

What do you get from this course?

SKILLS:

- Regression
- Classification
- Clustering
- Scikit Learn
- Scipy

PROJECTS:

- Cancer detection
- Predicting economic trends
- Predicting customer churn
- Recommendation engines
- Many more



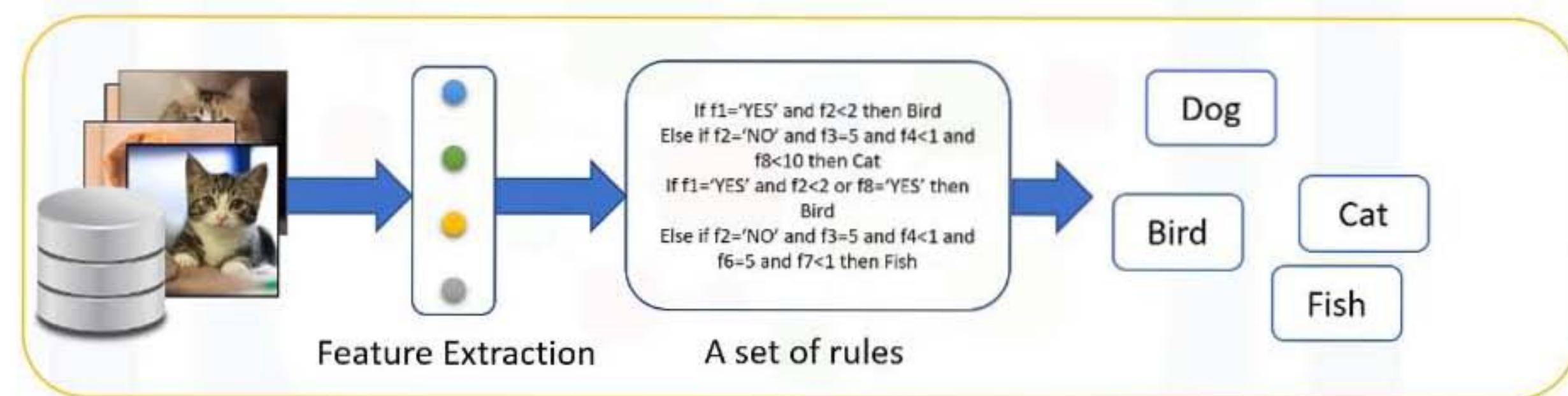
What is machine learning?

Machine learning is the subfield of computer science that gives “**computers the ability to learn without being explicitly programmed.**”

