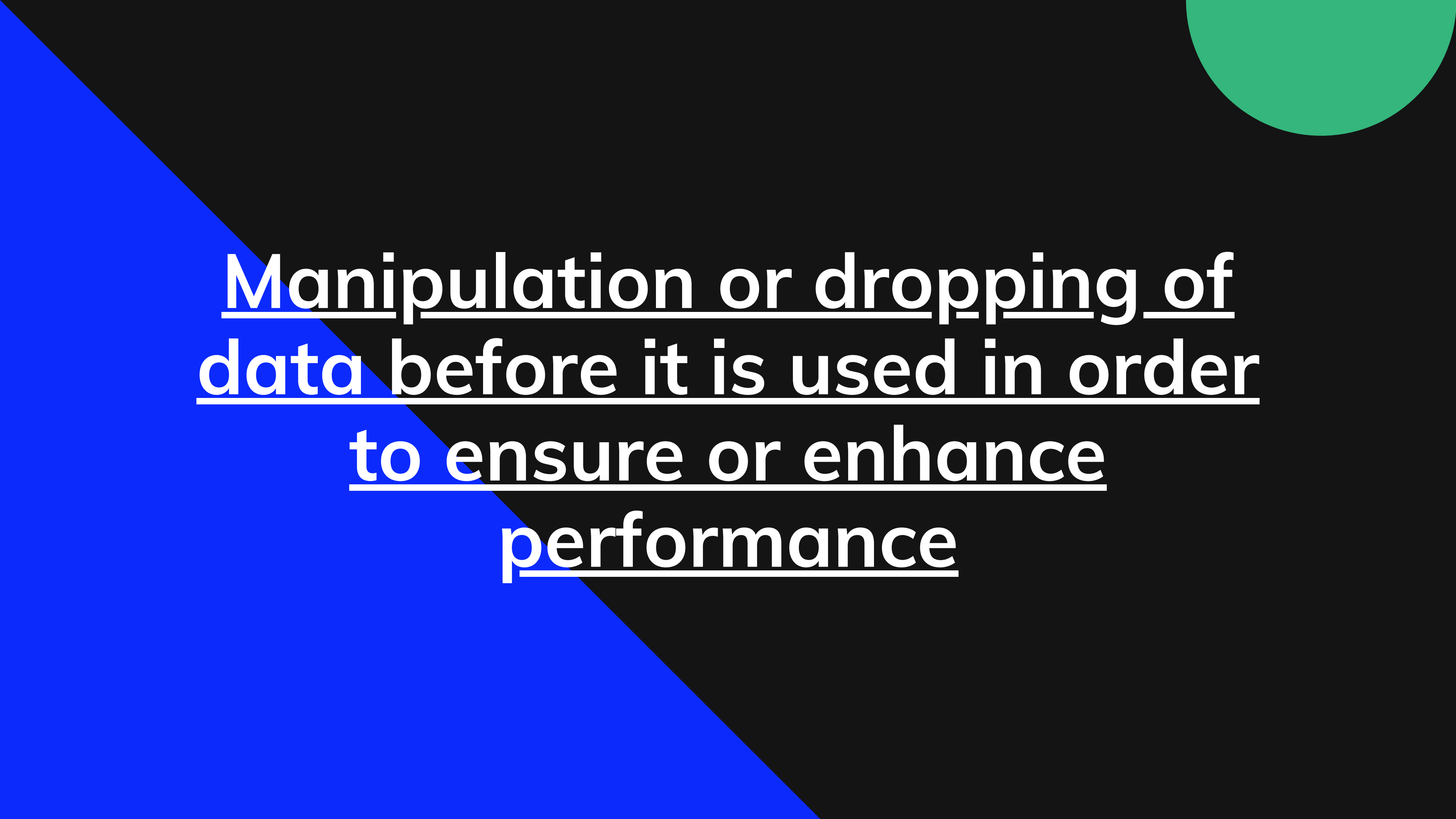




Data Pre-processing



What is Data Pre-Processing?



