

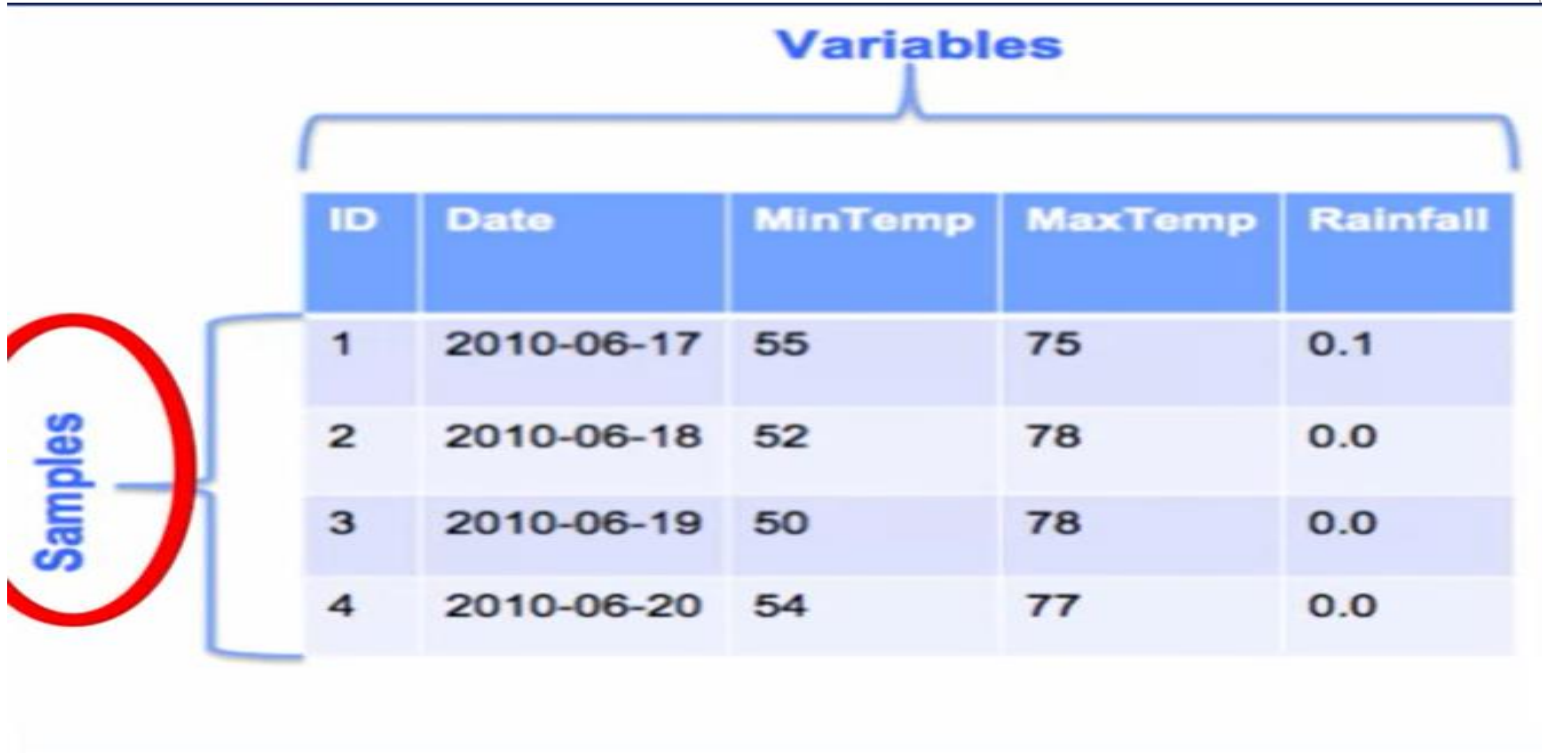
Unit -3

Presented By
Rajit Nair

Data exploration

- ▶ If you have been in a conversation in machine learning, you have probably heard terms like feature, sample, and variable. We will be defining some of those terms in this lecture.
- ▶ Let's discuss the feature and how it relates to a sample. Name some alternative terms for feature. Summarize how a categorical feature differs from a numerical feature. Before we delve into the methods for processing and analyzing data, let's first start with defining some term used to describe data, starting with sample and variable.
- ▶ A sample is an instance or example of an entity in your data.

Terms to describe data



The diagram shows a table with five columns: ID, Date, MinTemp, MaxTemp, and Rainfall. A blue bracket above the columns is labeled 'Variables'. A red circle on the left is labeled 'Samples', with a blue bracket grouping the four rows of data below it.

Variables				
ID	Date	MinTemp	MaxTemp	Rainfall
1	2010-06-17	55	75	0.1
2	2010-06-18	52	78	0.0
3	2010-06-19	50	78	0.0
4	2010-06-20	54	77	0.0

Other Names for 'Sample'

sample

row

instance

observation

record

example

Samples

ID	Date	MinTemp	MaxTemp	Rainfall
1	2010-06-17	55	75	0.1
2	2010-06-18	52	78	0.0
3	2010-06-19	50	78	0.0
4	2010-06-20	54	77	0.0

Other Names for 'Variable'

variable

feature

dimension

column

attribute

field

Variables

ID	Date	MinTemp	MaxTemp	Rainfall
1	2010-06-17	55	75	0.1
2	2010-06-18	52	78	0.0
3	2010-06-19	50	78	0.0
4	2010-06-20	54	77	0.0

- An important point to emphasize about variable is that, they are additional values with a data type. Each variable has a datatype associated with it. The most common data types are numeric and categorical.
- There are other data types as well such as string and date but we will focus on two of the more common data types, numeric and categorical. As the name implies, numeric variables are variables that take on number values. Numeric variables can be measured, and their values can be sorted in some way.
- Note that a numeric variable can take on just integer values or be continuous valued. It can also have just positive numbers, negative numbers or both.

Let's go over some examples of various numeric variables. A person's height is a positive, continuous valued number.

The score in an exam is a positive number that range between zero and a 100%.

The number of transactions per hour is a positive integer, whereas the change in a stock price can be either positive or negative. A variable with labels, names, or categories for values instead of numbers are called categorical variables.

For example a variable that describes the color of an item, such as the color of a car, can have values such as red, silver, blue, white and black. These are non-numeric values that describes some quality or characteristic of an entity.

These values can be thought of as names or labels that can be sorted into categories. Therefore, categorical variables are also referred to as qualitative variables, or nominal variables. Some examples of categorical variables are gender, marital status, type of customer, for example, teenager, adult, senior. Product categories, for example, electronics, kitchen, bathroom and color of an item.

To summarize, a sample is an instance or example of an entity in your data.

A variable captures a specific characteristic of each entity. So a sample has many variables to describe it. Data from real applications are often multidimensional, meaning that there are many dimensions or variables describing each sample.

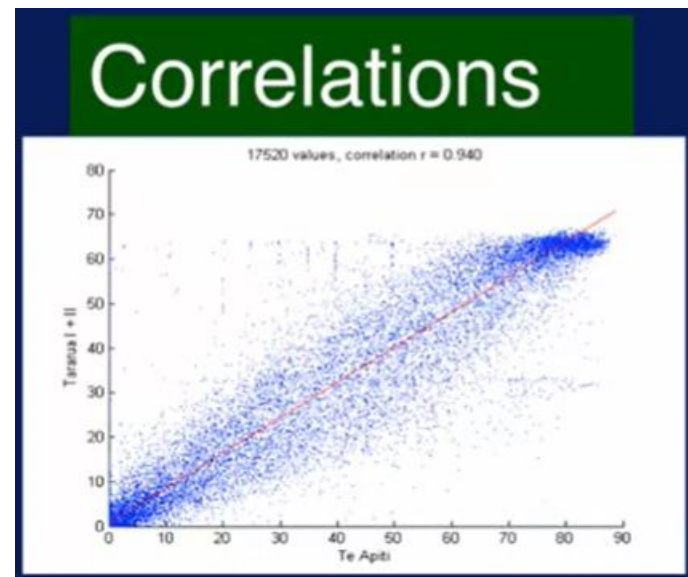
Each variable has a datatype associated with it, the most common data types are numeric and categorical. Note that there are many terms to describe these data related concepts.

Data Exploration

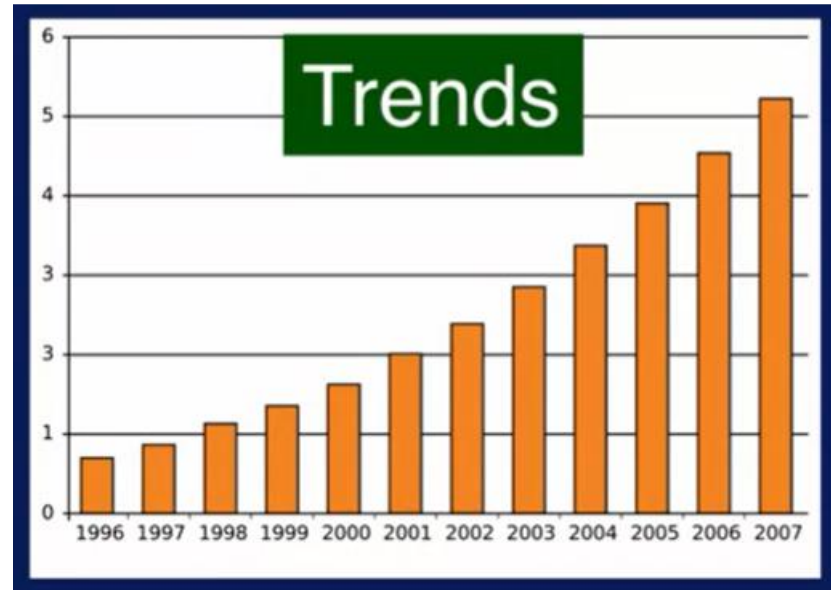
- ▶ Now that we are familiar with some commonly use terms to describe data, let's look at what data exploration is and why it's important.
- ▶ After this presentation, you will be able to explain why data exploration is necessary, articulate the objectives of data exploration, list the categories of techniques for exploring data. Data exploration means doing some preliminary investigation of your data set.
- ▶ The goal is to gain a better understanding of the data that you have to work with. If you understand the characteristics of your data, you can make optimal use of it in whatever subsequent processing and analysis you do with the data.
- ▶ Note that data exploration is also called exploratory data analysis, or EDA for short. How do you go about exploring data?

- There are two main categories of techniques to explore your data, one based on summary statistics and the other based on visualization methods. Summary statistics provide important information that summarizes a set of data values.
- There are many such statistics. Many of them you have probably heard of before, such as mean, median, and standard deviation. These are some very commonly used summary statistics. A summary statistic provides a single quantity that summarizes some aspects of the dataset.
- For example, the mean, is a single value that describes the average value of the dataset, no matter how large that dataset is. You can think of the mean as an indicator of where your dataset is centrally located on a number line, thus summary statistics provide a simple and quick way to summarize a dataset.
- Data visualization techniques allow you to look at your data, graphically. There are several types of plots that you can use to visualize your data.

- Some examples are histogram, line plot, and scatter plot. Each type of plot serves a different purpose, we will cover the use of plots to visualize your data in an upcoming lecture. What should you look for when exploring your data?
- You use statistics and visual methods to summarize and describe your dataset, and some of the things you'll want to look for are correlations, general trends and outliers.
- Correlations provide information about the relation took between variables in your data. By looking at correlations, you may be able to determine that two variables are very correlated. This means they provide the same or similar information about your data. Since this contain redundant information, this suggest that you may want to remove one of the variables to make the analysis simpler.

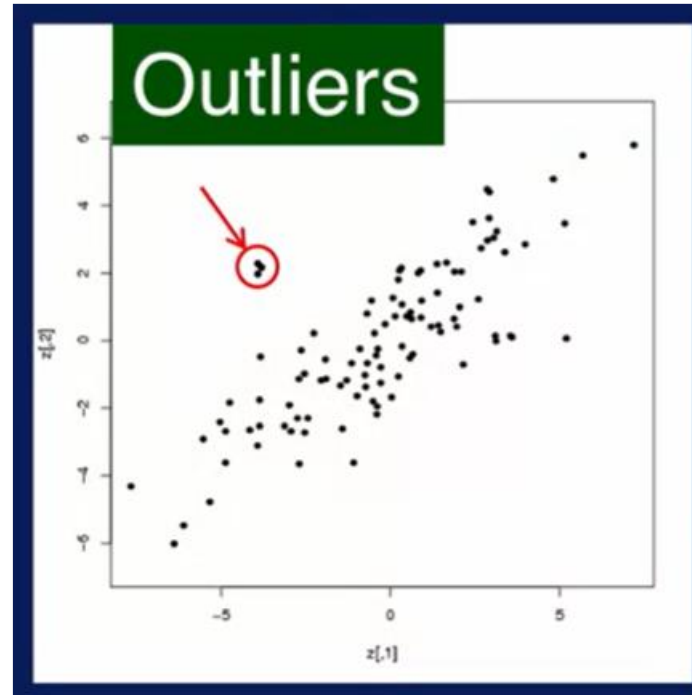


Trends in your data will reveal characteristics in your data. For example, you can see where the majority of the data values lie, whether your data is skewed or not, what the most frequent value or values are in a data set, etc. Looking at trends in your data can also reveal that a variable is moving in a certain direction, such as sales revenue increasing or decreasing over the years.



Calculating the minimum, the maximum and range of the data values are basic steps in exploring your data. Determining outliers is also very important. Outliers indicate potential problems with the data and may need to be eliminated in some applications.

In other applications, outliers represent interesting data points that should be looked at more closely. In either case, outliers usually require further examination. In summary, what you get by exploring your data is a better understanding of the complexity of the data so you can work with it more effectively.



Better understanding in turn will guide the rest of the process and lead to more informed analysis. Summary statistics and visualization techniques are essential in exploring your data. This should be used together to examine a dataset.