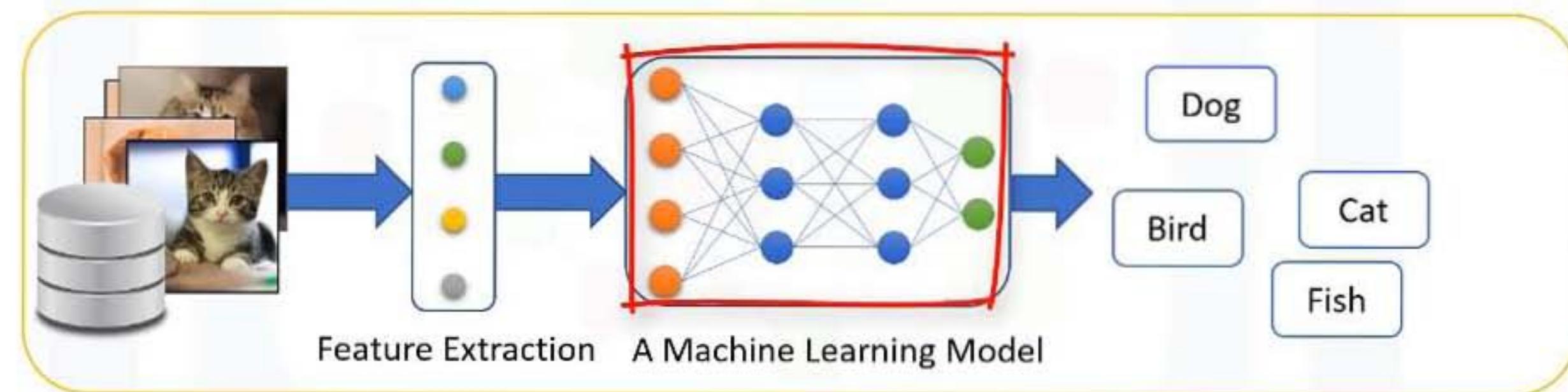
Arthur Samuel

American pioneer in the field of computer gaming and artificial intelligence, coined the term "machine learning" in 1959 while at IBM.

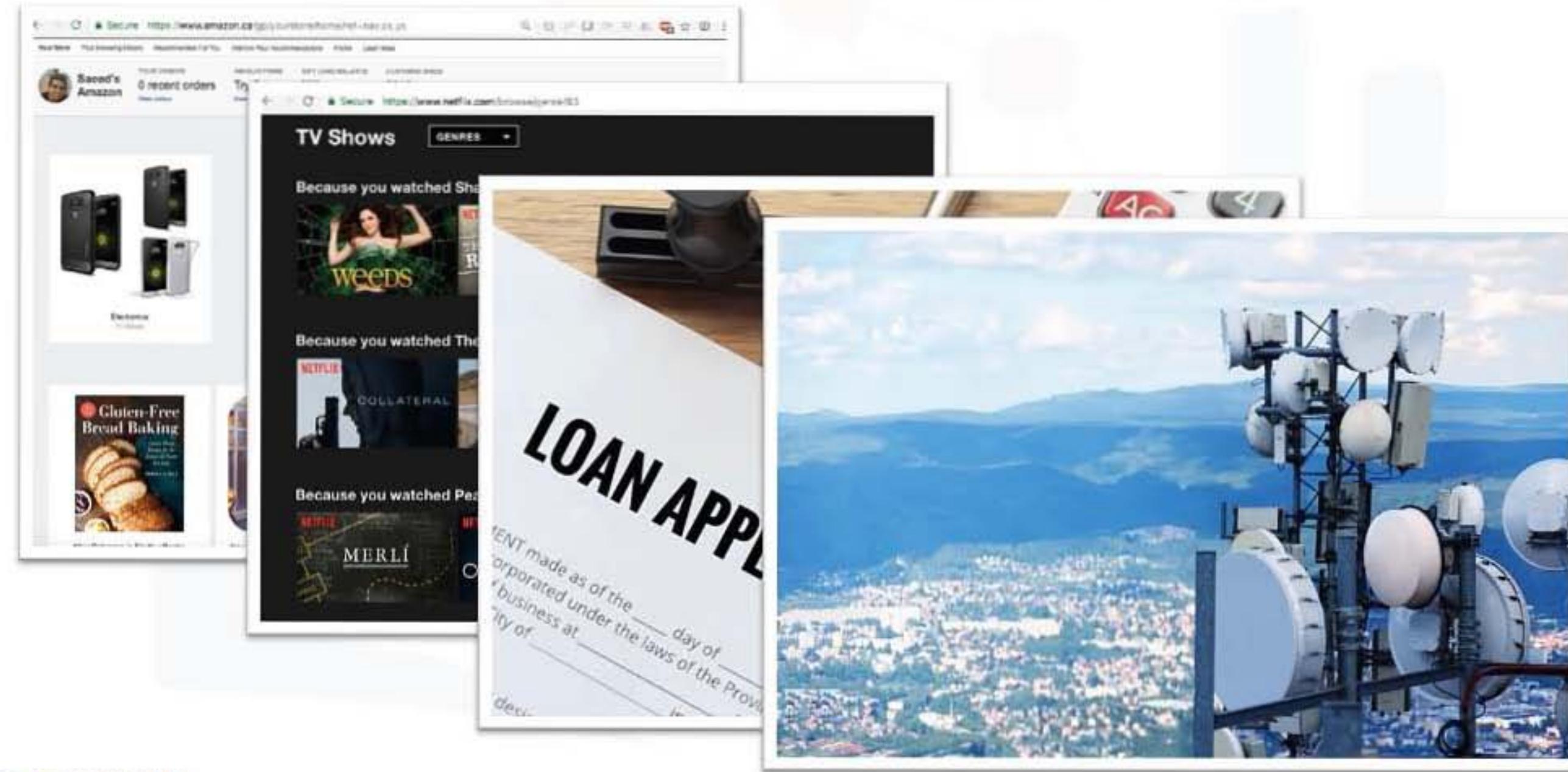
How machine learning works?