Manipulation or dropping of
data before it is used in order
to ensure or enhance
performance

The data can have many irrelevant and missing parts. To handle this part, data cleaning is done. It involves handling of missing data, noisy data etc.

You can do this in two ways, removal of entries or fill in missing values.



DATA CLEANING

HANDLING NOISY DATA

Noisy data is meaningless data that can't be interpreted by machines. It can be generated due to faulty data collection, data entry errors etc.



Binning

The whole data is divided into segments of equal size called bins. Each segmented is handled separately.



Regression

This is used to smooth the data and will help to handle data when unnecessary data is present



Clustering

This is used for finding the outliers and also in grouping the data. Clustering is generally used in unsupervised learning.

DATA INTEGRATION

This is usually used when compiling data from multiple sources, each of which would have different formats of storage and sources of information.

This commonly includes matching different names for the same values, and removal of unnecessary attributes.

Once data clearing has been done, we need to consolidate the quality data into alternate forms by changing the value, structure, or format of data.

This helps data to be better analysed by the developed models. This also sets the format in which the model receives data.

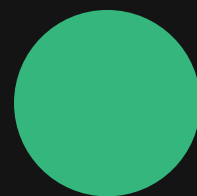
DATA TRANSFORMATION

NORMALIZATION

It involves scaling of numerical attributes, so that each attribute has nearly equal significance.

Normalization is one of the most widely used techniques to transform data

A few ways to normalise the data



Min-Max Normalization

Used for data having a range. It is used to transform the data to a range of 0 to 1 or -1 to 1.



Standardization (Z-Score)

The data is rescaled such that the mean is 0 and variance is 1.



Decimal Scaling

Scaling values by a power of 10, so as to eliminate the need for decimals. It is rarely used.



Clipping

Outlier values that greater/lesser than the maximum/minimum value are set to the maximum/minimum respectively.

MIN-MAX NORMALIZATION

$$y_i = \frac{x_i - x_{min}}{x_{max} - x_{min}}$$

It is used for data having a range.

The above formula transforms the data to a range of 0 to 1.

For transforming from -1 to 1, we can use

$$z_i = 2y_i - 1$$

STANDARDIZATION

Scales the mean to zero and variance (as well as standard deviation) to 1.

$$\bar{x} = \frac{1}{N} \sum_i^N x_i$$

Epsilon is an extremely small number to ensure that when variance is 0, there isn't an error.