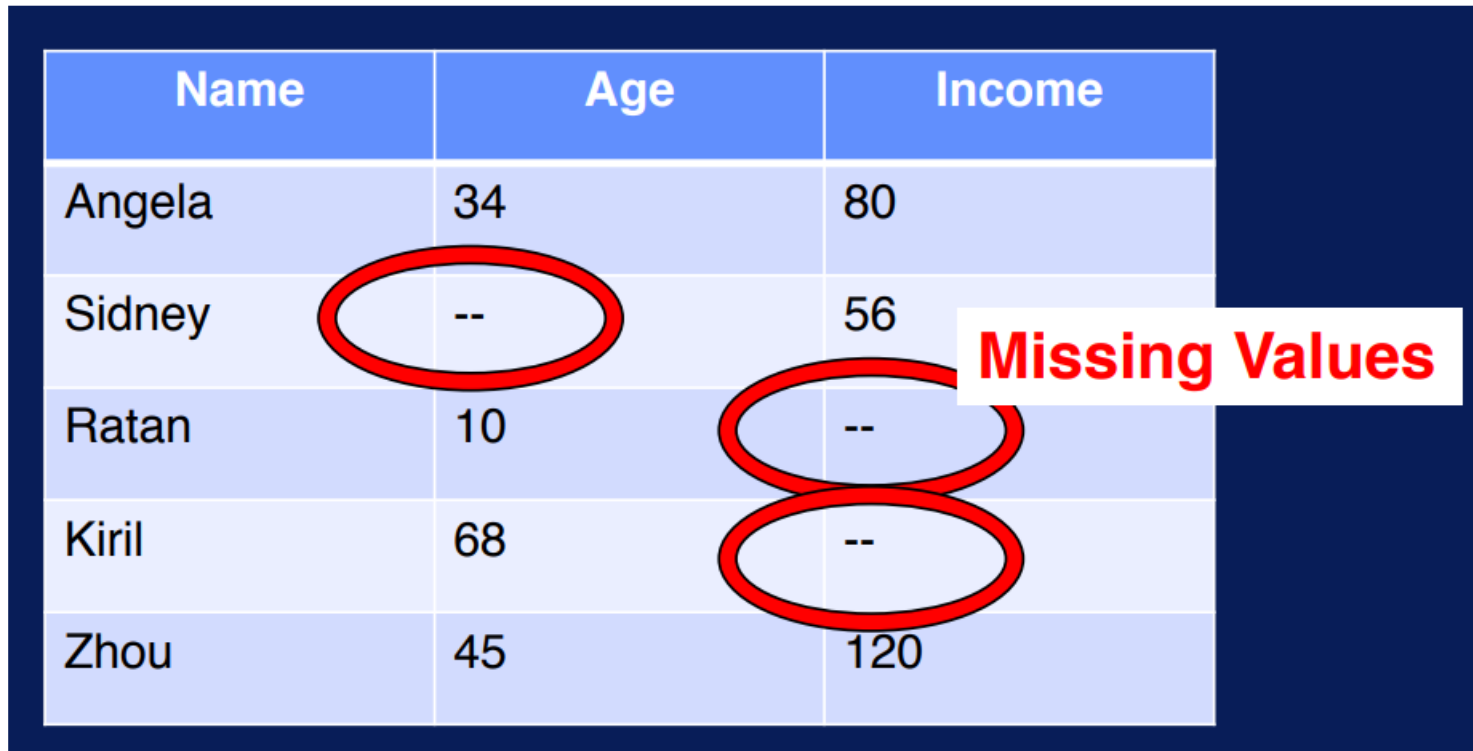
Data Quality

Data Quality Issues

1. Missing Data

Name	Age	Income
Angela	34	80
Sidney	--	56
Ratan	10	--
Kiril	68	--
Zhou	45	120

Missing Values



Duplicate Data

Name	Address
Angela	430 Park Drive
Sidney	7800 West View Street
Sid	7800 West View Street
Ratan	12442 Mountain Avenue
Kiril	45 East 5 th St
Kiril	1220 Mill Avenue
Zhou	4345 Apple Lane

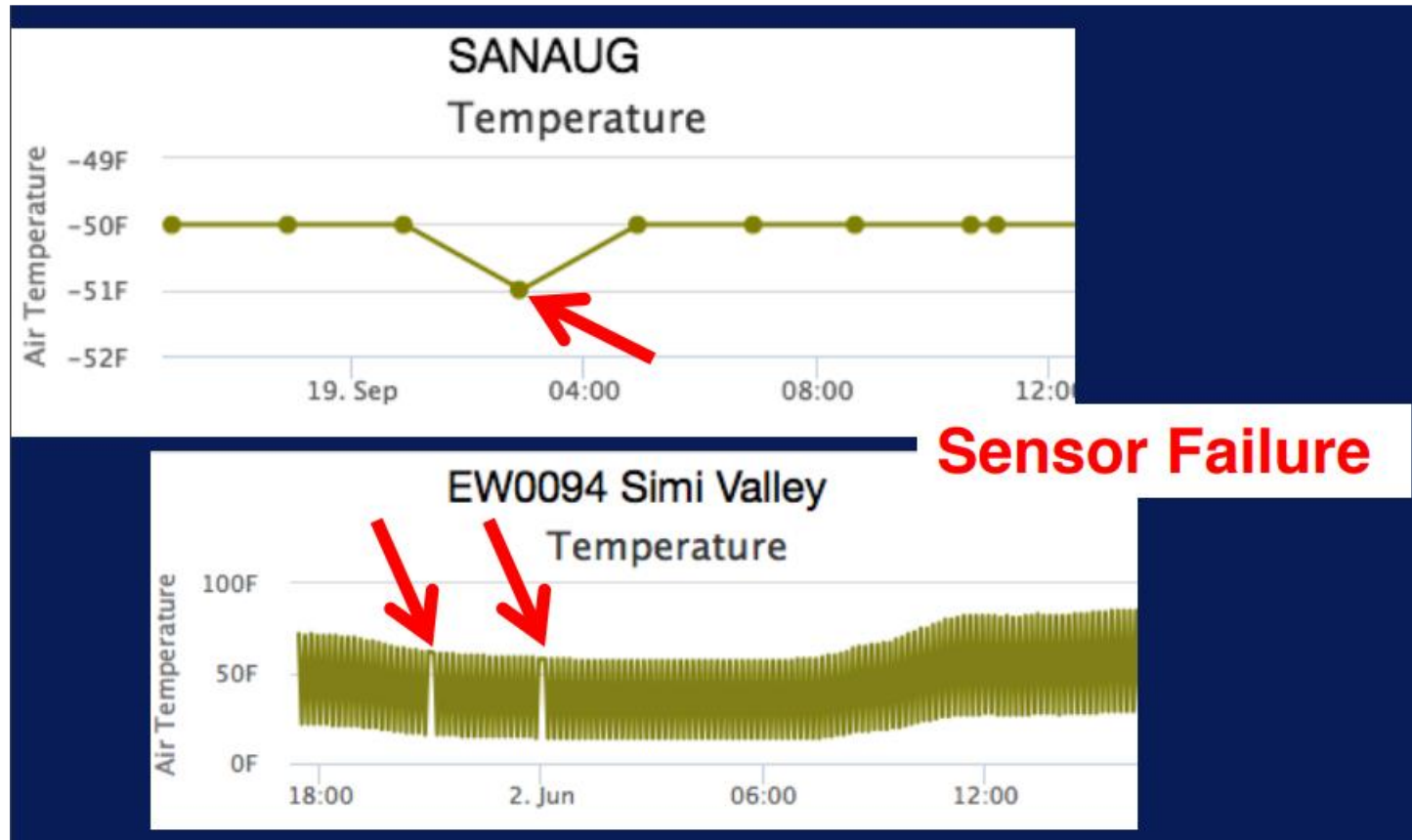
Invalid Data

Name	Zip Code
Angela	346412
Sidney	92618
Ratan	8033A
Kiril	11012
Zhou	59285

Noise

Name	Address
Angela	430 Park Drive
Sidney	780 ★❖©◆ View Street
Ratan	12443 Mountain Avenue
Kiril	1220 Mill Avenue
ZhČou	4345 Apple Lane

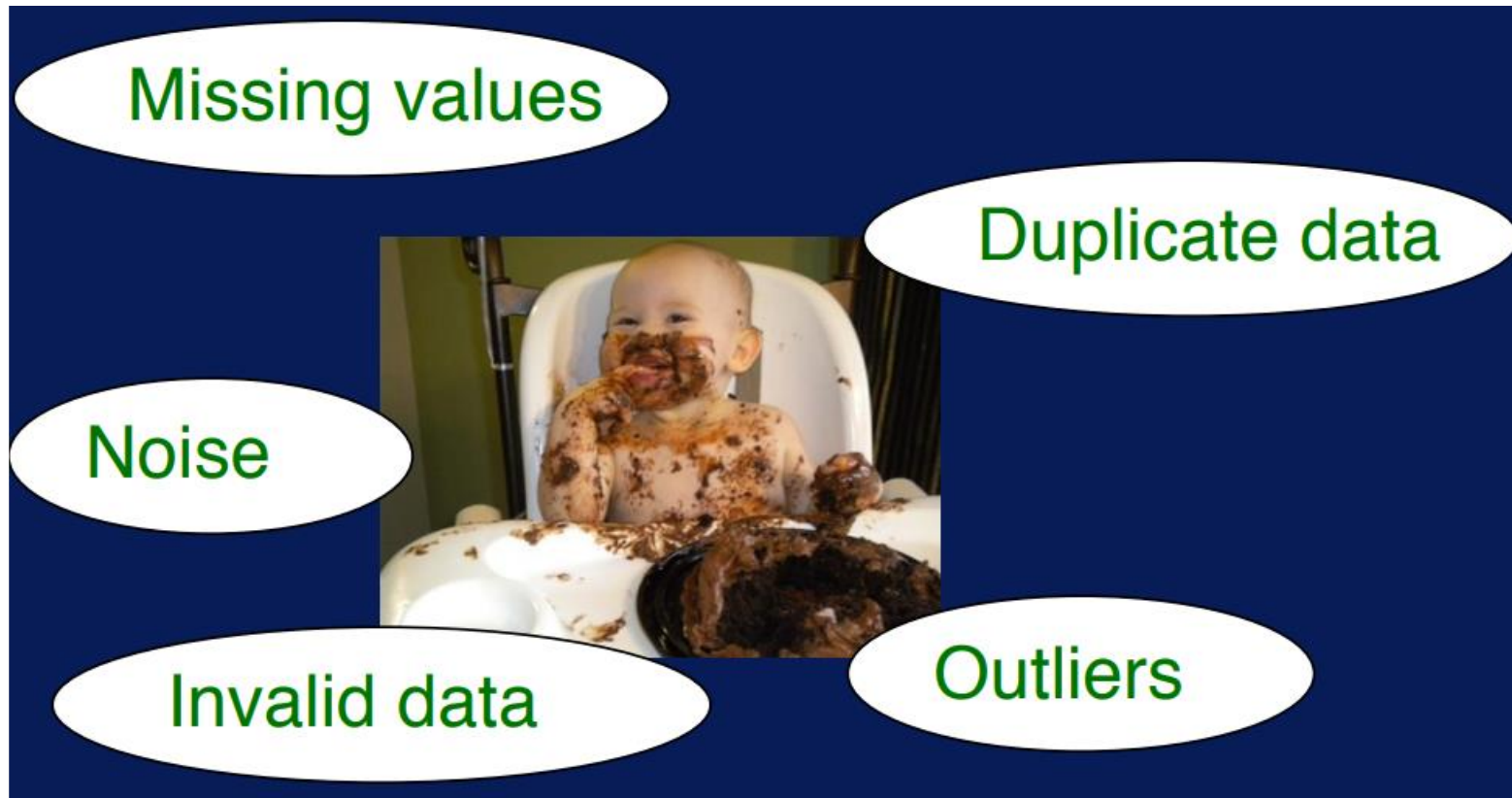
Outliers



Why Address Data Quality Issues?



Data Quality Issues



Removing Missing Data

Name	Age	Income
Angela	34	80
Sidney	--	56
Ratan	10	--
Kiril	68	--
Zhou	45	120



Name	Age	Income
Angela	34	80
<i>Sidney</i>	<i>--</i>	<i>56</i>
<i>Ratan</i>	<i>10</i>	<i>--</i>
<i>Kiril</i>	<i>68</i>	<i>--</i>
Zhou	45	120

Imputing Missing Data

Name	Age	Income
Angela	34	80
Sidney	--	56
Ratan	10	--
Kiril	68	--
Zhou	45	120



- Replace missing values with something reasonable

Name	Age	Income
Angela	34	80
Sidney	50	56
Ratan	10	50
Kiril	68	50
Zhou	45	120

Ways to Impute Missing Data

Replace missing value with

- ▶ Mean
- ▶ Median
- ▶ Most frequent
- ▶ Sensible value based on application

Duplicate Data

- Delete older record.
- Merge duplicate records

Name	Address
Sidney	7800 West View Street
Sid	7800 West View Street
Kiril	45 East 5 th St
Kiril	1220 Mill Avenue



Name	Address
Sidney	7800 West View Street
Sid	7800 West View Street
Kiril	45 East 5th St
Kiril	1220 Mill Avenue

Invalid Data

- ▶ Use external data source to get correct value
- ▶ Apply reasoning and domain knowledge to come up with reasonable value.



Name	Zip Code
Angela	346412
Ratan	8033A

Name	Zip Code
Angela	34641 2
Ratan	8033 1

Outliers

Remove outliers if they're not focus of analysis

- ▶ Analyze more closely if they are focus of analysis (e.g., fraud detection)


Domain Knowledge

- ▶ Required for addressing data quality issues effectively

Noise

- ▶ Filter out noise component.
- ▶ May also filter out part of data, so care must be taken.

