Data Science Life Cycle – Tools and Technologies

M.S.Mohan Sivam, B.E.,M.Tech.,M.B.A.,

Technical Lead – Bigdata Infrastructure

Tata Consultancy Services – Chennai

Agenda

- What is Data Science?
- Bigdata
- Data Science Life Cycle
- Tools and Technologies
- Top Algorithms in Data Science Use cases.
- Skillset required for Data scientist.

What is Data Science?

"Data science is an inter-disciplinary field that uses scientific methods, processes, algorithms and systems to extract knowledge and insights from many structural and unstructured data."

- Wikipedia

What is BigData?

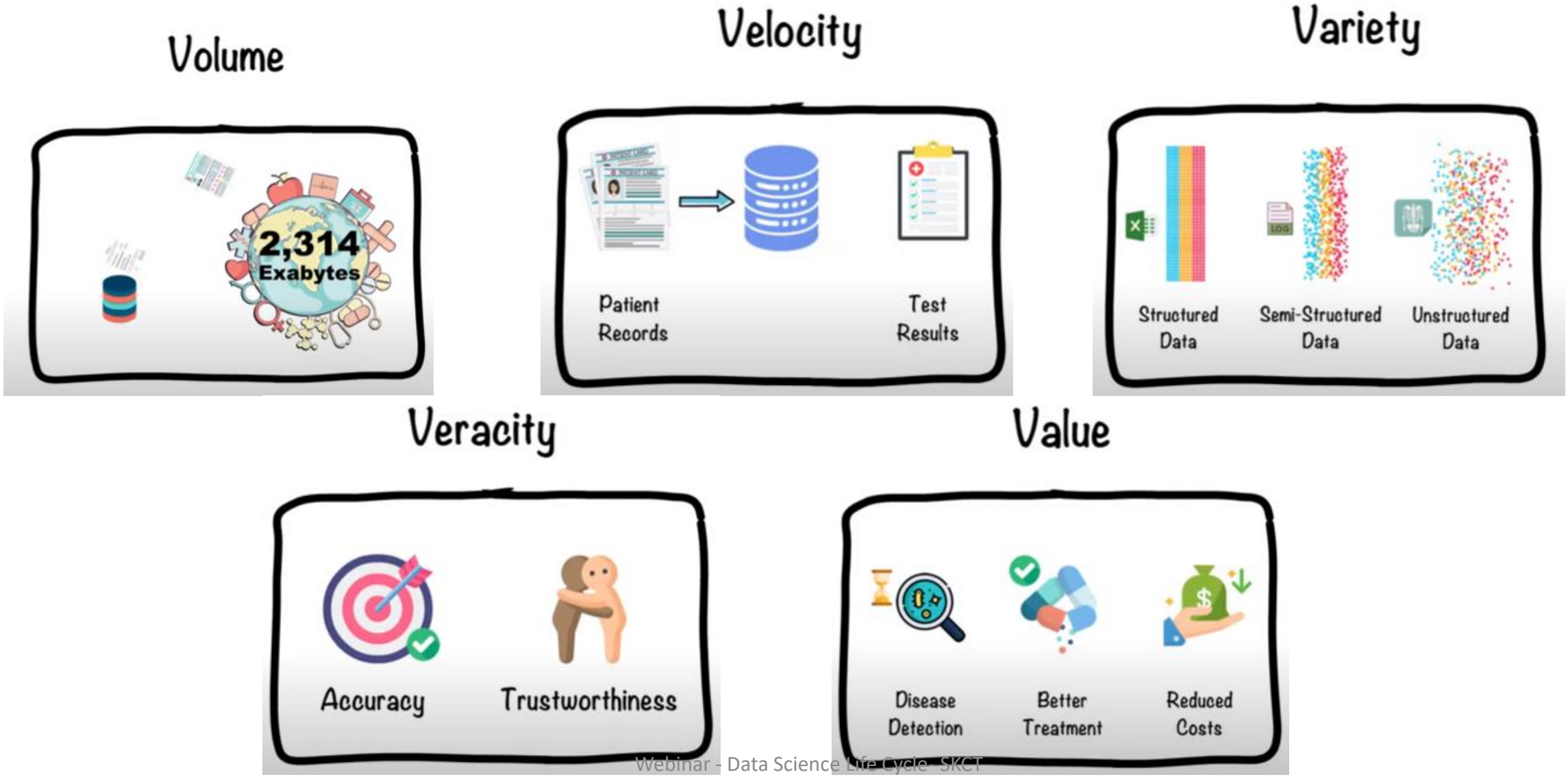
Put simply, big data is larger, more complex data sets, especially from new data sources.

