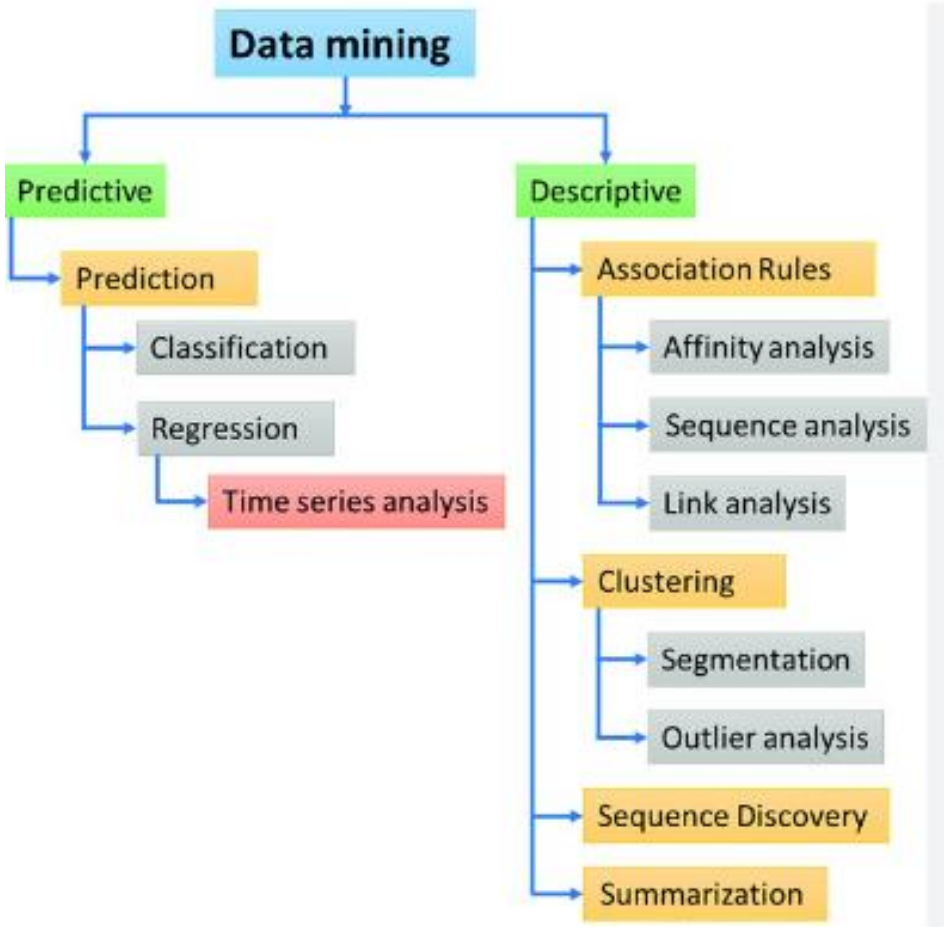


Data Mining Tasks

Data Mining Functionalities

Descriptive and Predictive Data Mining:

- Descriptive Mining is frequently used to provide Correlation, Cross-Tabulation, Frequency, and other types of information. It analyses stored data to determine what happened in the past.

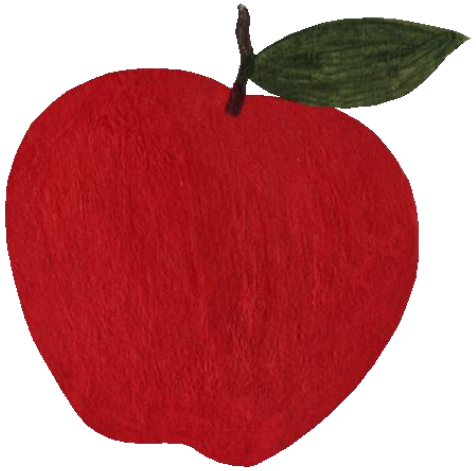
- Predictive Data Mining is the Analysis done to predict a future event or multiple data or trends. It explains what might happen in the future as a result of past Data Analysis.

- Descriptive Analytics is concerned with summarising and converting data into usable information for reporting and monitoring. Furthermore, it allows for a thorough examination of the data so that questions like “what happened?” and “what is happening?” can be easily answered.

- Predictive Data Mining is the Analysis done to predict a future event or other data or trends, as the term ‘Predictive’ means to predict something.
- Business Analysts can use Predictive Data Mining to make better decisions and add value to the analytics team’s efforts. Predictive Analytics is aided by Predictive Data Mining. Predictive Analytics, as we all know, is the use of data to predict outcomes.

1. Classification Machine Learning

What is this?



Apple

What is this?

Orange

What do we understand from this example?

Child will adopt and react to the objects given.