Name	Address
Sidney	7800 ★❖©◆ View Street
ZhČou	4345 Apple Lane

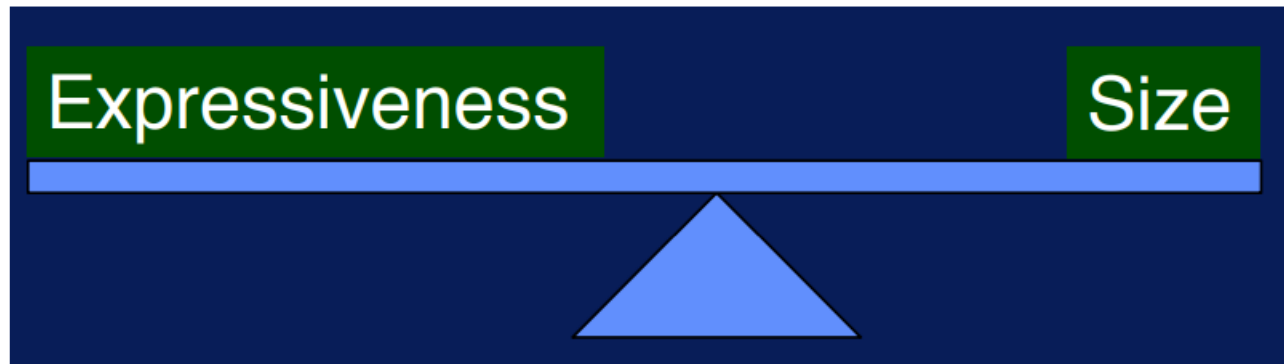


Name	Address
Sidney	7800 ★❖©◆ View Street
ZhČou	4345 Apple Lane

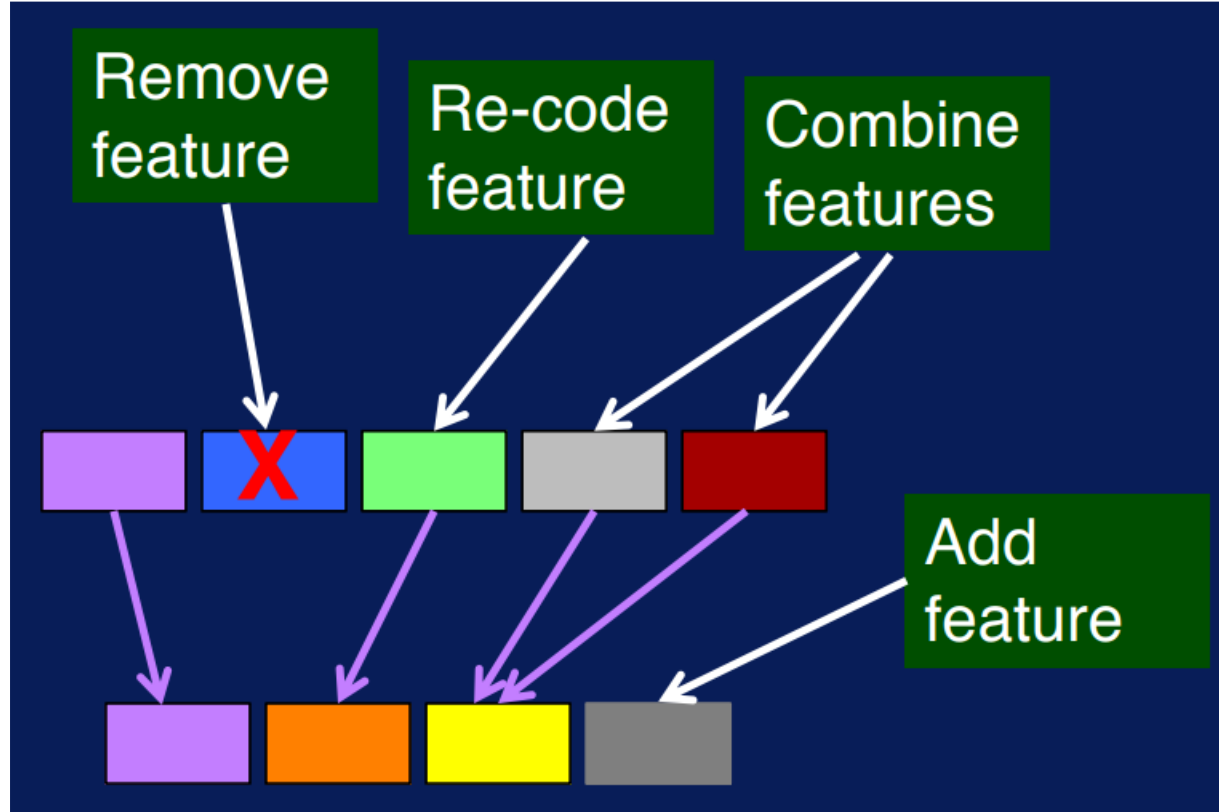
Feature Selection

What is Feature Selection?

- ▶ Characterize problem with smallest set of features




Feature Selection Methods



Adding Features

- New features derived from existing features

Name	State
Angela	AK
Sidney	CA
Ratan	WA
Kiril	OR
Zhou	CA



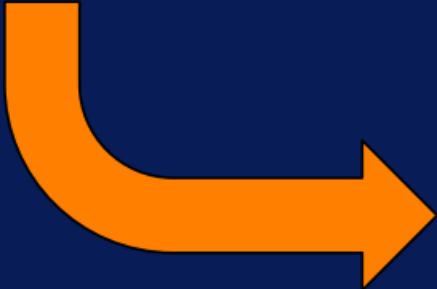
Name	State	<i>In-State</i>
Angela	AK	<i>F</i>
Sidney	CA	<i>T</i>
Ratan	WA	<i>F</i>
Kiril	OR	<i>F</i>
Zhou	CA	<i>T</i>

Removing Features

- Features that are very correlated
- Features with a lot of missing values
- Irrelevant features: ID, row number, etc.

Combining Features

Name	Height	Weight
Angela	1.8	68
Sidney	1.5	70
Ratan	2.0	84
Kiril	1.3	54
Zhou	2.0	61



Name	Height	Weight	<i>BMI</i>
Angela	180	68	21
Sidney	153	70	30
Ratan	204	84	20
Kiril	133	44	25
Zhou	208	81	19

Recoding Features

Examples

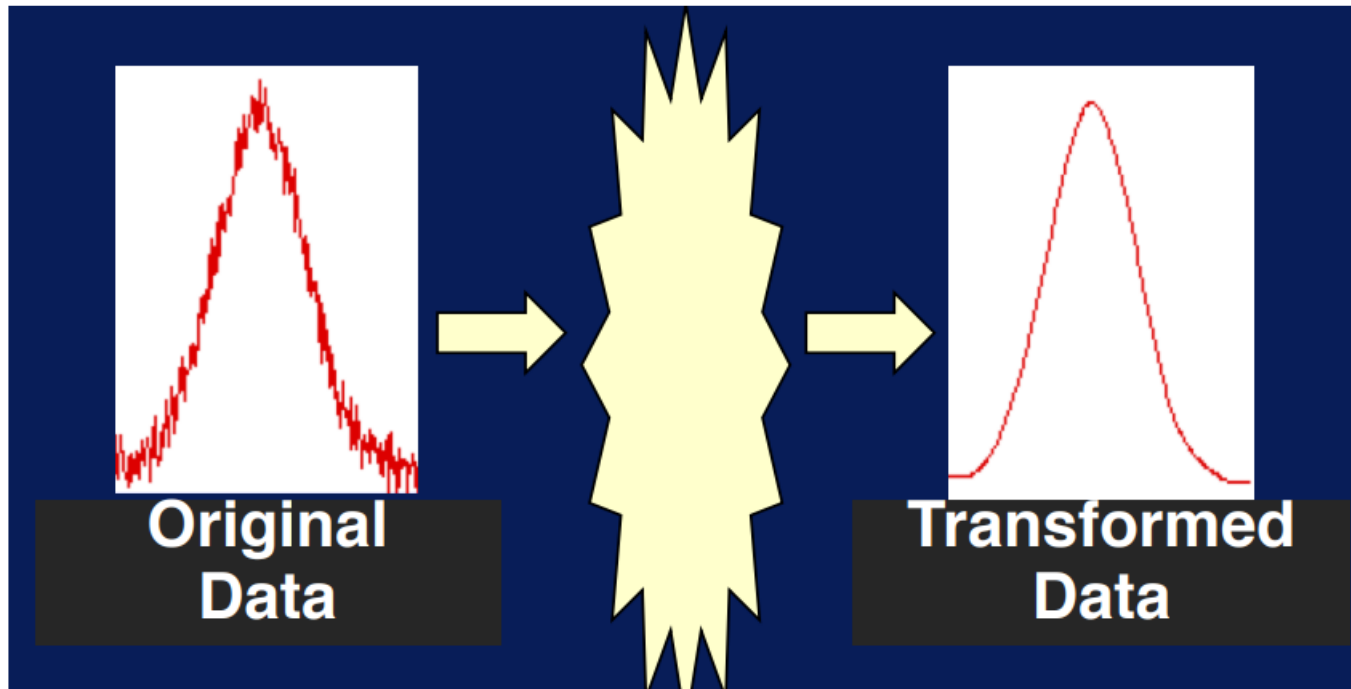
- ▶ Discretization: re-format continuous feature as discrete
- ▶ Customer's age => {teenager, young adult, adult, senior}

Breaking Up Features

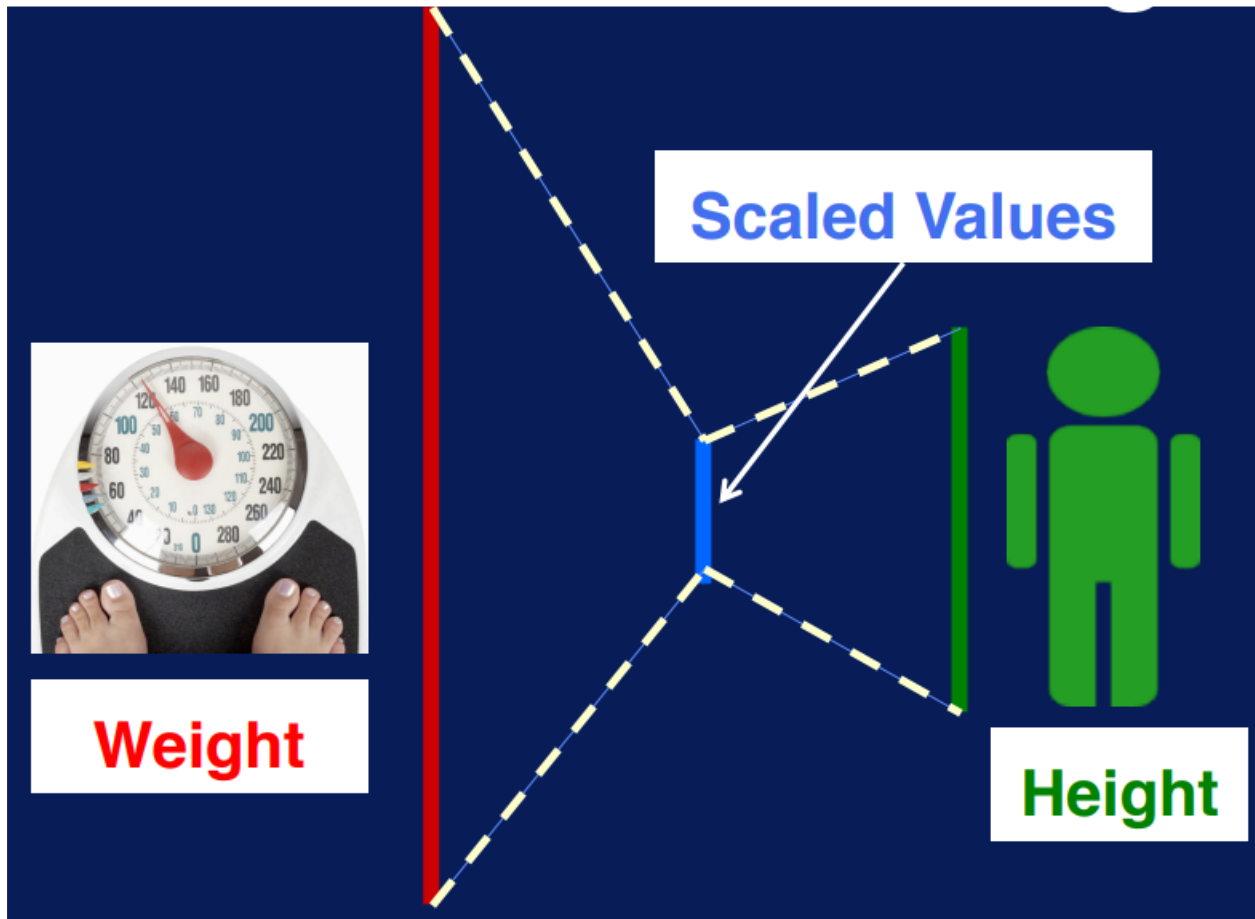
Address			
430 Park Drive, CA, 97283			
7800 W. View Street, FL, 34642			
1243 Mountain Ave., CO, 80334			
1220 Mill Avenue, IL 54622			
4345 Apple Lane, WA			
Address	State	Zip	
430 Park Drive	CA	97283	
7800 W. View Street	FL	34642	
1243 Mountain Ave.	CO	80334	
1220 Mill Avenue	IL	54622	
4345 Apple Lane	WA	98421	

