Understanding and Implementing Feature Selection

Janani Ravi

CO-FOUNDER, LOONYCORN

www.loonycorn.com

Overview

Understanding feature selection

Filter methods

Embedded methods

Wrapper methods

Different measures of correlation

Types of Data

Types of Data in Machine Learning

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

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

Categorical

Ratio scale

Interval scale

Ordinal

Nominal

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

Types of Data in Machine Learning

Numeric

Categorical

Ratio scale

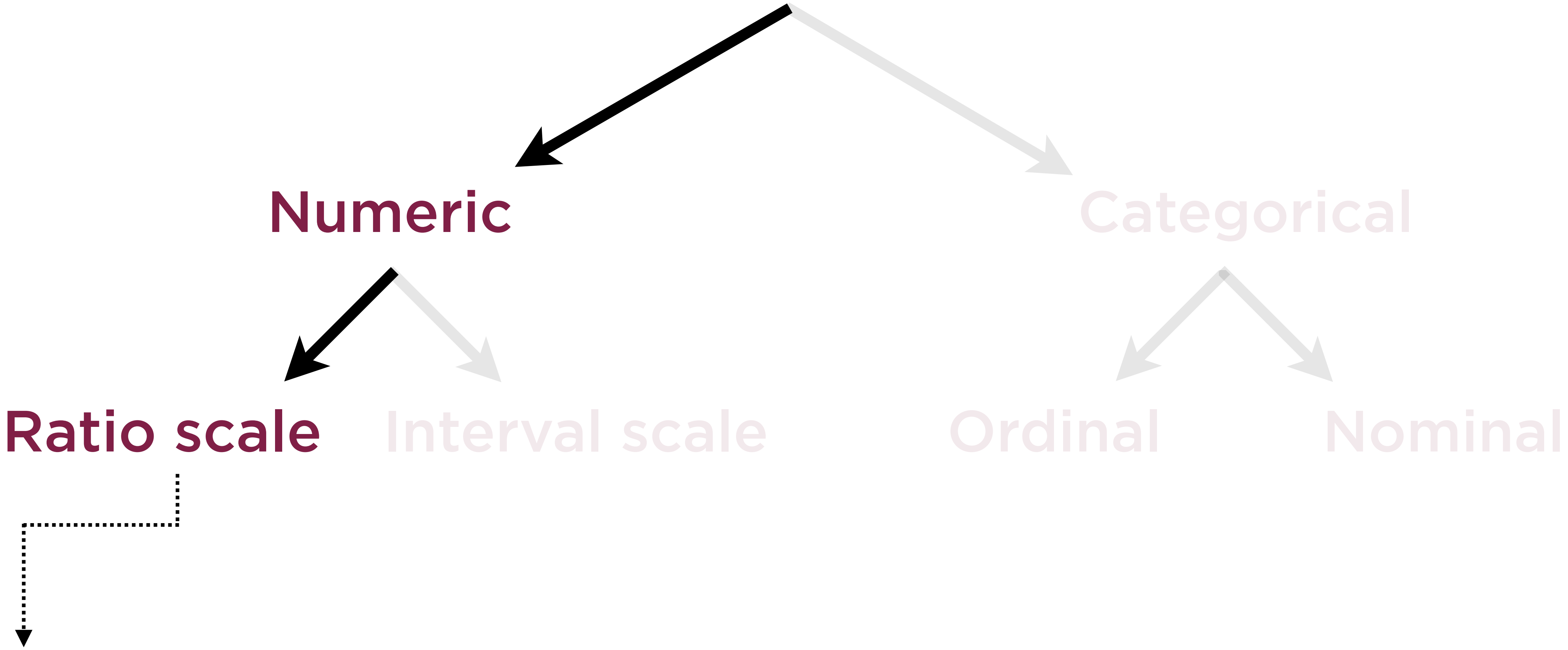
Interval scale

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

- Ratio scale
- **Interval scale**

Categorical

- Ordinal
- Nominal

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

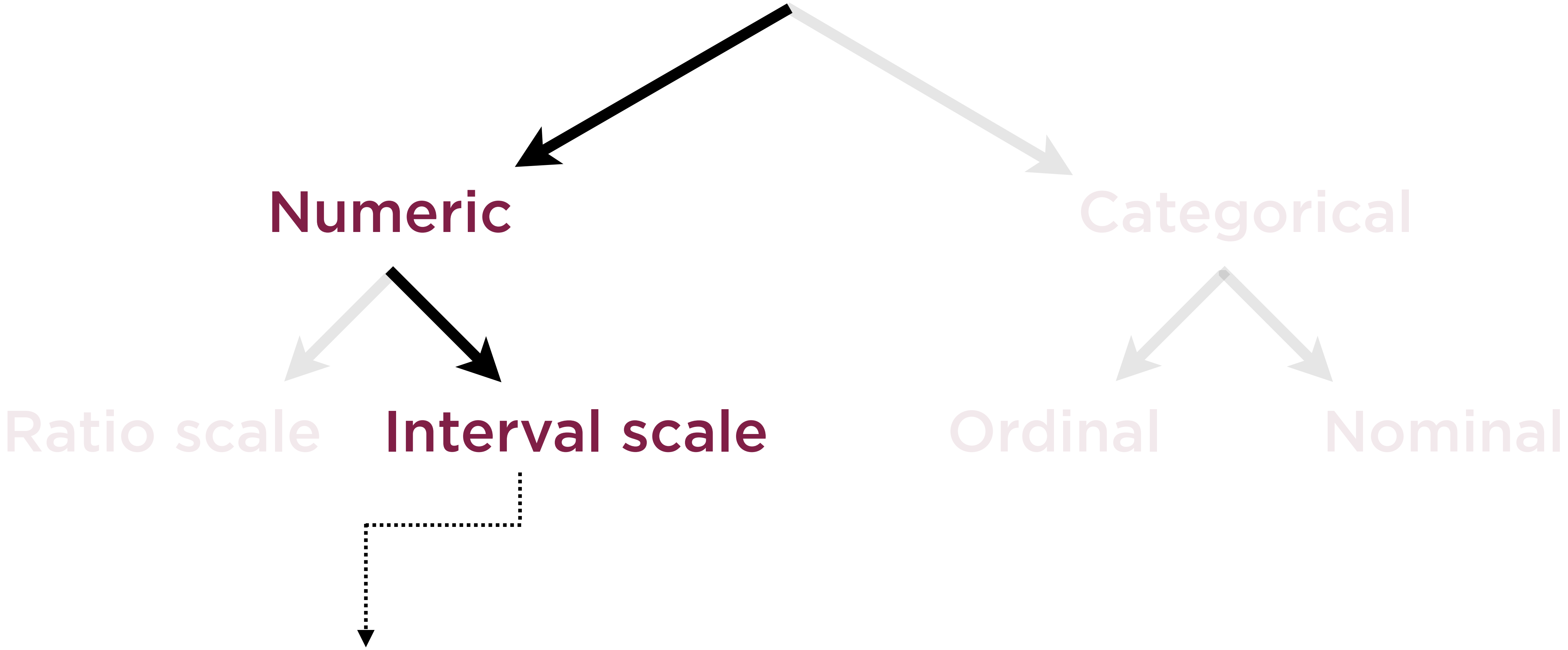
Types of Data in Machine Learning

Numeric | Categorical

Ratio scale | **Interval scale** | Ordinal | Nominal

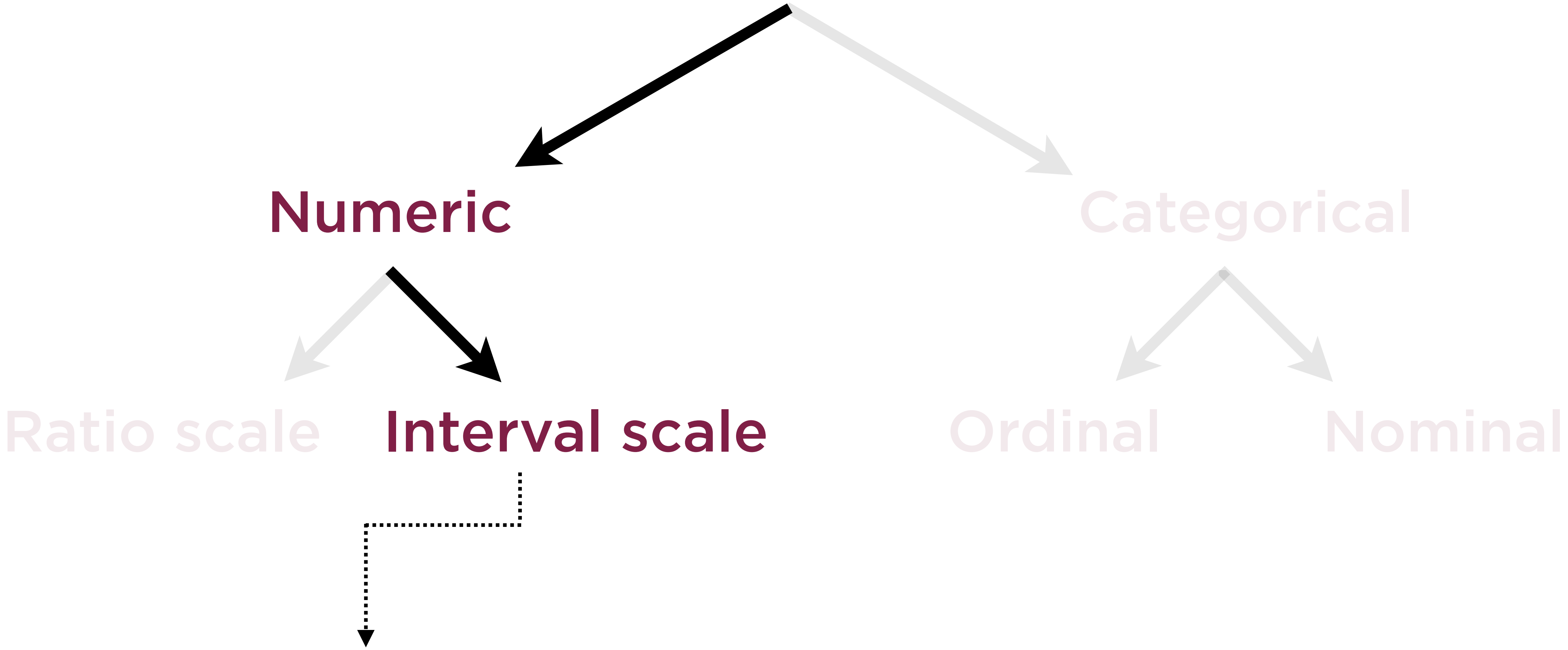
Data still numeric, but now multiplication and division no longer make sense, and zero point no longer meaningful

