Implementing Bin Counting and Feature Hashing

Janani Ravi

CO-FOUNDER, LOONYCORN

www.loonycorn.com

Overview

Converting continuous data into categorical data

Bucketing continuous data into bins

Bucketing data using Pandas and the KBinsDiscretizer

Hash nominal features to numeric features

Types of Data

Categorical

Male/Female, Month of year

Numeric (Continuous)

Weight in lbs, Temperature in °F

All other forms of data, such as text and image data, must be converted to one of these forms

Bucketing

Categorical

Male/Female, Month of year

Numeric (Continuous)

Weight in lbs, Temperature in °F

Bucketing techniques to convert continuous data to discrete categories

Hashing

Categorical

Male/Female, Month of year

Numeric (Continuous)

Weight in lbs, Temperature in °F

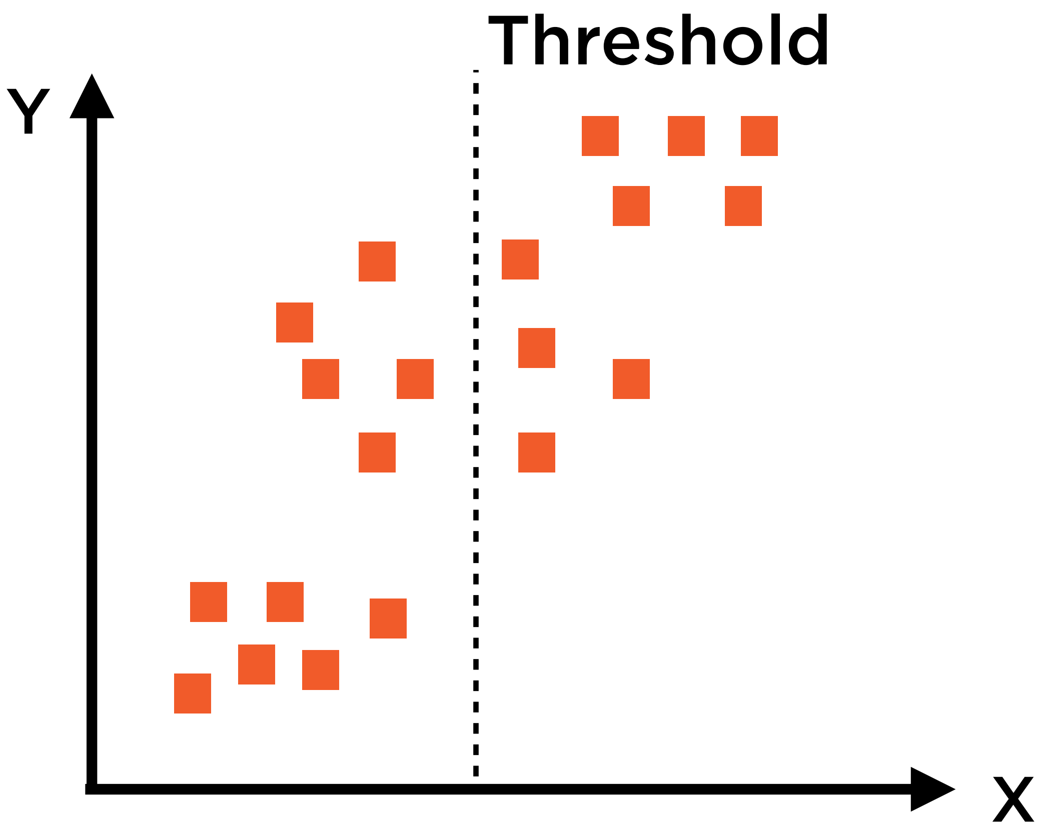
Converting data to numeric representations of lower dimensionality

Bucketing Continuous Data

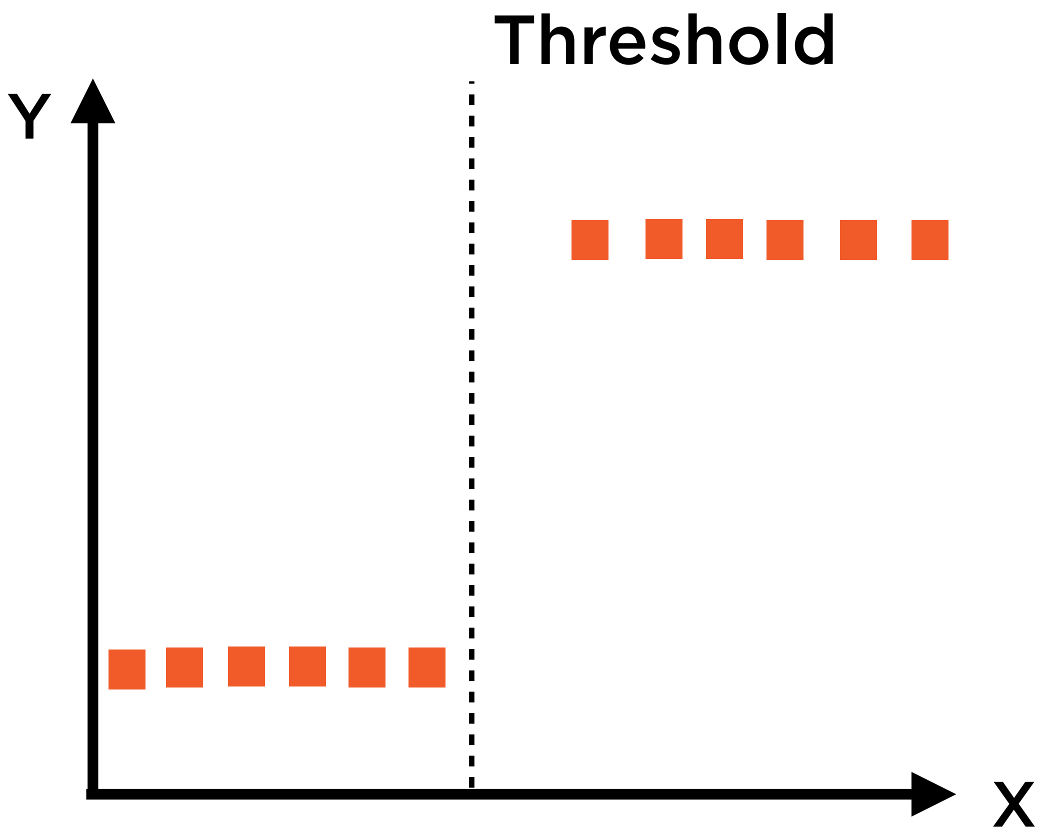
Binarizer

Converts continuous variable into a binary categorical variable based on a threshold specified by user.