How machine learning works?



Examples of machine learning



Major machine learning techniques

- Regression/Estimation
 - Predicting continuous values
- Classification
 - Predicting the item class/category of a case
- Clustering
 - Finding the structure of data; summarization
- Associations
 - Associating frequent co-occurring items/events

Major machine learning techniques

- Anomaly detection
 - Discovering abnormal and unusual cases
- Sequence mining
 - Predicting next events; click-stream (Markov Model, HMM)
- Dimension Reduction
 - Reducing the size of data (PCA)
- Recommendation systems
 - Recommending items

Difference between artificial intelligence, machine learning, and deep learning

- AI components:

- Computer Vision
- Language Processing
- Creativity
- Etc.

- Machine learning:

- Classification
- Clustering
- Neural Network
- Etc.

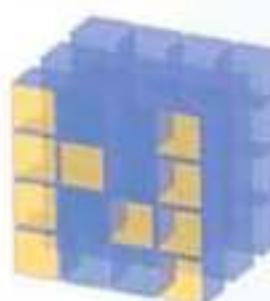
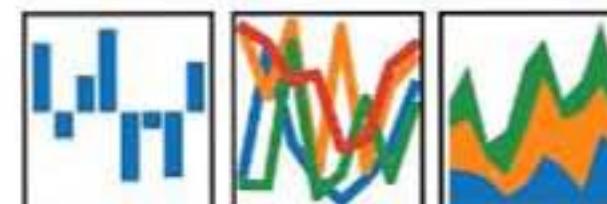
- Revolution in ML:

- Deep learning



Python libraries for machine learning

pandas
 $y_i t = \beta' x_{it} + \mu_i + \epsilon_{it}$



NumPy



SciPy

matplotlib



python

More about scikit-learn

- Free software machine learning library
- Classification, Regression and Clustering algorithms
- Works with NumPy and SciPy
- Great documentation
- Easy to implement



scikit-learn functions

```
from sklearn import preprocessing  
X = preprocessing.StandardScaler().fit(X).transform(X)
```

```
from sklearn.model_selection import train_test_split  
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.33)
```

```
from sklearn import svm  
clf = svm.SVC(gamma=0.001, C=100.)
```

```
clf.fit(X_train, y_train)
```

```
clf.predict(X_test)
```

```
from sklearn.metrics import confusion_matrix  
print(confusion_matrix(y_test, yhat, labels=[1,0]))
```

scikit-learn functions

```
from sklearn import preprocessing  
X = preprocessing.StandardScaler().fit(X).transform(X)
```

```
from sklearn.model_selection import train_test_split  
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.33)
```

```
from sklearn import svm  
clf = svm.SVC(gamma=0.001, C=100.)
```

