Building Features from Nominal Data

IMPLEMENTING APPROACHES TO WORKING WITH CATEGORICAL DATA

Janani Ravi
CO-FOUNDER, LOONYCORN
www.loonycorn.com

Overview

Categorical data vs. continuous data

Nominal vs. ordinal data

Represent categorical data using label encoding and one-hot encoding

Compare and contrast label encoding vs. one-hot encoding

Implementing categorical feature representations

Prerequisites and Course Outline

Prerequisites

Basic Python programming

Understanding of simple regression

Basic understanding of ML, features and targets

Prerequisite Courses

Understanding Machine Learning with Python

Building Your First scikit-learn Solution

Building Regression Models with scikit-learn

Course Outline

Working with categorical data

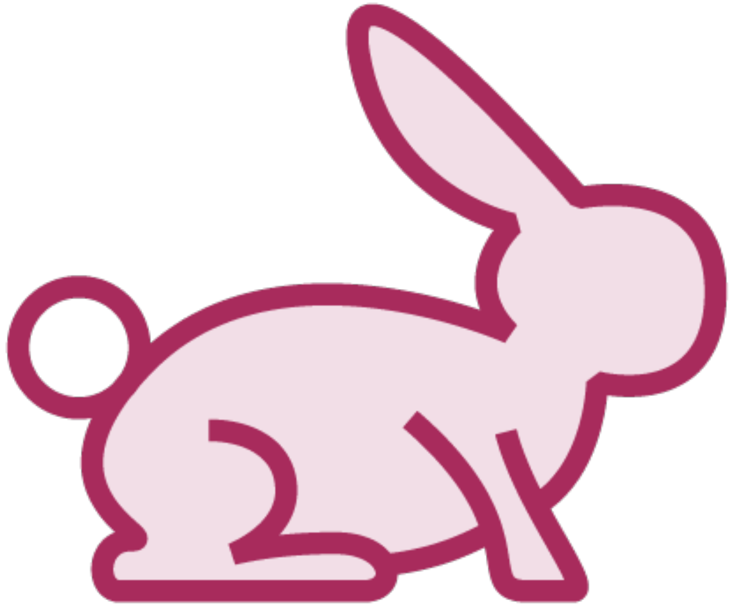
Dummy coding and one-hot coding

Contrast coding techniques

Discretizing data using bin counting and feature hashing

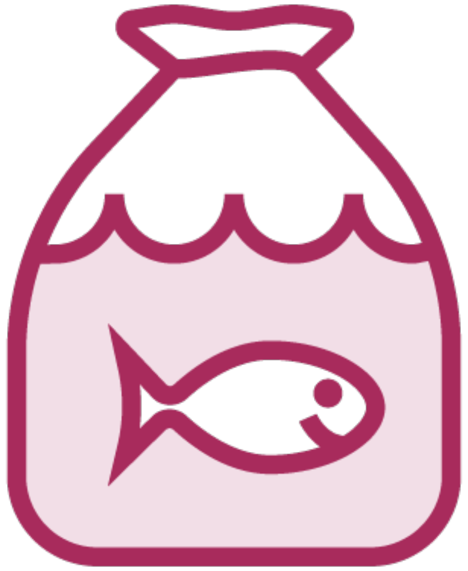
Types of Data Used in Machine Learning

Whales: Fish or Mammals?



Mammals

Members of the infraorder *Cetacea*

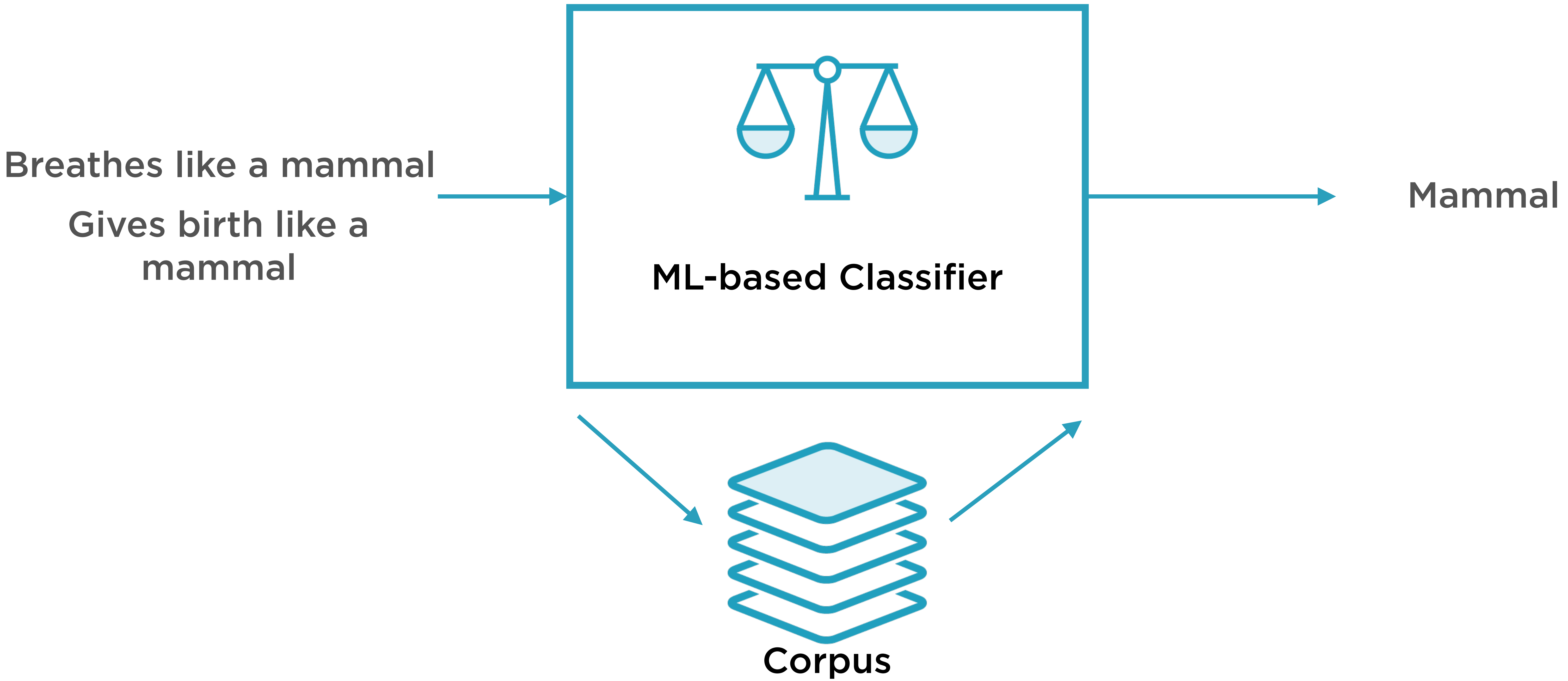


Fish

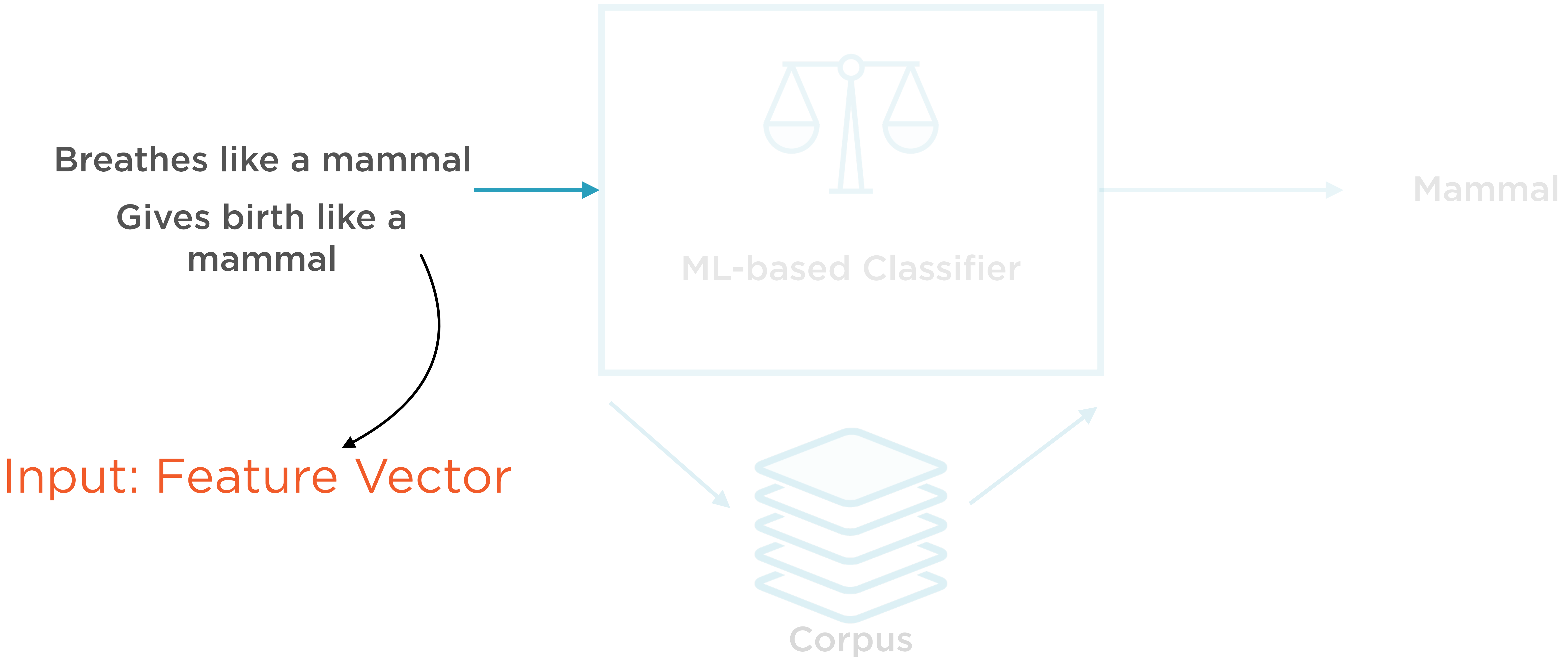
Look like fish, swim like fish, move with fish

Whales: Fish or Mammals?