These data sets are so voluminous that traditional data processing software just can’t manage them.

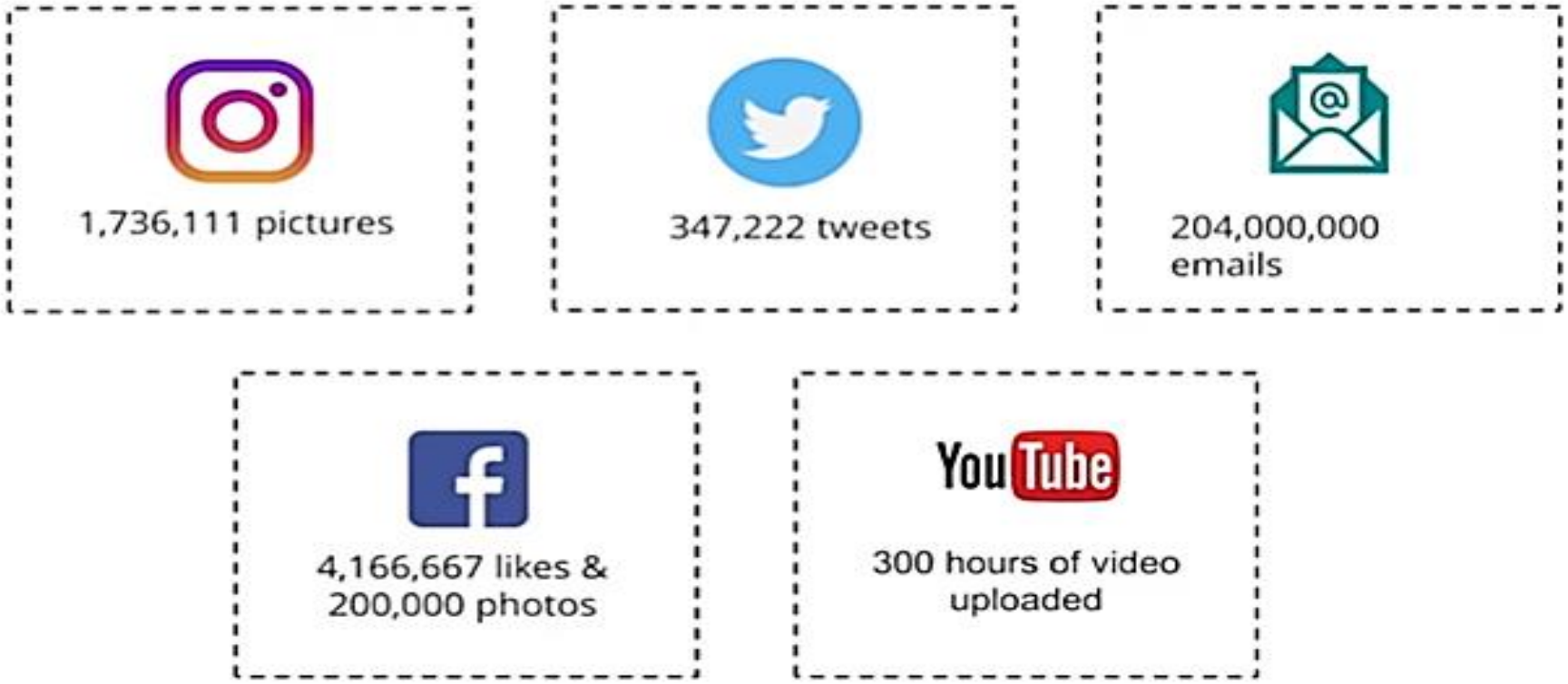
But these massive volumes of data can be used to address business problems you wouldn’t have been able to tackle before.

90% of the Data is Generated in last 2 year.