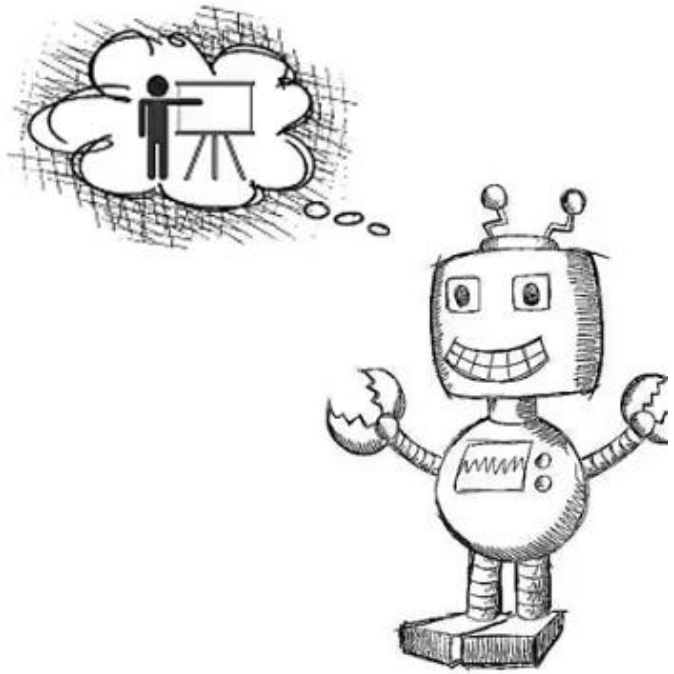
1. Classification

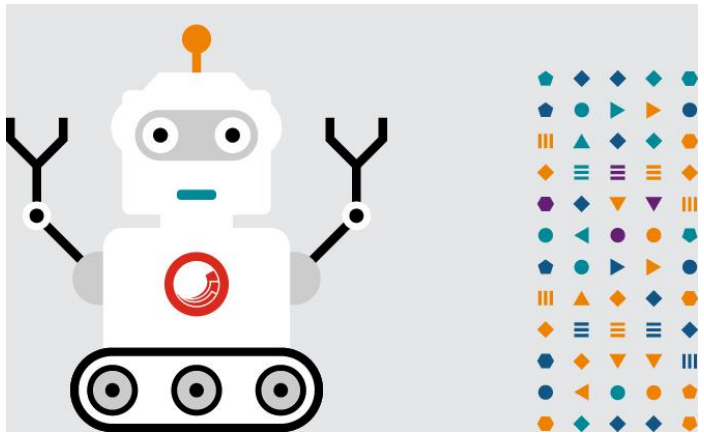
- It is often referred to as supervised learning.
- It maps data into predefined groups or classes.
- It uses if-then rule
- Predict class label

How human learning the environment/things?

How human learning?

HUMAN LEARN FROM PAST EXPERIENCES





How machine learning the environment

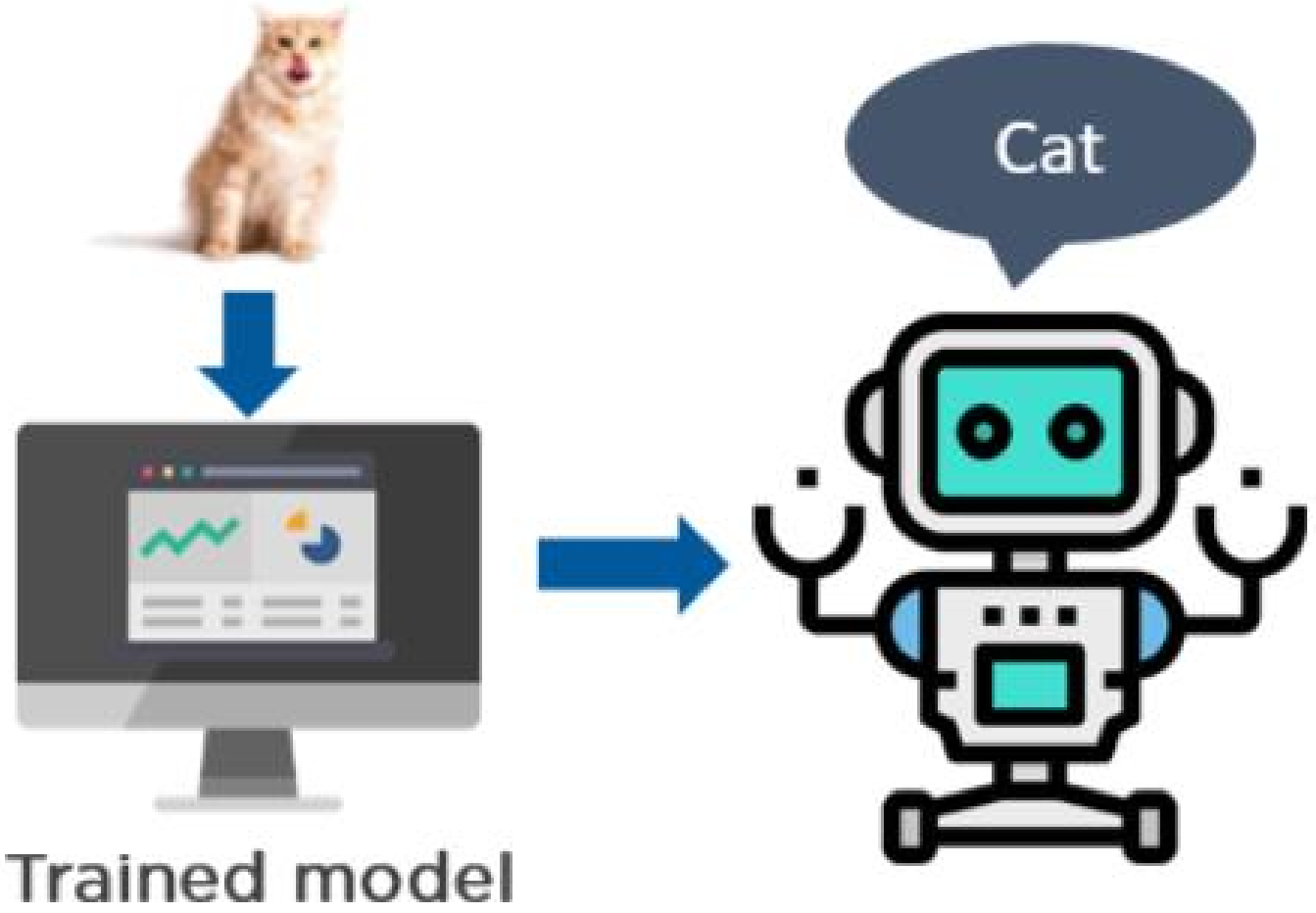
MACHINE WILL FOLLOW THE INSTRUCTIONS GIVEN BY PROGRAMMER

What if human can train the machines for past data!!!!