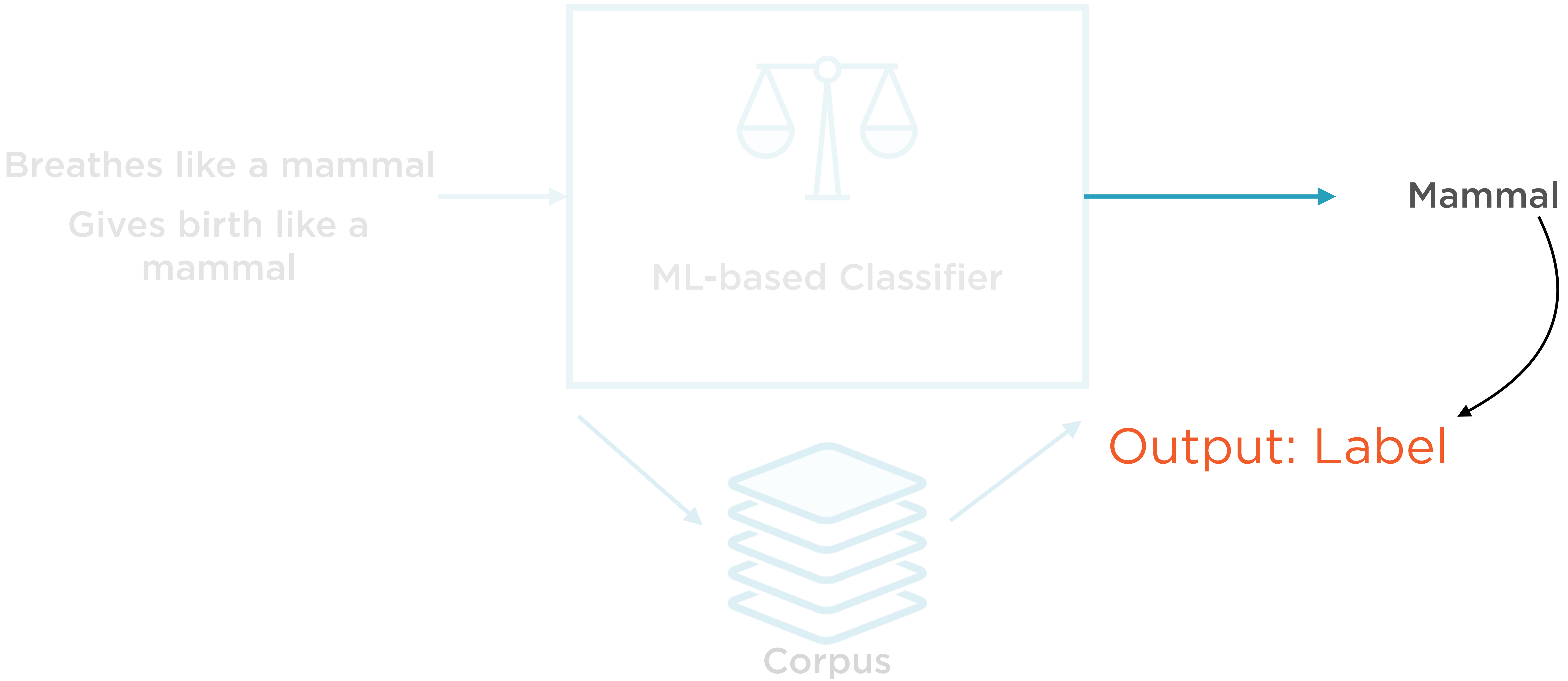
ML-based Binary Classifier



ML-based Binary Classifier



ML-based Binary Classifier



x Variables

The attributes that the ML algorithm focuses on are called features

Each data point is a list - or vector - of such features

Thus, the input into an ML algorithm is a feature vector

Feature vectors are usually called the x variables

y Variables

The attributes that the ML algorithm tries to predict are called labels

Labels are usually called the y variables

Types of labels

- categorical (classification)
- continuous (regression)

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

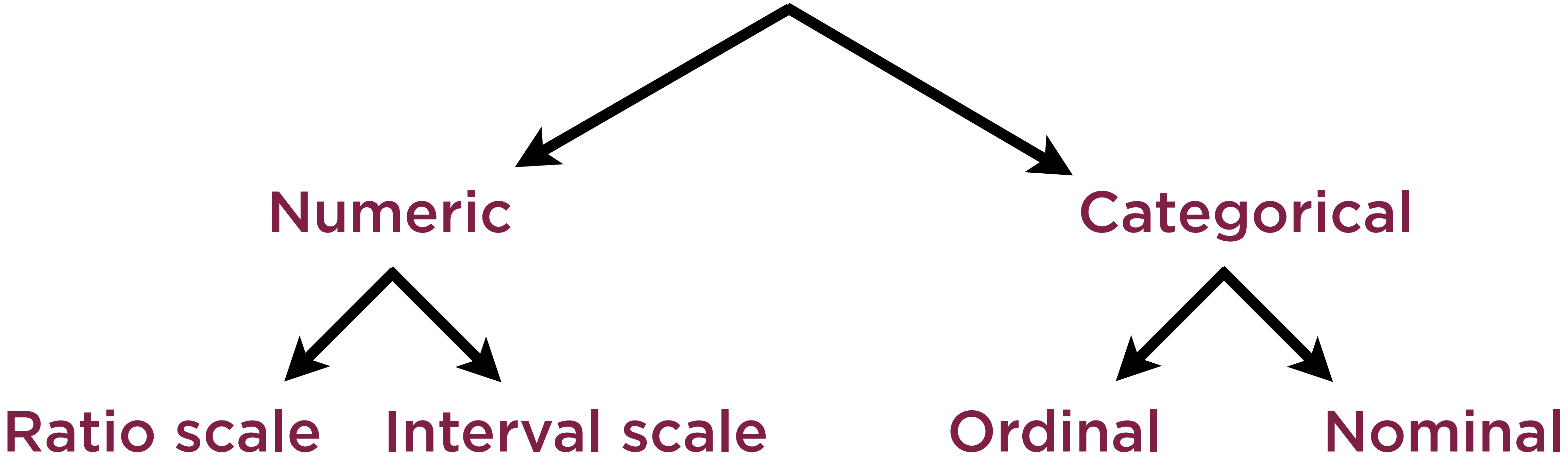
Numeric (Continuous) vs. Categorical Data

Numeric (Continuous)	Categorical
E.g. height or weight of individuals	E.g. day of week, month of year, gender, letter grade
Can take any value	Finite set of permissible values
Predicted using regression models	Predicted using classification models
Always can be sorted on magnitude	Categories may or may not be sortable

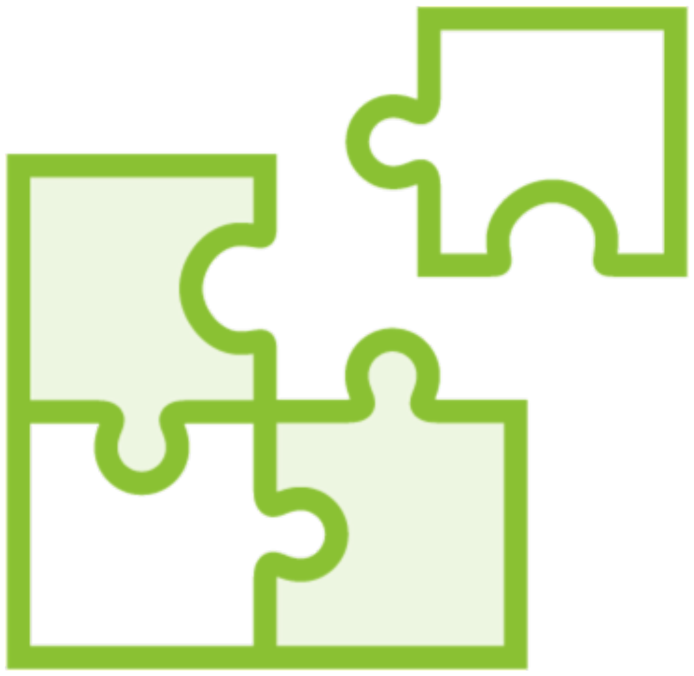
Use regression to predict numeric (continuous) y-variables

Use classification to predict categorical (discrete) y-variables

Types of Data in Machine Learning



Understanding Data Types is Important



Preprocessing of variables is different for numeric and categorical data

Certain statistical measurements may not apply for certain data types

Visualizations to convey information will be different in exploratory data analysis

Numeric Data

Types of Data in Machine Learning

- Numeric
  - Ratio scale
  - Interval scale
- Categorical
  - Ordinal
  - Nominal

Numerical Data

Discrete

Cannot be measured but can be counted

Continuous

Cannot be counted but can be measured

Numerical Data

Discrete

Cannot be measured but can be counted

Continuous

Cannot be counted but can be measured

Number of visitors in an hour, number of heads when a coin is flipped 100 times

Numerical Data

Discrete

Cannot be measured but can be counted

Continuous

Cannot be counted but can be measured

Height of an individual, home prices, stock prices

Types of Data in Machine Learning

Numeric

Categorical

Ratio scale

Interval scale

Ordinal

Nominal

Have an intrinsic order, can always be sorted on magnitude

Types of Data in Machine Learning

Numeric

Categorical

Ratio scale

Interval scale

Ordinal

Nominal

"Usual" numeric data,
expressed as ratio to 1
e.g. 7 == 7:1

Types of Data in Machine Learning

- **Numeric**
  - **Ratio scale**
  - Interval scale
- Categorical
  - Ordinal
  - Nominal

All arithmetic operations apply: addition, subtraction, multiplication and division