BigData (Eg – Health care)

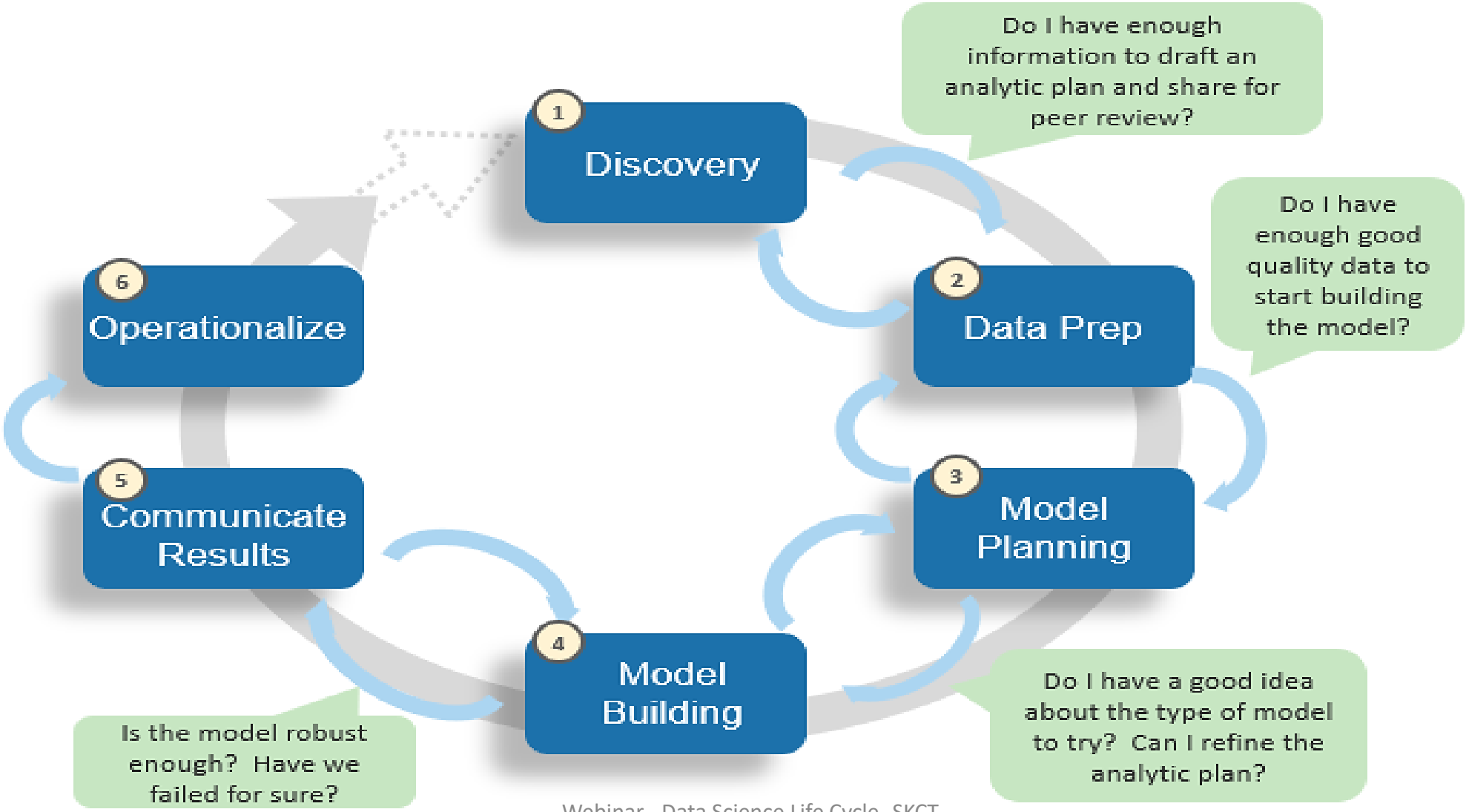


Every minute of the Day!



But How do we store and process this BigData!?

Data Science Life Cycle



Data Science Tools

Contd..

HADOOP

NOSQL DATA BASES

ETL TOOLS

DEVOPS

MACHINE LEARNING

ARTIFICIAL INTELLIGENCE

DATA VISUALIZATION

ETC.,

Top Algorithms in Data Science Use cases.