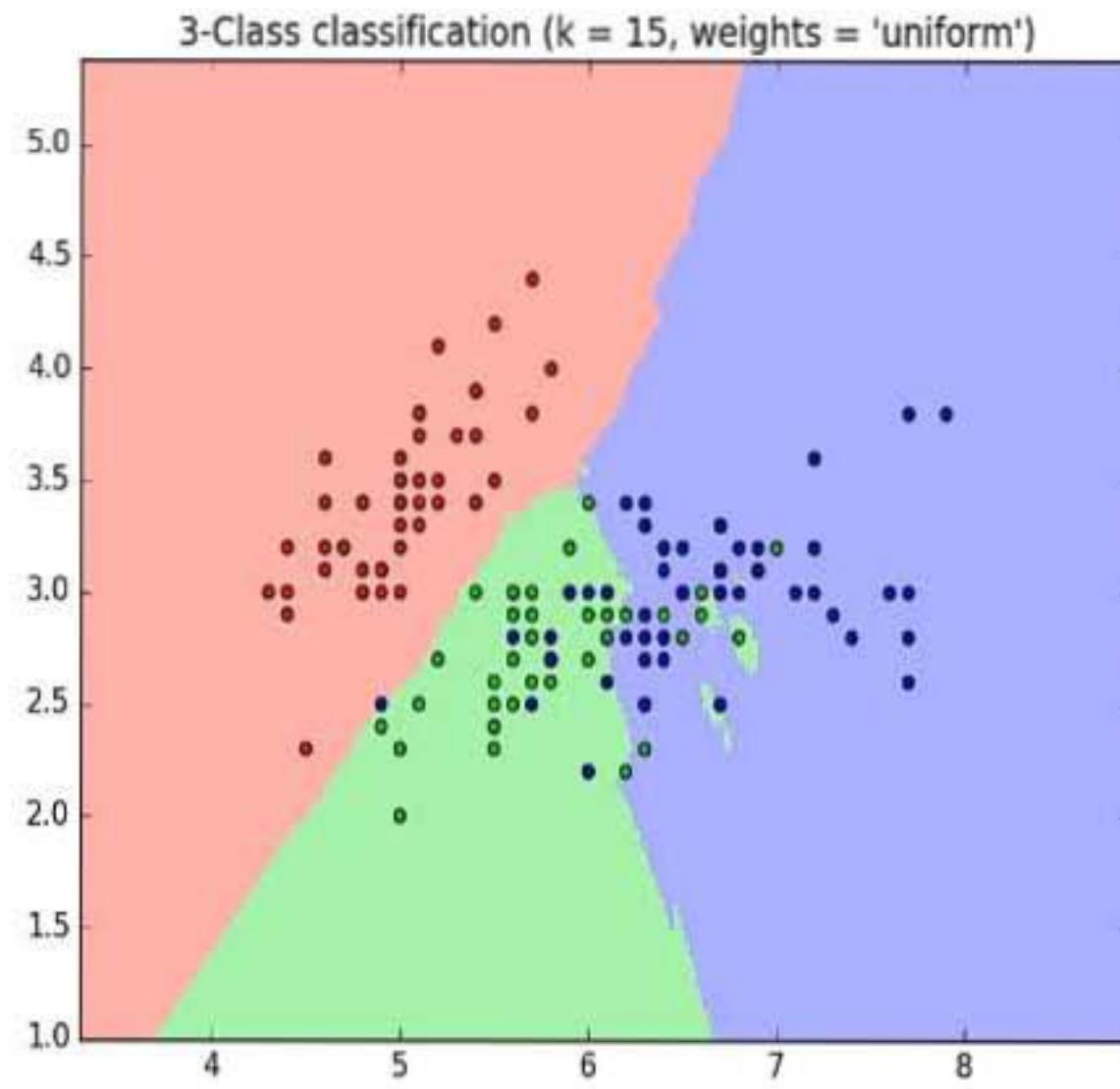
```
clf.fit(X_train, y_train)
```

```
clf.predict(X_test)
```

```
from sklearn.metrics import confusion_matrix  
print(confusion_matrix(y_test, yhat, labels=[1,0]))
```

```
import pickle  
s = pickle.dumps(clf)
```

What is supervised learning?