$$y_i = \frac{x_i - \bar{x}}{\sqrt{\sigma^2 + \epsilon}}$$

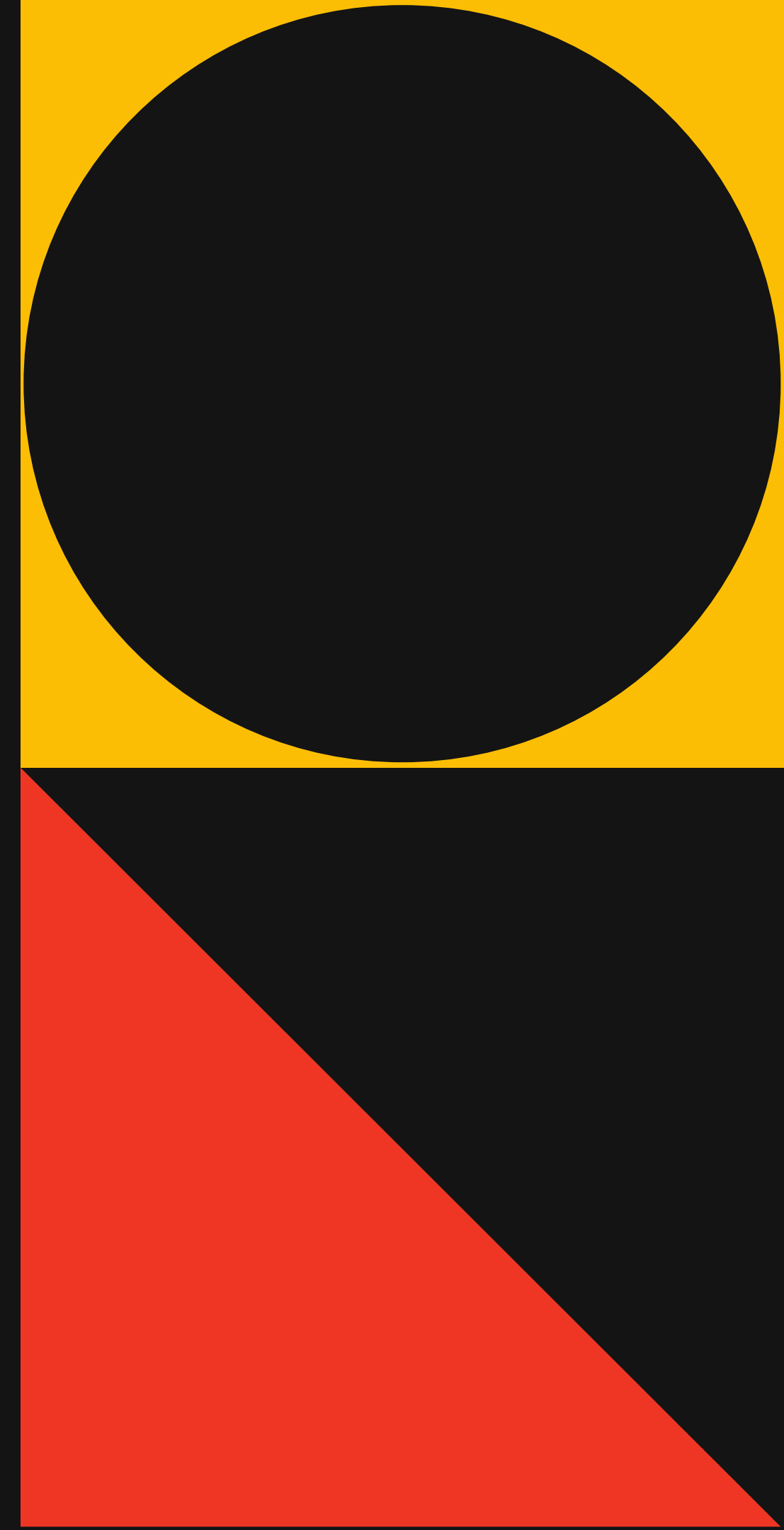
$$\sigma^2 = \frac{1}{N} \sum_i^N (x_i - \bar{x})^2$$