Feature Transformation

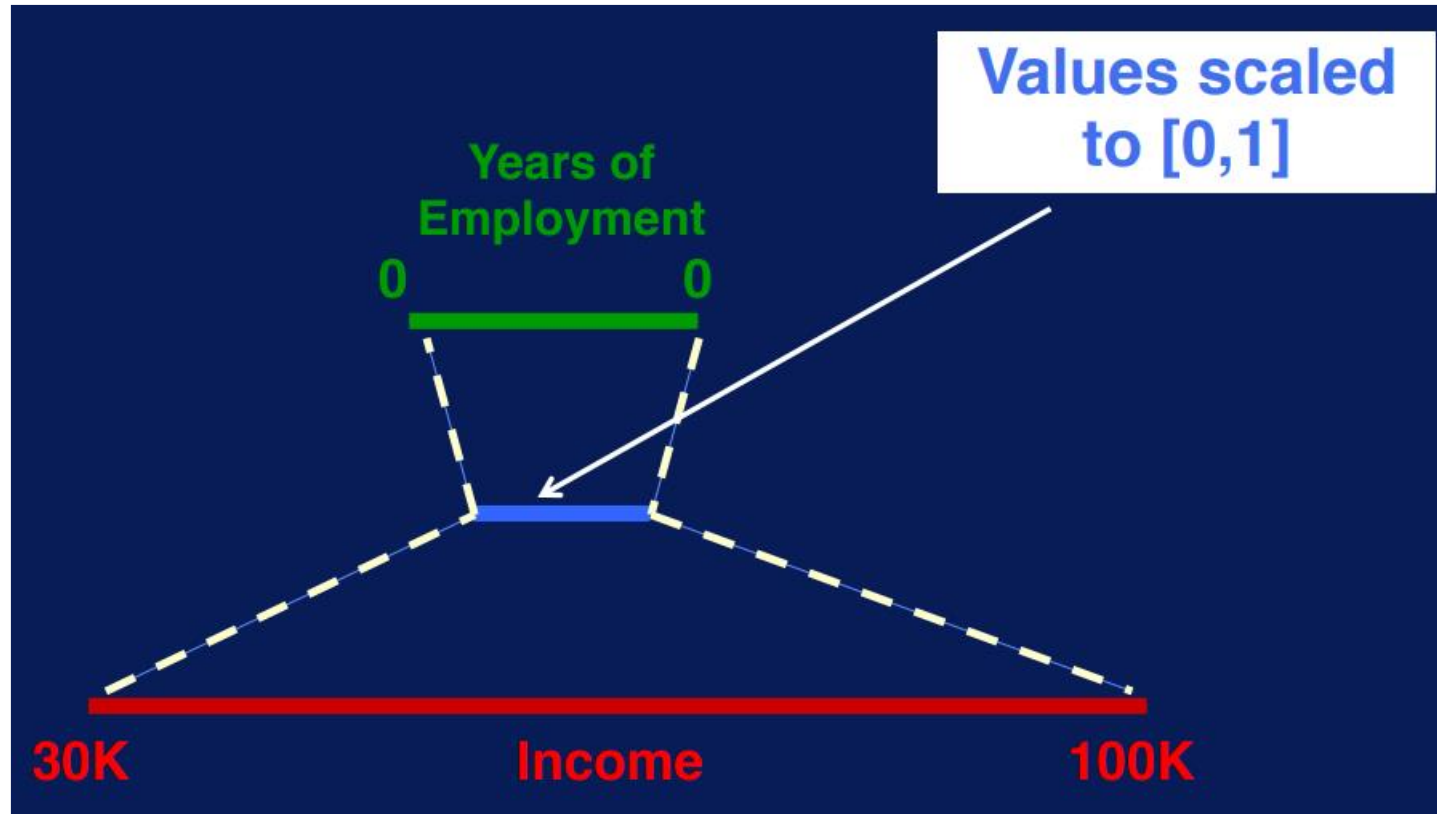
- Feature Transformation



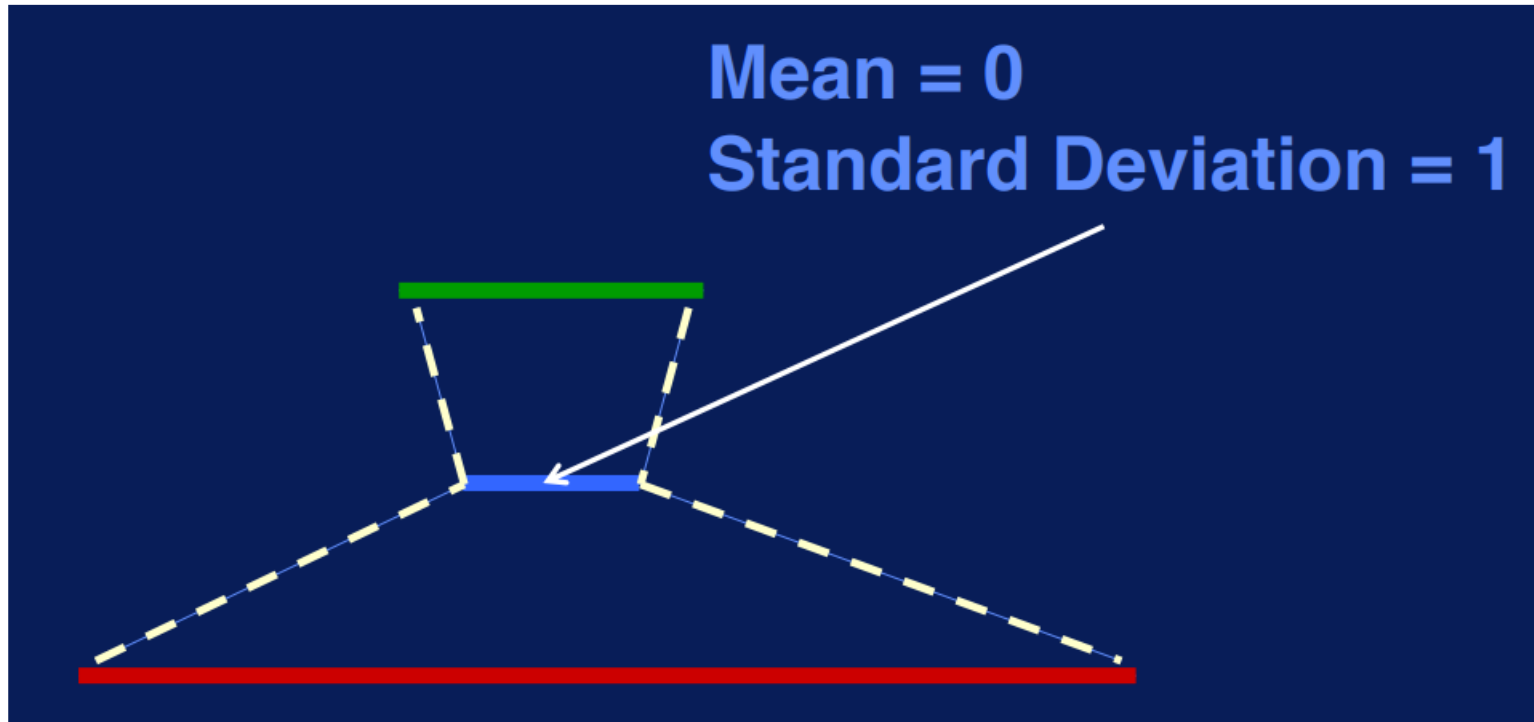
Scaling



Scaling to a Range



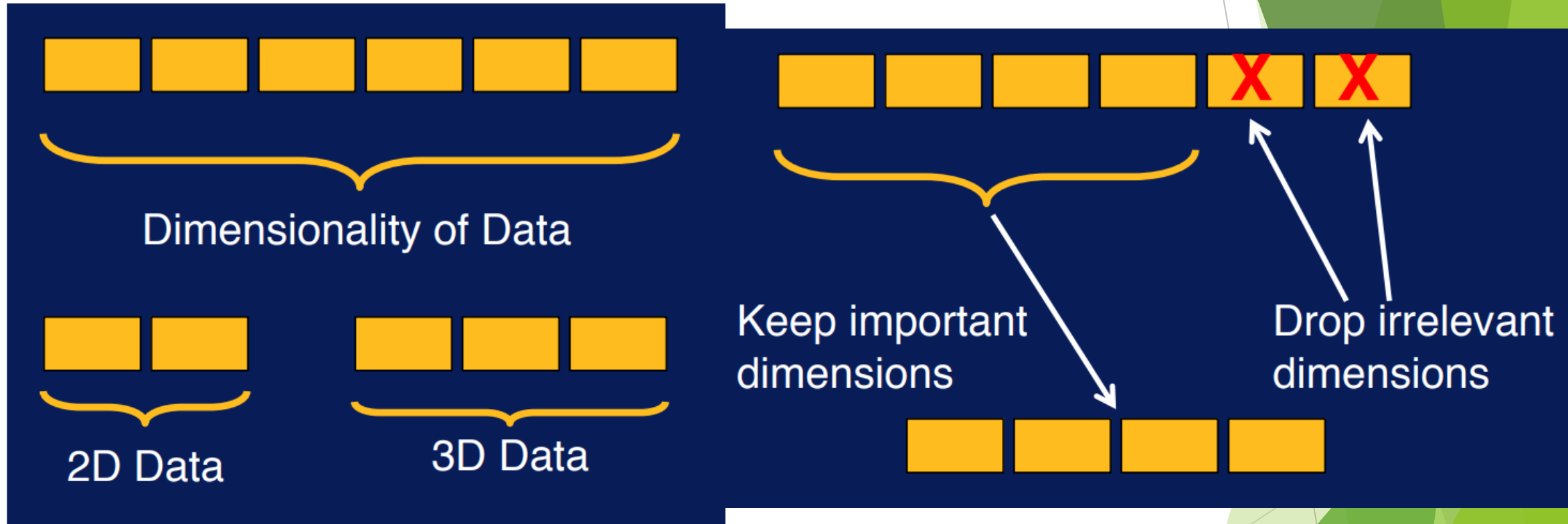
Zero-Normalization / Standardization



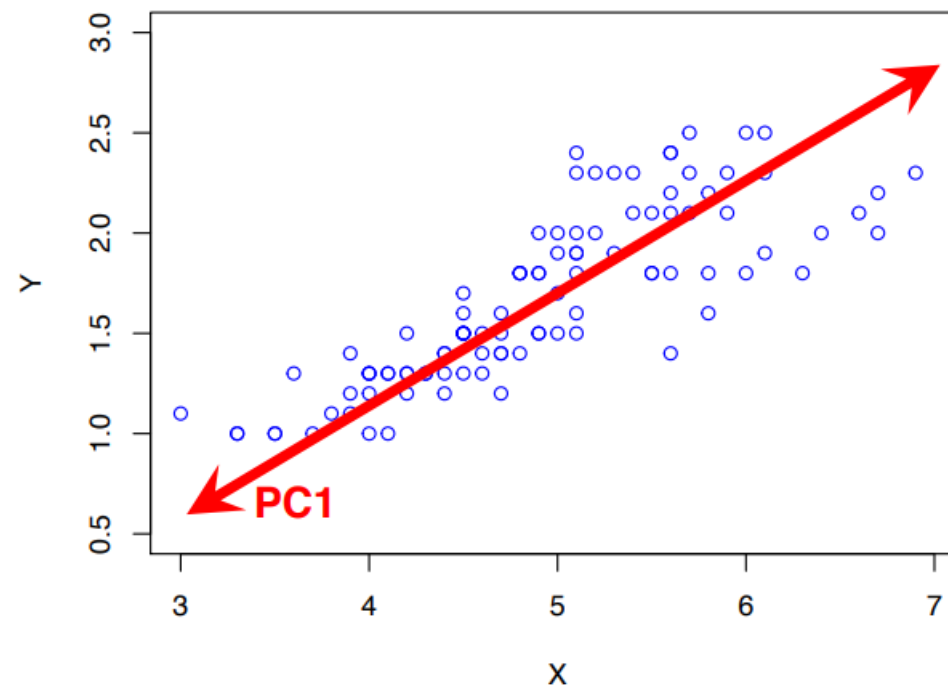
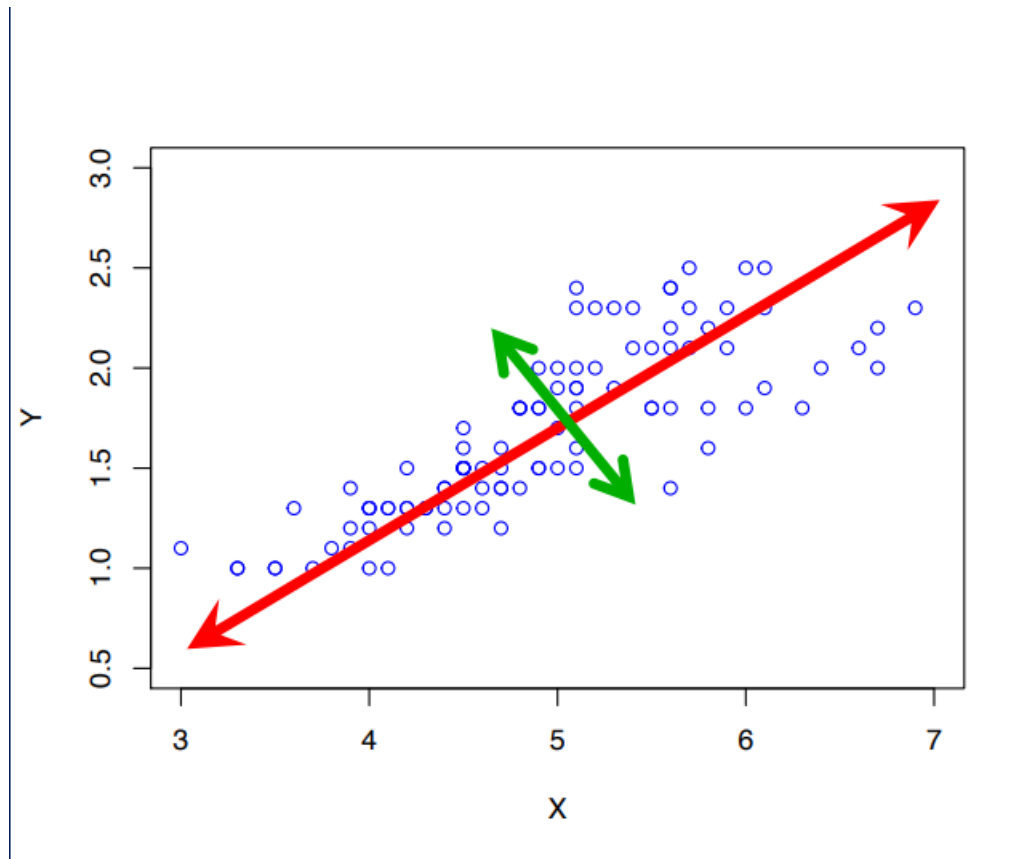
Feature Transformation