Decision Tree

Random Forest

Association Rule Mining

Linear Regression

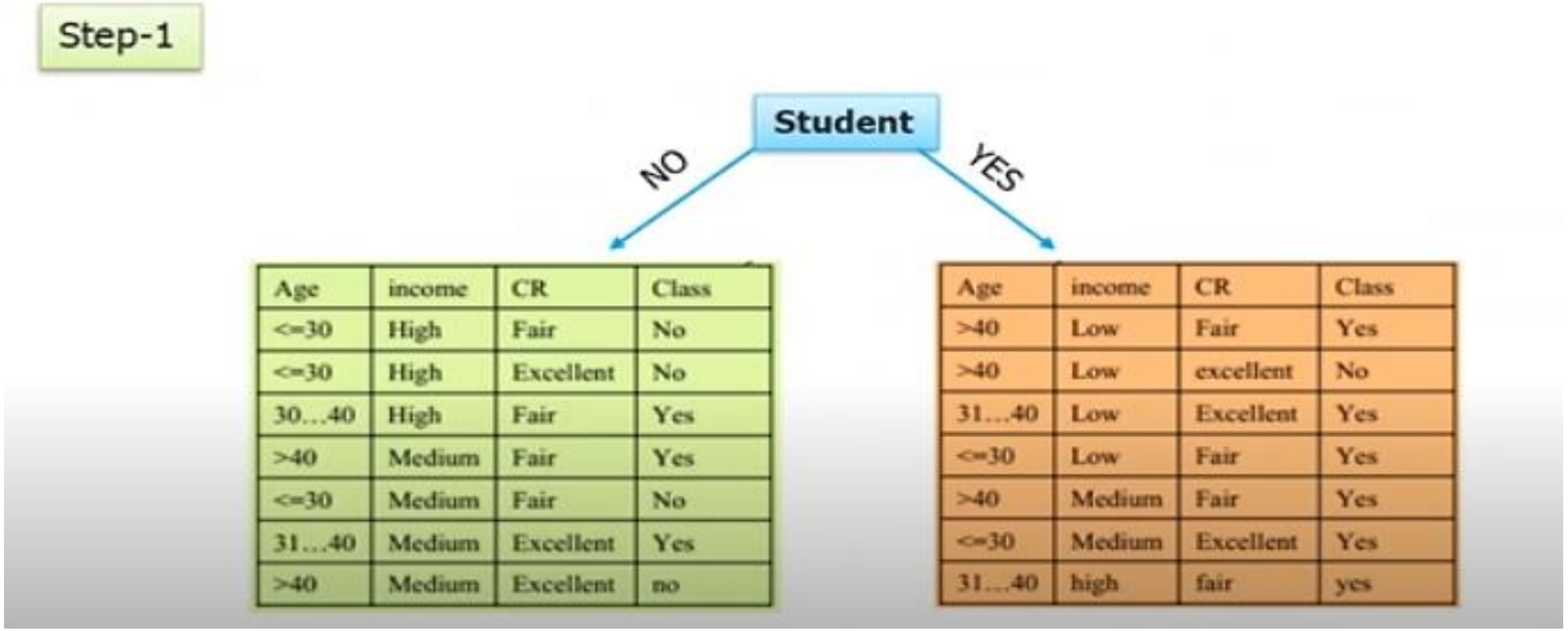
K-Means Clustering

Decision Tree

rec	Age	Income	Student	Credit_rating	Buys_computer
r1	<=30	High	No	Fair	No
r2	<=30	High	No	Excellent	No
r3	31…40	High	No	Fair	Yes
r4	>40	Medium	No	Fair	Yes
r5	>40	Low	Yes	Fair	Yes
r6	>40	Low	Yes	Excellent	No
r7	31…40	Low	Yes	Excellent	Yes
r8	<=30	Medium	No	Fair	No
r9	<=30	Low	Yes	Fair	Yes
r10	>40	Medium	Yes	Fair	Yes
r11	<-=30	Medium	Yes	Excellent	Yes
r12	31…40	Medium	No	Excellent	Yes
r13	31…40	High	Yes	Fair	Yes
r14	>40	Medium	No	Excellent	No

Contd..