Binarizer



Continuous
Input

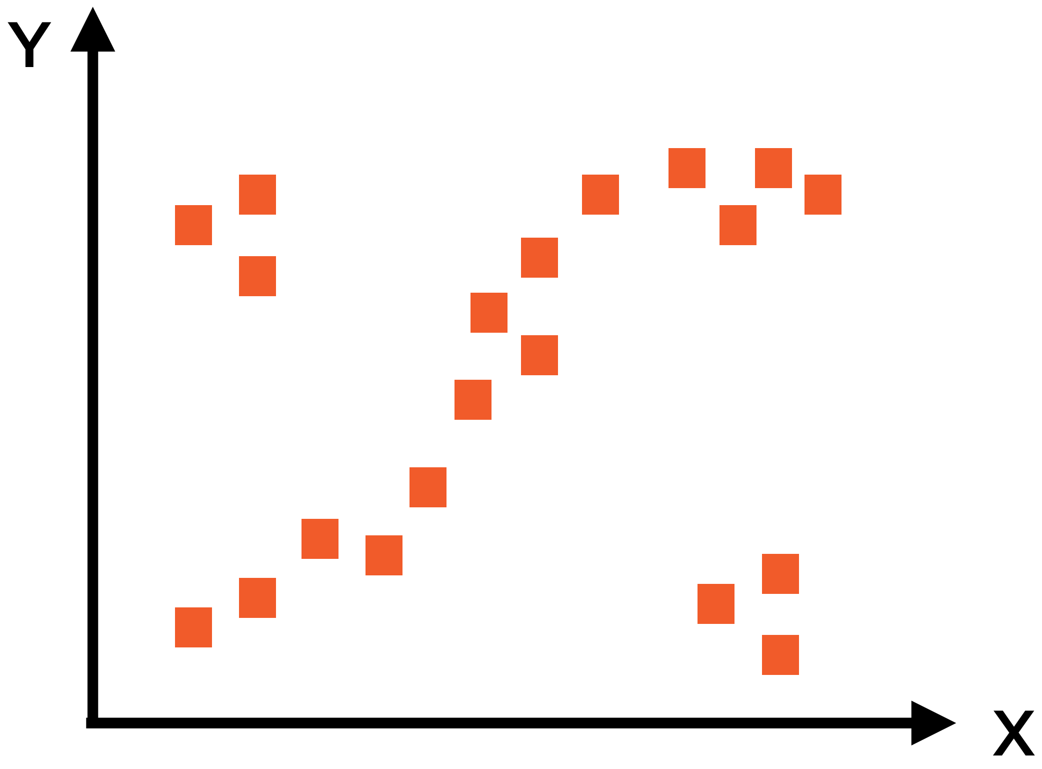


Binary Categorical
Output

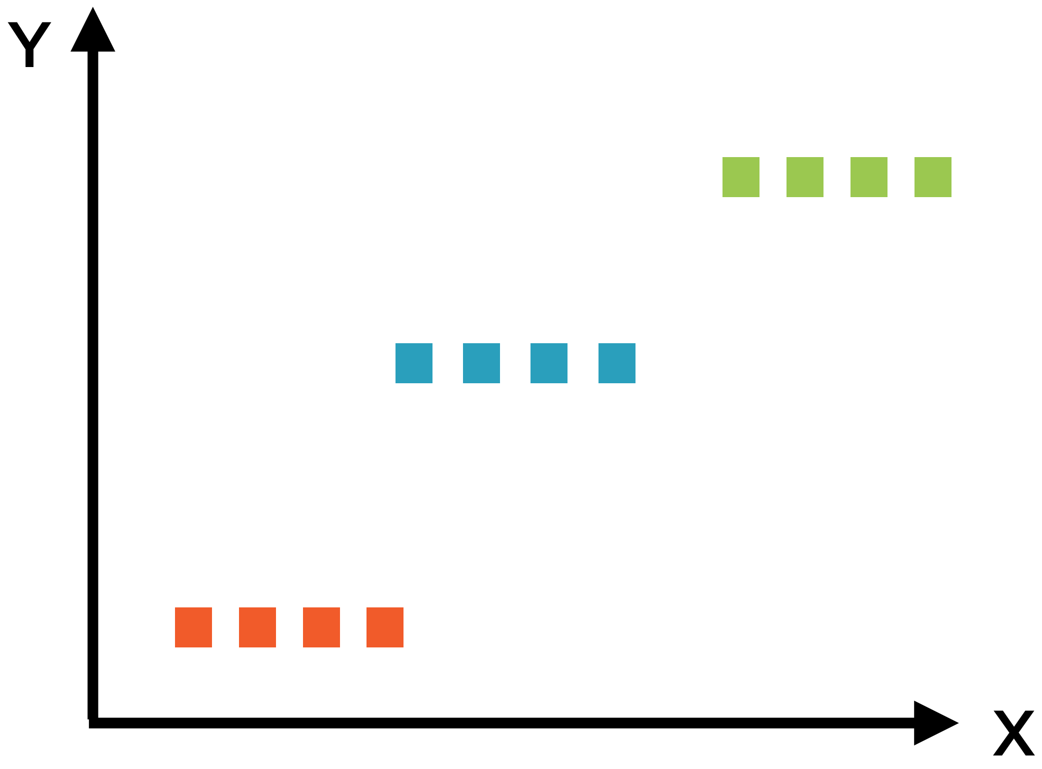
KBinsDiscretizer

Generalizes idea of binarizer; converts continuous data into categorical data arranged into a specified number of bins.

KBinsDiscretizer



Before



After (3 Bins)

KBinsDiscretizer Strategies

Uniform

Bin widths are constant in each feature

Quantile

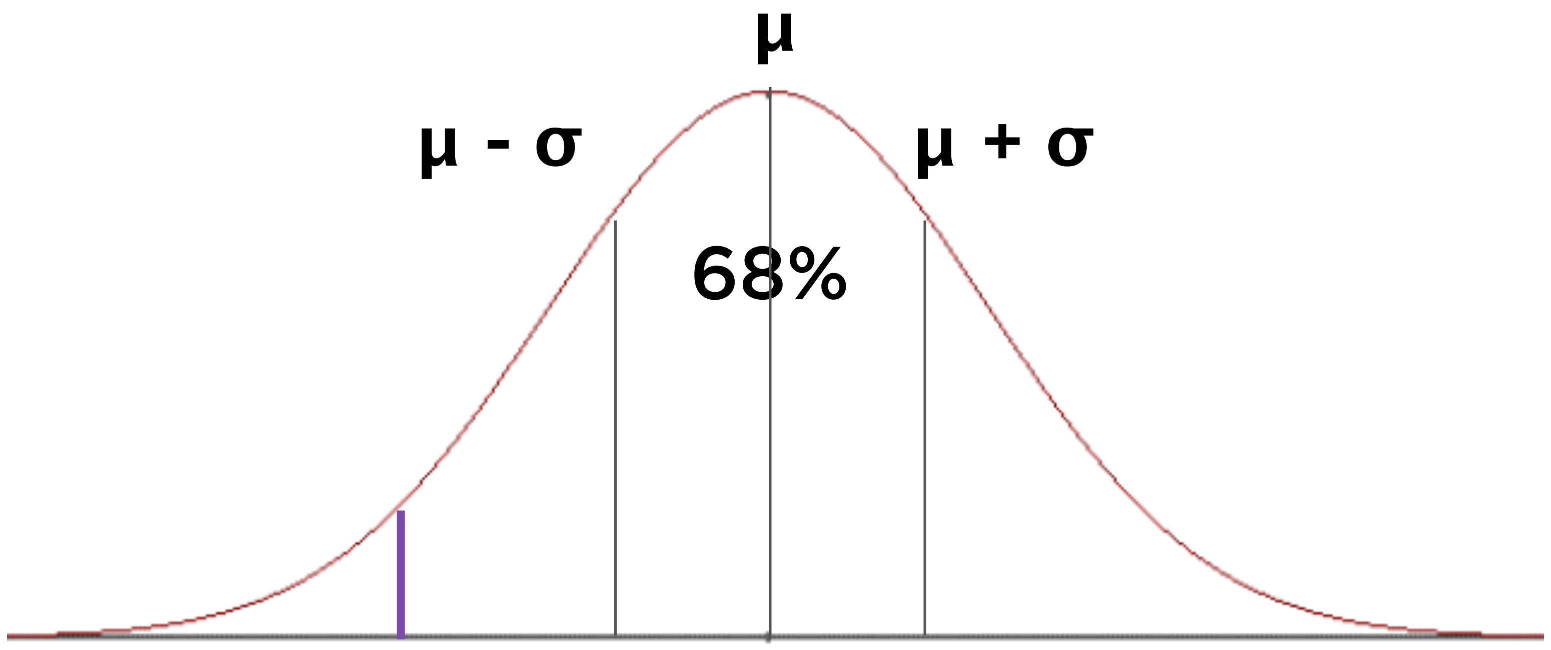
All bins in each feature have approximately the same number of samples