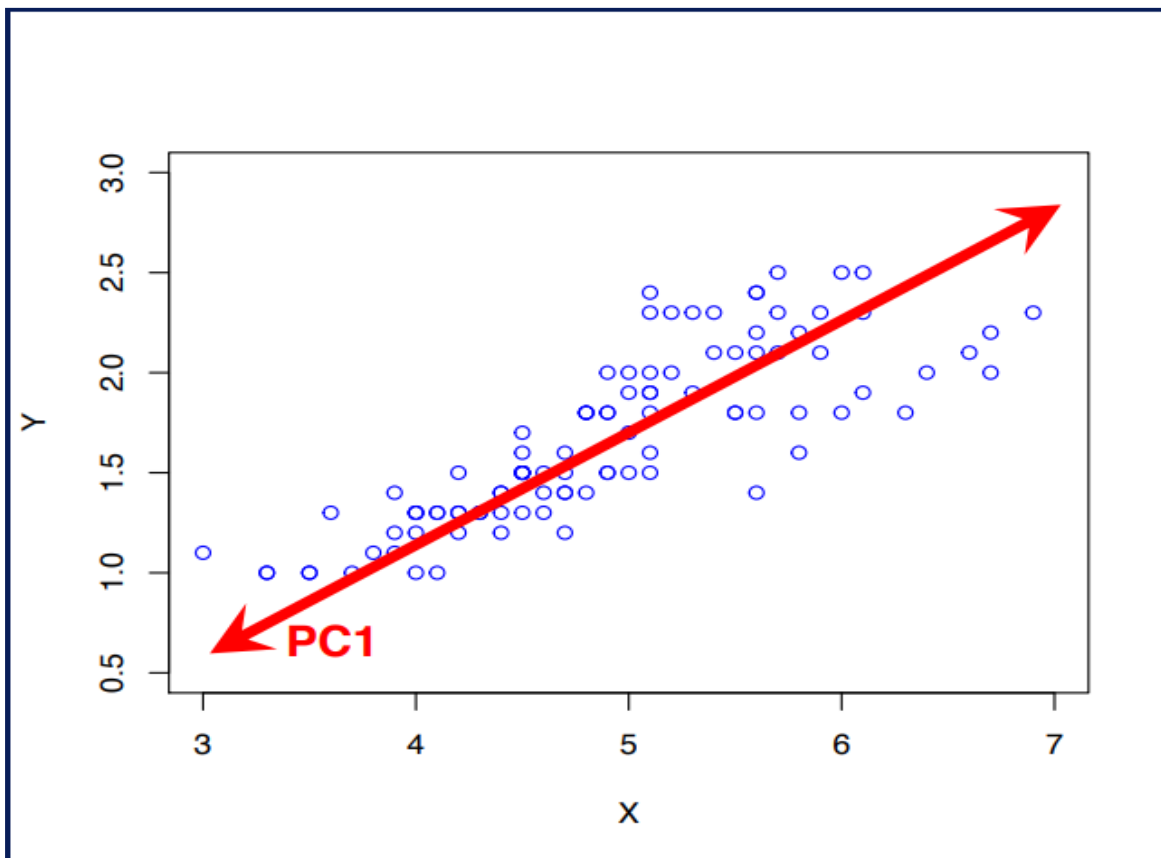
- ▶ What: Map feature values to new set of values
- ▶ Why: Have data in format suitable for analysis
- ▶ Caveat: Take care not to filter out important characteristics of data

Dimensionality Reduction



Principal Component Analysis





PCA Main Points

- ▶ Finds a new coordinate system such that
- ▶ PC1 captures greatest variance
- ▶ PC2 captures second greatest variance, etc.
- ▶ First few PCs capture most of variance
- ▶ Define lower-dimensional space for data.

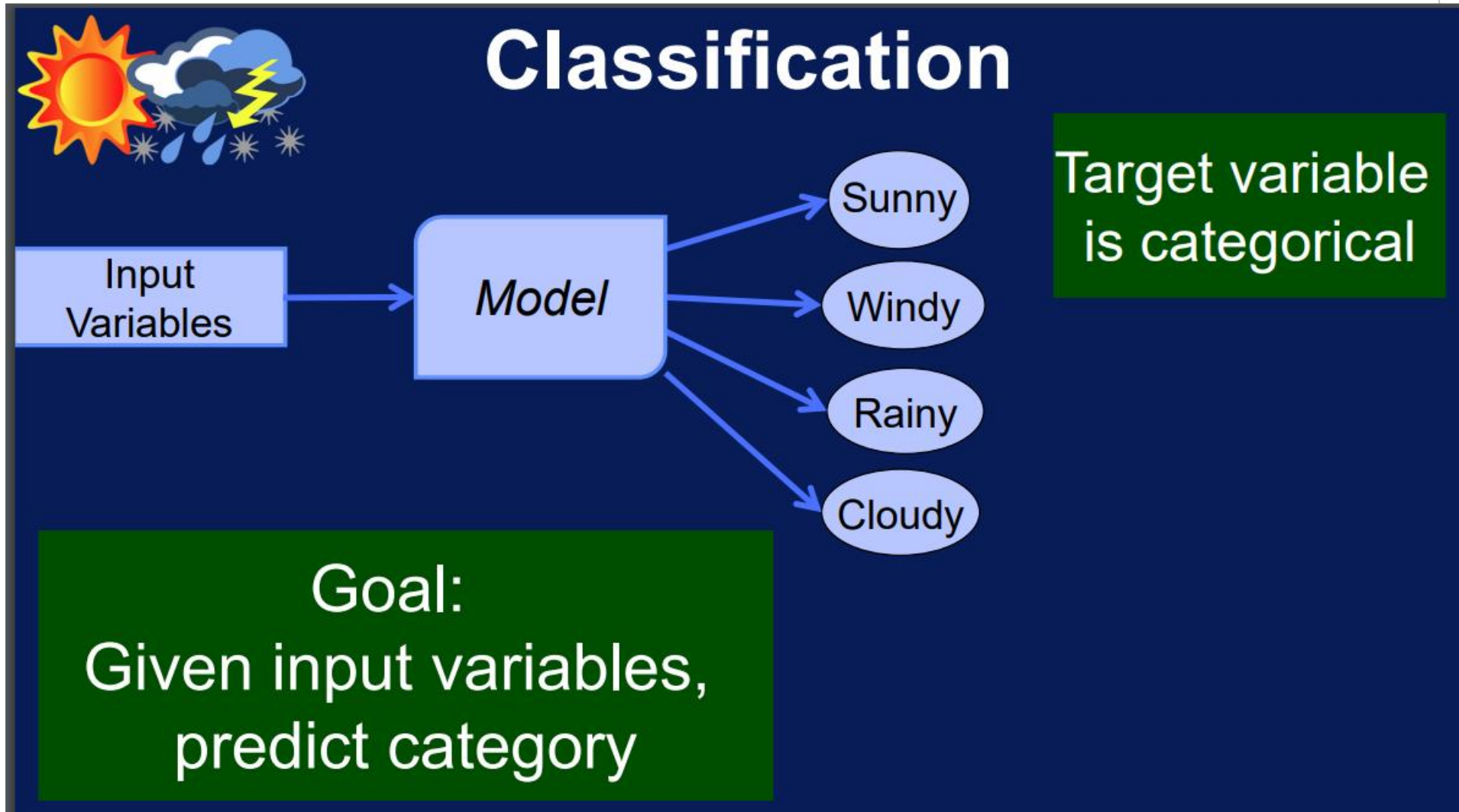
Example

- ▶ Original dimensions

Income, age, occupation, etc.

- ▶ New dimensions PC1, PC2, PC3, etc.
- ▶ More difficult to interpret!

Classification Overview

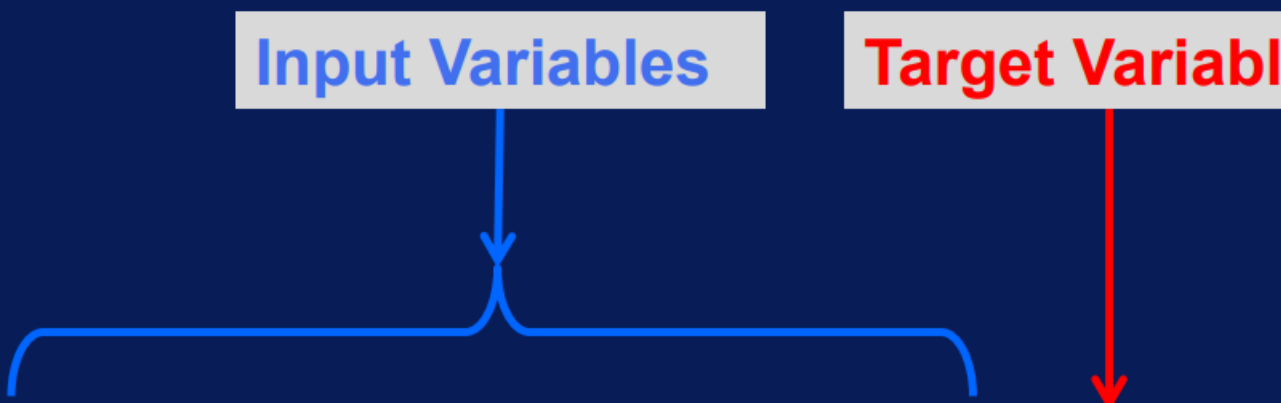


Data for Classification

Data for Classification

Input Variables

Target Variable



The diagram illustrates the relationship between input and target variables. A blue bracket groups the first three columns of the table (Temperature, Humidity, Wind Speed) under the 'Input Variables' label. A red arrow points from the 'Target Variable' label to the 'Weather' column.

Temperature	Humidity	Wind Speed	Weather
79	48	2.7	Sunny
60	80	3.8	Rainy
68	45	17.9	Windy
57	77	4.2	Cloudy

Classification is Supervised

Target


Label

Output

Class Variable

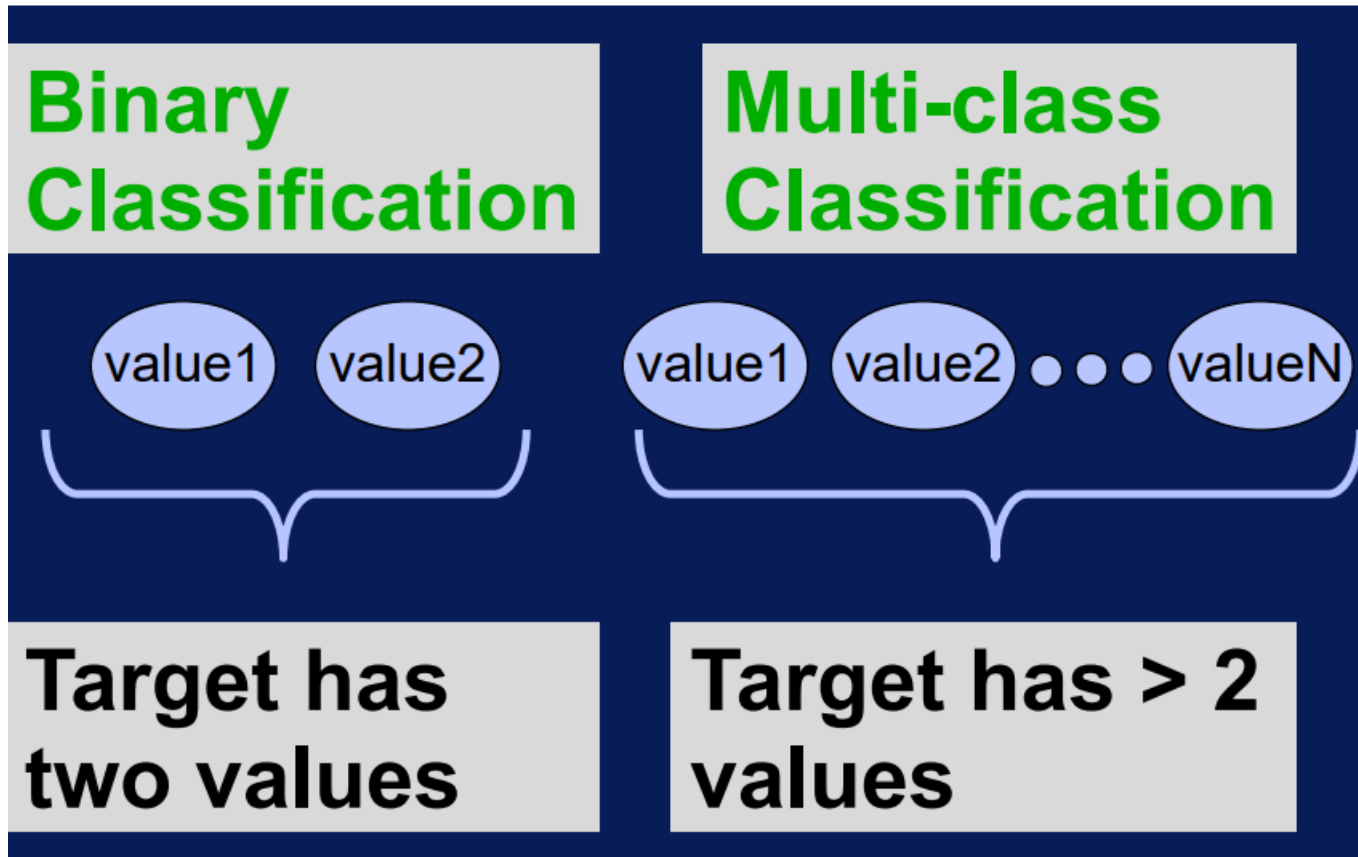
Class

Category



Temperature	Humidity	Wind Speed	Weather
79	48	2.7	Sunny
60	80	3.8	Rainy
68	45	17.9	Windy
57	77	4.2	Cloudy