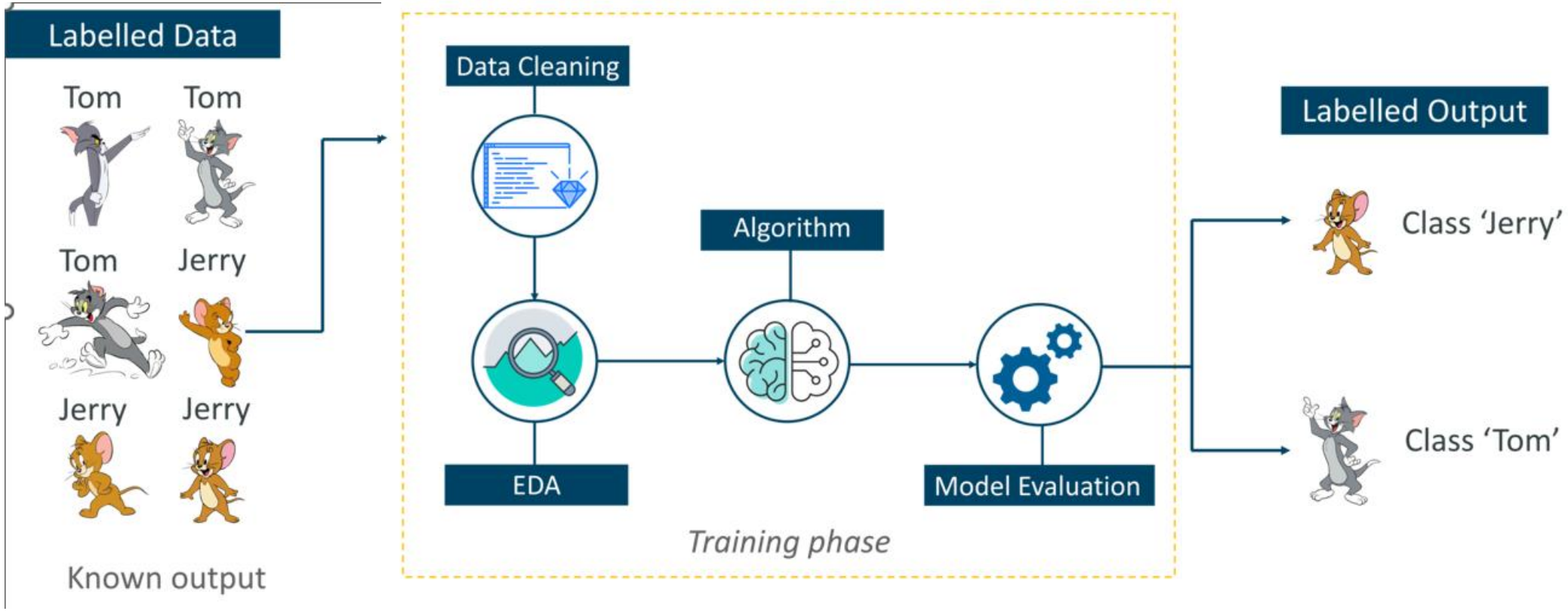
MACHINE WILL DO IT MUCH FASTER….ML…..UNDERSTANDING AND REASONING

Types of classification

- 1. Binary classification
- 2. Multi-class classification -N+1 class i.e more than 2


Cat
Trained model

Supervised learning



Diameter	Weight	Red	Green	Blue	Name
2.96	86.76	172	85	2	Orange
3.91	88.05	166	78	3	Orange
5.43	108.54	157	98	2	Apple
5.51	109.49	150	98	5	Grape
11.06	191.08	151	57	6	Apple
11.06	191.08	151	57	6	Grape
11.06	191.08	151	57	6	Grape
11.06	191.08	151	57	6	Orange
13.17	223.49	162	79	13	Grape
13.17	223.51	163	74	23	Apple
13.17	223.52	140	66	22	Orange
13.17	223.55	165	75	26	Orange
13.17	223.56	125	69	24	Apple

Applications

- **Face Recognition**
- **Spam Classification:**
- **Advertisement Popularity**

Example

- An airport security screening station is used to determine if passengers are customers or criminals or terrorists.

- To do this, the face of each passenger is scanned and its basic pattern(Distance between eyes, size and shape of mouth, shape of head etc) is identified.
- The pattern is compared to entries in a database to see the matches.