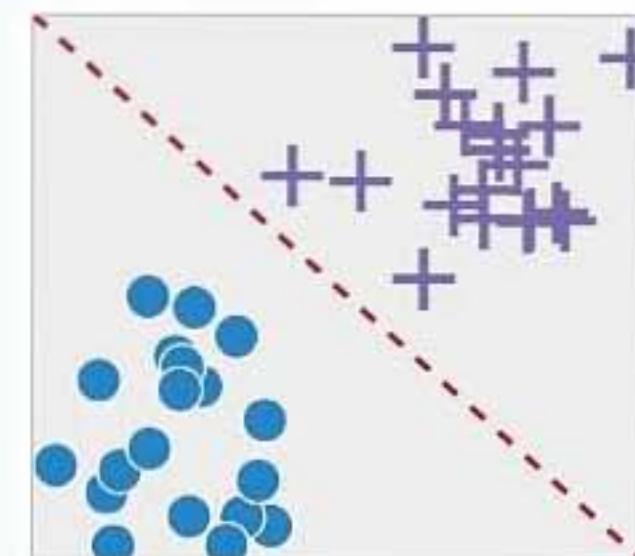


We “teach the model,”
then with that knowledge,
it can predict unknown or
future instances.

What is classification?

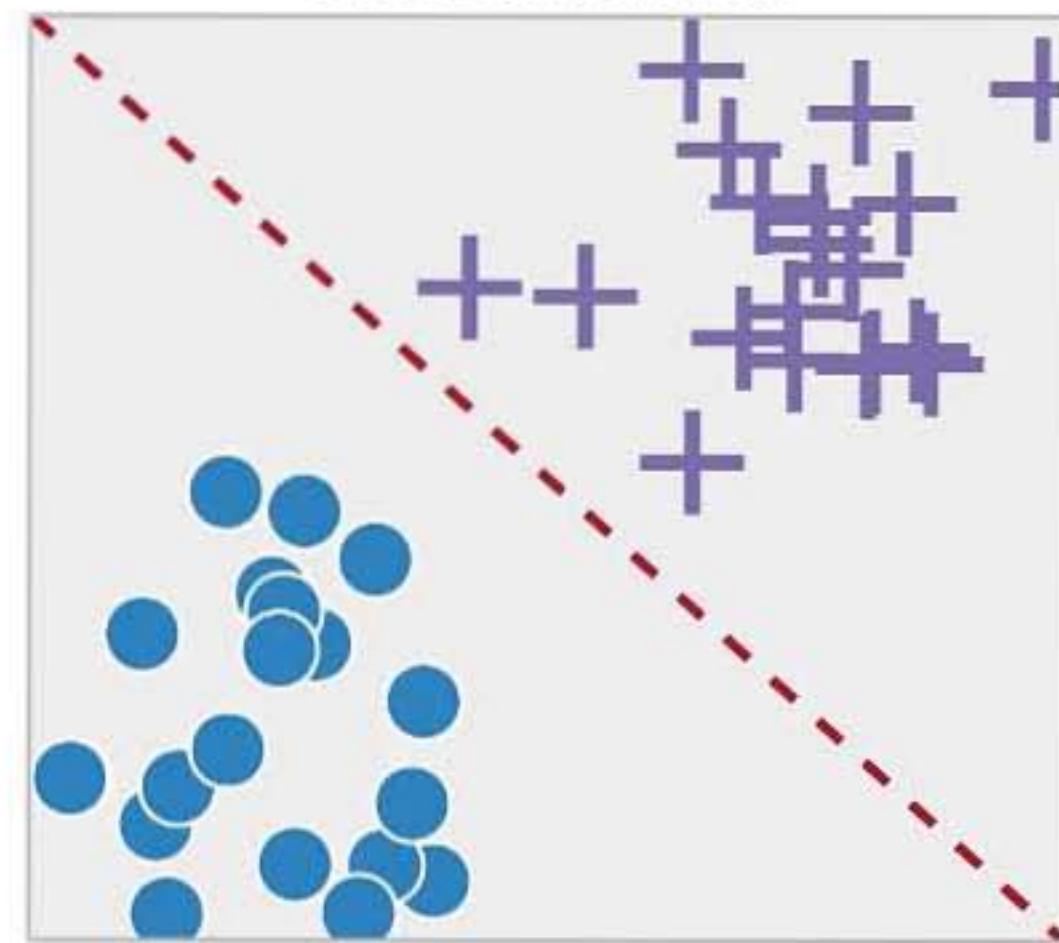
Classification is the process of predicting discrete class labels or categories.