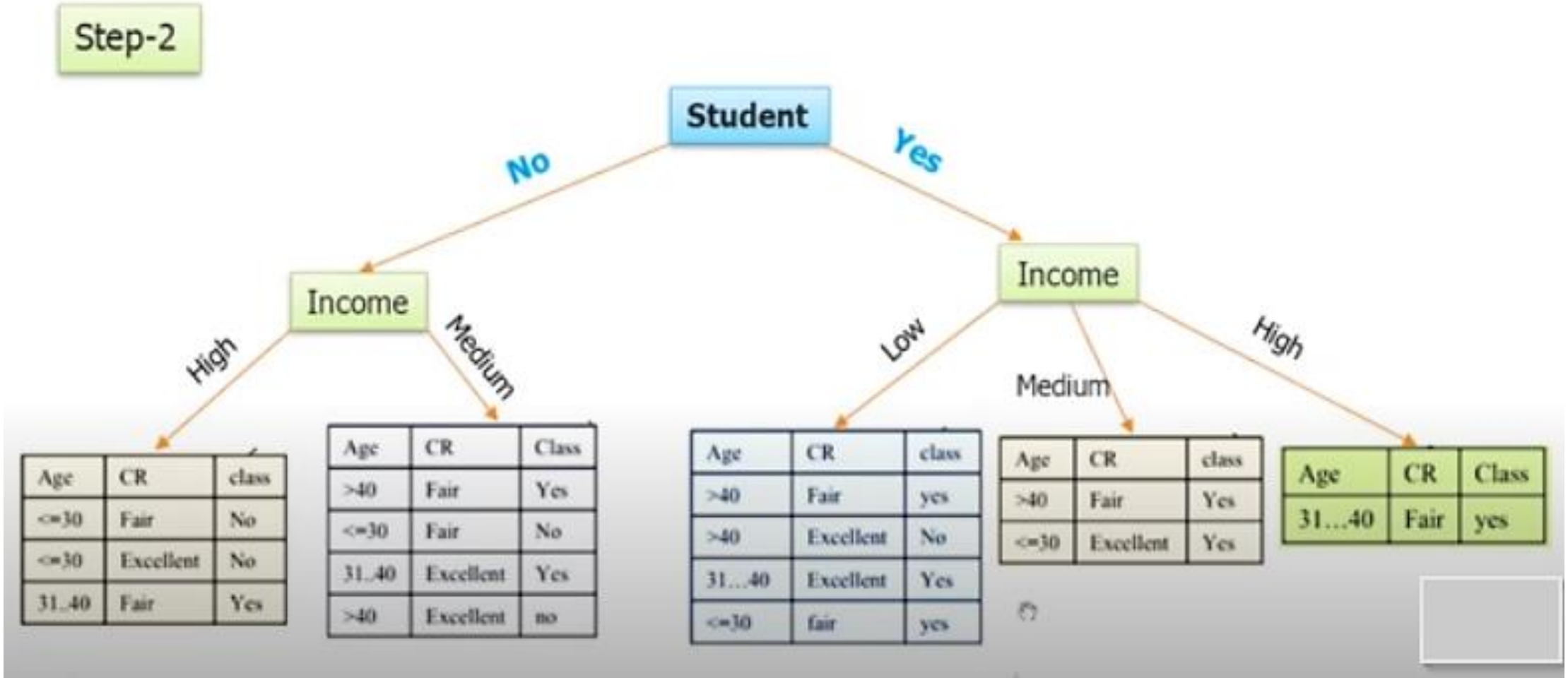
- Step 1



Age	income	CR	Class
<=30	High	Fair	No
<=30	High	Excellent	No
30…40	High	Fair	Yes
>40	Medium	Fair	Yes
<=30	Medium	Fair	No
31…40	Medium	Excellent	Yes
>40	Medium	Excellent	no

Age	income	CR	Class
>40	Low	Fair	Yes
>40	Low	excellent	No
31…40	Low	Excellent	Yes
<=30	Low	Fair	Yes
>40	Medium	Fair	Yes
<=30	Medium	Excellent	Yes
31…40	high	fair	yes

Contd..