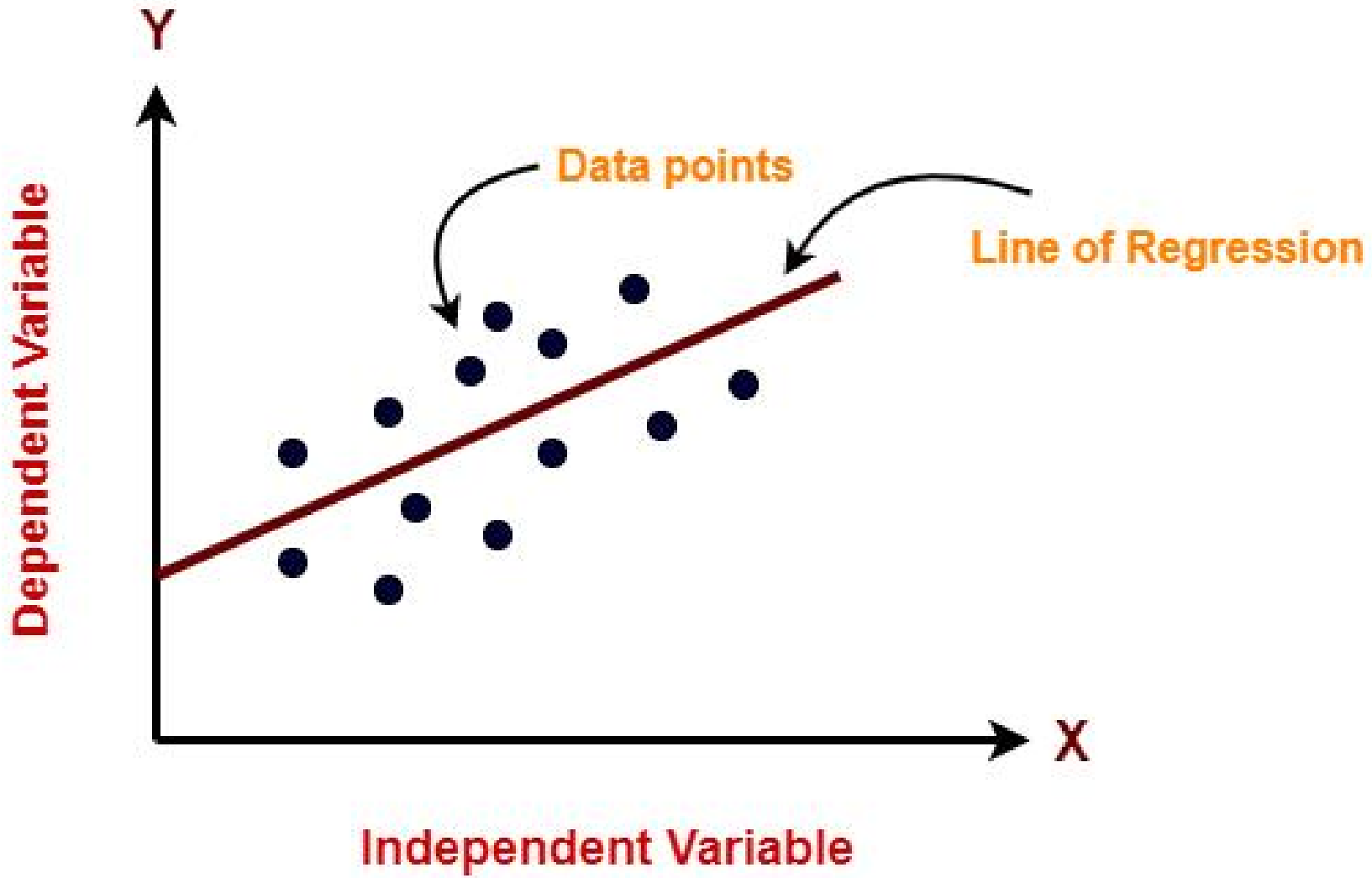
2. Regression

- Regression is used to map a data item to a real valued prediction variable.
- contains real numbers

- It assumes that the target data fit into some known type of function(linear, logistic, etc) and then determines the best function of this type that models the given data.
- Some error analysis is used ot determine which function is “best”

List of regression algorithms in Machine Learning

- Linear Regression
- Ridge Regression
- Neural Network Regression
- Lasso Regression
- Decision Tree Regression
- Random Forest
- KNN Model
- Support Vector Machines (SVM)
- Gausian Regression
- Polynomial Regression