ID	Clump	UnifSize	UnifShape	MargAdh	SingEpiSize	BareNuc	BlandChrom	NormNucl	Mit	Class
1000025	5	1	1	1	2	1	3	1	1	benign
1002945	5	4	4	5	7	10	3	2	1	benign
1015425	3	1	1	1	2	2	3	1	1	malignant
1016277	6	8	8	1	3	4	3	7	1	benign
1017023	4	1	1	3	2	1	3	1	1	benign
1017122	8	10	10	8	7	10		7	1	malignant
1018099	1	1	1	1	2	10	3	1	1	benign
1018561	2	1	2	H	2	1	3	1	1	benign
1033078	2	1	1	1	2	1	1	1	5	benign
1033078	4	2	1	1	2	1	2	1	1	benign

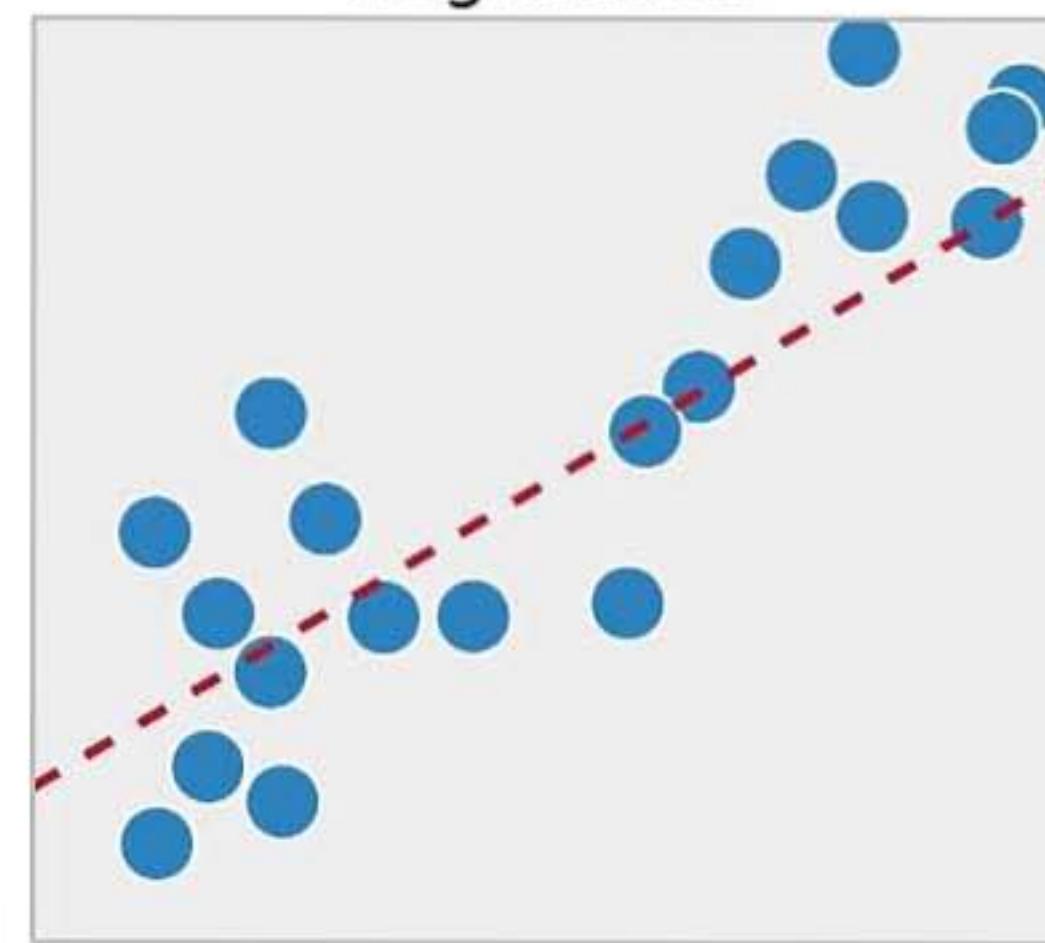


Types of supervised learning

Classification



Regression

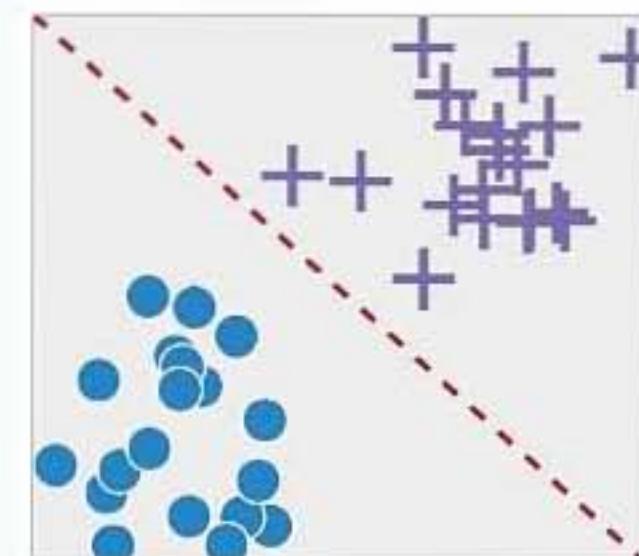


What is classification?