Types of Classification



Classification Examples

Binary Classification

- Will it rain tomorrow or not?
- Is this transaction legitimate or fraudulent

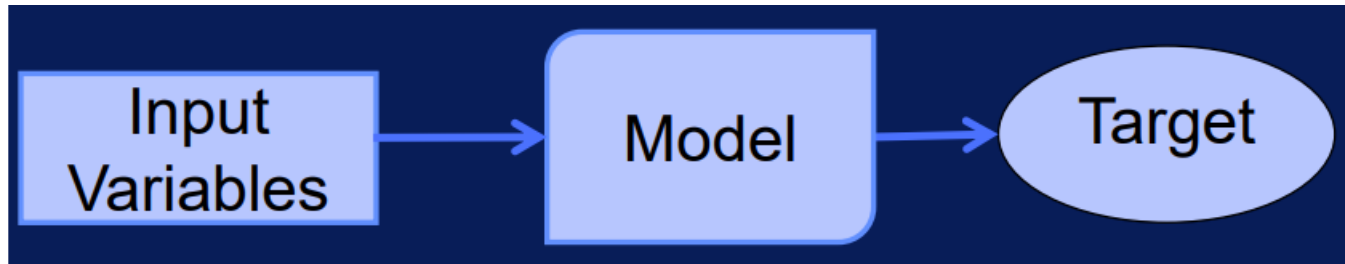
Multi-Class Classification

- What type of product will this customer buy?
- Is this tweet positive, negative, or neutral



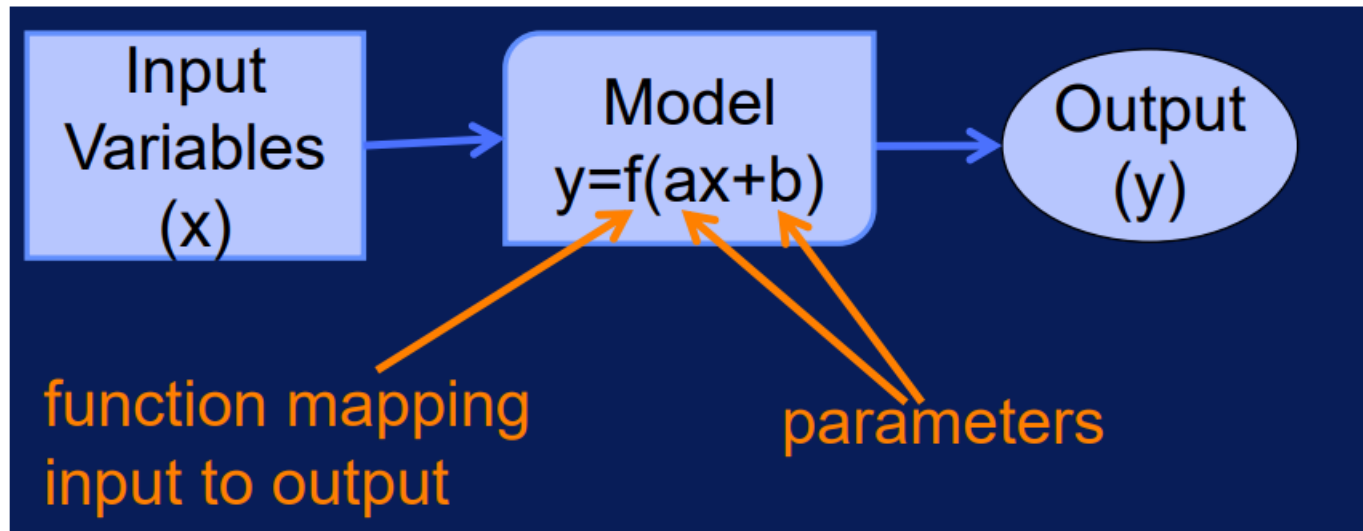
Classification Main Points

- Predict category from input variables
- Classification is a supervised task
- Target variable is categorical

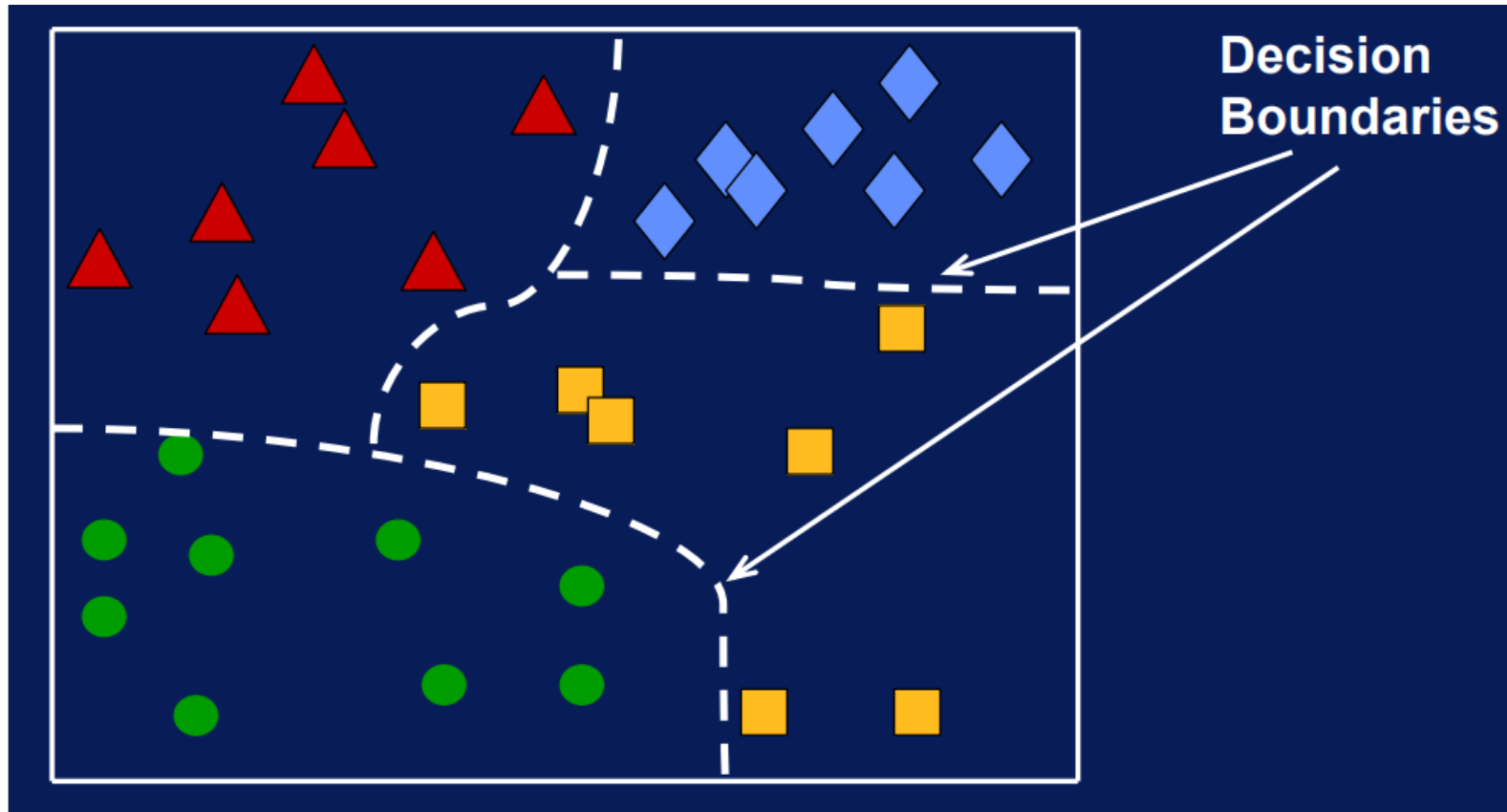


What is a Machine Learning Model?

- A mathematical model with parameters that map input to output



Building Classification Model



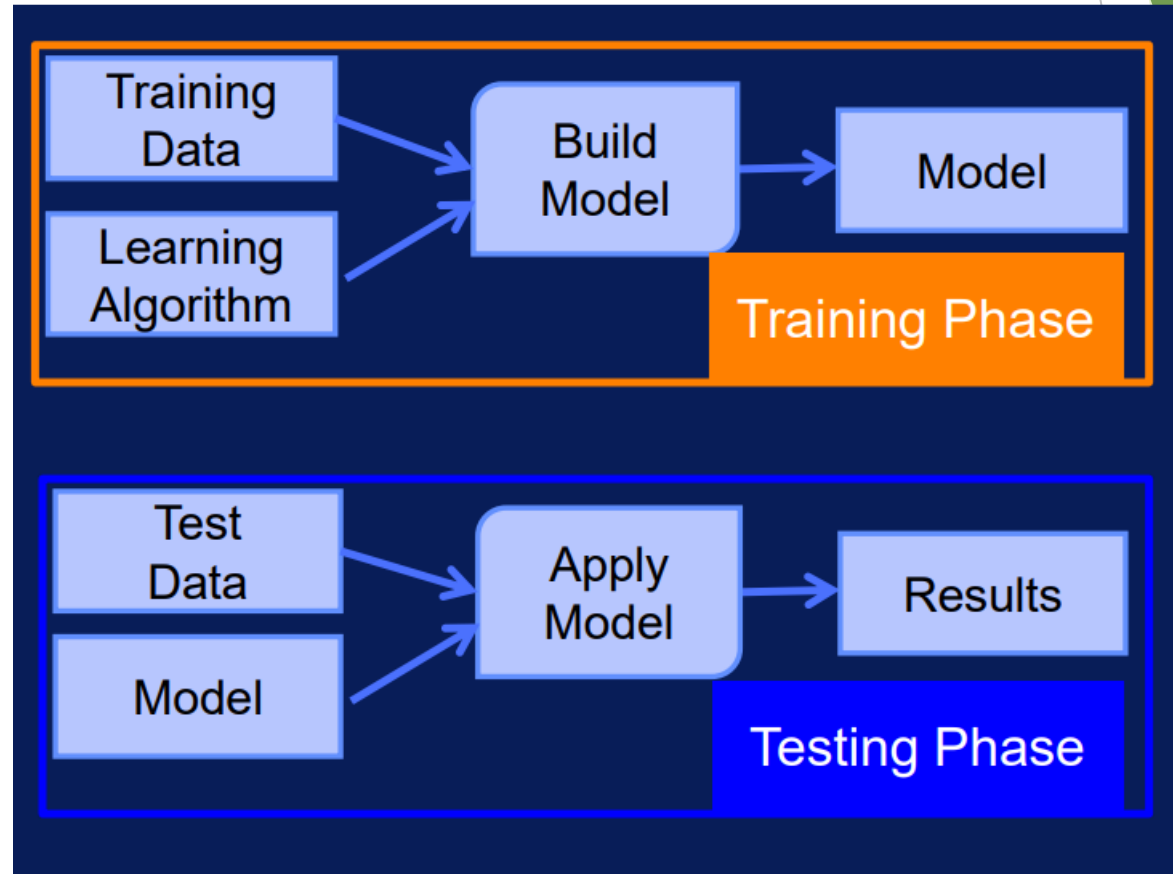
Building vs. Applying Model

► Training Phase

- Adjust model parameters
- Use training data

► Testing Phase

- Apply learned model
- Use new data

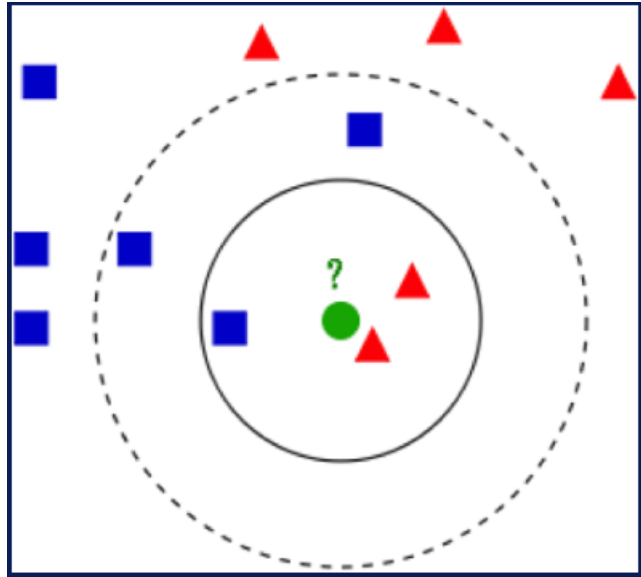


Classification Algorithms

- ▶ Common basic classification algorithms
- ▶ kNN (K nearest neighbors)
- ▶ Decision tree
- ▶ Naïve bayes

K-NN (K Nearest Neighbor)