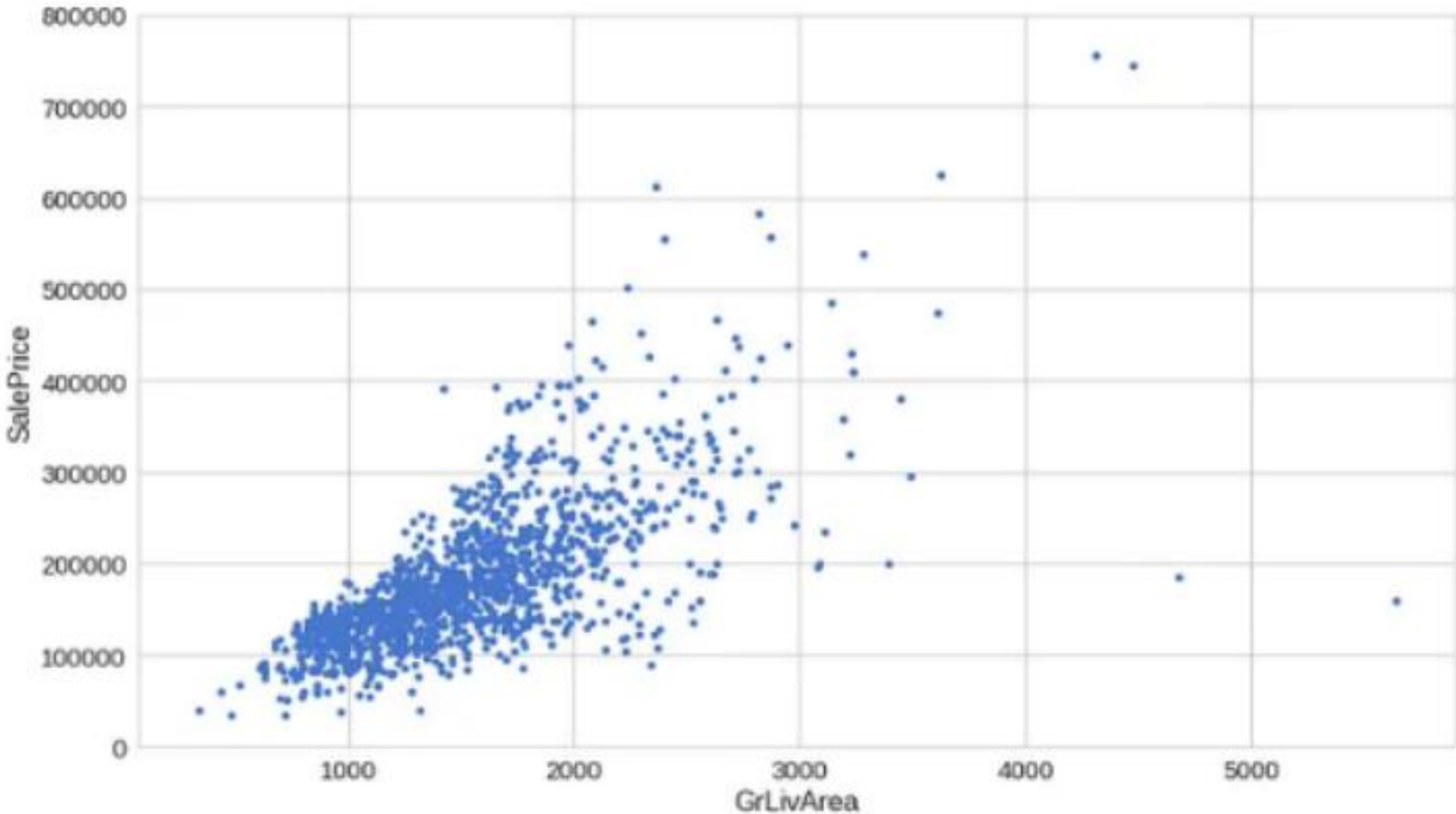

Y
Data points
Line of Regression
Dependent Variable
X
Independent Variable

Independent Vs Dependent variable

- Independent variables (IVs) are the ones that you include in the model to explain or predict changes in the dependent variable(continuous)
- Independent indicates that they stand alone and other variables in the model do not influence them.
- Independent variables are also known as predictors, factors, treatment variables, explanatory variables, input variables, x-variables, and right-hand variables

- The dependent variable (DV) is what you want to use the model to explain or predict.
- The values of this variable depend on other variables.

- It is the outcome that you're studying
- It's also known as the response variable, outcome variable, and left-hand variable. Statisticians commonly denote them using a **Y.**


800000
700000
600000
500000
400000
300000
200000
100000
0
SalePrice
1000
2000
3000
4000
5000
GrLivArea

bathrooms	floors	condition	city	price
1.5	3	3	Shoreline	313000
2.5	2	5	Seattle	2384000
2	1	4	Kent	342000
2.25	1	4	Bellevue	420000
2.5	1	4	Redmond	550000
1	1	3	Seattle	490000
2	1	3	Redmond	335000
2.5	2	3	Maple Valley	482000
2.5	1	4	North Bend	452500
2	3	3	Seattle	640000
1.75	1	3	Lake Forest Park	463000
2.5	3	5	Seattle	1400000
1.75	1	3	Sammamish	588500
1	1	4	Seattle	365000

Integer

Example

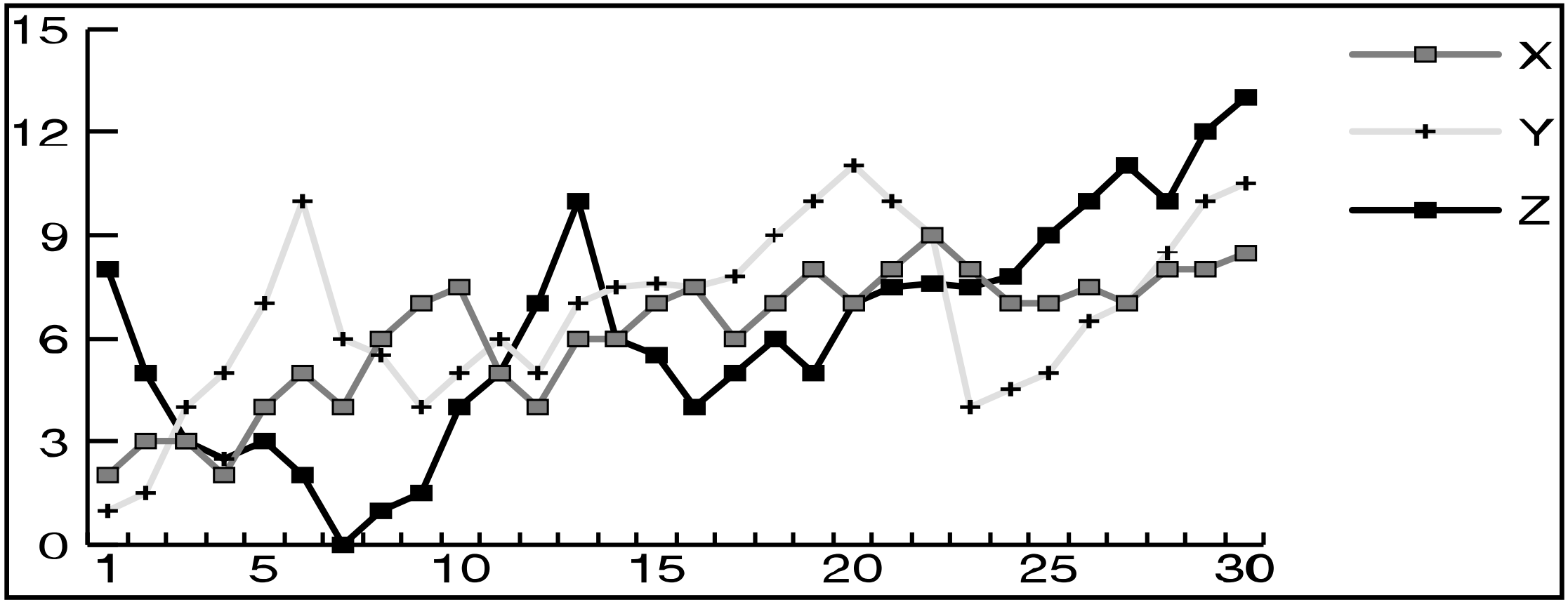
- A college professor wishes to reach a certain level of savings before his retirement.