K-means

Bins based on the centroids of a K-means clustering procedure

A graph showing the count of values in each bin is called a Histogram

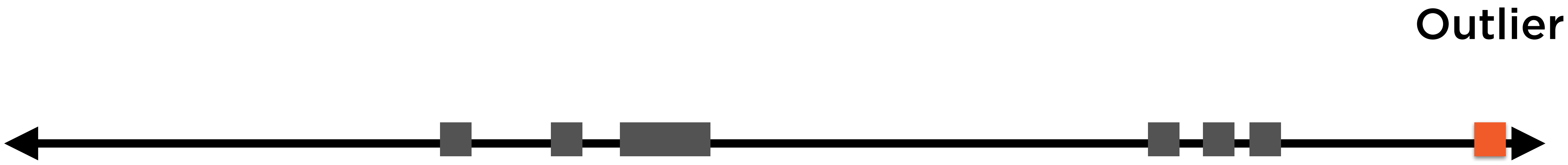
Continuous Distribution



N(μ,σ)

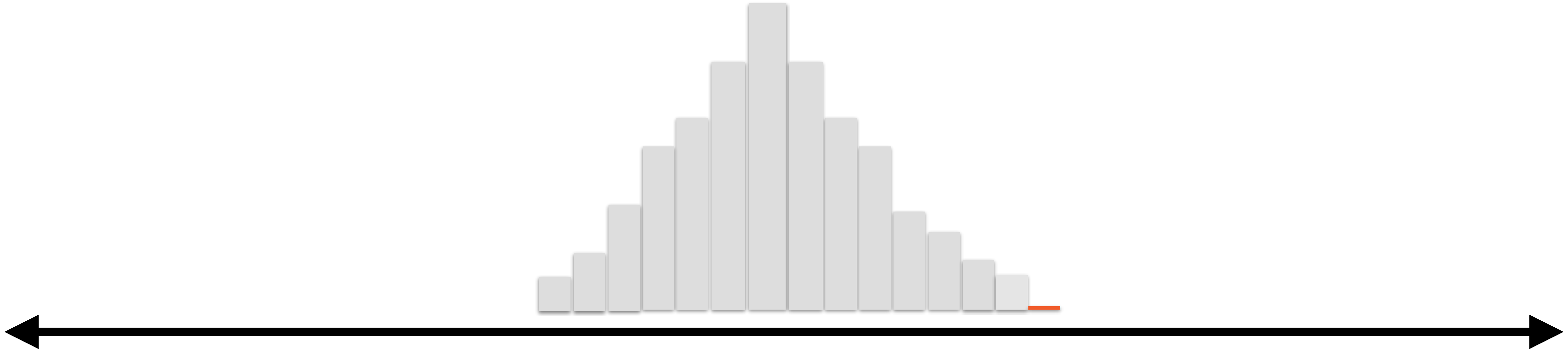
$$N(\mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Data Drawn from Distribution



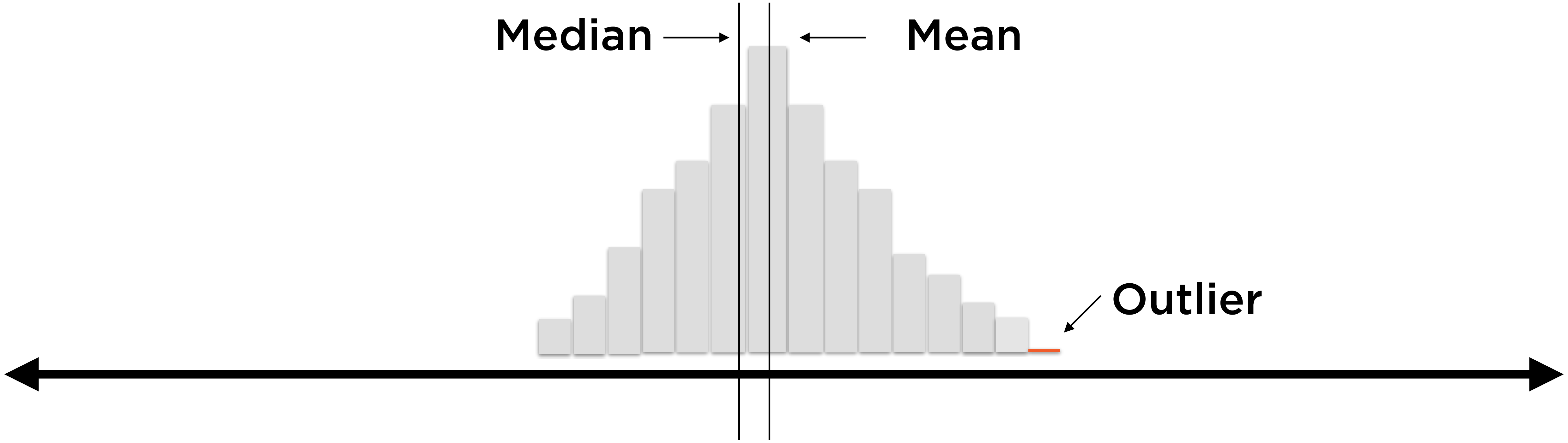
Outliers might represent data errors, or genuinely rare points legitimately in dataset

