Technique to standardize the independent features present in the data in a fixed range.
- Done to handle highly varying magnitudes or values or units.
- If feature scaling is not done, then a machine learning algorithm tends to weigh greater values, higher and consider smaller values as the lower values, regardless of the unit of the values.

Data Preprocessing for Machine Learning



STANDARDIZATION

a.k.a. Z-Score Normalization

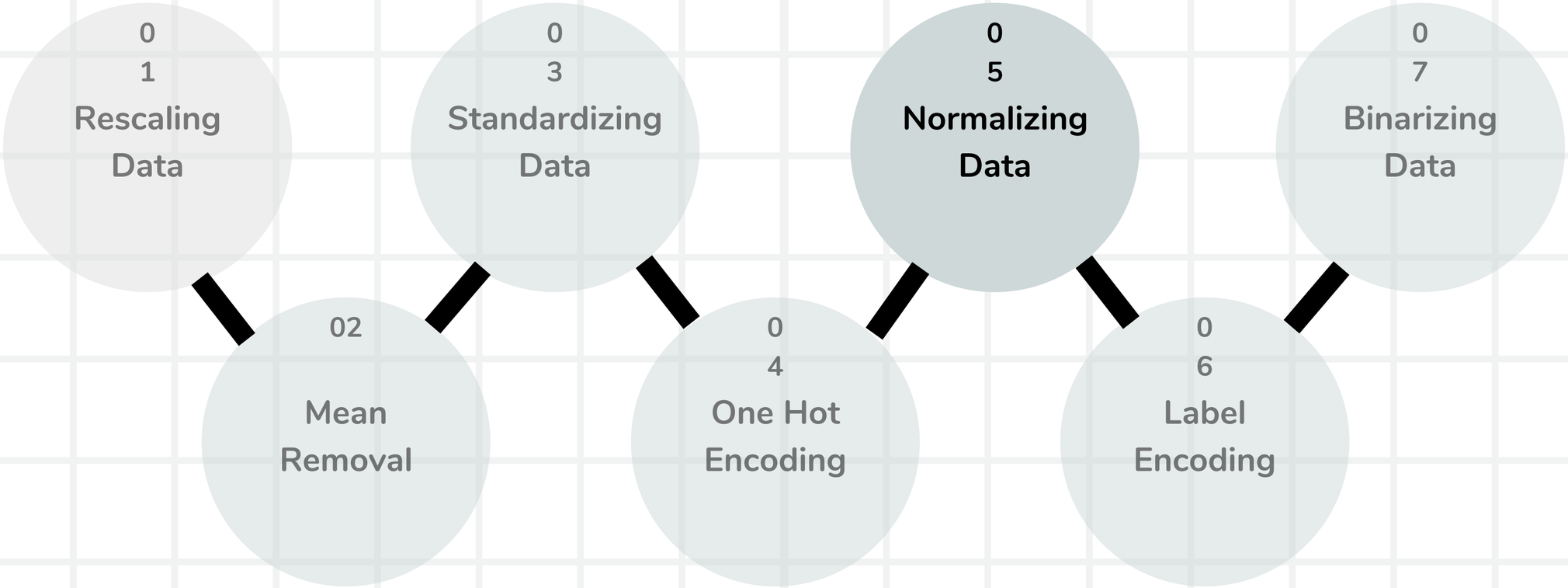
$$X_{new} = \frac{X_i - X_{mean}}{\text{Standard Deviation}}$$

- Transforms the data distribution to a mean of 0 and standard deviation of 1.
- Helpful in cases where the data follows a Gaussian distribution.

Example Applications:

- Gradient-based Optimization Algorithms
- Principal Component Analysis (PCA)
- Distance-Based Algorithms

Data Preprocessing for Machine Learning



NORMALIZATION

a.k.a. Min-Max Scaling

$$X_{normalized} = \frac{(X - X_{minimum})}{(X_{maximum} - X_{minimum})}$$

- Scales the range to [0, 1] or sometimes [-1, 1].
- Useful when there are no outliers as it cannot cope up with them.

Example Applications:

- Neural Networks (Inputs)
- Image Processing (Pixel brightness)
- Algoriths with bounded inputs

ROBUST SCALING

$$X_{new} = \frac{X - X_{median}}{IQR}$$

- Utilizes median and interquartile range to scale the data (Q3-Q1)
- Improved normalization method on data with several outliers.

Benefits:

- Resistent to outliers
- Preserves facts integrity
- Robust to outliers
- Handles skewed data

Normalize or standardize?

1. Exam Scores

2. Temperature Readings

3. Stock Market Data

4. Customer Reviews

5. Health Metrics

6. Housing Prices

7. Sensor Data

8. Sales Records

9. IQ Test Scores

10. SocMed Engagement Metrics

1. Exam Scores

2. Temperature Readings

3. Stock Market Data

4. Customer Reviews

5. Health Metrics

6. Housing Prices

7. Sensor Data

8. Sales Records

9. IQ Test Scores

10. SocMed Engagement Metrics

Normalize

Standardize


Normalization
Standardization


DATA SCIENTIST

NORMALIZATION VS STANDARDIZATION

Normalization	Standardization
Minimum and maximum value of features are used for scaling	Mean and standard deviation is used for scaling
It is used when features are of different scales	It is used when we want to ensure zero mean and unit standard deviation
Scales values between [0, 1] or [-1, 1]	It is not bounded to a certain range
It is really affected by outliers	It is much less affected by outliers
Scikit-Learn provides a transformer called MinMaxScaler for Normalization	Scikit-Learn provides a transformer called StandardScaler for standardization
This transformation squishes the n-dimensional data into an n-dimensional unit hypercube	It translates the data to the mean vector of original data to the origin and squishes or expands
It is useful when we don’t know about the distribution	It is useful when the feature distribution is Normal or Gaussian
It is a often called as Scaling Normalization	It is a often called as Z-Score Normalization

RECAP

- Data preprocessing is crucial due to challenges posed by raw data such as noise, outliers, and differing scales.
- Standardization brings data to a common scale with mean 0 and standard deviation 1.
- Normalization scales data to a range of [0, 1], often used with neural networks.
- Choosing the right preprocessing technique depends on data characteristics and algorithm requirements.
- Standardization and normalization have distinct purposes: centering vs. scaling data.
- Python libraries like NumPy and scikit-learn provide tools for implementing these techniques.
- Proper data preprocessing enhances model performance, but improper preprocessing can lead to issues like data leakage.
- Data preprocessing is an essential step toward building accurate and reliable machine learning models.

Hands-on Demonstration

READ MORE...

- https://www.geeksforgeeks.org/normalization-vs-standardization/
- https://www.geeksforgeeks.org/ml-feature-scaling-part-2/
- https://scikit-learn.org/stable/modules/preprocessing.html#normalization
- https://www.kaggle.com/getting-started/159643 https://www.kdnuggets.com/2020/04/data-transformation-standardization-normalization.html
- https://www.kaggle.com/code/durgancegaur/a-guide-to-any-classification-problem
- https://subscription.packtpub.com/book/data/9781789808452/1/ch01lvl1sec05/data-preprocessing-using-mean-removal
- https://www.analyticsvidhya.com/blog/2020/04/feature-scaling-machine-learning-normalization-standardization/
- https://towardsdatascience.com/scale-standardize-or-normalize-with-scikit-learn-6ccc7d176a02

FIN

DAYATA | YAP

CS 4103 | Intelligent Systems

September 6, 2023