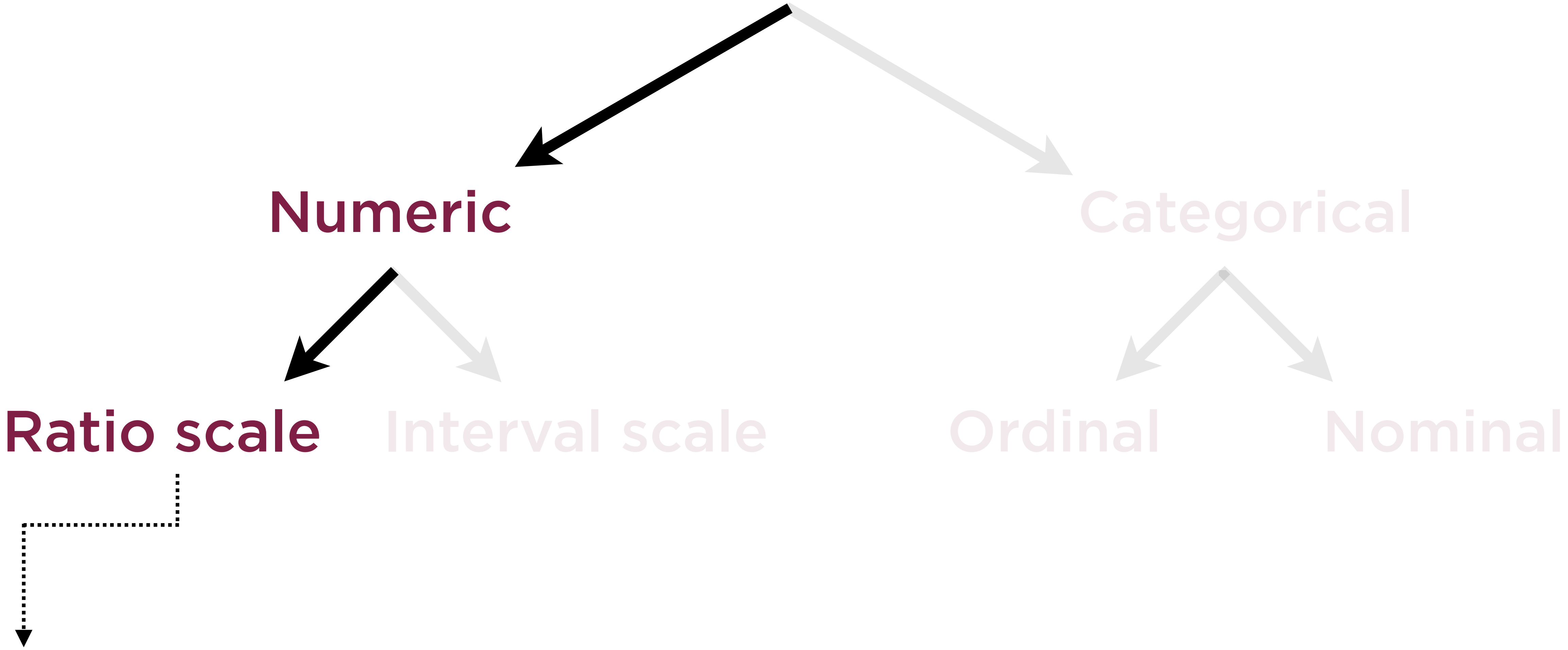
Types of Data in Machine Learning

Numeric | Categorical

Ratio scale | Interval scale | Ordinal | Nominal

E.g. weight of 20 lbs is twice as much as a weight of 10 lbs

Types of Data in Machine Learning



Ratio scale data has a meaningful zero point
(the only type of data in this chart that does)

Types of Data in Machine Learning

Numeric

Categorical

Ratio scale

Interval scale

Ordinal

Nominal

Weight of 0 lbs is equivalent to “no weight”

Types of Data in Machine Learning

Numeric

Categorical

Ratio scale

Interval scale

Ordinal

Nominal

Ordered units that have the same difference i.e. the interval

Types of Data in Machine Learning

Numeric

Categorical

Ratio scale

Interval scale

Ordinal

Nominal

Data still numeric, but now multiplication and division no longer make sense, and zero point no longer meaningful

Types of Data in Machine Learning

Numeric

Categorical

Ratio scale

Interval scale

Ordinal

Nominal

Difference between 90 Fahrenheit and 30 Fahrenheit is equal to 60 Fahrenheit

Types of Data in Machine Learning

Numeric Categorical

Ratio scale **Interval scale** Ordinal Nominal

But temperature of 90 Fahrenheit is not
thrice temperature of 30 Fahrenheit

Types of Data in Machine Learning

Numeric

Categorical

Ratio scale

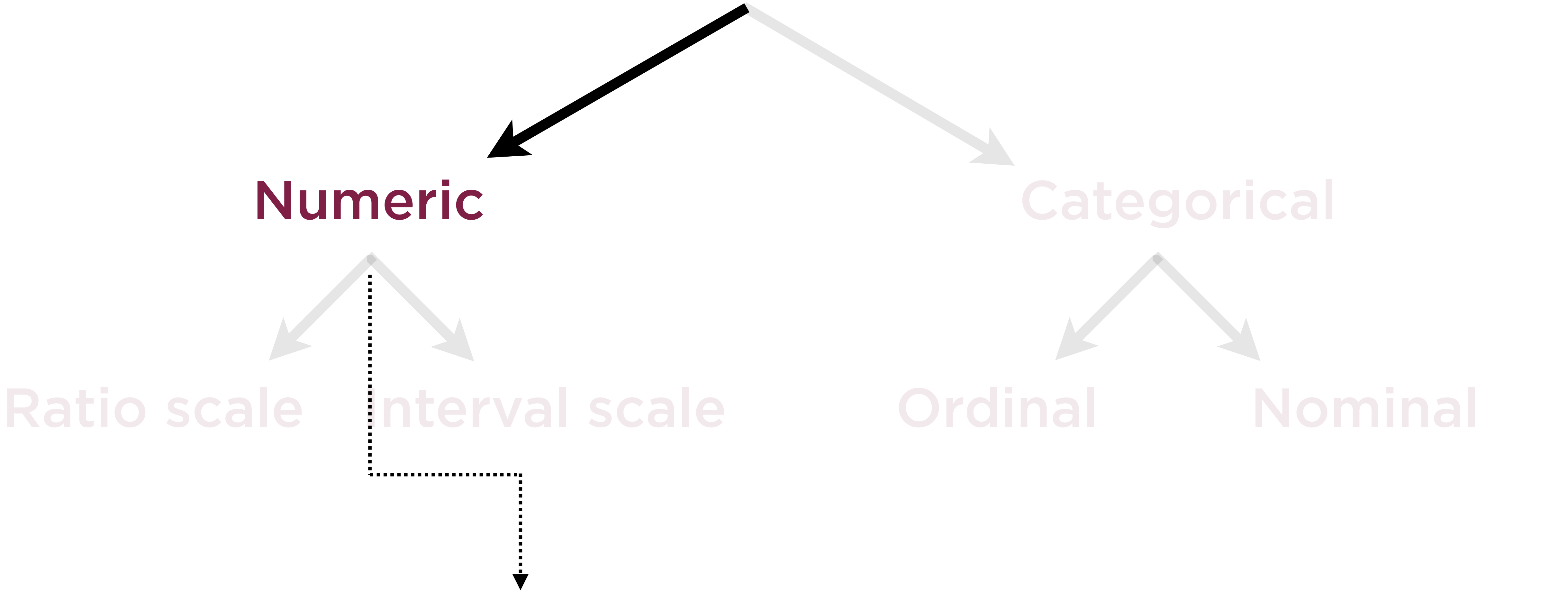
Interval scale

Ordinal

Nominal

0 Fahrenheit is not equivalent to
“no temperature”

Types of Data in Machine Learning



Numeric data can draw from an unrestricted range of continuous values

Types of Data in Machine Learning

Numeric

- Ratio scale
- Interval scale

Categorical

- Ordinal
- Nominal

Can calculate mean, standard deviation, correlation etc.

Visualizing Continuous Data

Histograms for univariate data

Box plots for statistical distributions

Scatter plots for relationships

Numeric Features

- Can represent any kind of information
- The range of each feature will be different
- The average and dispersion of features will also be different
- Comparing different features is hard

Machine learning algorithms typically do not work well with numeric data with **different scales**

Feature Scaling

Scaling

Standardization

Feature Scaling

Scaling

Standardization

Numeric values are **shifted and rescaled** so all features have the same scale i.e. within the same minimum and maximum values

Feature Scaling

Scaling

Standardization

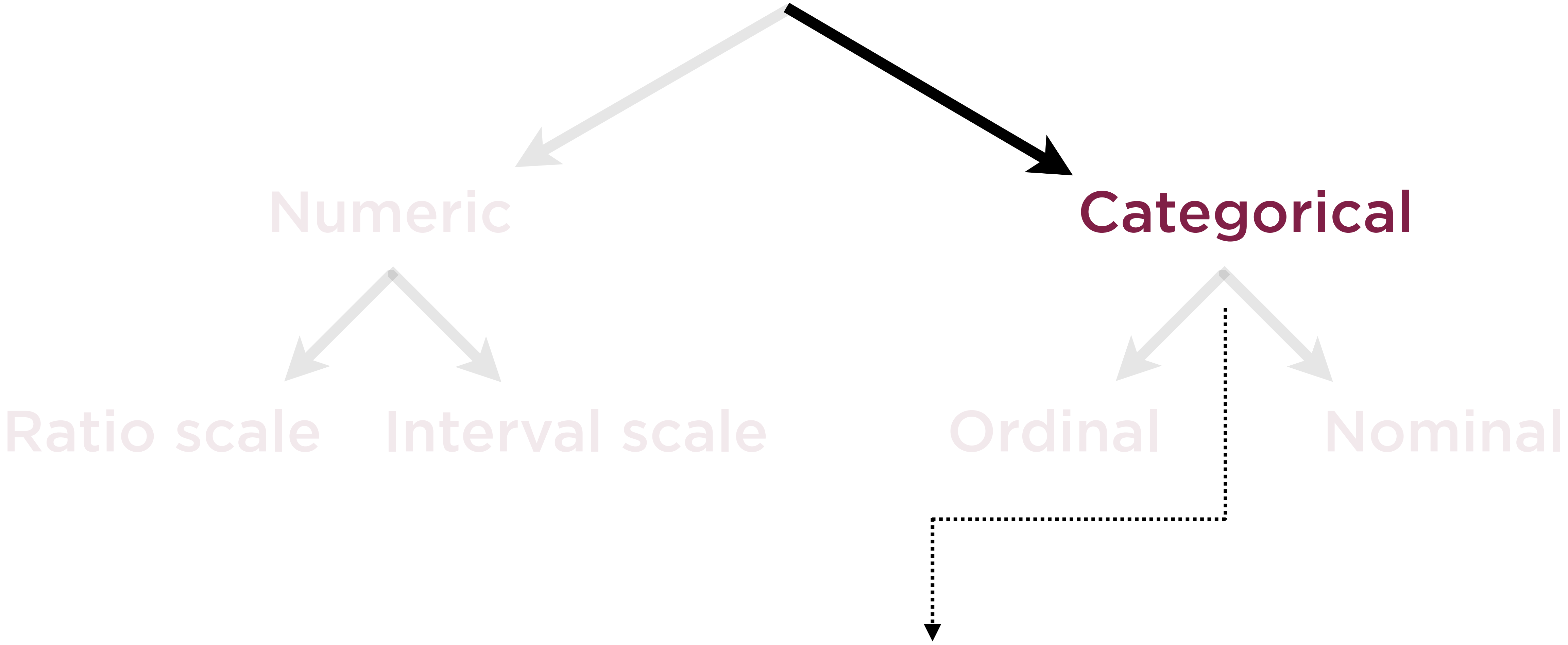
Centers data round the mean and divides each value by the variance so all features have 0 mean and unit variance