- ▶ Simple classification technique
- ▶ Label sample based on its neighbors



kNN Assumption

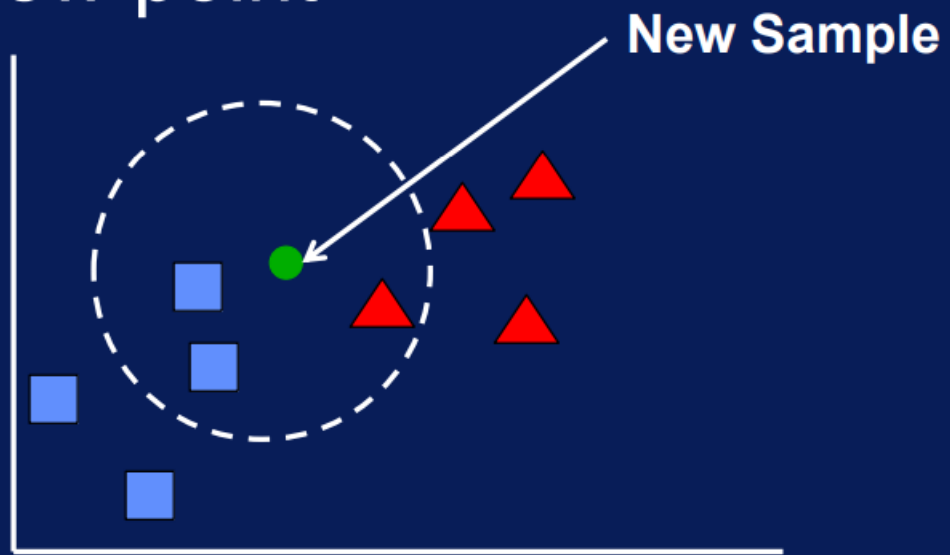
- Duck test

Quack



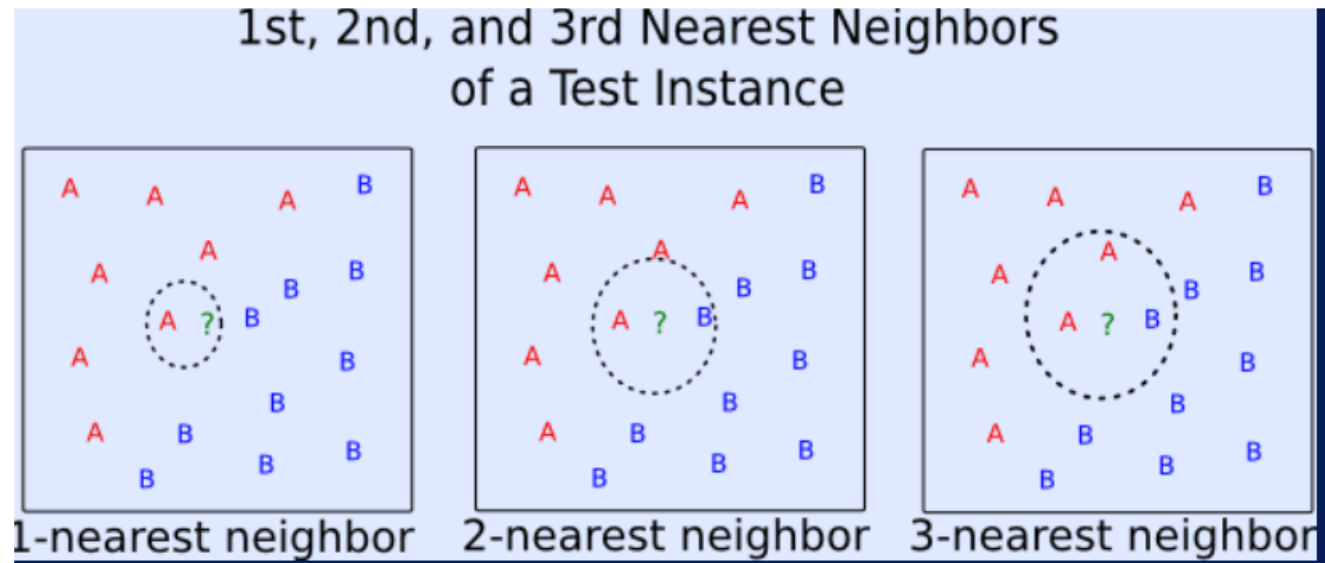
How kNN Works

- Use labels of neighboring samples to determine label for new point

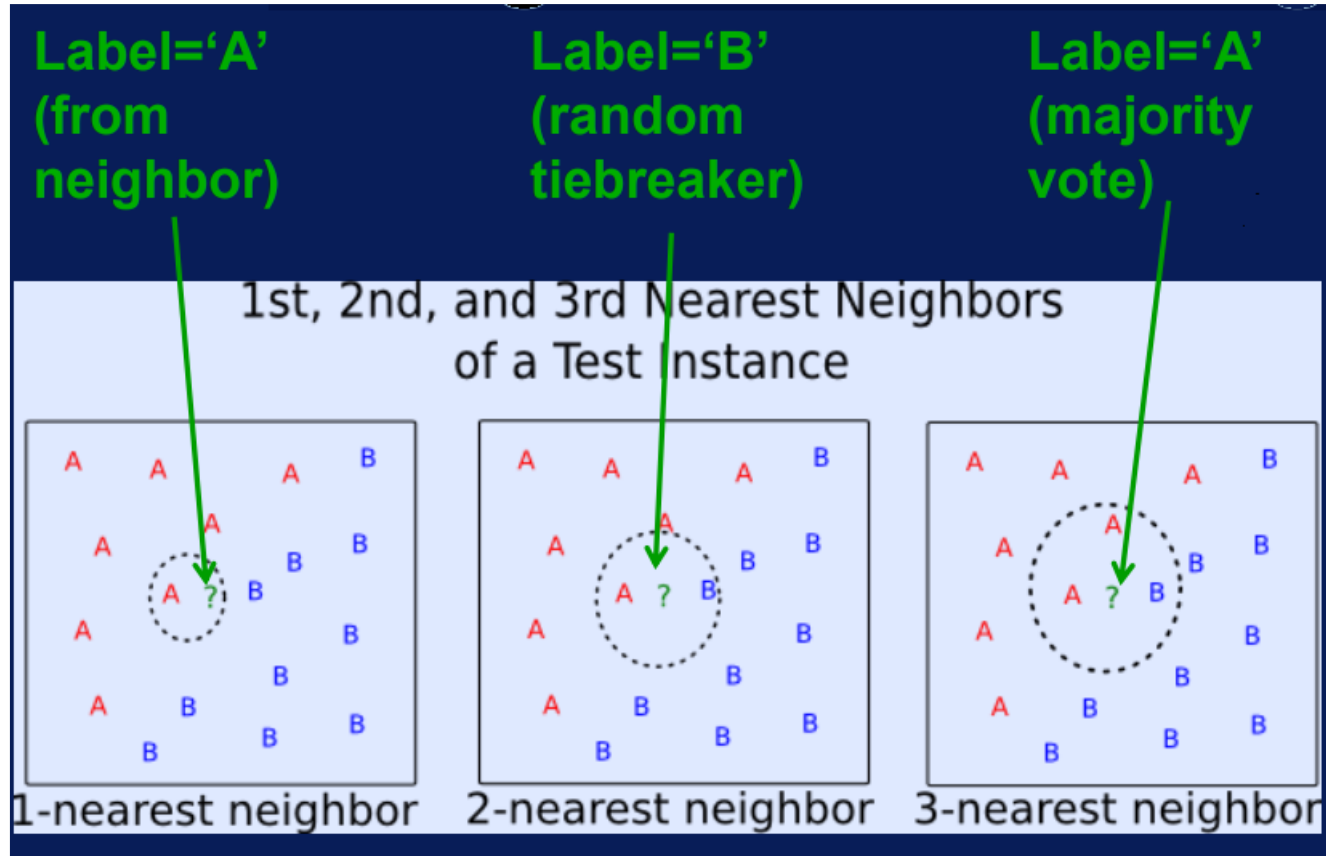


What is k?

- Value of k determines number of closest neighbors to consider

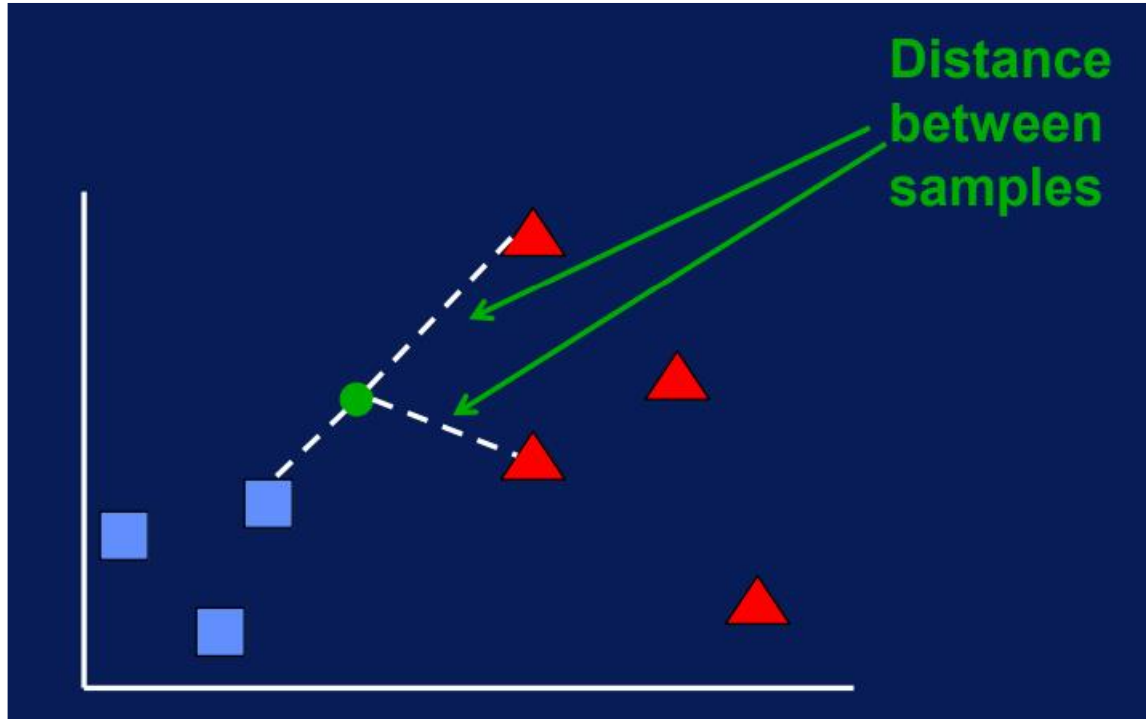


Using k Nearest Neighbors



Distance Measure

Need measure to determine “closeness”



kNN Classification

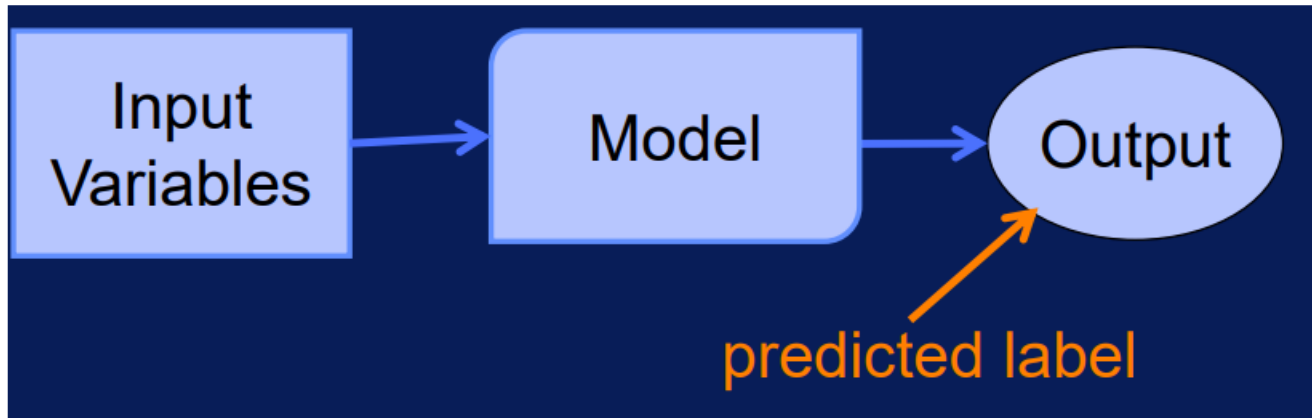
- ▶ No separate training phase
- ▶ Can generate complex decision boundaries
- ▶ Can be slow
- ▶ Distance between new sample and all samples must be computed to classify new sample

Generalization & Overfitting

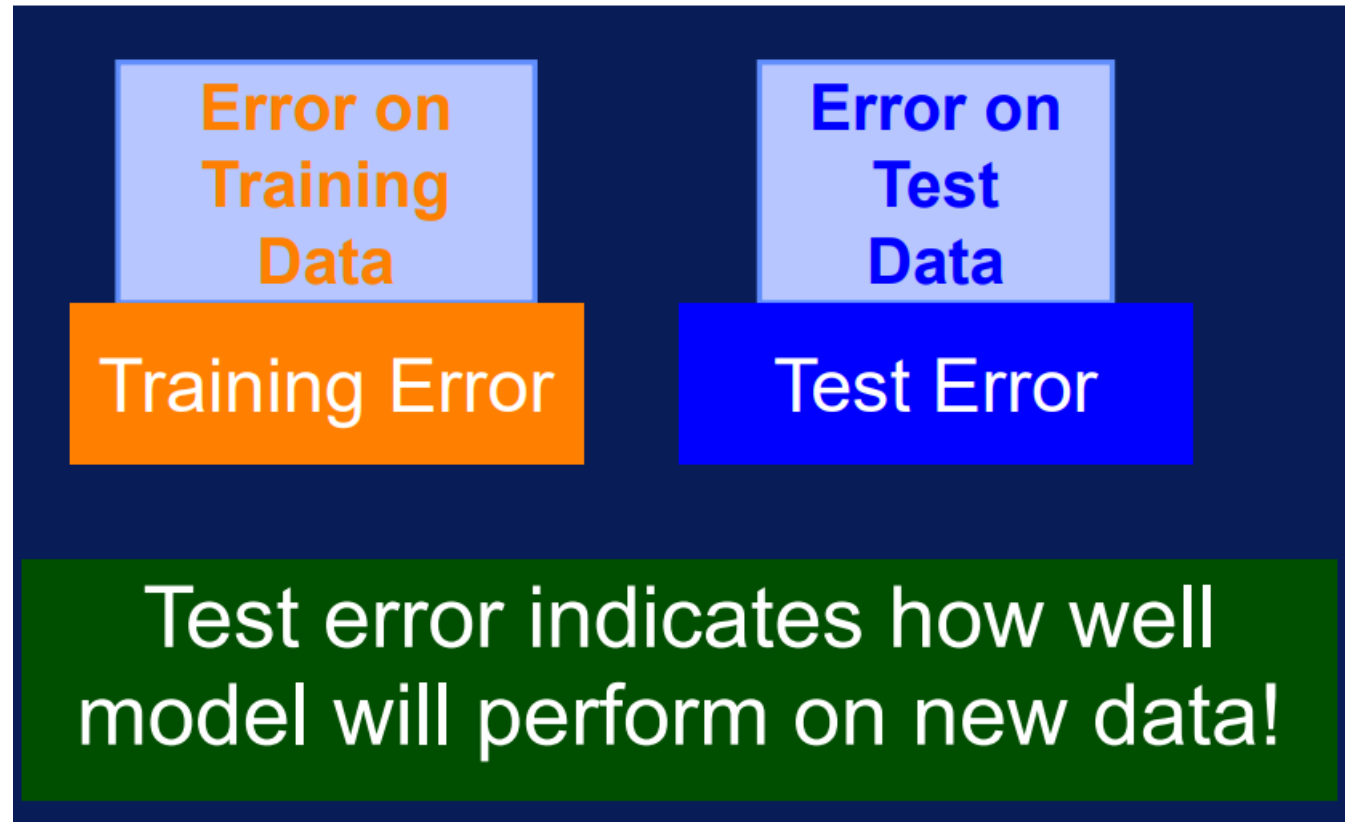
Errors in Classification

Success: Output = Target

- Error: Output \neq Target
- Error rate = Error = Misclassification Error
- $\# \text{ errors} / \# \text{ samples} = \% \text{ error}$