ATTRIBUTE SELECTION

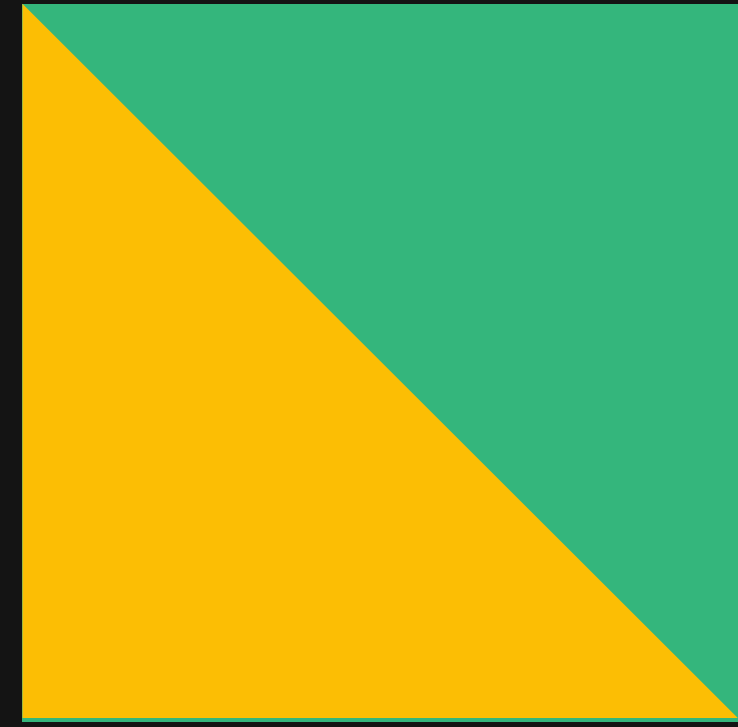
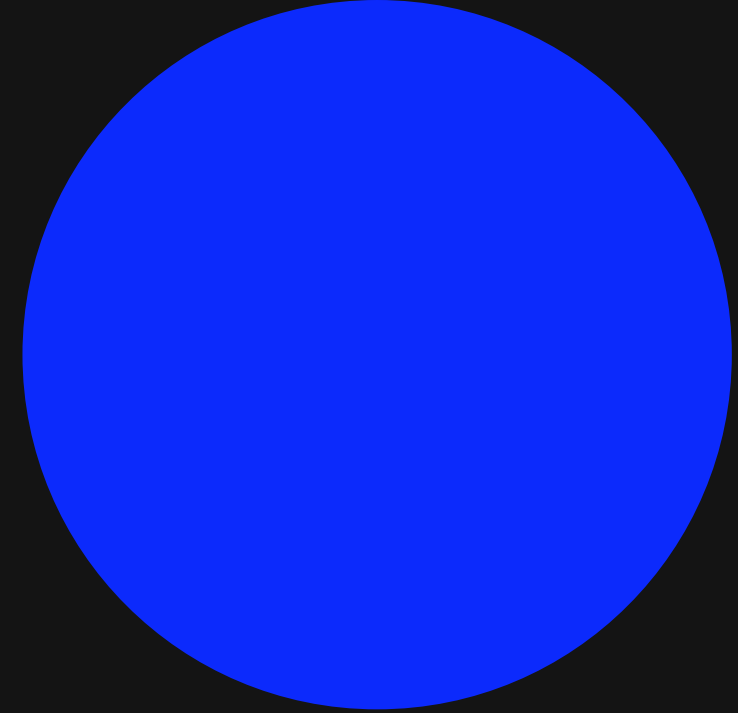
New attributes are introduced in the data based on evaluation of earlier attribute(s).

AGGREGATION

Presenting the data in summary format. Used mainly to check operations done on previous data and their overall effect.



Regularization





TYPES OF REGULARIZATION

Modifying the loss function

Modifying the Sampling method

Modifying training algorithms

MODIFYING THE LOSS FUNCTION

L1 (Lasso) Regularisation

Penalty of sum of absolute weights scaled by a hyper parameter is added to the loss function.

L2 (Ridge) Regularisation

Penalty of sum of square of weights scaled again by a hyper parameter is added to the loss function.



L1 REGULARIZATION

Promotes Sparsity

L1 regularization promotes sparsity in the model by encouraging some coefficients to become exactly zero, effectively performing feature selection.

Feature Importance Ranking

It can provide a feature importance ranking based on the magnitude of the non-zero coefficients. Features with larger non-zero coefficients are considered more important.

$$loss = Error + \lambda \sum_i |weight_i|$$

When should it be used?

It works much better when your data has many correlated features. This also helps when you have low amount of data or high number of features

L2 REGULARIZATION

$$loss = Error + \lambda \sum_i (weight_i)^2$$

Encourages non-zero values

L2 regularization encourages small but non-zero coefficient values, distributing the impact of features across all variables.

When is it more useful?

It works much better when your data has many correlated features.

Feature Importance Ranking

It can provide a feature importance ranking based on the magnitude of the non-zero coefficients. Features with larger non-zero coefficients are considered more important.