Classification is the process of predicting discrete class labels or categories.

ID	Clump	UnifSize	UnifShape	MargAdh	SingEpiSize	BareNuc	BlandChrom	NormNucl	Mit	Class
1000025	5	1	1	1	2	1	3	1	1	benign
1002945	5	4	4	5	7	10	3	2	1	benign
1015425	3	1	1	1	2	2	3	1	1	malignant
1016277	6	8	8	1	3	4	3	7	1	benign
1017023	4	1	1	3	2	1	3	1	1	benign
1017122	8	10	10	8	7	10		7	1	malignant
1018099	1	1	1	1	2	10	3	1	1	benign
1018561	2	1	2	H	2	1	3	1	1	benign
1033078	2	1	1	1	2	1	1	1	5	benign
1033078	4	2	1	1	2	1	2	1	1	benign

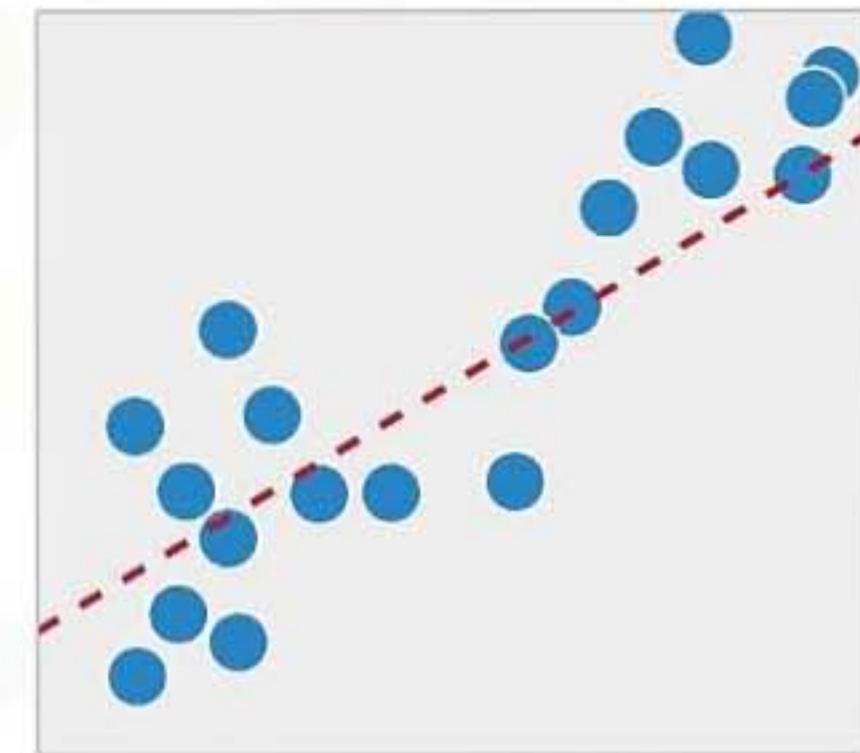
Categorical Values



What is regression?