Errors in Classification



Generalization

*Performs well
on new data*

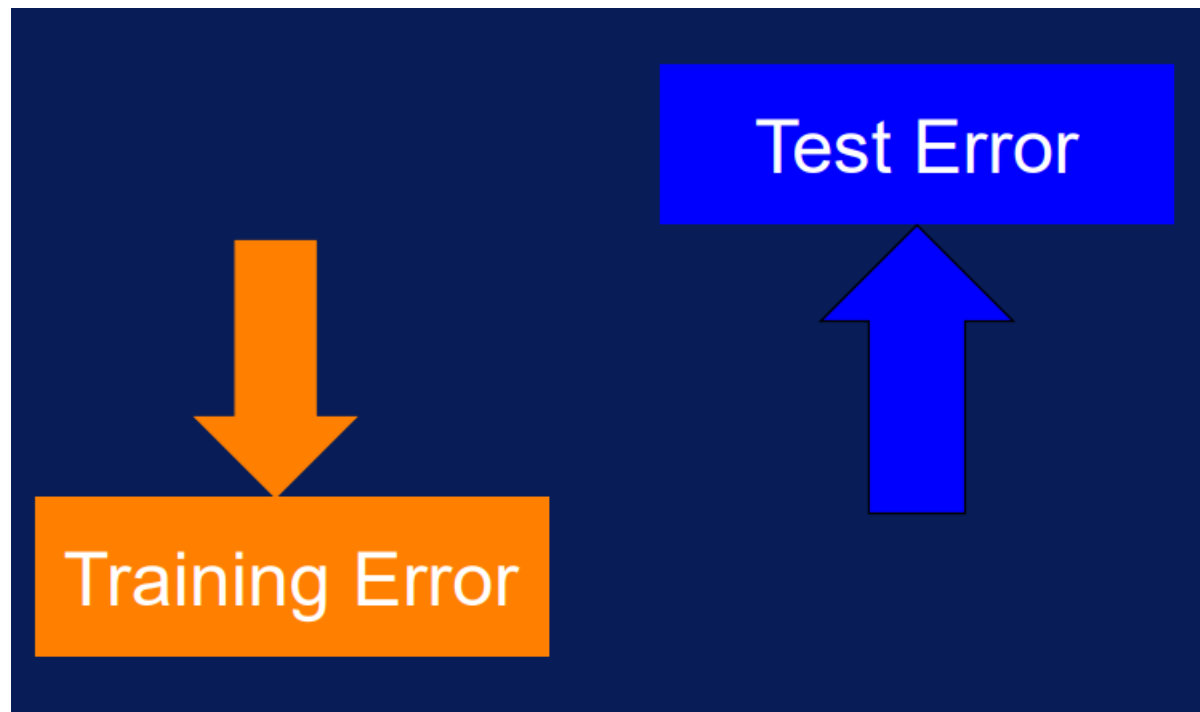


*Good
Generalization*

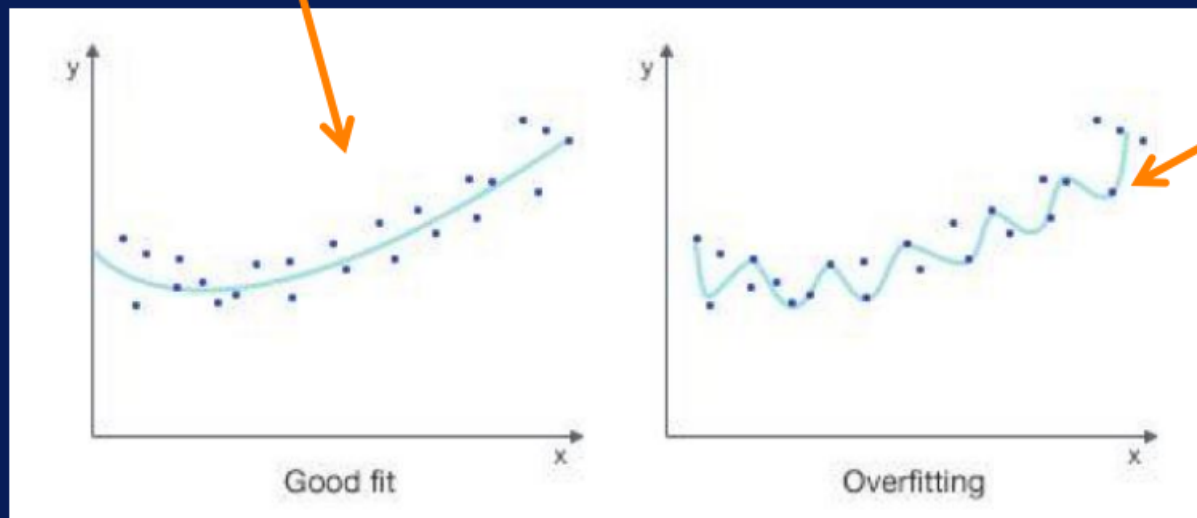


Test Error = Generalization Error

Overfitting



Model is fitting to
structure of data

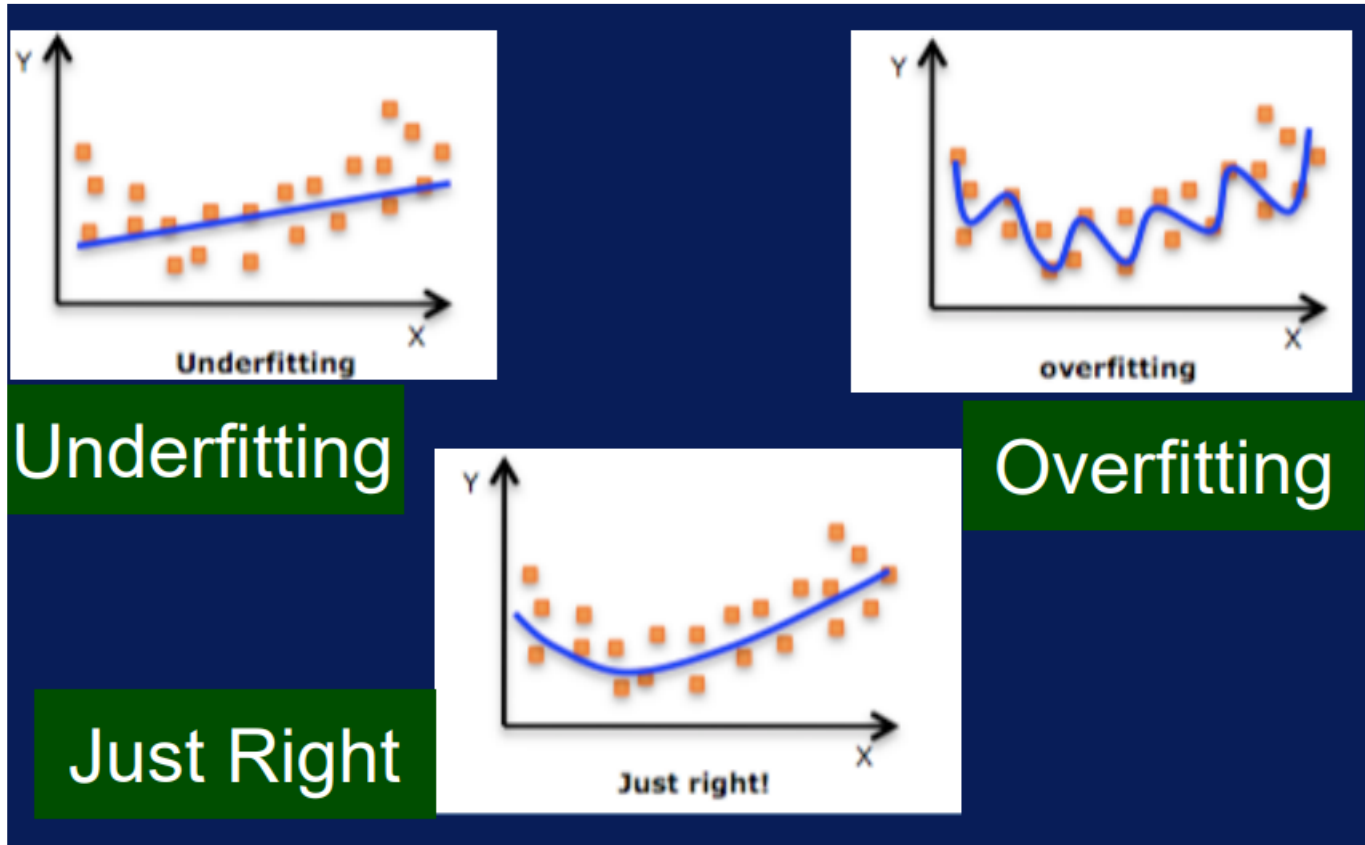


Model is fitting
to noise in data

Overfitting & Generalization

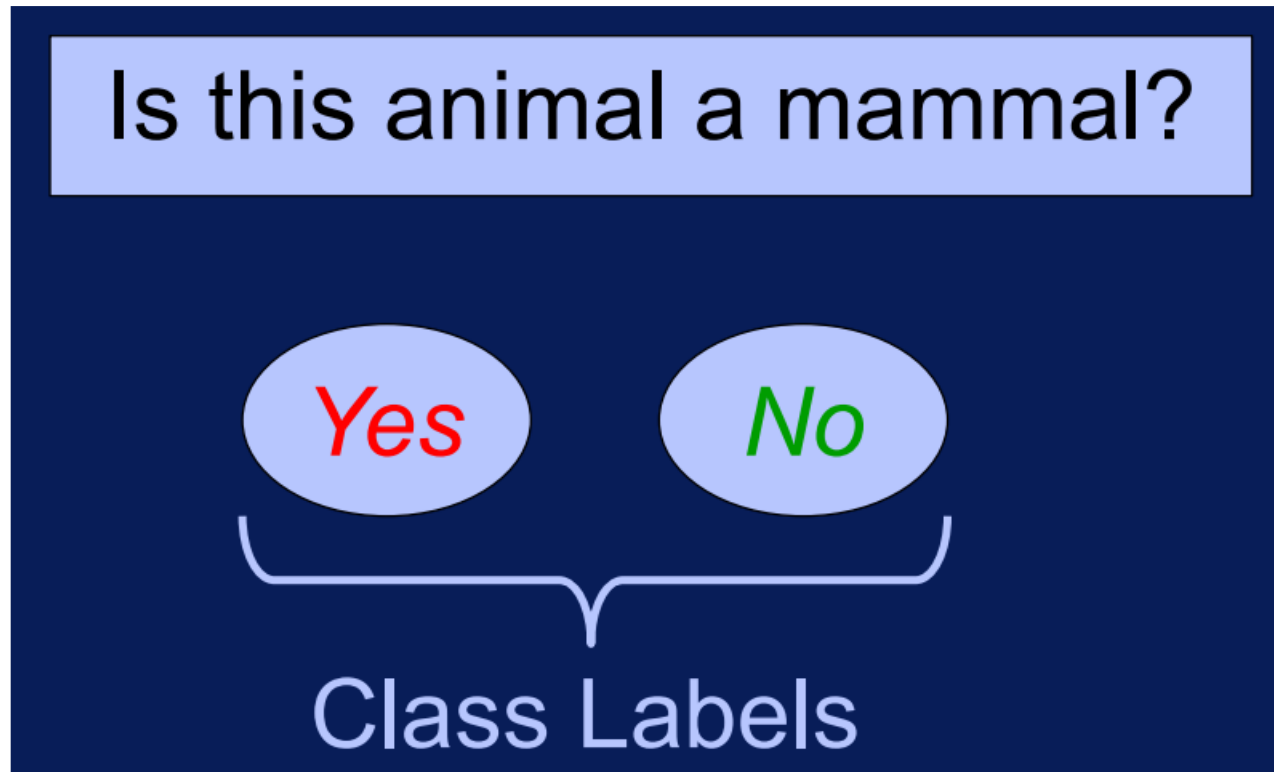
Overfitting → Poor
Generalization

Overfitting & Underfitting



Metrics to Evaluate Model Performance

► Classification



Types of Classification Errors

True Label	Predicted Label	Error Type
Yes	Yes	True Positive (TP)
No	No	True Negative (TN)
No	Yes	False Positive (FP)
Yes	No	False Negative (FN)

Accuracy Rate

True	Predicted	Error
Yes	Yes	True Positive (TP)
No	No	True Negative (TN)
No	Yes	False Positive (FP)
Yes	No	False Negative (FN)

$$\begin{aligned}\text{Accuracy Rate} &= \frac{\# \text{ correct predictions}}{\# \text{ total predictions}} \\ &= \frac{TP + TN}{TP + TN + FP + FN}\end{aligned}$$

The background features abstract, overlapping green geometric shapes, primarily triangles and polygons, in various shades of green, creating a modern and dynamic visual effect. The shapes are layered, with some appearing more prominent than others, and they extend towards the corners of the frame.

Thank You!