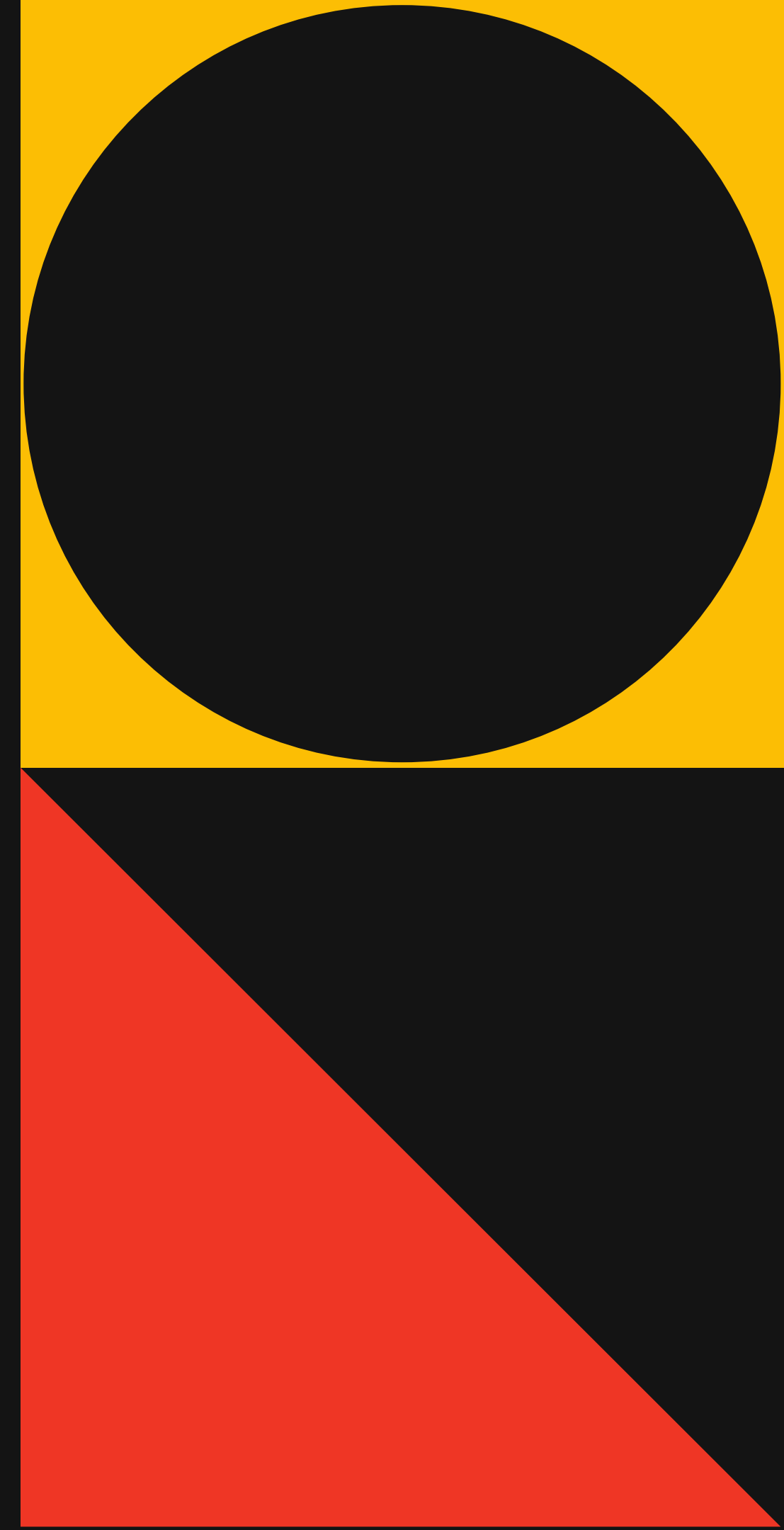
MODIFYING THE TRAINING ALGORITHM

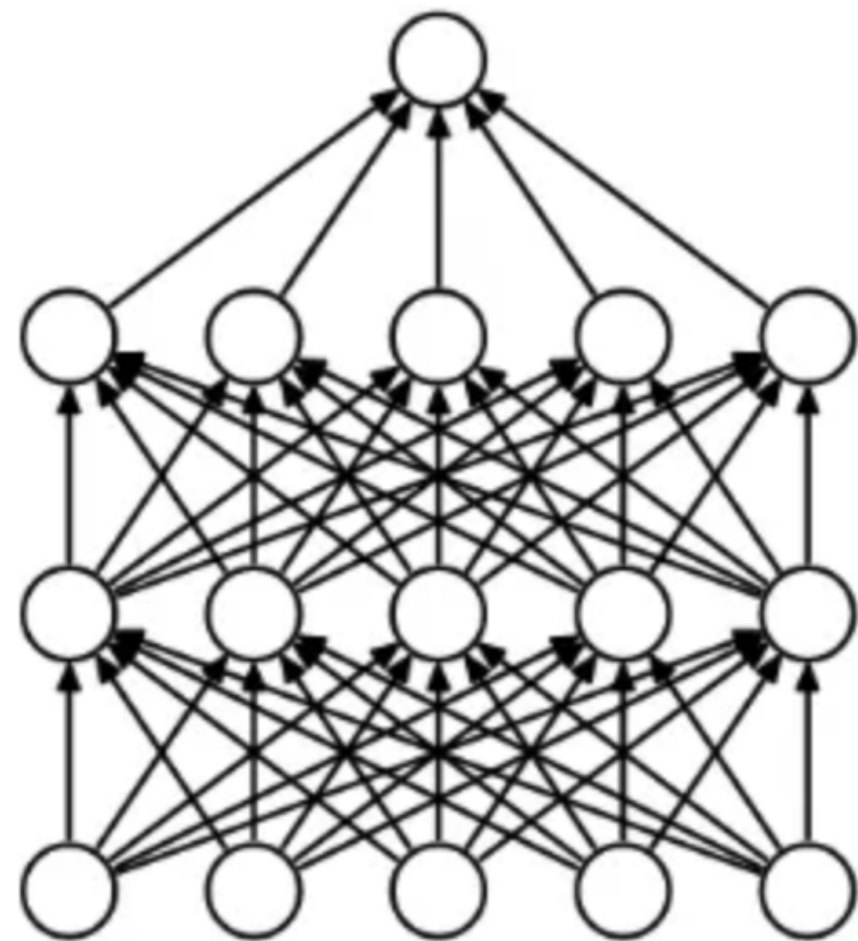
Dropout

In each training iteration, some connections are randomly dropped and the resultant output rescaled.