- Peroidically, he predicts what his retirement savings will be based on its current value and several past values.
- Simple linear regression formula he might uses to fit the past value into the function and predict the future.

- Try to predict home price
- How many numbers of T-shirt will produce in certain time period

3. Time series Analysis

- The value of an attribute is examined as it varies over time.
- The values usually are obtained as evenly spaced time points(Daily, weekly, hourly etc)

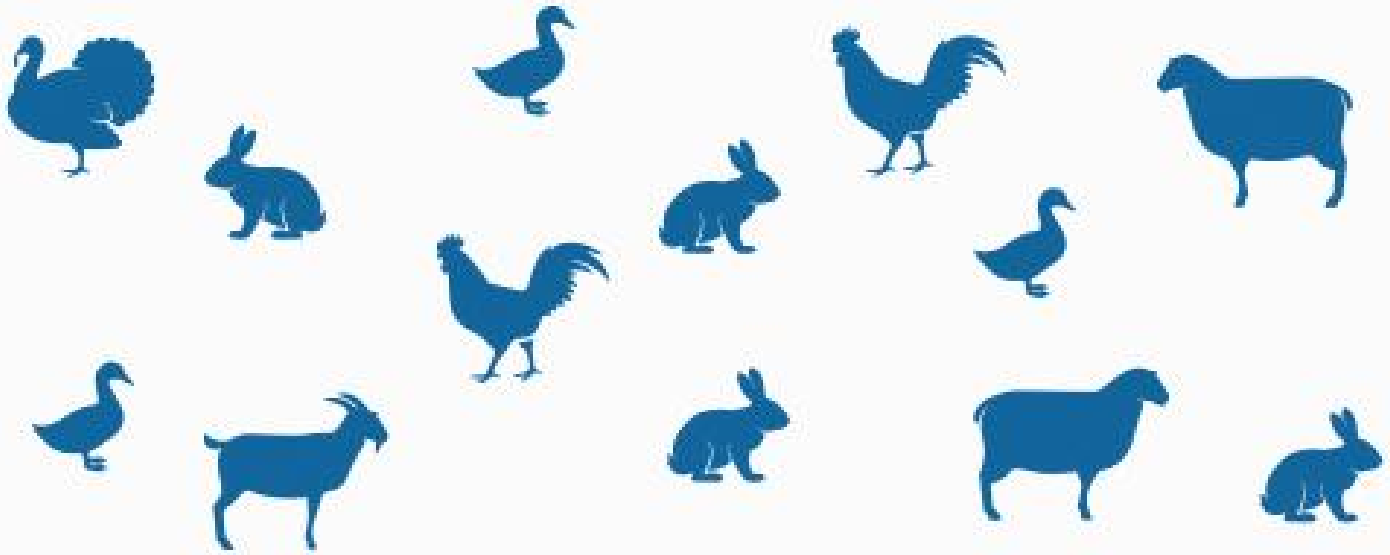

15
12
9
6
3
0
1
5
10
15
20
25
30
X
Y
Z

Example

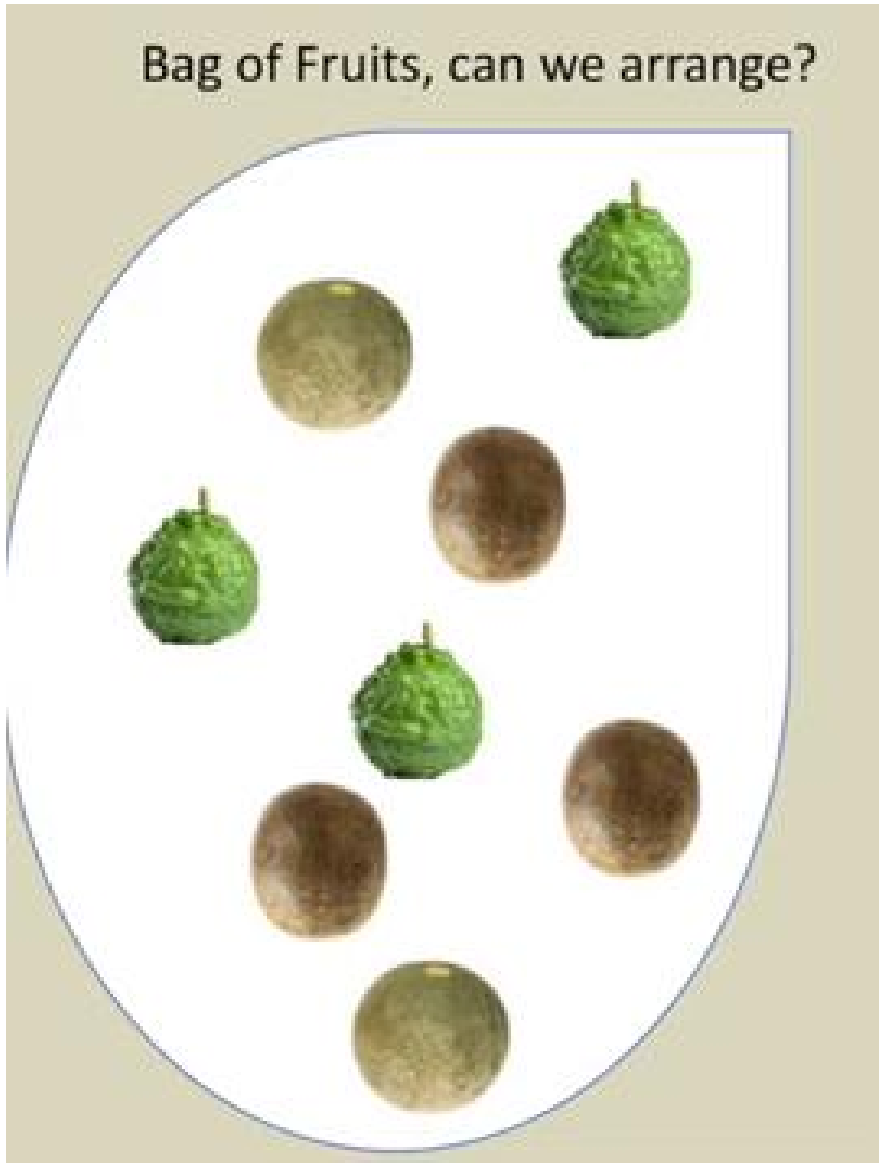
- Mr. suresh is trying to determine whether to purchase stock from companies like x,y and z.
- For a period of one month he charts the daily price for each company is presented in figure.
- Then, suresh decides to purchase stock X because it is less volatile.
- The behavior of Y between days 6 and 20 is identical to theat for Z between days 13 and 27.