Histogram of Bin Counts



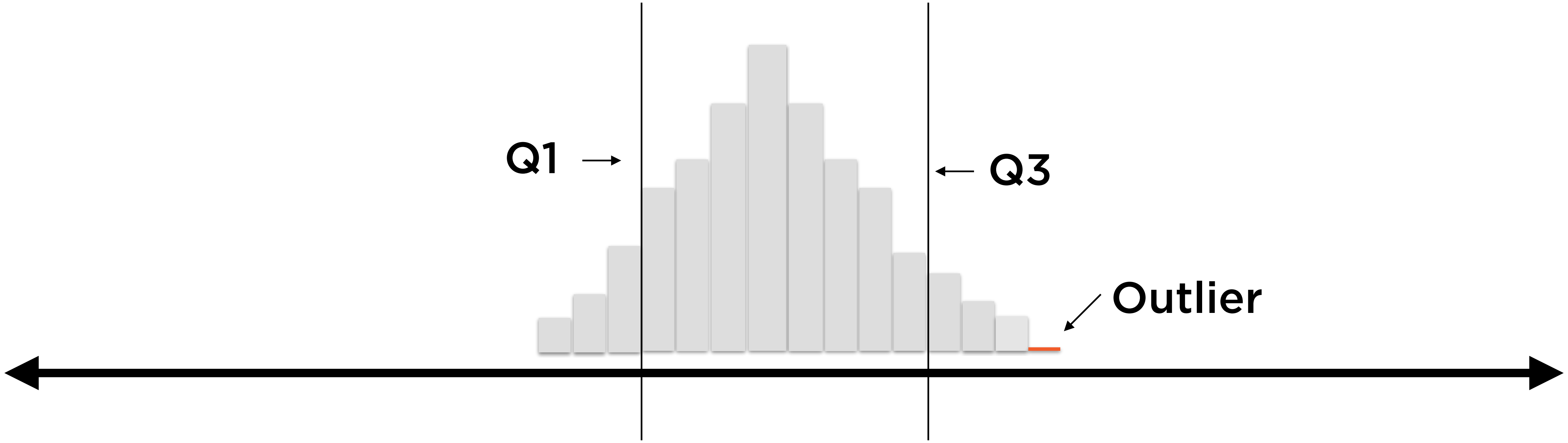
Bucketize data and count how many data points fall within each bucket

Median



Median = 50th percentile: 50% of points on either side

Histogram of Bin Counts



Q3 = 75th percentile: 75% of points smaller than this

Q1 = 25th percentile: 25% of points smaller than this

Inter-quartile Range (IQR) = 75th percentile - 25th percentile

Demo

Bucketing continuous data using Pandas

Demo

Discretizing continuous data using the KBinsDiscretizer

Hashing

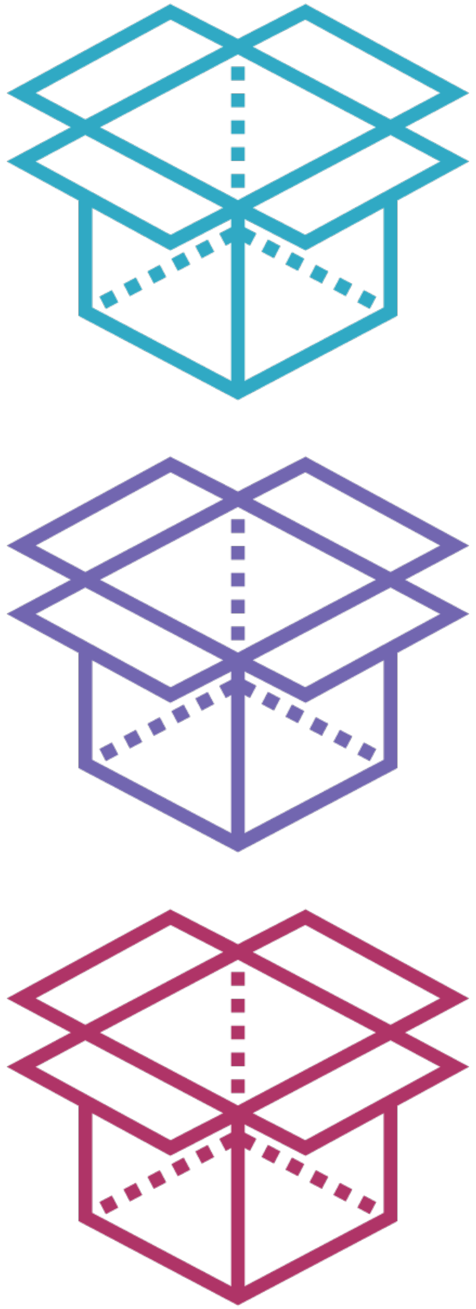
Hashing

A technique that allows you to lookup specific values very quickly

Hashing

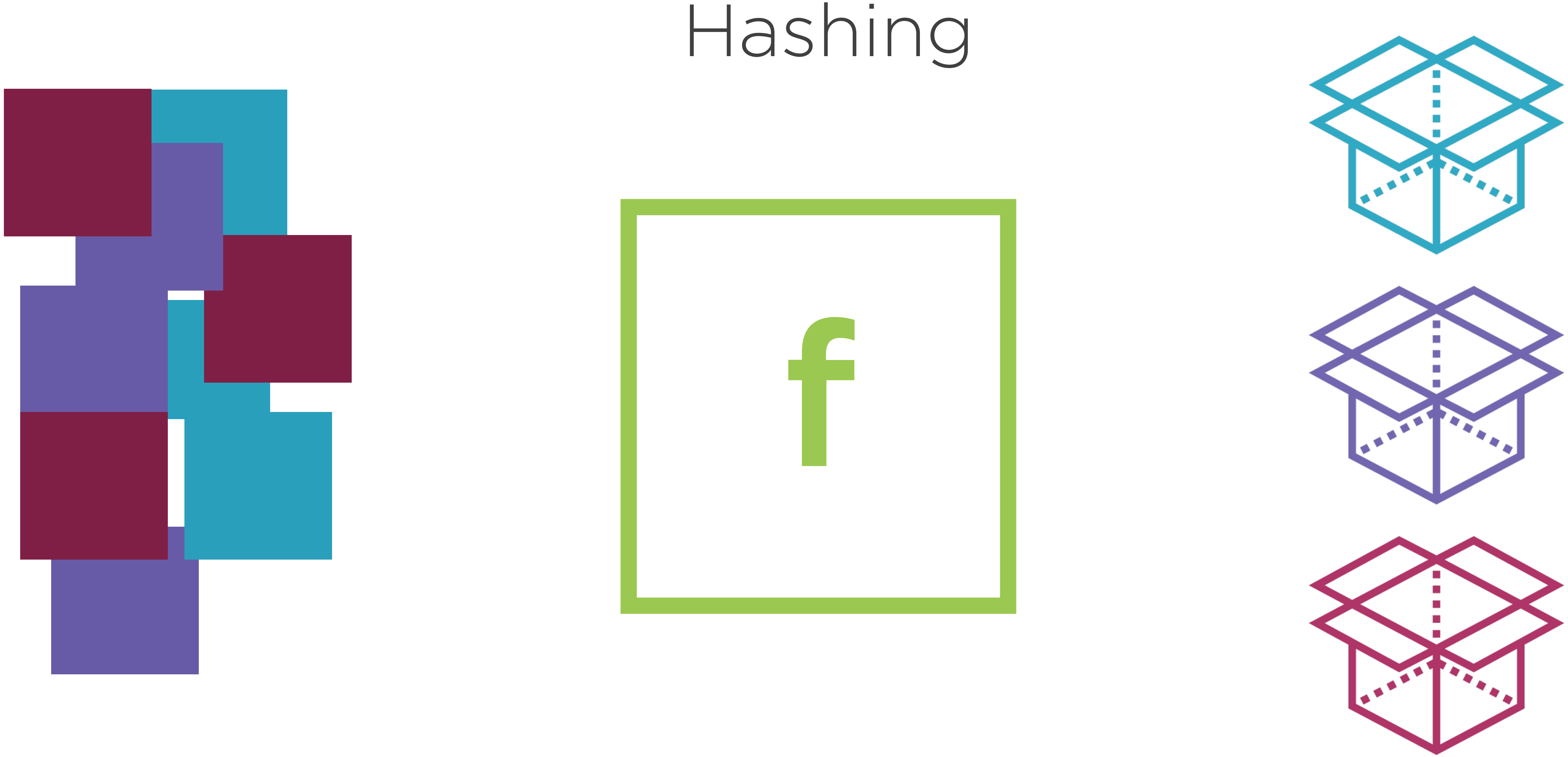
Also can be used to perform dimensionality reduction