Age	CR	class
<=30	Fair	No
<=30	Excellent	No
31..40	Fair	Yes

Age	CR	Class
>40	Fair	Yes
<=30	Fair	No
31..40	Excellent	Yes
>40	Excellent	no

Age	CR	class
>40	Fair	yes
>40	Excellent	No
31…40	Excellent	Yes
<=30	fair	yes

Age	CR	class
>40	Fair	Yes
<=30	Excellent	Yes

Age	CR	Class
31…40	Fair	yes

Contd..

Classification rules :

1. student(no)^income(high)^age(<=30) => buys_computer(no)
2. student(no)^income(high)^age(31...40) => buys_computer(yes)
3. student(no)^income(medium)^CR(fair)^age(>40) => buys_computer(yes)
4. student(no)^income(medium)^CR(fair)^age(<=30) => buys_computer(no)
5. student(no)^income(medium)^CR(excellent)^age(>40) => buys_computer(no)
6. student(no)^income(medium)^CR(excellent)^age(31..40) =>buys_computer(yes)
7. student(yes)^income(low)^CR(fair) => buys_computer(yes)
8. student(yes)^income(low)^CR(excellent)^age(31..40) => buys_computer(yes)
9. student(yes)^income(low)^CR(excellent)^age(>40) => buys_computer(no)
10. student(yes)^income(medium)=> buys_computer(yes)
11. student(yes)^income(high)=> buys_computer(yes)