4. Prediction

- Prediction is type of classification.
- The difference is that prediction is predicting a future state rather than a current state.
- Prediction application including flooding, speech recognition, machine learning and pattern recognition.
- Moreover, future values may be predicted using time series analysis or regression techniques.


Classification
Clustering

Clustering

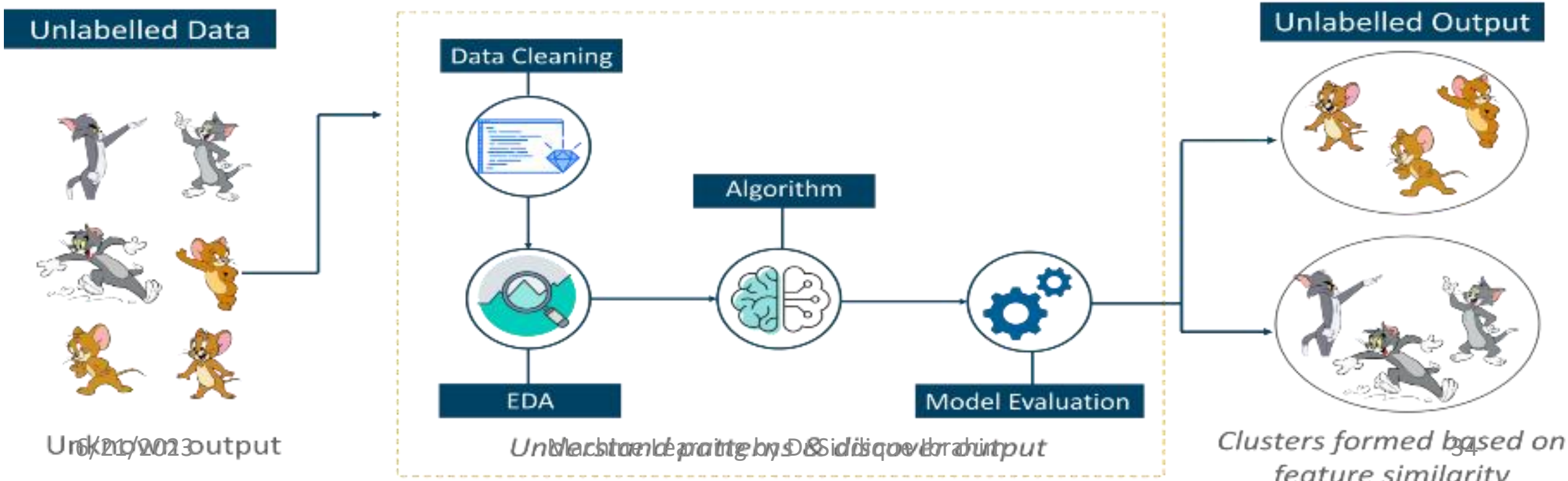

Bag of Fruits, can we arrange?


We want to arrange
similar fruits together
We found 3 bags

5. Clustering

- Unsupervised learning or segmentation.
- it can be thought of as partitioning or segmenting the data into groups.
- The clustering is usually accomplished by determining the similarity among the data on predefined attributes.
- Most similar attributes are grouped into cluster.

- **Unsupervised learning** involves training by using unlabeled data and allowing the model to act on that information without guidance.
- Think of unsupervised learning as a smart kid that learns without any guidance.
- In this type of Machine Learning, the model is not fed with labeled data, as in the model has no clue that 'this image is Tom and this is Jerry',
- it figures out patterns and the differences between Tom and Jerry on its own by taking in tons of data.



Example

- A certain national department store chain creates special catelogs targeted to various demographic groups based on attributes such as income, age, location, weight, height, etc.
- To determine the target mailing of the various catelogs, the company performs a clustering of potential customers based on the determined attribute values.

6. Outlier Analysis

- A database may contain data objects do not comply with the generatl behavior.
- These are outlier.
- Most data mining methods disascard outlier as noise or exceptions.
- However, in some applications such fraud detection, rare events can be more interesting than the more regular one

Example

- It will helps to identify fraudulent uage of credit cards by detecting purchase of extremely large amount will be compared to regulare amount.
- Also detected with location and type of purchase or purchase frequency.