Categorical Data

Types of Data in Machine Learning

- **Numeric**
  - **Ratio scale**
  - **Interval scale**
- **Categorical**
  - **Ordinal**
  - **Nominal**

Types of Data in Machine Learning



Categorical data can only draw from a specific, restricted set of values

Types of Data in Machine Learning

Numeric

Categorical

Ratio scale

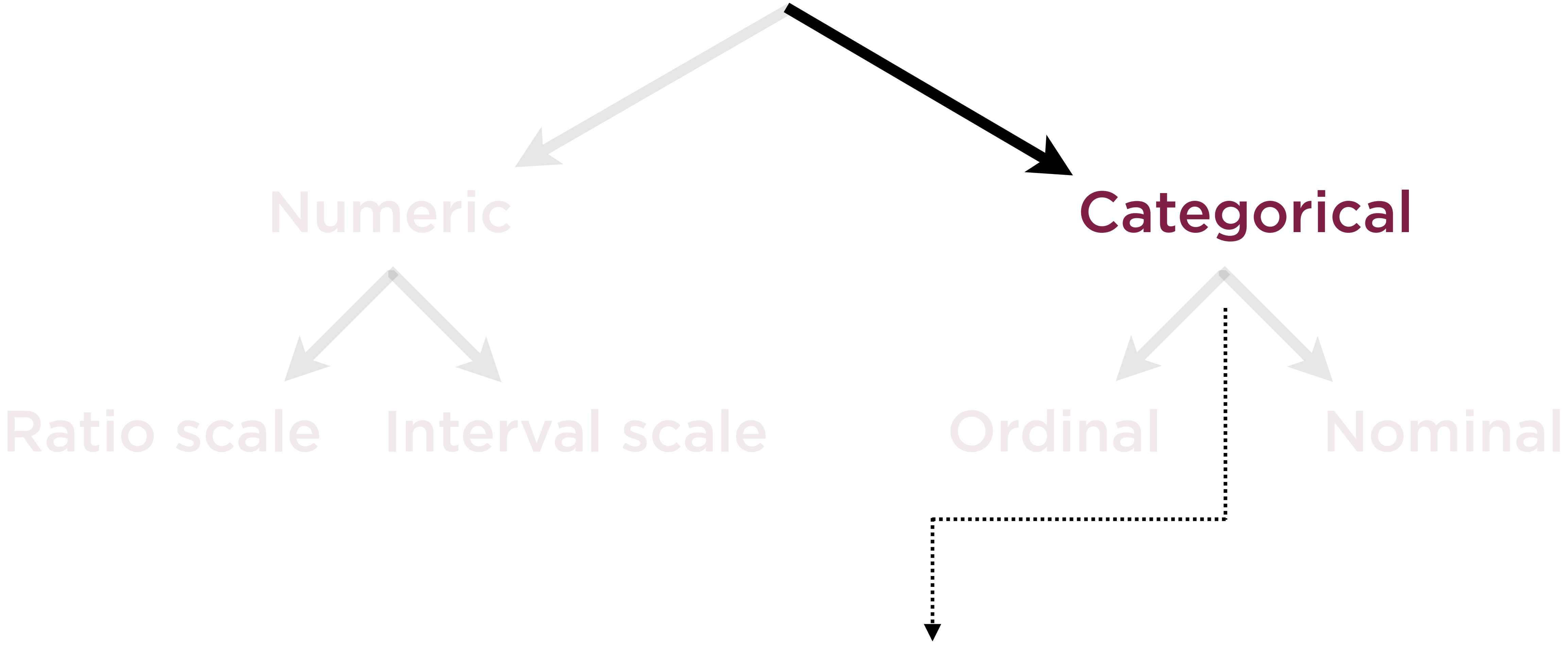
Interval scale

Ordinal

Nominal

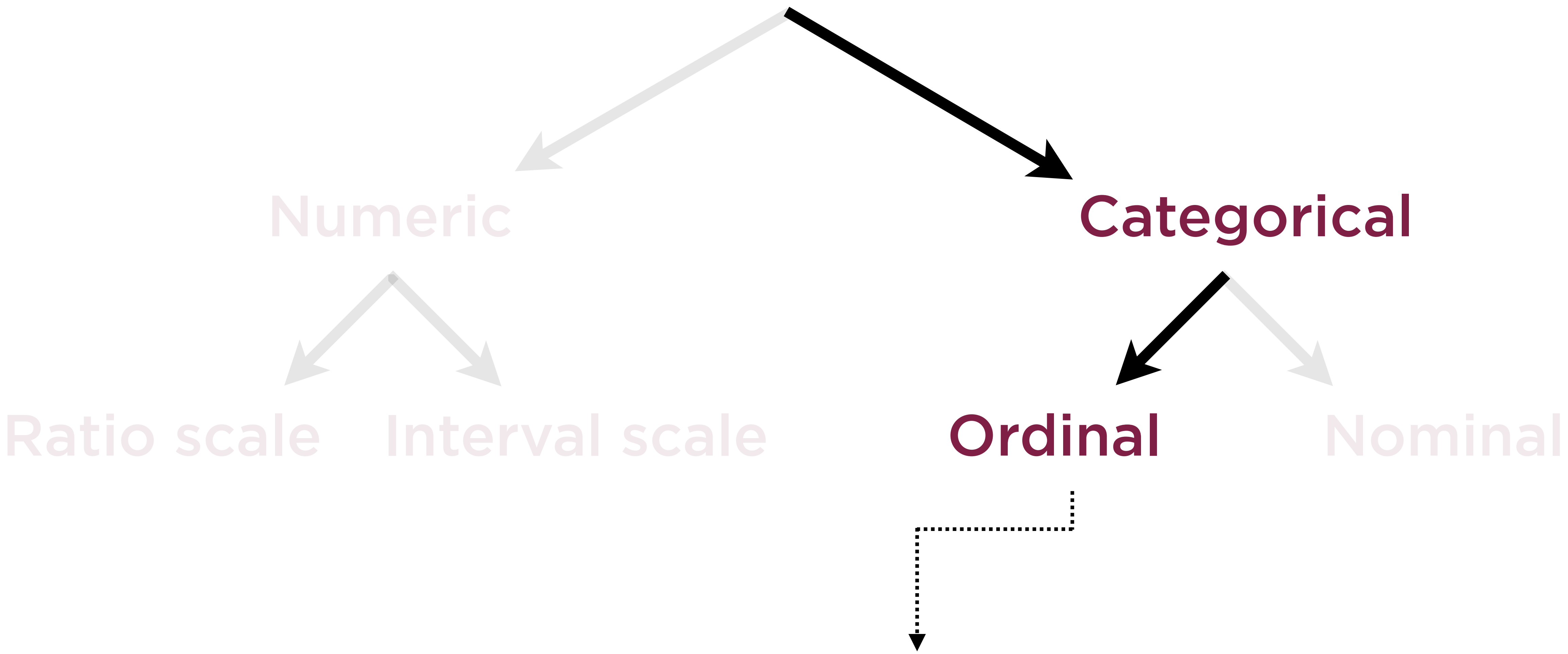
Not meaningful to calculate mean, standard deviation, correlation

Types of Data in Machine Learning



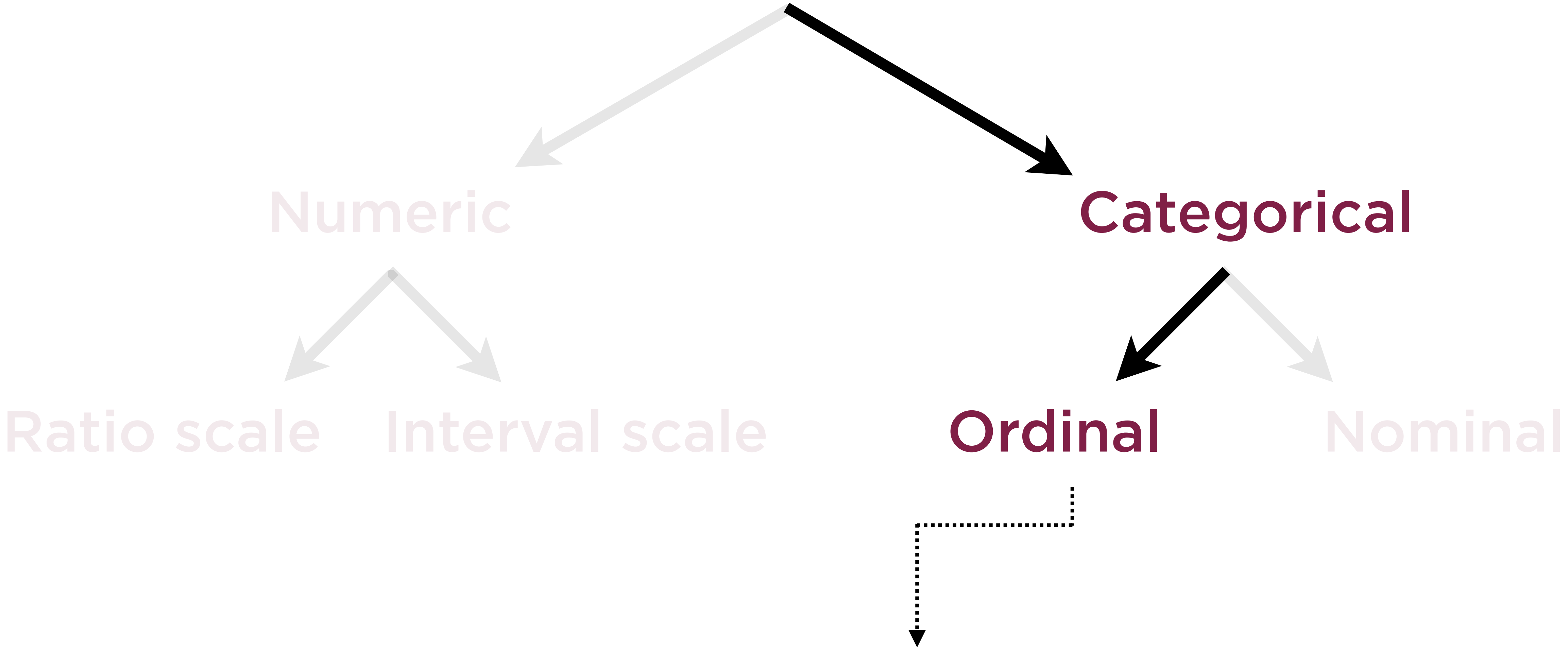
Fine to tabulate categorical data using count frequencies and percentages