Hashing



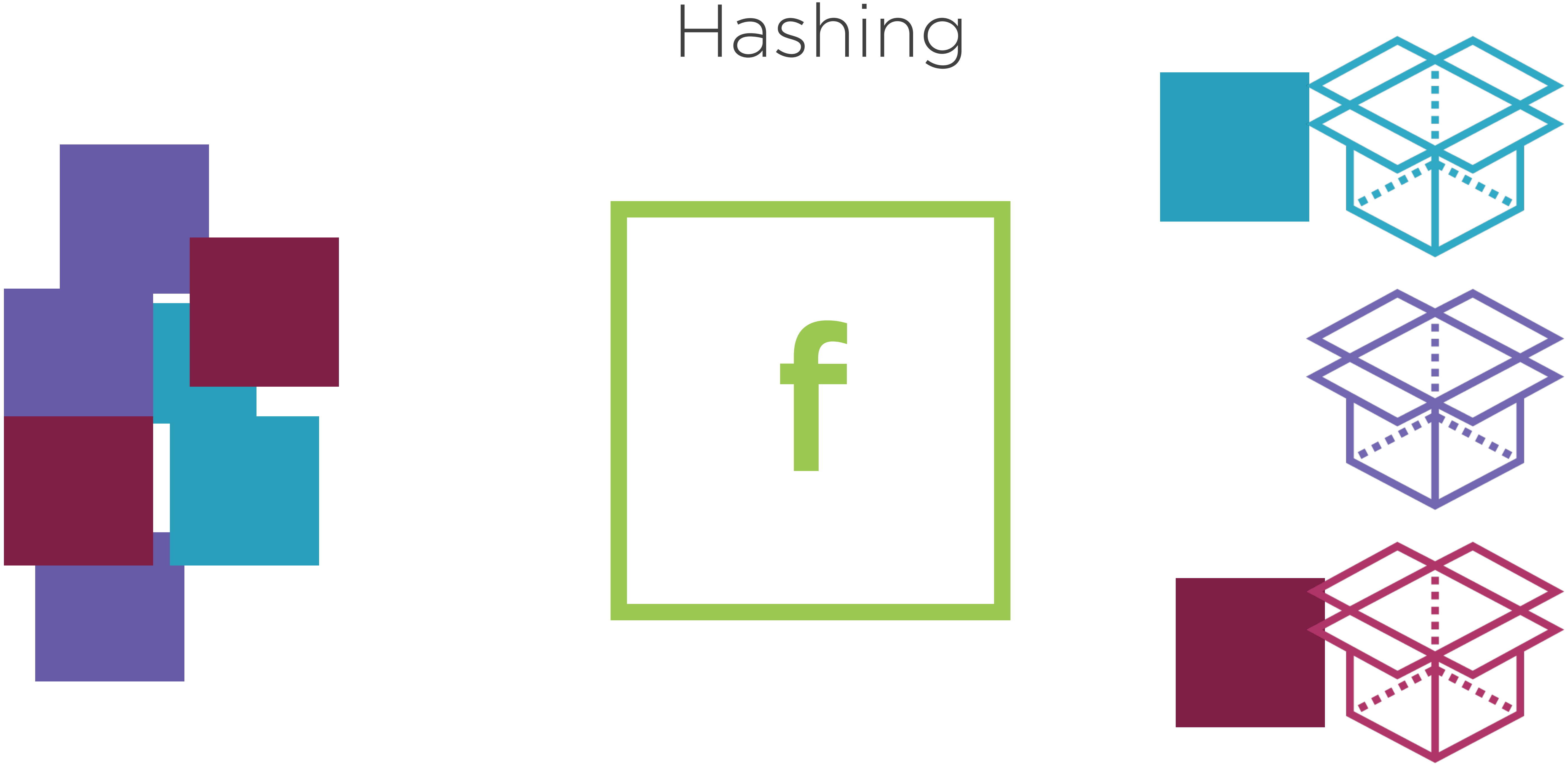
Have a fixed number of categories or buckets