Types of Data in Machine Learning



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

Types of Data in Machine Learning



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

Types of Data in Machine Learning

Numeric

Categorical

Ratio scale

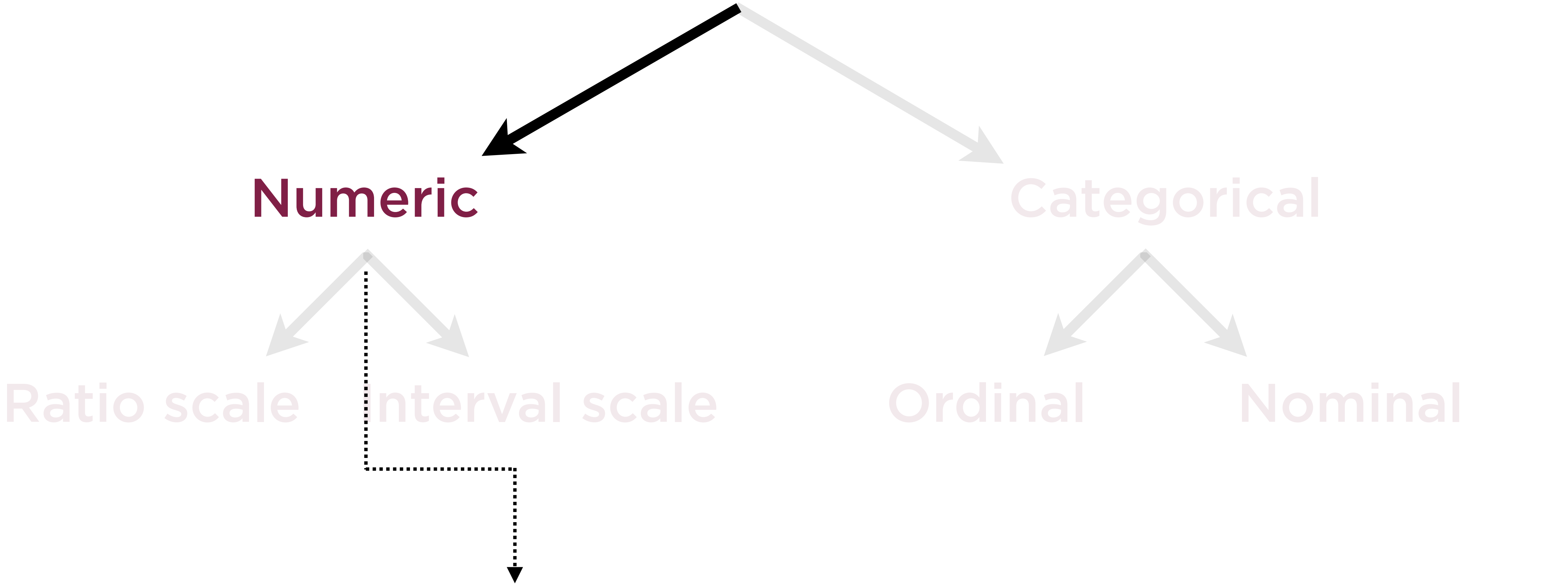
Interval scale

Ordinal

Nominal

0 Fahrenheit is not equivalent to "no temperature"

Types of Data in Machine Learning



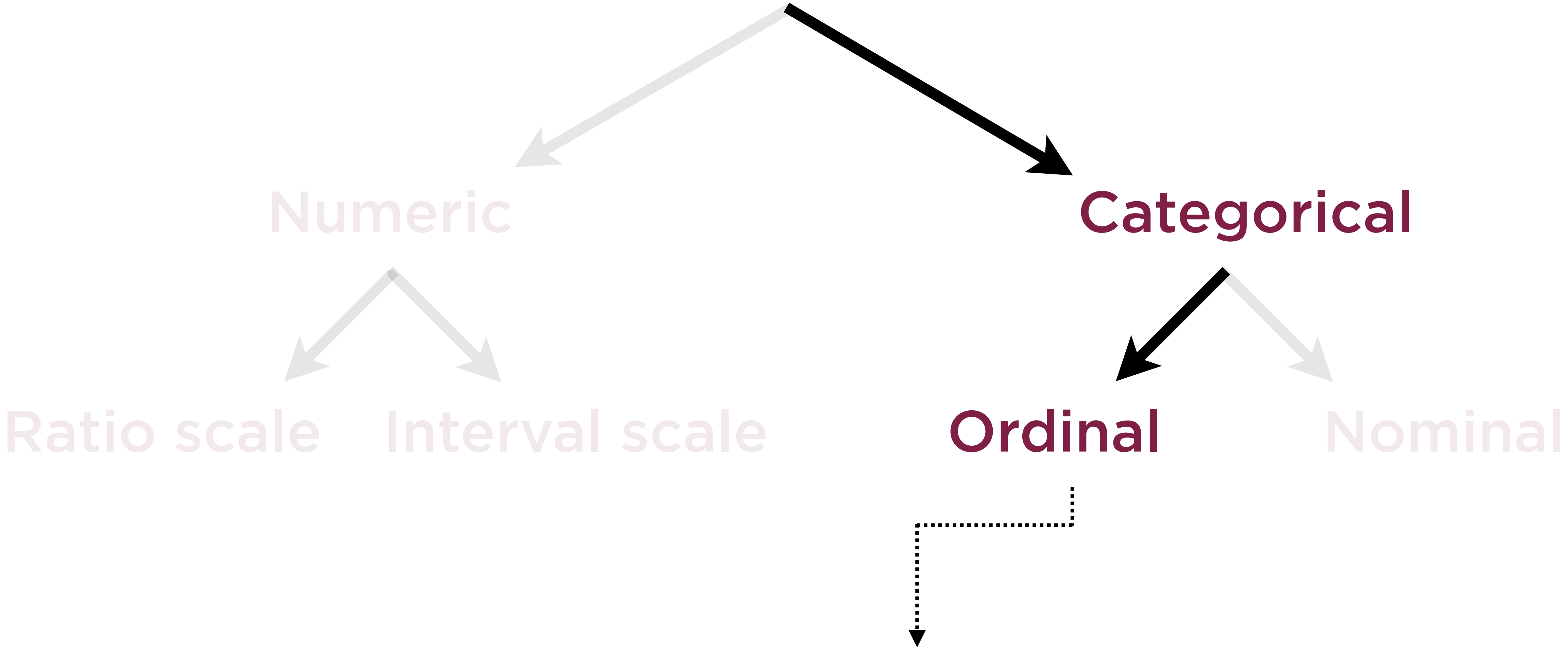
Numeric data can draw from an unrestricted range of continuous values

Types of Data in Machine Learning

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

Can calculate mean, standard deviation, Pearson correlation etc.