Types of Data in Machine Learning



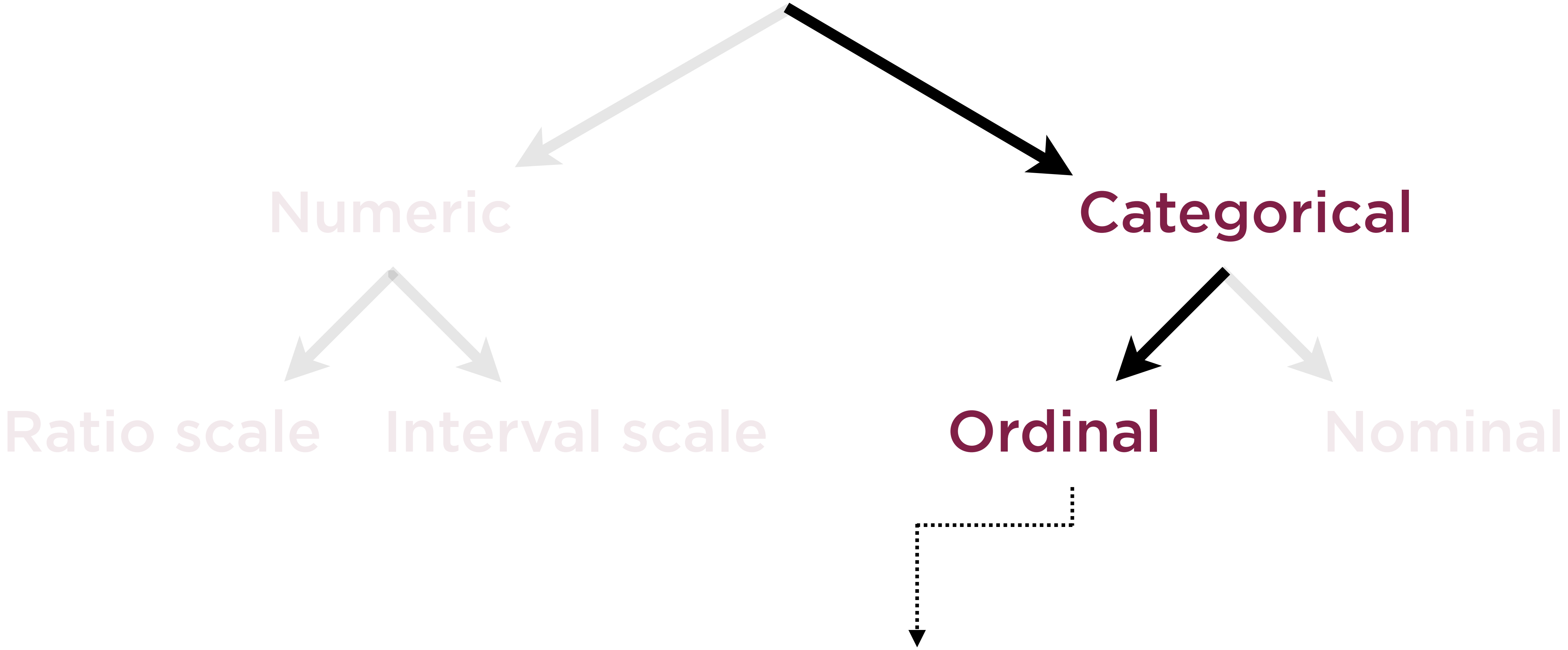
Ordinal data is categorical, but can still be ordered

Types of Data in Machine Learning



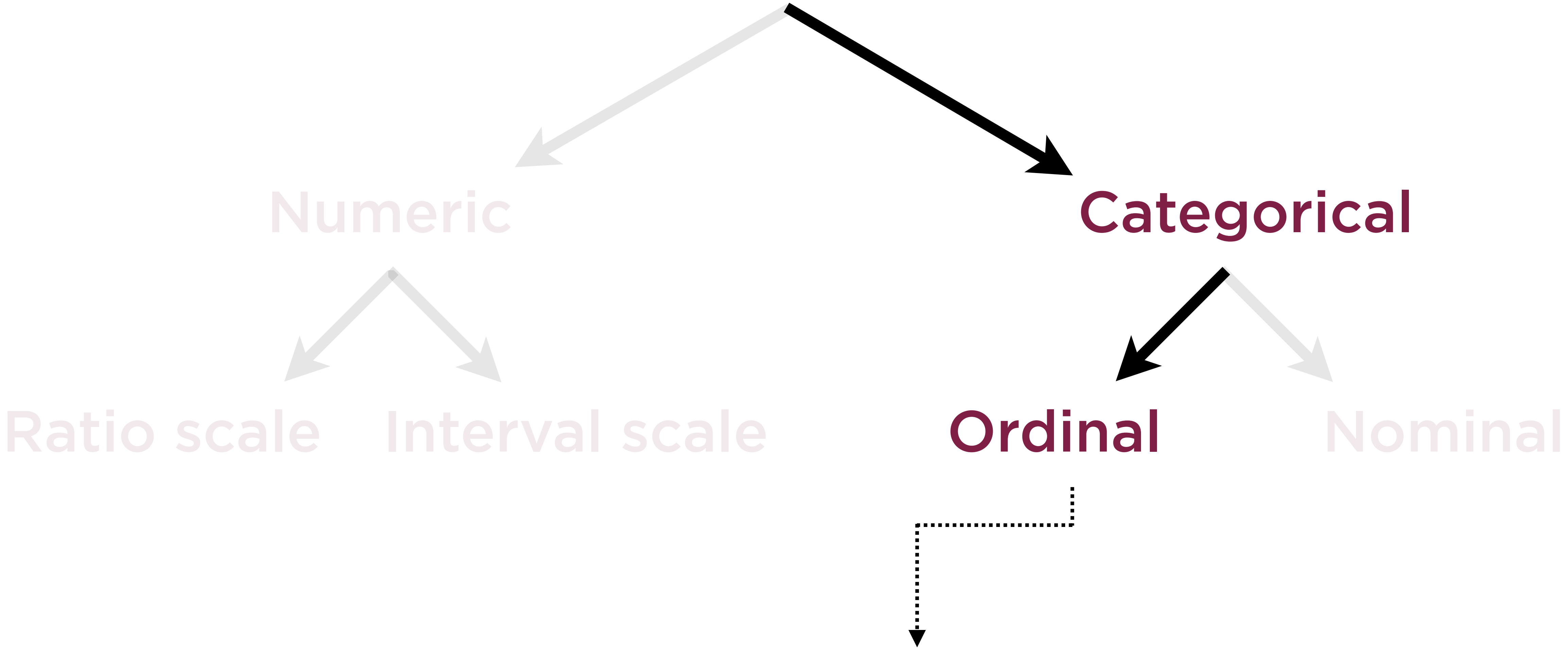
E.g. month of the year,
ratings on a scale of 1 to 5

Types of Data in Machine Learning



Order exists, but differences are not necessarily meaningful

Types of Data in Machine Learning



E.g. Differences in quality between three, two, one, and no Michelin stars for a restaurant are not uniform