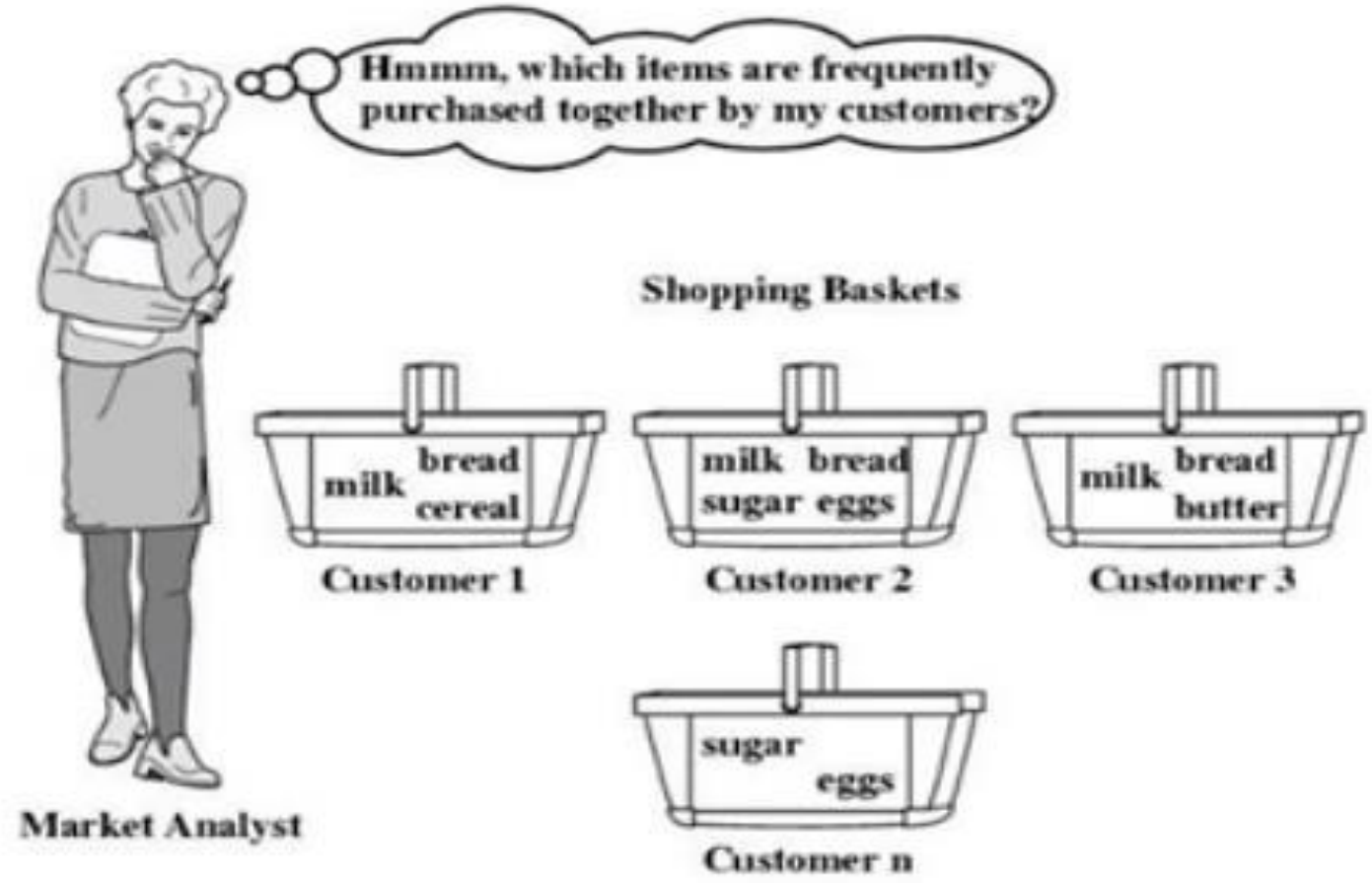
Random Forest

- Suppose you are indecisive about watching a movie, “Kaatru Veliyidai”
- You can do one of the following

1. Either you ask your best friend, whether you will like the movie or not.

2. Or you can as your Group of Friends.

Association Rule Mining



Contd..

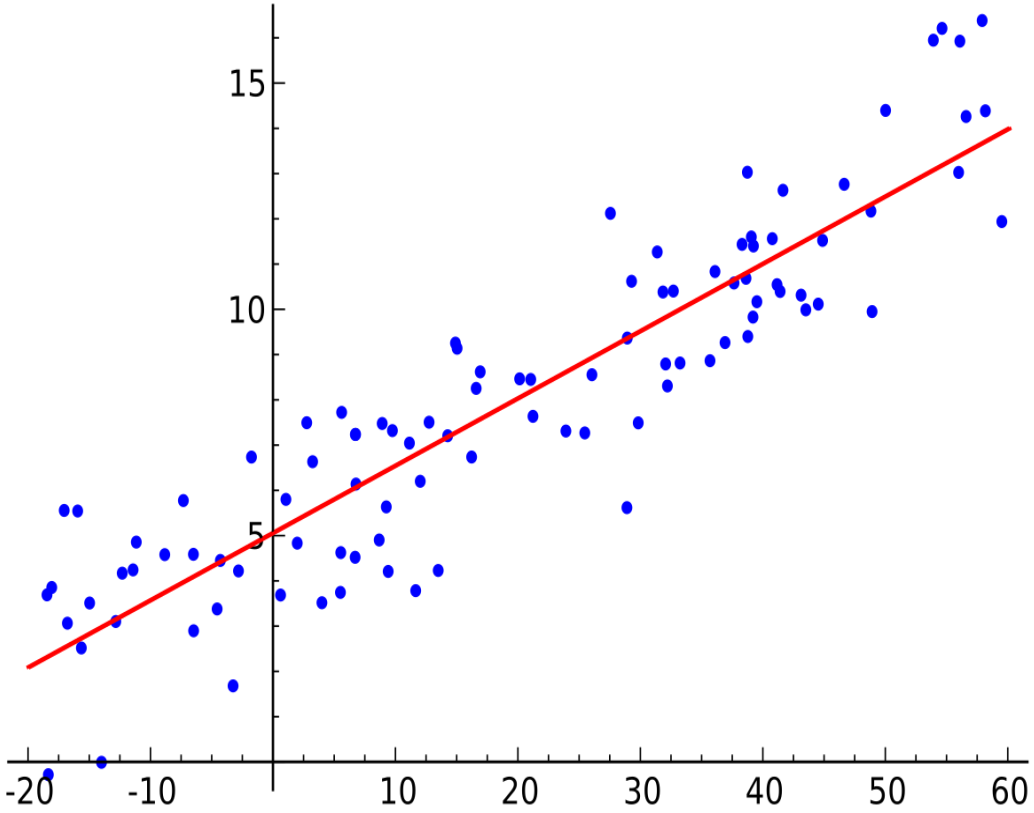
- **Association Rule Mining** is a Popular and well Researched method for discovering interested relations between variables in large Data.
- The Rule found in the sales data of a Super Market would indicate that if a Customer buys Onions and Tomato together, he or she is likely to also buy a Chicken.