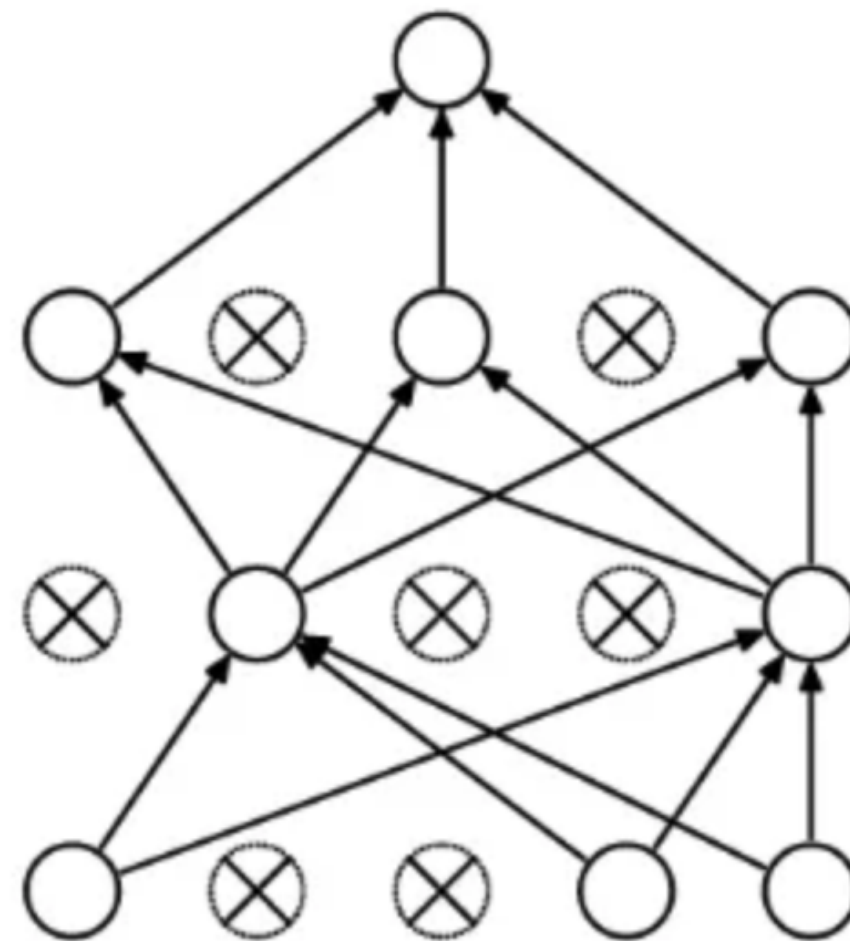
Injecting Noise

Introducing random variation while updating weights





(a) Standard Neural Net



(b) After applying dropout.

Dropout

Some nodes are randomly dropped and the resultant is rescaled to compensate for the dropped values.

By applying dropout during training, the network effectively trains multiple sub-networks, as different subsets of neurons are dropped out at each update step. This ensemble of sub-networks helps in reducing overfitting, as the network learns to generalize from a variety of different architectures. Dropout also acts as a form of regularization, as it discourages complex co-adaptations of neurons and encourages the learning of more robust features.



MODIFYING THE SAMPLING METHOD

Data Augmentation

Introduction of more synthetic data with noise, which makes the model more resistant to variations.

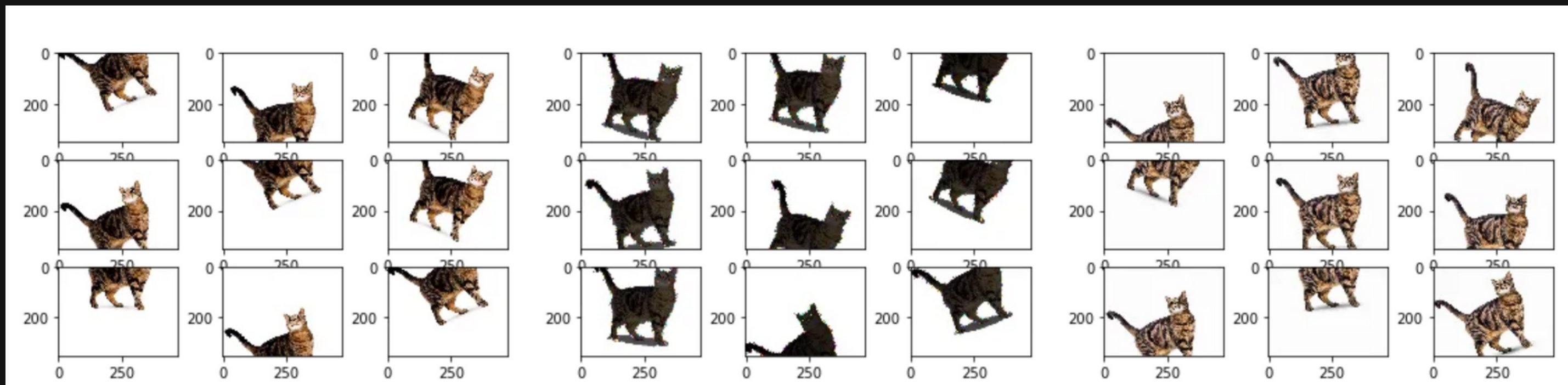
K-Fold Cross Validation

Dataset is divided into k equally sized subsets, and the model is trained and evaluated k times, each time using a different subset as the validation set and the remaining subsets as the training set.

Data Augmentation

Introduction of more synthetic data with noise, which makes the model more resistant to variations.

Hence, to smoothen out the entire feature space, we can generate artificial data based on the original data, like in images, we can flip and rotate, convert to grey scale, add noise to the images, crop, resize, change contrast, brightness, or introduce deformations.



K-Fold Cross Validation

Dataset is divided into k equally sized subsets, and the model is trained and evaluated k times, each time using a different subset as the validation set and the remaining subsets as the training set.



The purpose of training multiple models in k-fold cross-validation is to obtain a more reliable estimate of the model's performance by evaluating it on different subsets of the data. It helps in assessing the model's generalization ability and reducing the impact of data variability.