Types of Data in Machine Learning

Numeric

Categorical

Ratio scale

Interval scale

Ordinal

Nominal

Even less in common with numeric data - cannot even be ordered

Types of Data in Machine Learning

Numeric

Categorical

Ratio scale

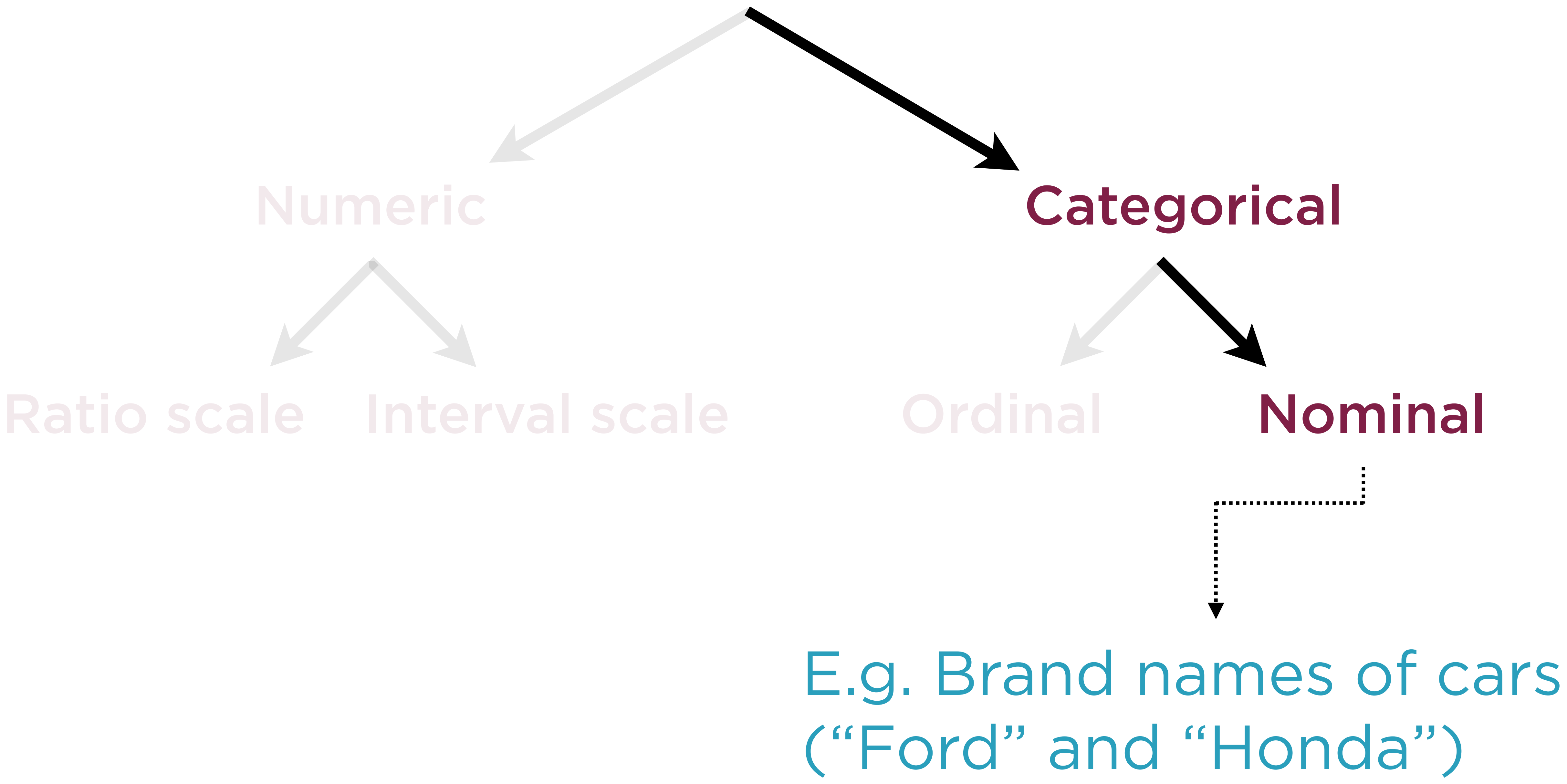
Interval scale

Ordinal

Nominal

Ordinal data can at least be ordered; nominal data are simply names

Types of Data in Machine Learning



Visualizing Categorical Data

Pie chart for proportions

Bar chart for frequency counts in categories

Categorical data has to be **numerically encoded** before it can be used in ML models

Representing Categorical Data

```
['New York', 'London','Paris','Bangalore']
```

Categorical Data

Classes often represented in string format

Categories as Nominal Data

Label encoding

Numeric id for each category; single column suffices

One-hot encoding

Separate column with 1 or 0 for presence/absence of each category

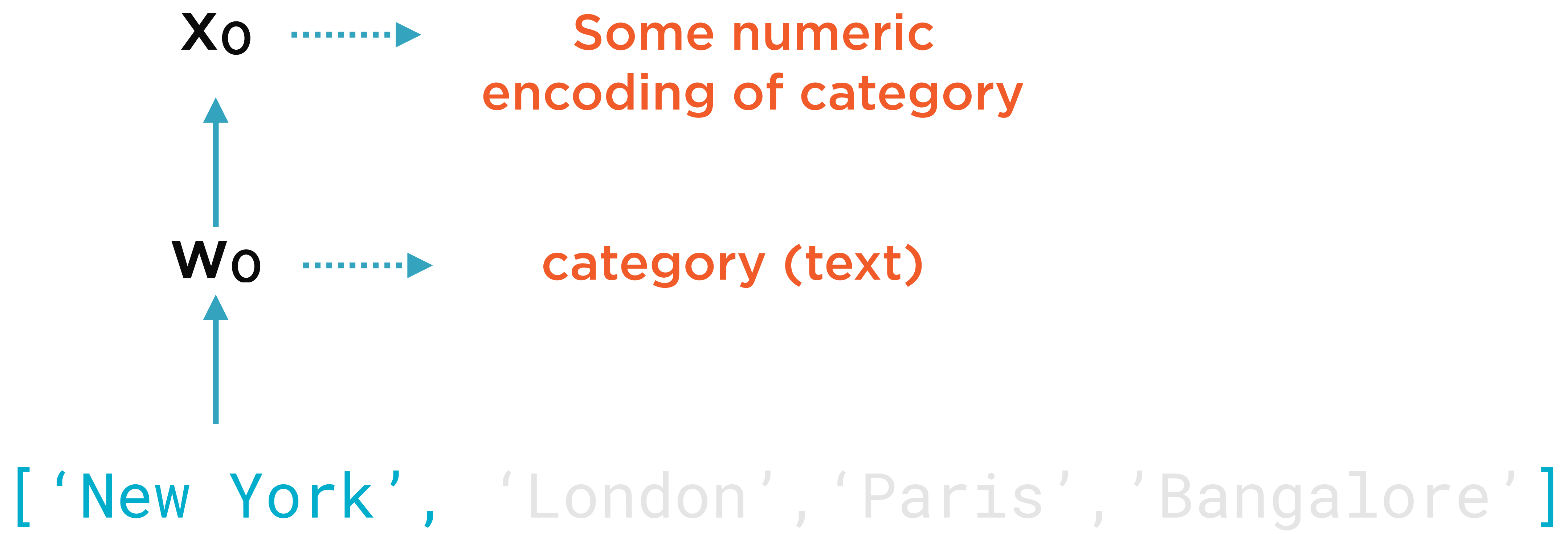
Categories as Nominal Data

Label encoding

Numeric id for each category; single column suffices

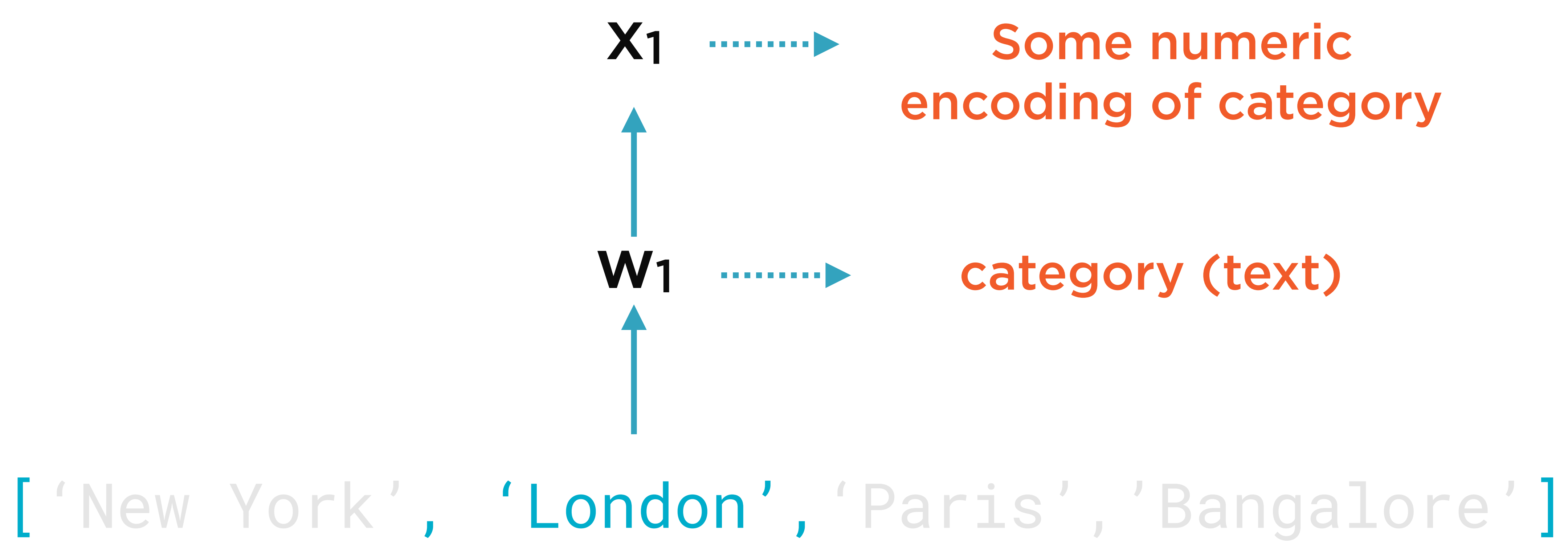
One-hot encoding

Separate column with 1 or 0 for presence/absence of each category



Categorical Data

Represent each category using some numeric encoding



Categorical Data

Represent each category using some numeric encoding

Categorical Data

Represent each category using some numeric encoding

Represent Each Category as a Number

Represent Each Category as a Number

1056

w_3

['New York', 'London','Paris','Bangalore']

Represent Each Category as a Number

Categories as Nominal Data

Label encoding

Numeric id for each category; single column suffices

One-hot encoding

Separate column with 1 or 0 for presence/absence of each category

```
[‘New York’, ‘London’,‘Paris’,’Bangalore’]
```

Categorical Data

Classes often represented in string format

$x_i = 0 \text{ or } 1$

One-hot Encoding of 1 Category

Represent each category with a binary variable

$x_i = 0 \text{ or } 1$

One-hot Encoding of 1 Category

Need as many columns as categories in the data

One-hot Encoded Cities

New York	London	Paris	Bangalore

One-hot Encoded Cities

Category	New York	London	Paris	Bangalore
New York				
London				
Paris				
Bangalore				

One-hot Encoded Cities

Category	New York	London	Paris	Bangalore
New York	1	0	0	0
London				
Paris				
Bangalore				

One-hot Encoded Cities

Category	New York	London	Paris	Bangalore
New York	1	0	0	0
London	0	1	0	0
Paris				
Bangalore				

One-hot Encoded Cities

Category	New York	London	Paris	Bangalore
New York	1	0	0	0
London	0	1	0	0
Paris	0	0	1	0
Bangalore				

One-hot Encoded Cities

Category	New York	London	Paris	Bangalore
New York	1	0	0	0
London	0	1	0	0
Paris	0	0	1	0
Bangalore	0	0	0	1

One-hot Encoded Cities

Category	New York	London	Paris	Bangalore
New York	1	0	0	0
London	0	1	0	0
Paris	0	0	1	0
Bangalore	0	0	0	1