Types of Data in Machine Learning



Ordinal data is categorical, but can still be ordered

Types of Data in Machine Learning

Numeric

Categorical

Ratio scale

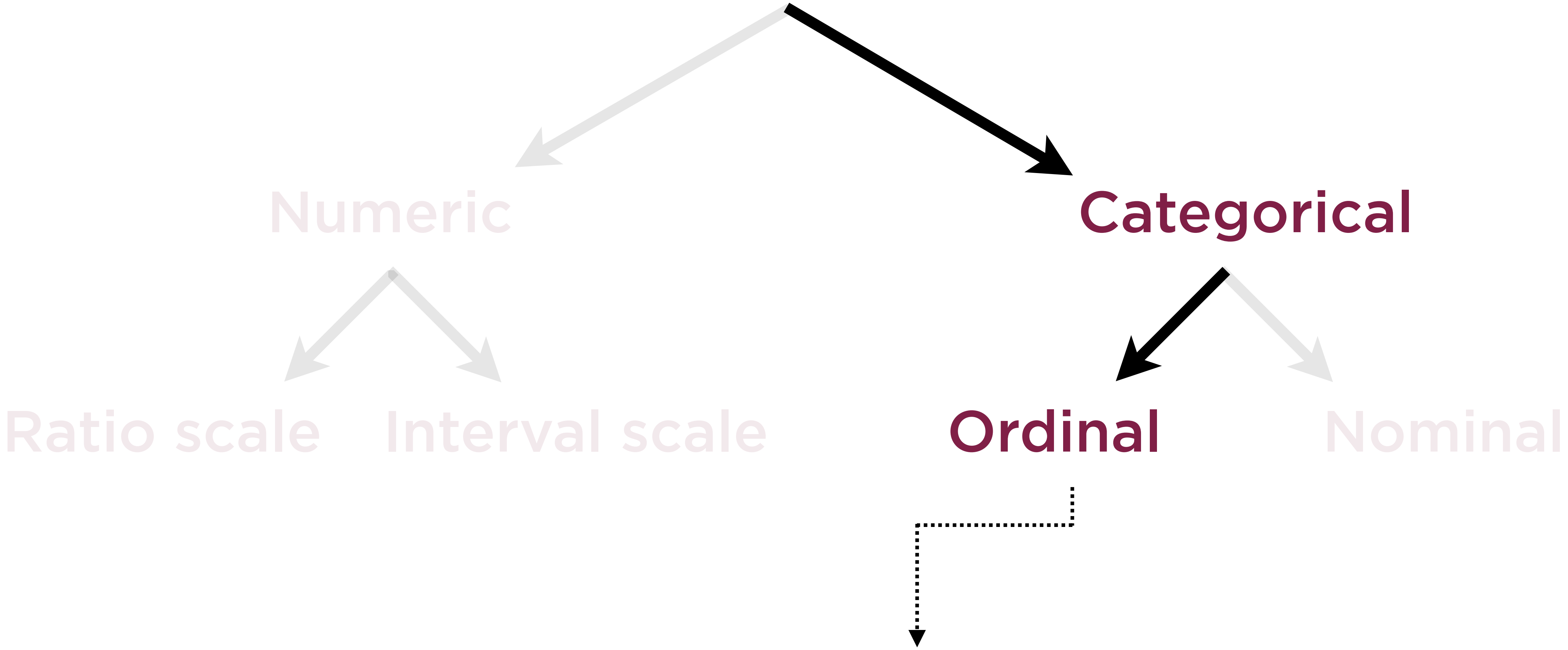
Interval scale

Ordinal

Nominal