7. Summarization

- Summarization maps data into subsets with associated simple descriptions.
- it also called characterization or generalization.
- summary type information(such as mean, median etc) can be derived from the data.

Example

- One of the many criteria used to compare universities by the Indian goverment is NIRF ranking is avg score.

National Institutional Ranking Framework
Ministry of Education
Government of India

Home Ranking

India Rankings 2023: University

Rank-band: 101-150 | Rank-band: 151-200

Show 100 entries

Search:

Institute ID	Name	City	State	Score	Rank
IR-O-U-0220	Indian Institute of Science More Details | |	Bengaluru	Karnataka	83.16	1
IR-O-U-0109	Jawaharlal Nehru University More Details | |	New Delhi	Delhi	68.92	2
IR-O-U-0108	Jamia Millia Islamia More Details | |	New Delhi	Delhi	67.73	3
IR-O-U-0575	Jadavpur University More Details | |	Kolkata	West Bengal	66.07	4
IR-O-U-0500	Banaras Hindu University More Details | |	Varanasi	Uttar Pradesh	65.85	5
IR-O-U-0234	Manipal Academy of Higher Education-Manipal More Details | |	Manipal	Karnataka	64.98	6
IR-O-U-0436	Amrita Vishwa Vidyapeetham More Details | |	Coimbatore	Tamil Nadu	64.67	7
IR-O-U-0490	Vellore Institute of Technology More Details | |	Vellore	Tamil Nadu	64.33	8
IR-O-U-0496	Aligarh Muslim University More Details | |	Aligarh	Uttar Pradesh	63.88	9
IR-O-U-0042	University of Hyderabad More Details | |	Hyderabad	Telangana	62.09	10
IR-O-U-0120	University of Delhi More Details | |	Delhi	Delhi	61.45	11

Cleanest City in India

Summary

The Ministry of Housing & Urban Affairs ranks cities based on cleanliness index. This list summarises the cities topping those lists annually.

Year	First			Runner up	
	City		State	City	State
2021	Indore		Madhya Pradesh	Surat	Gujarat
2020	Metropolis	Indore	Madhya Pradesh	Surat	Gujarat
	Big city	Ahmedabad	Gujarat		
	Medium city	Mysore	Karnataka		
	Small city	Ambikapur	Chhattisgarh		
2019	Metropolis	Indore	Madhya Pradesh	Ambikapur	Chhattisgarh
	Big city	Ahmedabad	Gujarat		
	Medium city	Ujjain	Madhya Pradesh	Mysore	Karnataka
	Small city	New Delhi (Municipal Council)	Delhi		
2018	Metropolis	Indore	Madhya Pradesh	Bhopal	Madhya Pradesh
	Big city	Vijayawada	Andhra Pradesh		
	Medium city	Mysore	Karnataka		
	Small city	New Delhi (Municipal Council)	Delhi		
2017	Indore		Madhya Pradesh	Bhopal	Madhya Pradesh
2016	Mysore		Karnataka	Chandigarh	Chandigarh Territory
2015				Tiruchirapalli	Tamil Nadu

8. Association Rules

- Association analysis is the discovery of association rules showing attribute value conditions that occur frequently together in a given set of data.
- Widely used in market basket analysis or transaction data analysis.
- Support and confidence measures

Example

- A grocery store retailer is trying to decide whether to put bread on sale.
- To help determine the impact of this decision, the retailer generates association rules that show what other products are frequently purchased with bread.
- He find that 60% sold alone and 80% of the time it sold with jam.

Example of discovered patterns

- Association rules:
- "80% of customers who buy cheese and milk also buy bread, and 5% of customers buy all of them together"
- Cheese, Milk→ Bread [sup =5%, confid=80%]

9. sequence Discovery

- It is used to determine sequential pattern in data.
- These patterns are similar to associations in that data, but the relationship is based on time.
- Unlike association, which requires the items to be purchased at the same time, in sequence discovery the items are purchased over time in some order.

Example

- A similar type of discovery can be seen in the sequence within which data are purchased.
- Most people who purchase house may be found to purchase sofa within a month.

- The web development company wants to analysis the user log details to determine the web page analysis.
- They are interested in determing what seqences of pages are freqently accessed.
- They determined 80% of the users of page A follow one the following pattern:
- {A,B,C} { A,D,B,C} {A,E,B,D,C} {A,F,B,C}

10. Characterization and Discrimination

Data characterization is a summarization of the general characteristics or features of a target class of data. The data corresponding to the user-specified class are typically collected by a query.

ex: Description of all users who spent more than $10,000 a year at *AllElectronics?* A general profile of all customers, such as age, salary, location and credit ratings. Among all the customers meeting target condition (spent > $10,000), 10% are "Youth", 60% are "Adults" and 30% are "Seniors".

The output of data characterization can be presented in pie charts, bar charts, multidimensional data cubes, and multidimensional tables. They can also be presented in rule form.

Characterization and Discrimination (1)

- **Data Characterization**
  - → Summarize the general features of a target class of data
  - → Tools: statistical measures, data cube-based OLAP roll-up, etc.
  - → Output: charts, curves, multidimensional data cubes, etc.
  - → Example

Summarize the characteristics of costumers who spend more than 1000€ →

Costumers profile

- 40-50 years old
- Employed
- excellent credit ratings

- **Data Discrimination**
  - → Comparison of the general features of a target class with the general features of contrasting classes
  - → Output: similar to characterization + comparative measures
  - → Example

Compare customers who shop for computer products regularly(more than 2 times a month) with those who rarely shop for such products(less then three times a year) →

Comparative profile

Frequent costumers	Rare costumers
80% •Are between 20 and 40 •Have university education	60% •Are senior or youths •Have no university degree