Label Encoding vs. One-hot Encoding

Words as Nominal Data

Label encoding

Numeric id for each word; single column suffices

One-hot encoding

Separate column with 1 or 0 for presence/absence of each word

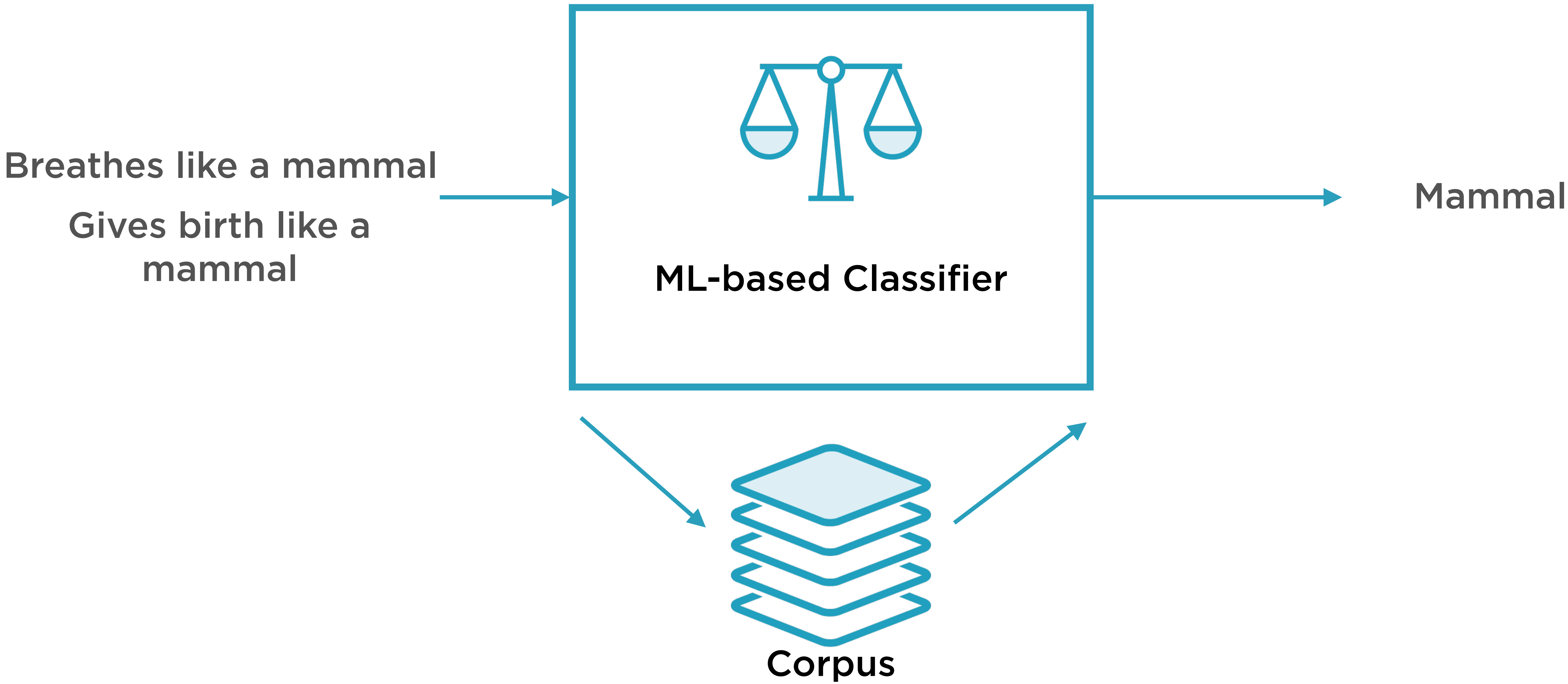
Label Encoding vs. One-hot Encoding

Label Encoding	One-hot Encoding
Single column to represent categories	Need as many columns as categories in the data
Each category takes numeric value	Each category is a row with single 1 rest 0s
More concise	Verbose - especially as number of categories grows

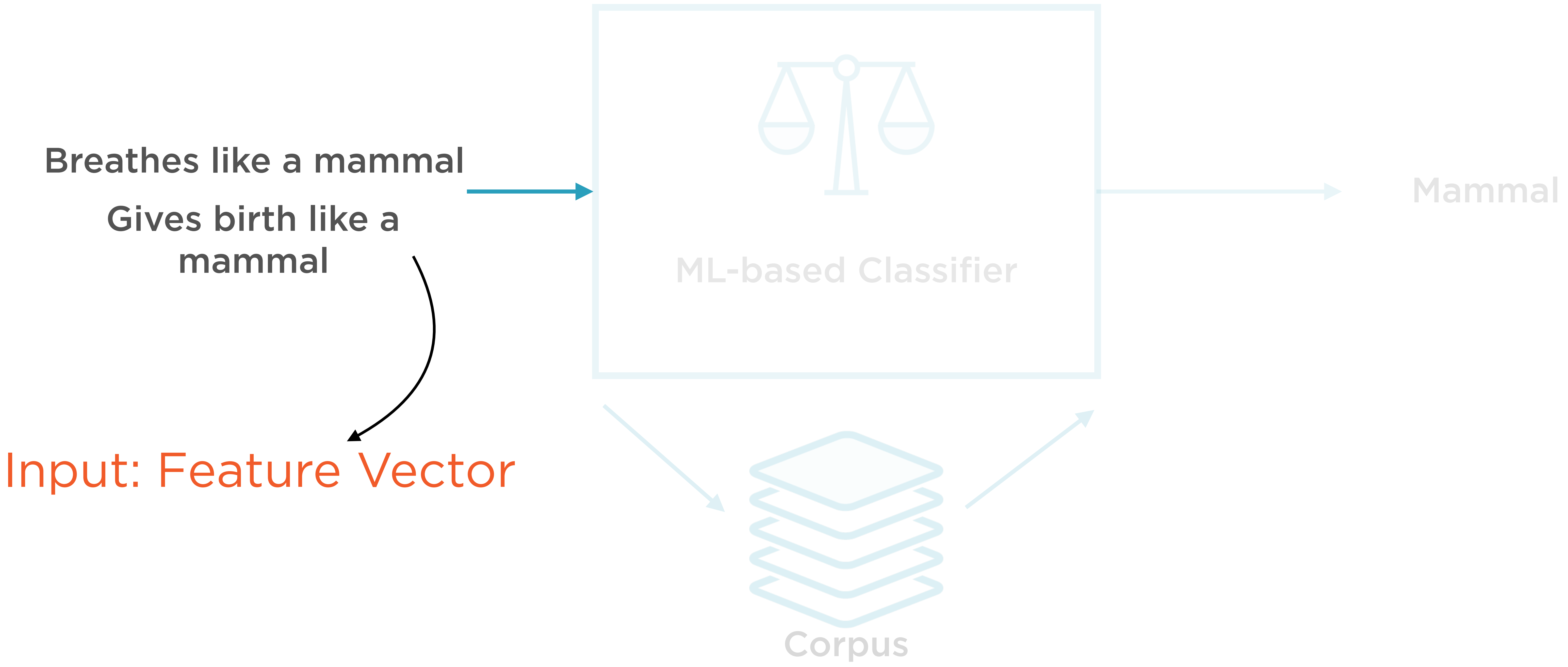
Label Encoding vs. One-hot Encoding

Label Encoding	One-hot Encoding
Numeric ids present illusion of sortability	One-hot encoded vectors are clearly not sortable
Ideally should use only for ordinal categorical data	Can use for both nominal and ordinal categorical data

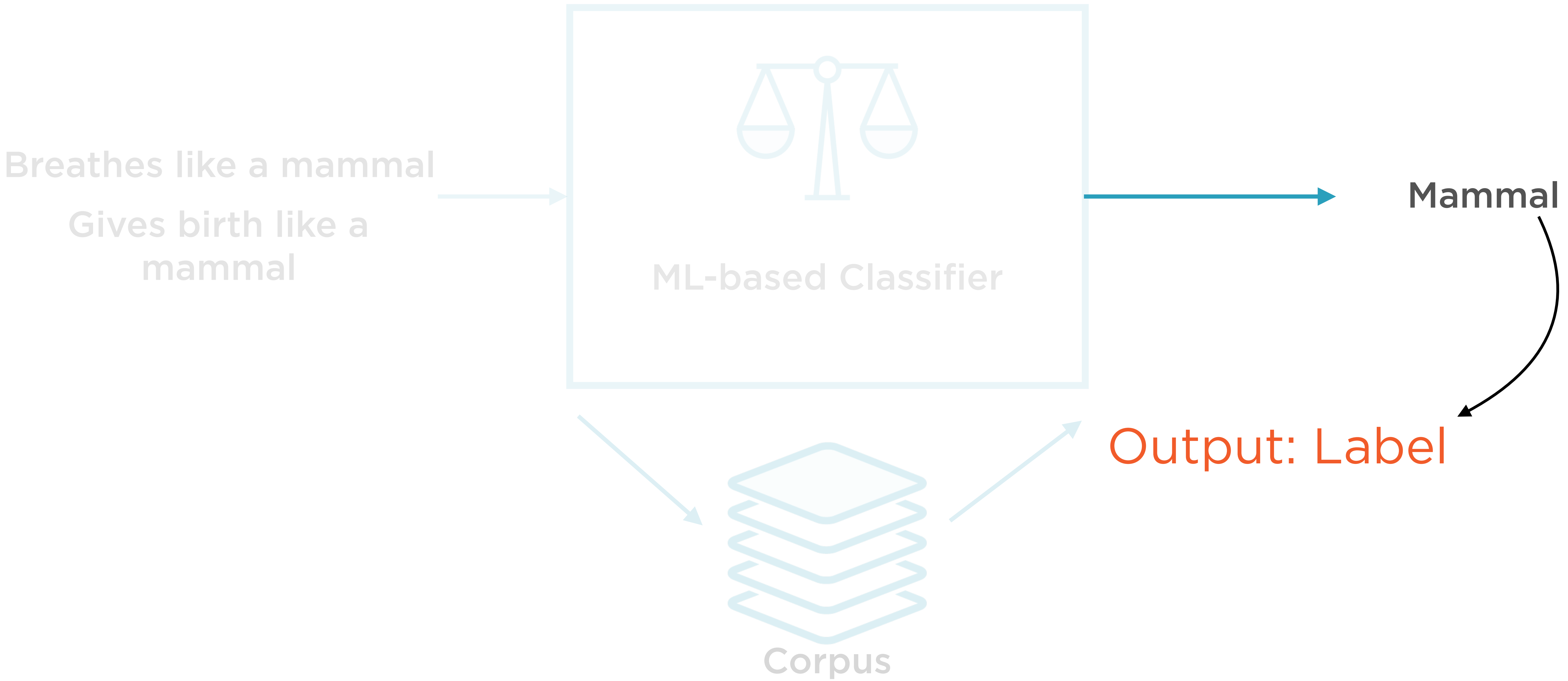
ML-based Binary Classifier



ML-based Binary Classifier



ML-based Binary Classifier



Label Encoding vs. One-hot Encoding

Label Encoding	One-hot Encoding
Often used for labels, even with nominal data	Usually used for features, not labels
Usually for y-variables (labels)	Usually for x-variables
Prevent classification from becoming multi-label problem	Would lead to overly complex multi-label problem if used for y-variables