Hashing

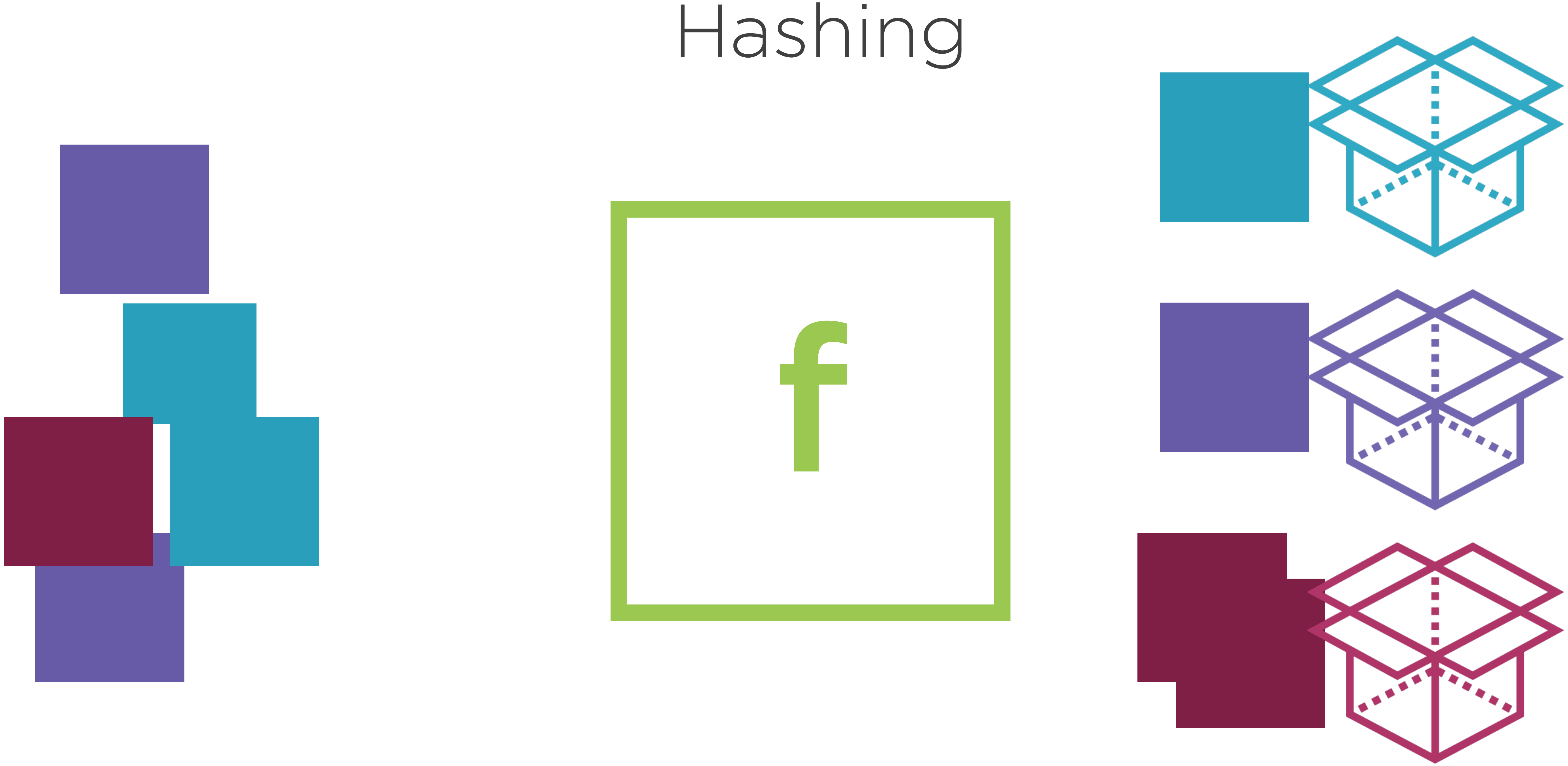


A hash function determines which bucket each value belongs to

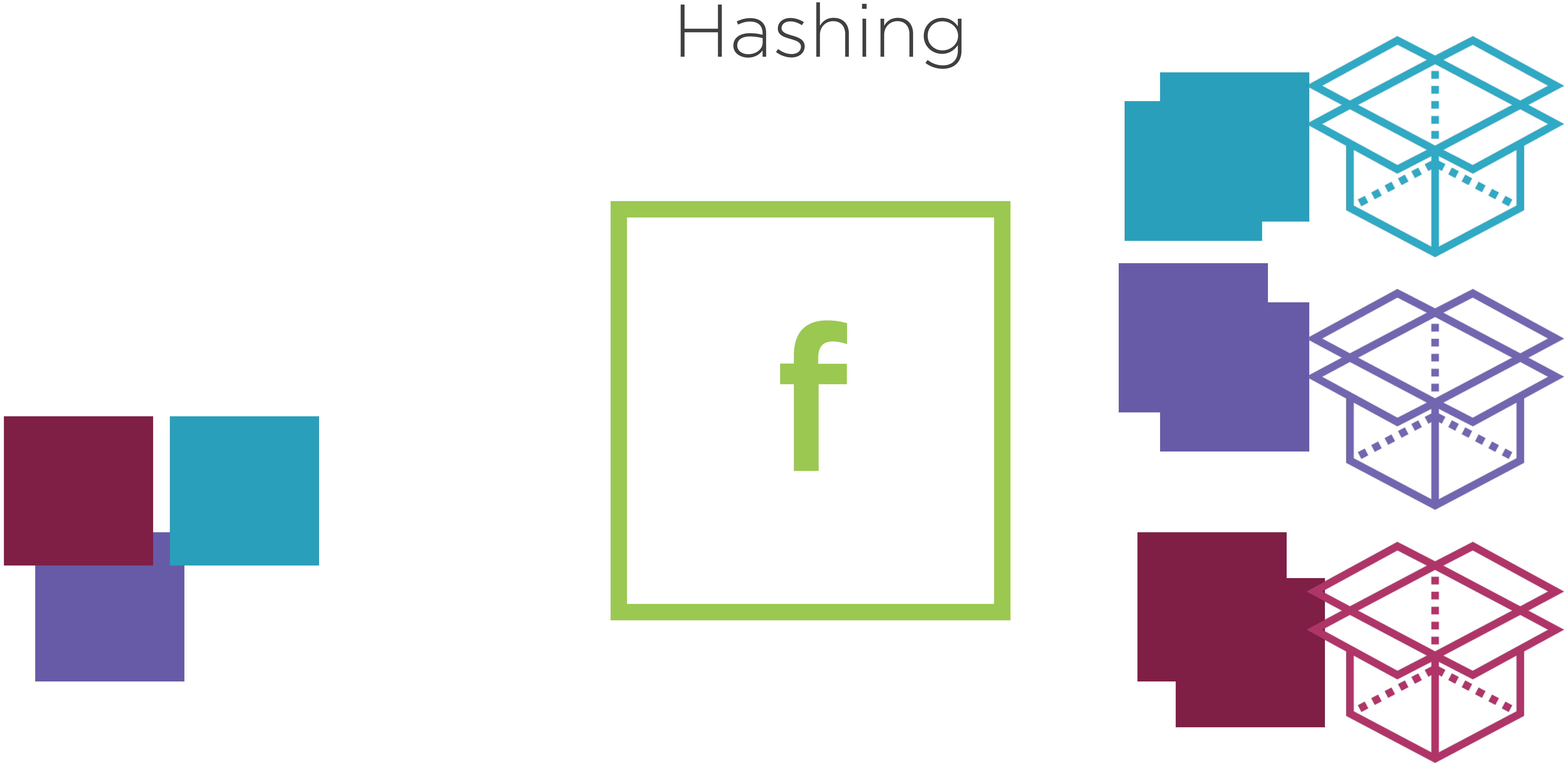
Hashing



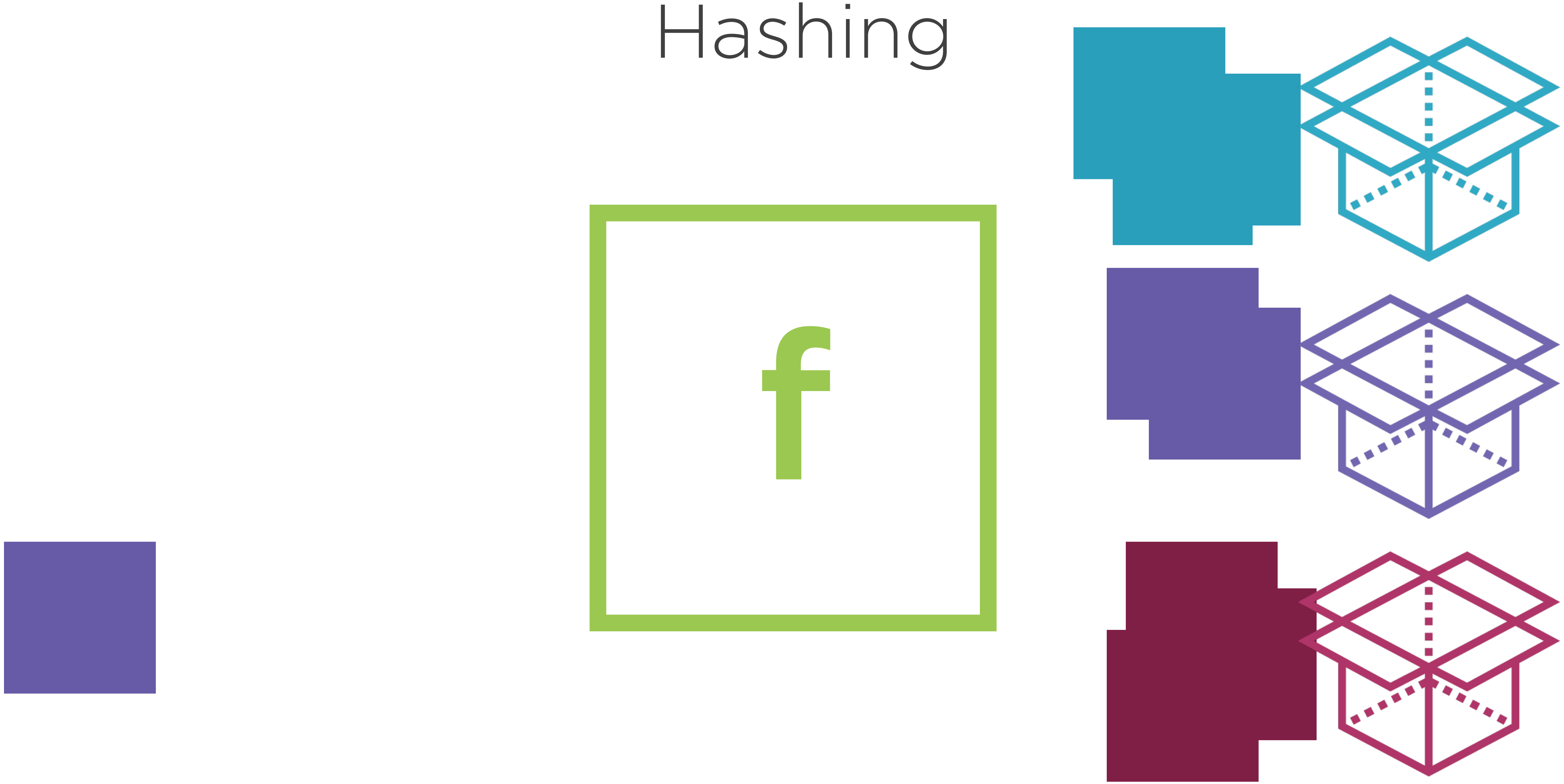
Hashing



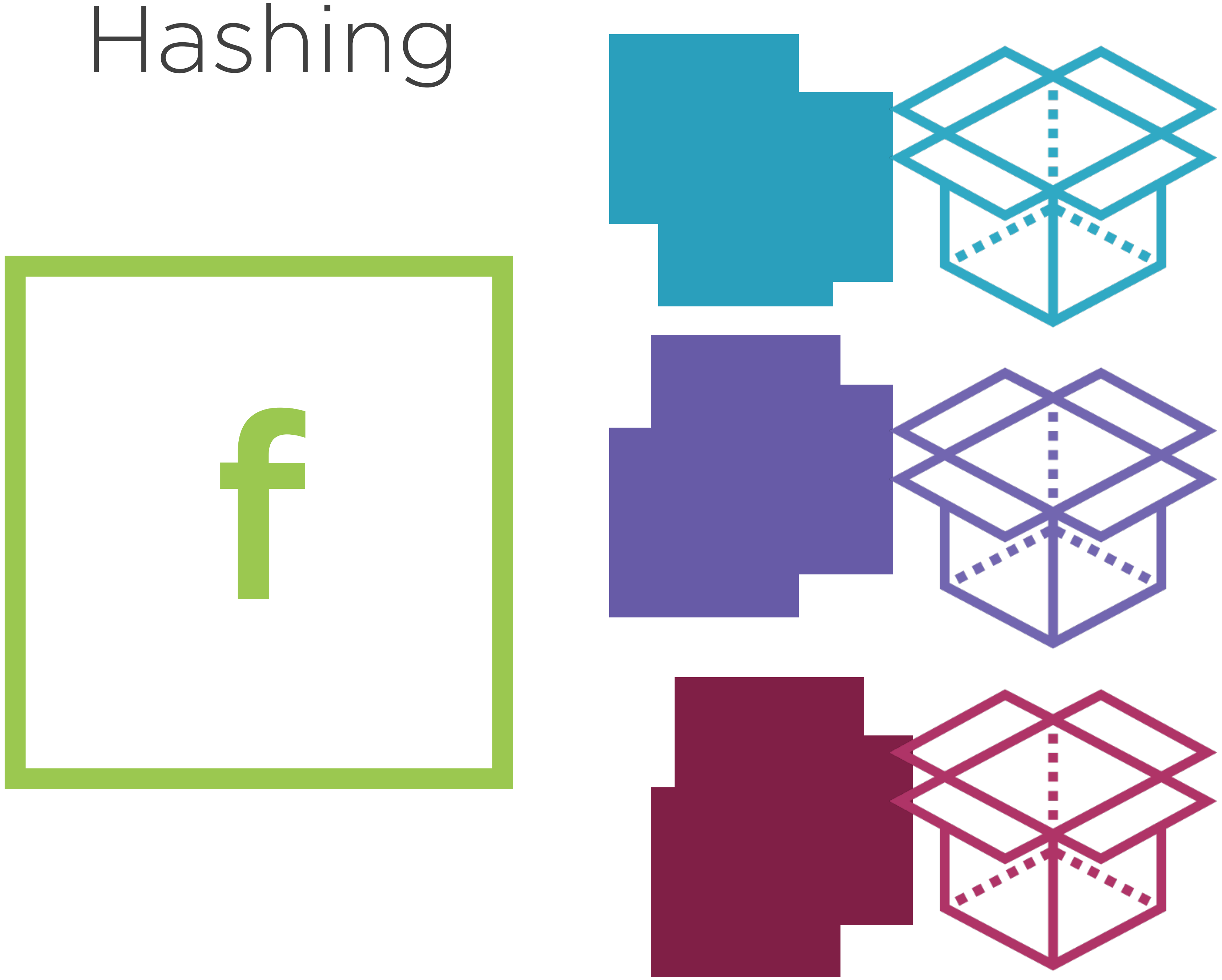
Hashing



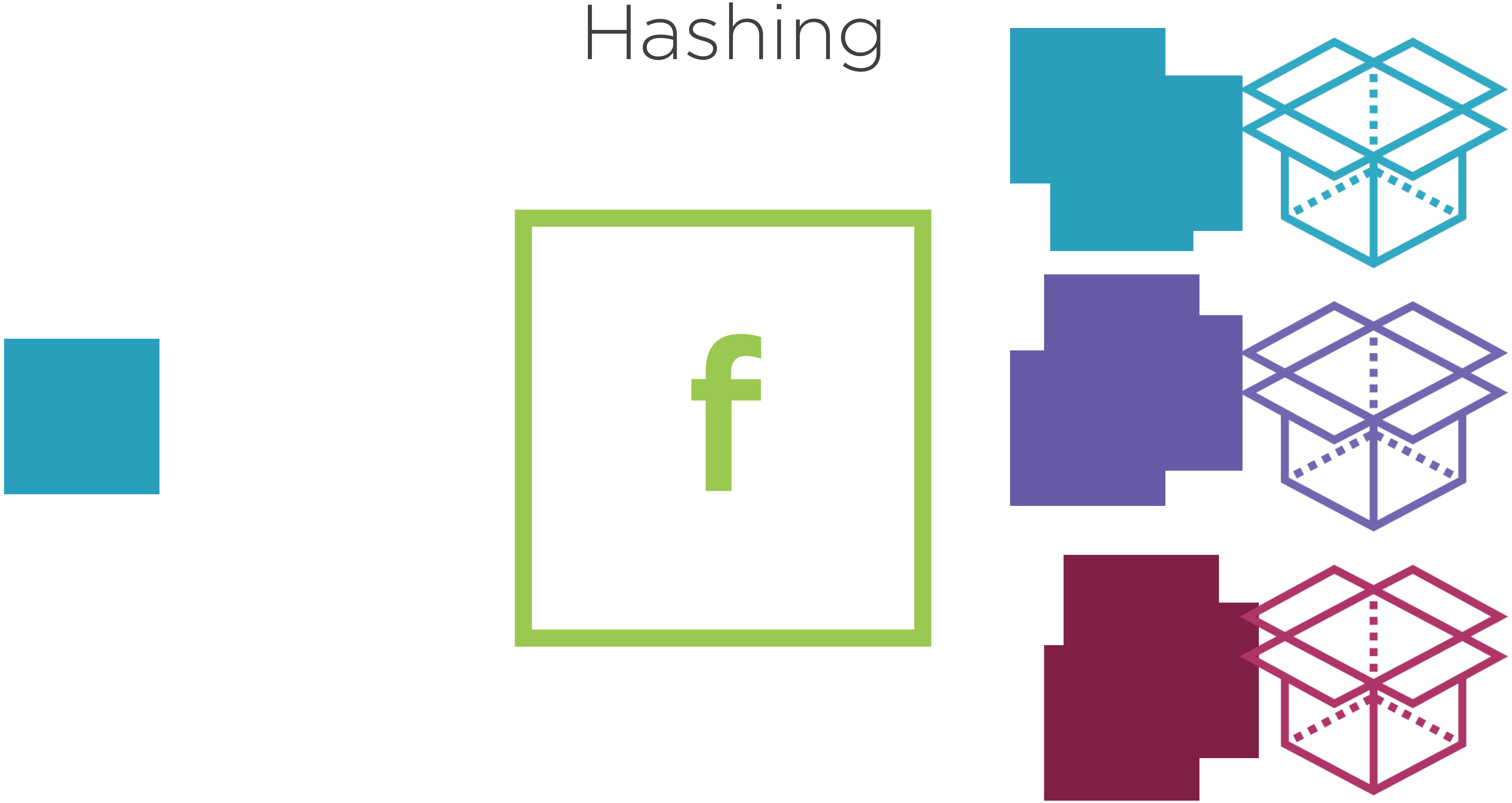
Hashing



Hashing

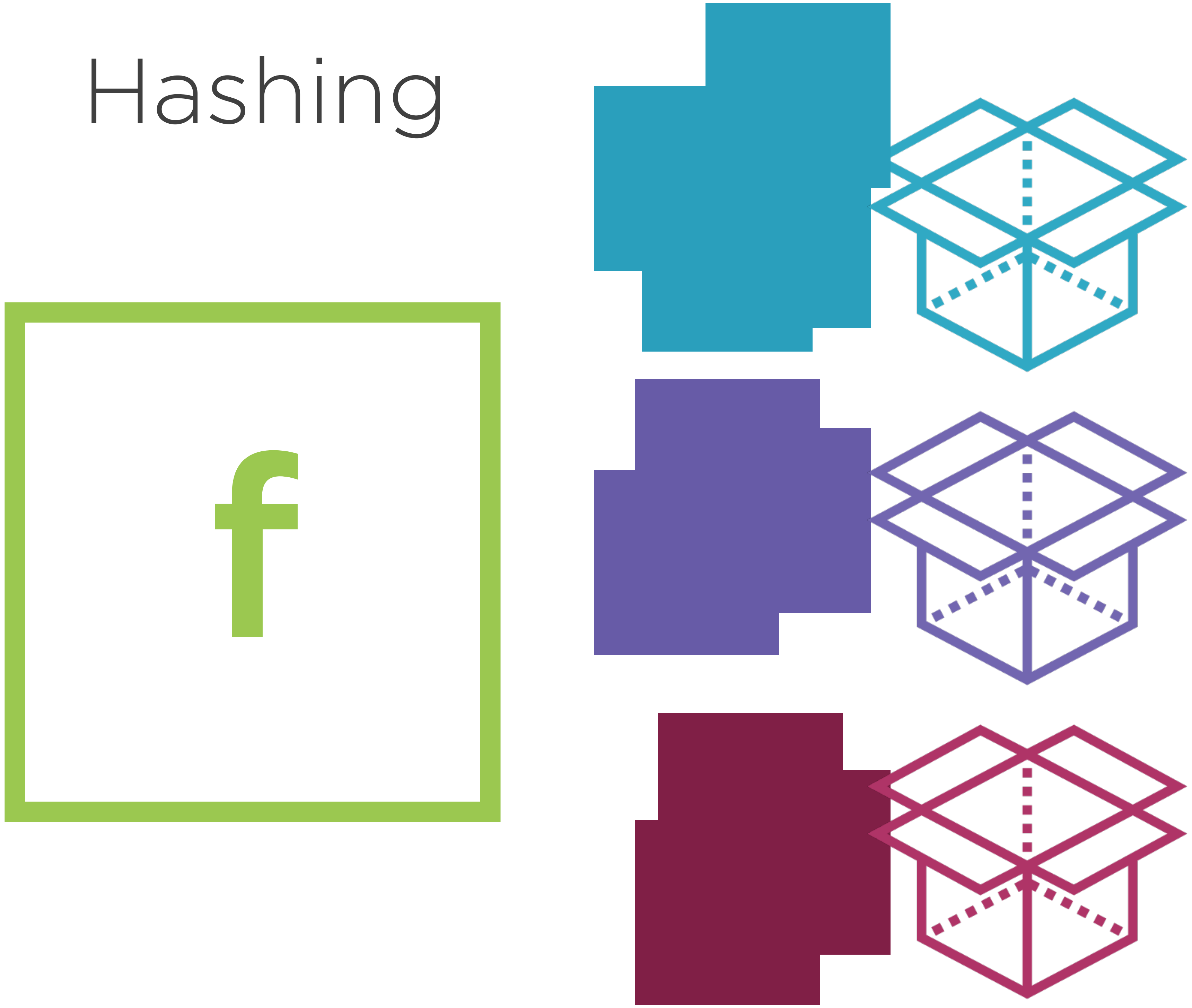


Hashing



For any new value we know immediately which bucket it belongs to

Hashing



For any new value we know immediately which bucket it belongs to

Hashing

Each value is hashed so it falls in one of these buckets

Hashing

A value can only belong to one bucket and always belongs to the same bucket

Feature Hashing in Text

Apply a hash function to words to determine their location in the feature vector representing a document. Fast and memory efficient but has no inverse transform.

Dimensionality Reduction

Input: N-dimensional data

Output: k-dimensional data

Where k < N

Hashing

Input: N-dimensional data

Output: 1-dimensional data

Output is the hash bucket the data maps to

Hashing

Input: N-dimensional data

Output: k-dimensional data

Can easily extend hashing to output desired dimensionality

Demo

Converting nominal data to numeric form using feature hashing

Summary

Converting continuous data into categorical data

Bucketing continuous data into bins

Bucketing data using Pandas and the KBinsDiscretizer

Hash nominal features to numeric features

Related Courses

Building Features from Numeric Data

Building Features from Image Data

Building Features from Text Data