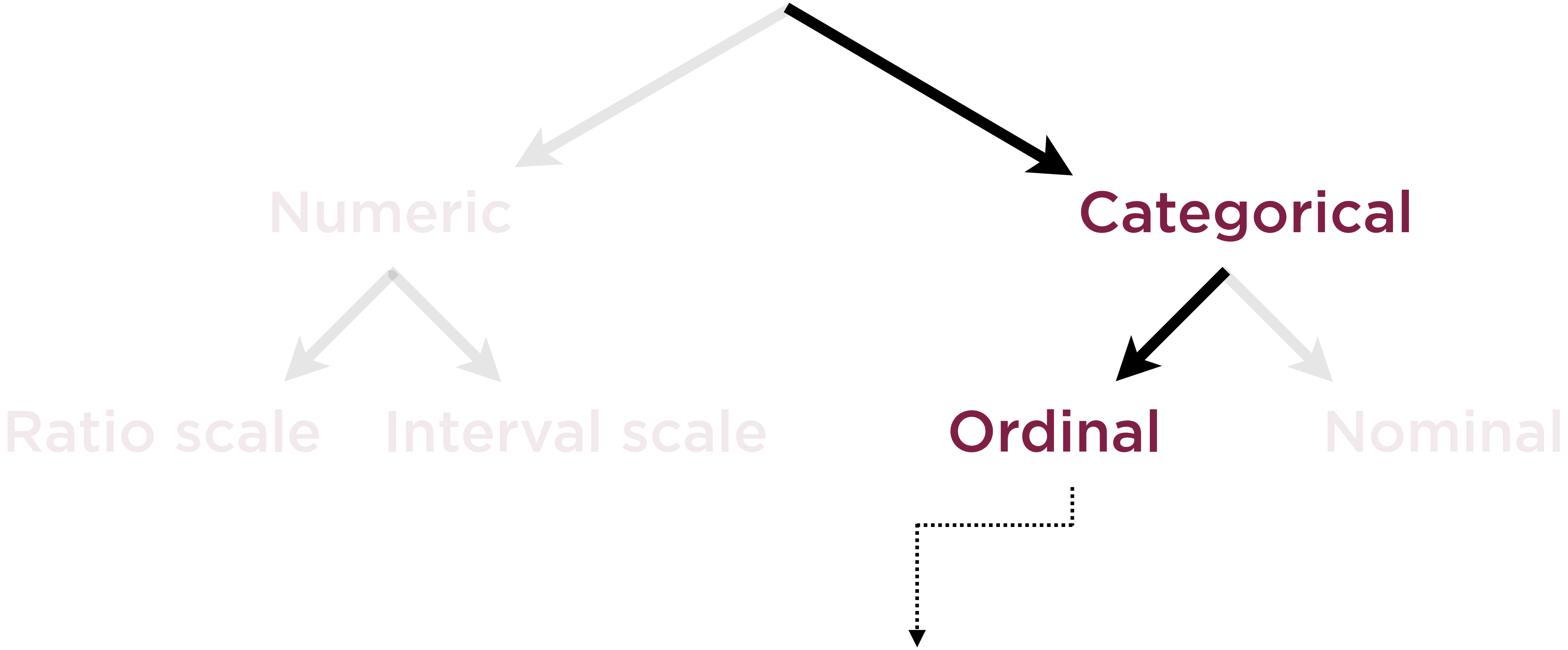
E.g. month of the year,
ratings on a scale of 1 to 5

Types of Data in Machine Learning



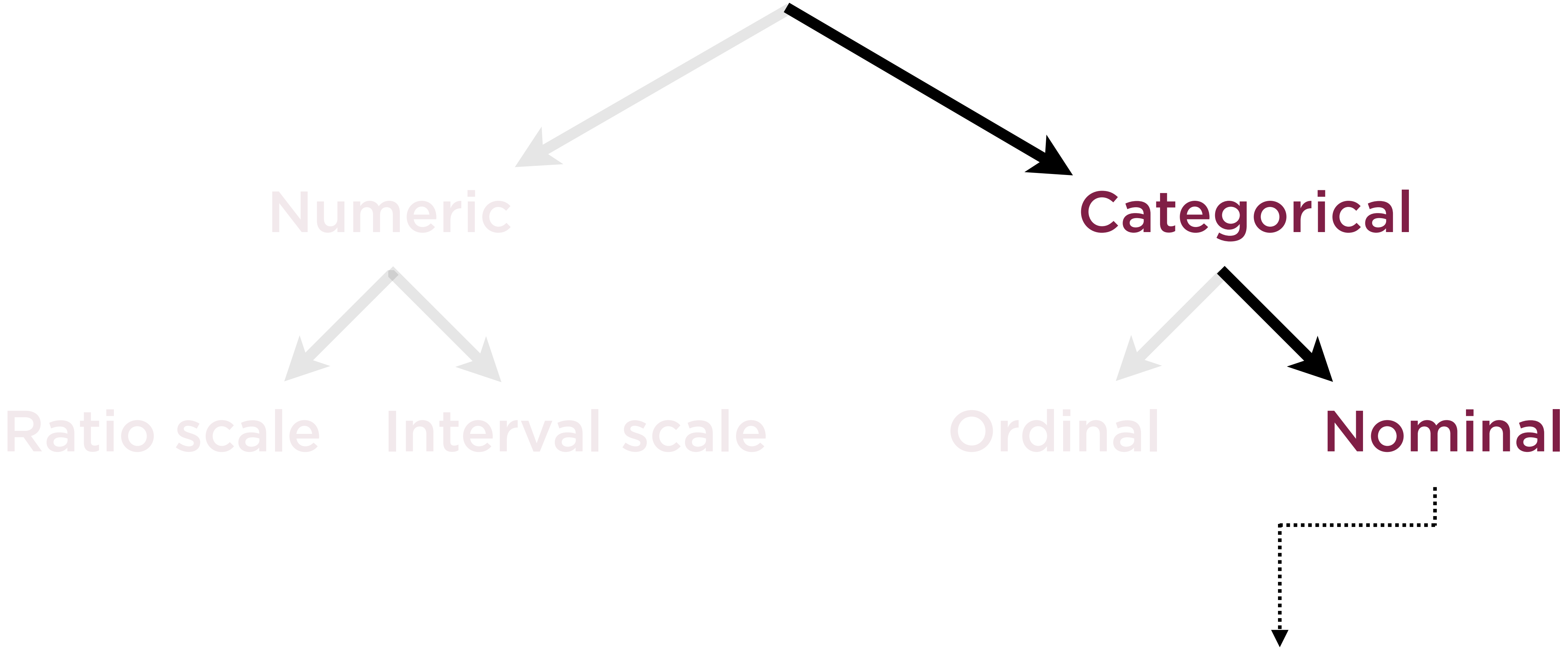
Order exists, but differences are not necessarily meaningful