$$\text{Support(A->B)}=\frac{Number\ of\ Records\ containing\ both\ A\ and\ B}{Amount\ of\ Records}$$

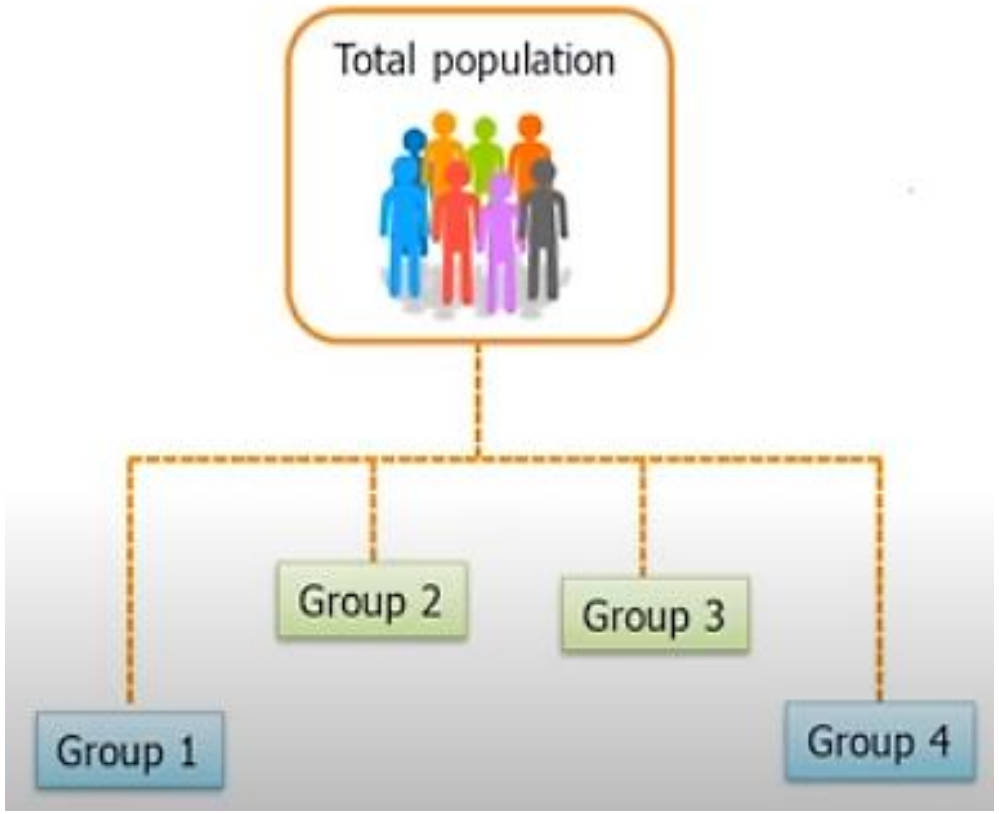
$$\text{Confidence(A->B)}=\frac{Number\ of\ Records\ containing\ both\ A\ and\ B}{Number\ of\ Records\ containing\ A}$$

Regression Analysis – Linear Regression

- Regression Analysis helps to understand how value of Dependent variable Changes when any one of Independent variable Changes, while other dependent variables are kept fixed.
- Linear Regression is the most Popular Algorithm used for Prediction and Forecasting



K-Means Clustering



- Kmeans algorithm is an iterative algorithm that tries to partition the dataset into Kpre-defined distinct non-overlapping subgroups (clusters) where each data point belongs to only one group. It tries to make the inter-cluster data points as similar as possible while also keeping the clusters as different (far) as possible.
- The Objects Group 1 should be as similar as possible.
- But there should be much difference between objects in different Groups.
- The attributes of the Objects are allowed to determine which objects should be grouped together.

https://github.com/msmohansivam/SKCT_Webinar_Hadoop

Thank You

https://in.linkedin.com/in/m-s-mohan-sivam-281305a7

+91 9790748175