Type of Classification

Types of Classification Tasks

Binary

"Yes/No", "True/False", "Up/Down"

Output is binary categorical variable

Multi-class

Digit classification

Output variable takes 1 of N (>2) values

Multi-label

("True", "Female"), ("False", "Female")

Output is tuple of multiple binary variables (not disjoint)

Multi-output

("Sunday", "January")

Multiclass + multilabel

Multi-class Classification

Many classification algorithms are inherently binary

- Logistic regression
- Support Vector Machines

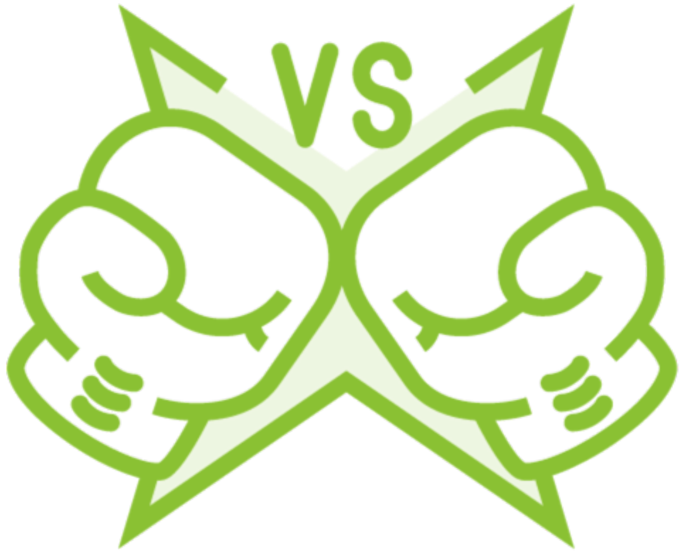
Inherently binary classifiers can be generalized for multi-class classification

Multi-class Classification

Some other algorithms are inherently multi-class

- Naive Bayes

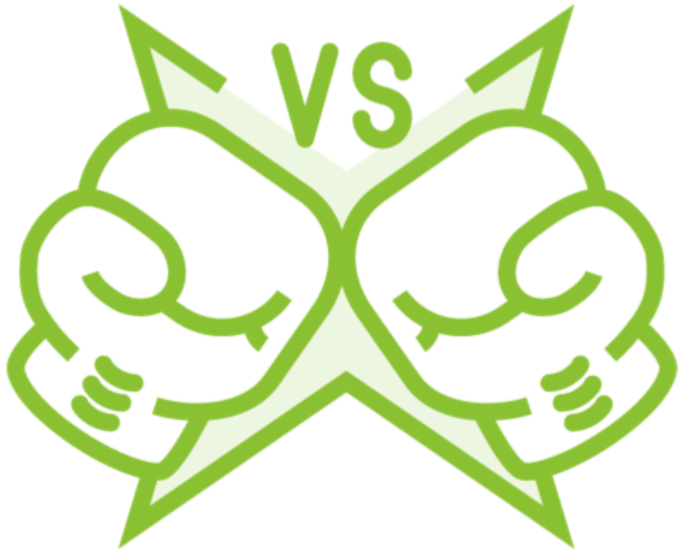
Multi-class Digit Classification



One-versus-all: Train 10 binary classifiers

- 0 or not 0
- 1 or not 1
- 2 or not 2
- Predicted label = output of detector with highest score

Multi-class Digit Classification



One-versus-one: Train 45 binary classifiers

One detector for each pair of digits

- 0 vs 1, 0 vs 2, 0 vs 3 and so on
- 1 vs 2, 1 vs 3 and so on

For N labels, need N(N-1)/2 classifiers

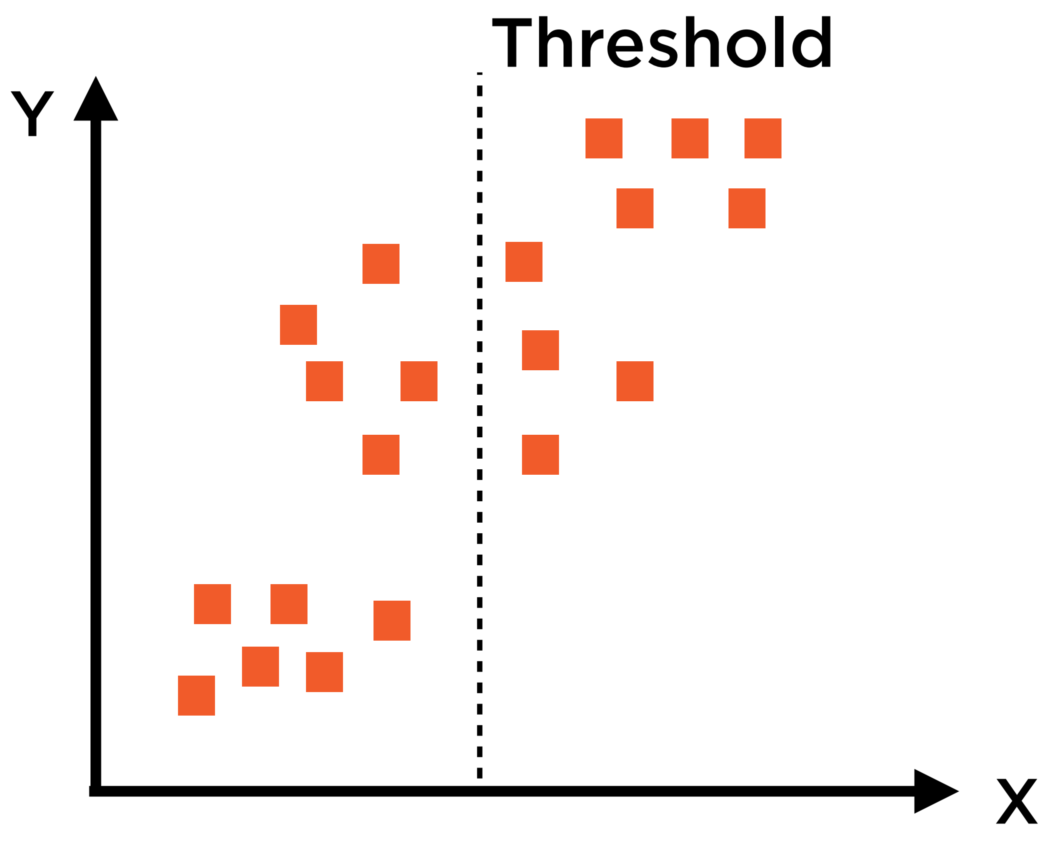
- Predicted label = output of digit that wins most duels

If you would like to one-hot encode your labels in scikit-learn - use LabelBinarizer, not OneHotEncoder

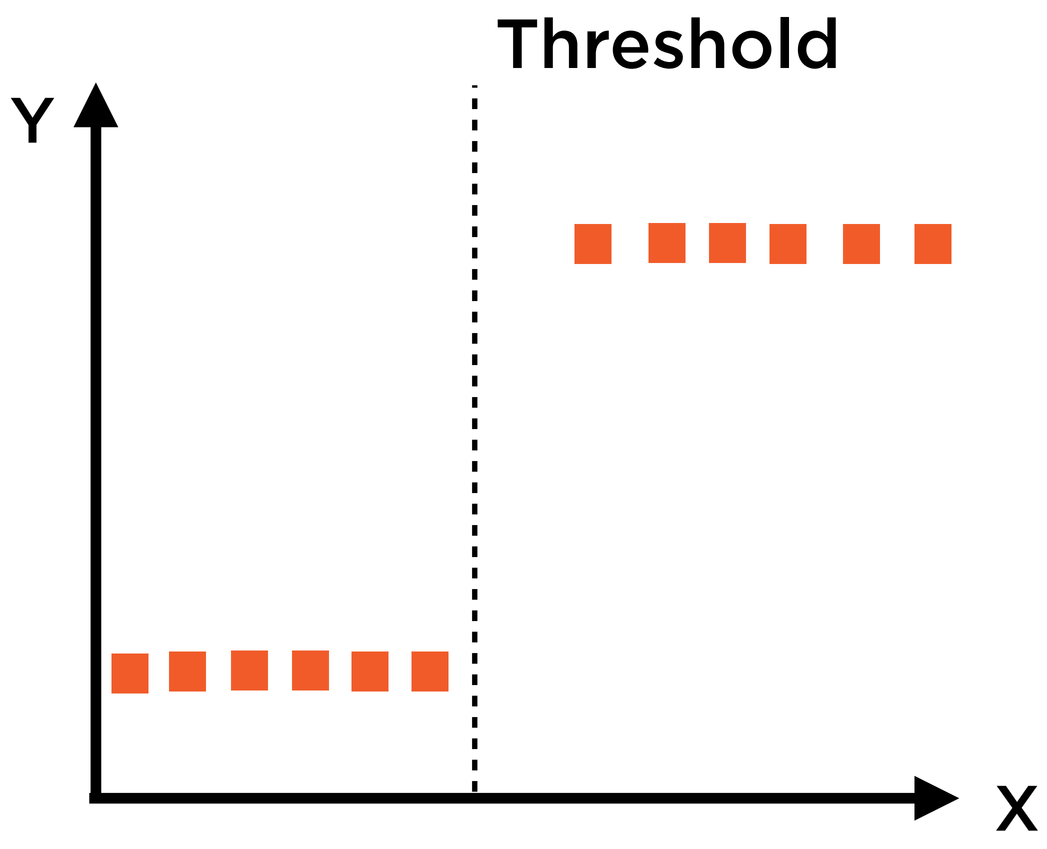
Binarizer

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

Binarizer



Continuous input



Binary categorical output

Label Binarizer

Binarize labels in one-vs-all fashion; convert multi-class labels to binary labels

Label Binarizer

E.g. to binarize days of week

- Create seven binary variables
- Variable 1: Is it Sunday? Yes or no
- Variable 2: Is it Monday? Yes or no
- ...

Inter-operates with all regression and binary classification algorithms

Demo

Converting categorical data to numeric data using one-hot-encoding

Demo

Converting categorical data to ordinal data using label encoding

Demo

Using the label binarizer to binarize labels

Demo

Using the multi-label binarizer to represent multiple categories

Summary

Categorical data vs. continuous data

Nominal vs. ordinal data

Represent categorical data using label encoding and one-hot encoding

Compare and contrast label encoding vs. one-hot encoding

Implementing categorical feature representations