Types of Data in Machine Learning



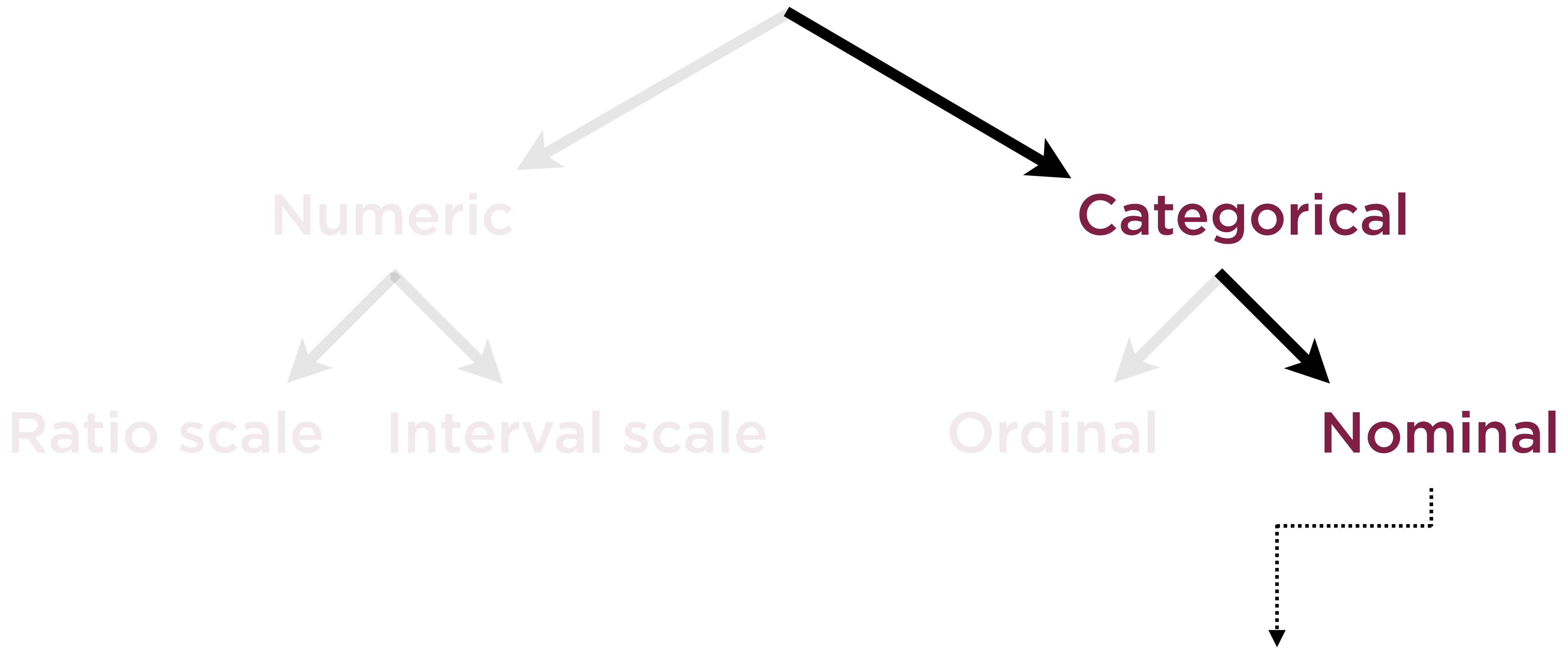
E.g. Differences in quality between three, two, one, and no Michelin stars for a restaurant are not uniform

Types of Data in Machine Learning



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

Types of Data in Machine Learning



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

Types of Data in Machine Learning

Numeric

Categorical

Ratio scale

Interval scale

Ordinal

Nominal

E.g. Brand names of cars
(“Ford” and “Honda”)

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

Use regression to predict numeric (continuous) y-variables

Use classification to predict categorical (discrete) y-variables

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

Measures of Correlation

Correlation

Any statistical relationship, whether causal or not, between two random variables.

Correlation Coefficient

Numerical measure of the correlation between two random variables.

Measures of Correlation

Pearson

Kendall

Spearman

Measures of Correlation

Pearson	Kendall	Spearman

Works with only numeric data - most restrictive, most common

Measures of Correlation

Pearson	**Kendall**	Spearman

Rank correlation measure - works with interval, ratio, and ordinal data

Measures of Correlation

Pearson

Kendall

Spearman

Rank correlation measure - works with interval, ratio, and ordinal data

Shared Properties

Each of these metrics satisfy some properties

- Maximum value of +1
- Minimum value of -1
- Uncorrelated data has 0 correlation

Pearson Correlation Coefficient

Commonest metric

Measure linear relationship

Works with numeric data

Assumes normally distributed data

Kendall Rank Correlation

Rank correlation measure

Works even with ordinal data

- Will not work with nominal data

Used to measure whether ranked orderings are similar or not

No linear relationship posited

Spearman Rank Correlation

Works even with ordinal data

- Will not work with nominal data

No linear relationship posited

Assumes monotonic relationship i.e. either increasing or decreasing

Evaluating Correlations

Cohen’s standard to measure strength of association

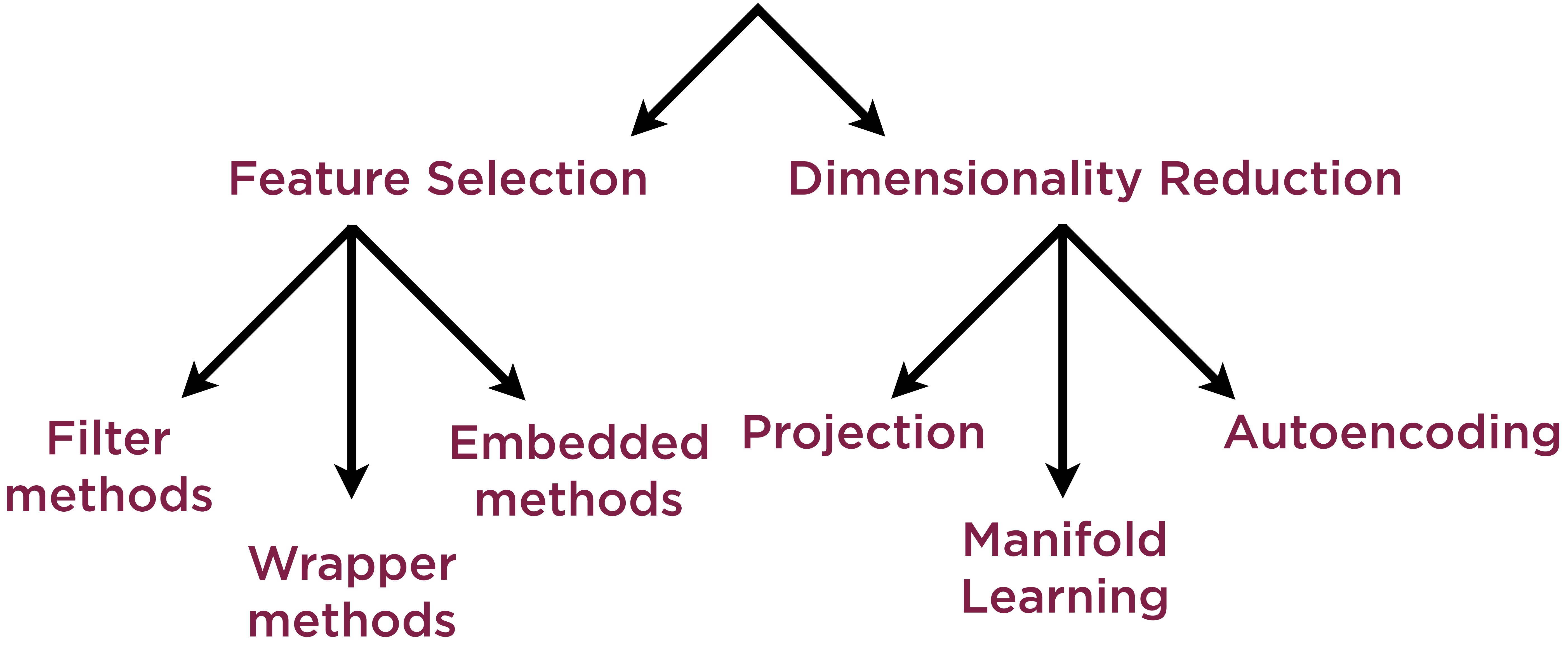
Magnitude	Association
0 to 0.1	None
0.1 to 0.3	Small/weak
0.3 to 0.5	Moderate
> 0.5	Strong

Feature Selection

Problems with Data

Insufficient data	Too much data	Non-representative data
Missing data	Duplicate data	Outliers

Reducing Complexity



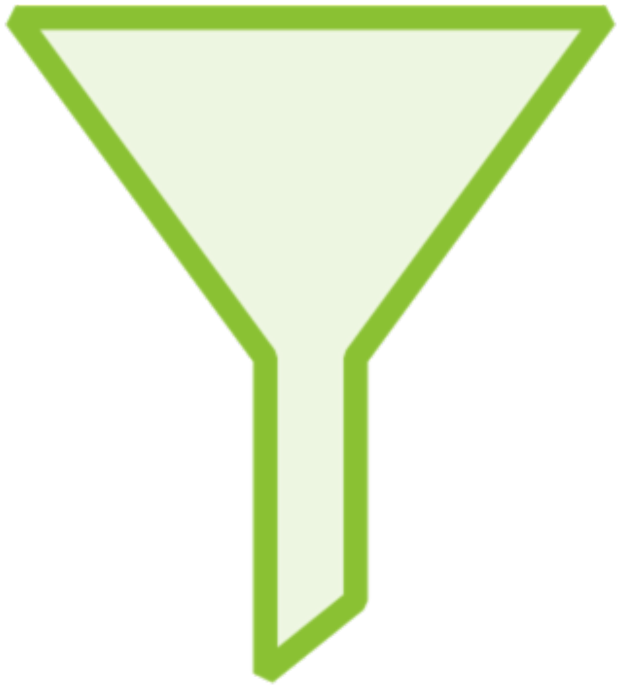
Reducing Complexity

- Feature Selection
  - Filter methods
  - Wrapper methods
  - Embedded methods
- Dimensionality Reduction
  - Projection
  - Manifold Learning
  - Autoencoding

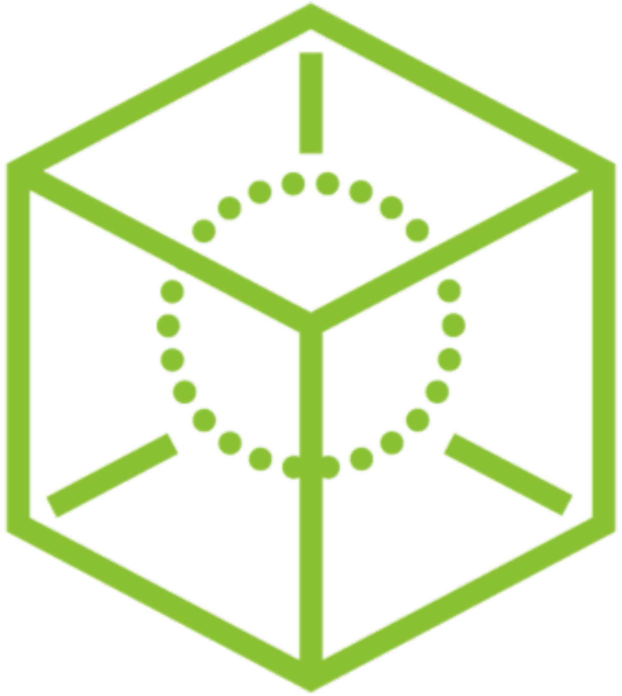
Choosing Feature Selection

Use Case	Possible Solution
Many X-variables Most of which contain little information Some of which are very meaningful Meaningful variables are independent of each other	Feature selection

Feature Selection Techniques



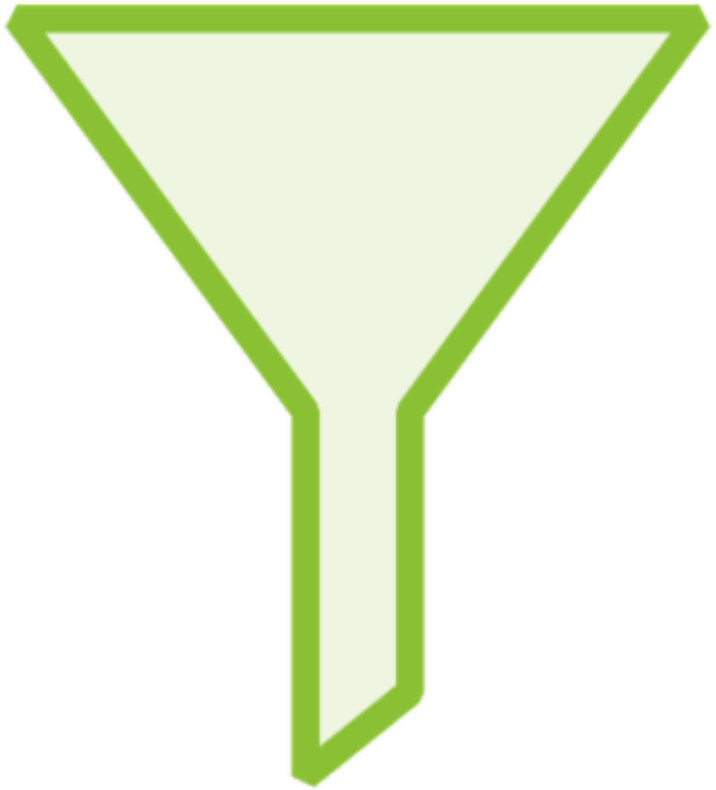
Filter methods



Embedded methods

Wrapper methods

Filter Methods

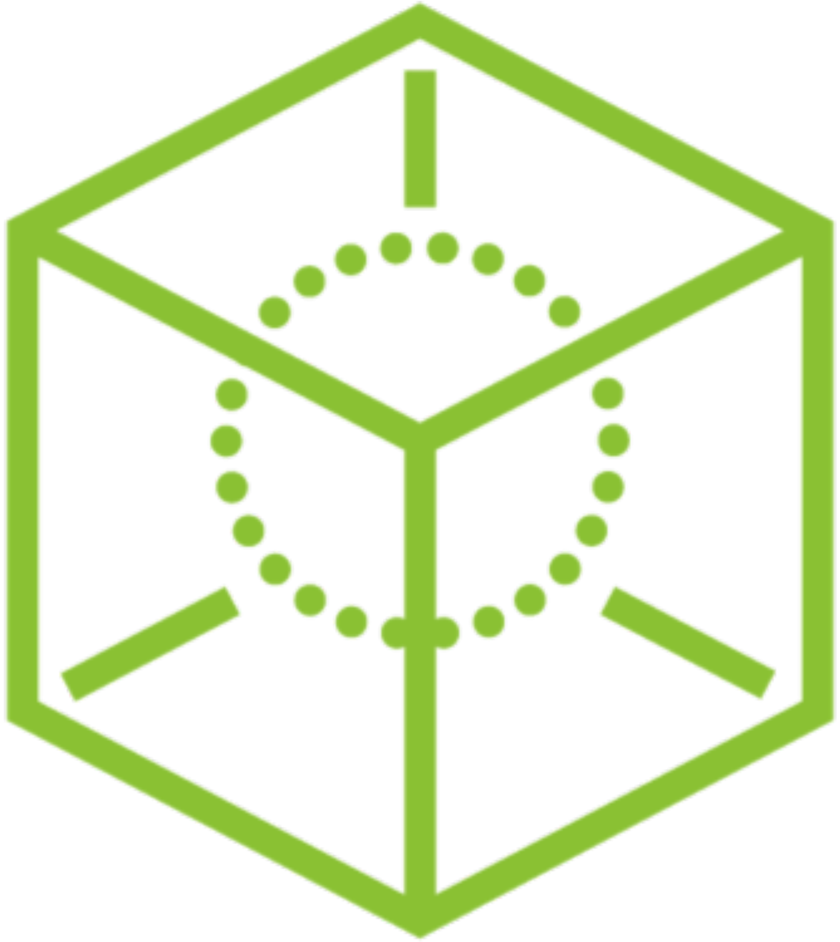


Features (columns) selected independently of choice of model

Rely on statistical properties of features

Either individually (univariate) or jointly (multi-variate)

Embedded Methods



Features (columns) selected during model training

Feature selection effectively embedded within modeling

Only specific types of models perform feature selection

Wrapper Methods

Somewhere between filter and embedded feature selection

Features are chosen by building different candidate models

Forward and backward stepwise regression are examples

Wrapper Methods

Each candidate model has different subset of features

However all candidate models are similar in structure

Features may be added or dropped to see whether the model improves

Demo

Performing feature selection using the missing value ratio

Demo

Computing feature correlations using different techniques

Visualizing feature correlations

Demo

Performing feature selection using filter, wrapper, and embedded techniques

Summary

Understanding feature selection

Filter methods

Embedded methods

Wrapper methods

Different measures of correlation