Regression is the process of predicting continuous values.

	ENGINESIZE	CYLINDERS	FUELCONSUMPTION_COMB	CO2EMISSIONS
0	2.0	4	8.5	196
1	2.4	4	9.6	221
2	1.5	4	5.9	136
3	3.5	6	11.1	255
4	3.5	6	10.6	244
5	3.5	6	10.0	230
6	3.5	6	10.1	232
7	3.7	6	11.1	255
8	3.7	6	11.6	267
9	2.4	4	9.2	?



What is unsupervised learning?

Customer Id	Age	Edu	Years Employed	Income	Card Debt	Other Debt	Address	DebtIncomeRatio
1	41	2	6	19	0.124	1.073	NBA001	6.3
2	47	1	26	100	4.582	8.218	NBA021	12.8
3	33	2	10	57	6.111	5.802	NBA013	20.9
4	29	2	4	19	0.681	0.516	NBA009	6.3
5	47	1	31	253	9.308	8.908	NBA008	7.2
6	40	1	23	81	0.998	7.831	NBA016	10.9
7	38	2	4	56	0.442	0.454	NBA013	1.6
8	42	3	0	64	0.279	3.945	NBA009	6.6
9	26	1	5	18	0.575	2.215	NBA006	15.5
10	47	3	23	115	0.653	3.947	NBA011	4
11	44	3	8	88	0.285	5.083	NBA010	6.1
12	34	2	9	40	0.374	0.266	NBA003	1.6

What is unsupervised learning?

Customer Id	Age	Edu	Years Employed	Income	Card Debt	Other Debt	Address	DebtIncomeRatio
1	41	2	6	19	0.124	1.073	NBA001	6.3
2	47	1	26	100	4.582	8.218	NBA021	12.8
3	33	2	10	57	6.111	5.802	NBA013	20.9
4	29	2	4	19	0.681	0.516	NBA009	6.3
5	47	1	31	253	9.308	8.908	NBA008	7.2
6	40	1	23	81	0.998	7.831	NBA016	10.9
7	38	2	4	56	0.442	0.454	NBA013	1.6
8	42	3	0	64	0.279	3.945	NBA009	6.6
9	26	1	5	18	0.575	2.215	NBA006	15.5
10	47	3	23	115	0.653	3.947	NBA011	4
11	44	3	8	88	0.285	5.083	NBA010	6.1
12	34	2	9	40	0.374	0.266	NBA003	1.6

The model works on its own
to discover information.

What is unsupervised learning?

Customer Id	Age	Edu	Years Employed	Income	Card Debt	Other Debt	Address	DebtIncomeRatio
1	41	2	6	19	0.124	1.073	NBA001	6.3
2	47	1	26	100	4.582	8.218	NBA021	12.8
3	33	2	10	57	6.111	5.802	NBA013	20.9
4	29	2	4	19	0.681	0.516	NBA009	6.3
5	47	1	31	253	9.308	8.908	NBA008	7.2
6	40	1	23	81	0.998	7.831	NBA016	10.9
7	38	2	4	56	0.442	0.454	NBA013	1.6
8	42	3	0	64	0.279	3.945	NBA009	6.6
9	26	1	5	18	0.575	2.215	NBA006	15.5
10	47	3	23	115	0.653	3.947	NBA011	4
11	44	3	8	88	0.285	5.083	NBA010	6.1
12	34	2	9	40	0.374	0.266	NBA003	1.6

Unsupervised learning techniques:



ALL OF THIS DATA
IS UNLABELED

The model works on its own
to discover information.

What is unsupervised learning?

Customer Id	Age	Edu	Years Employed	Income	Card Debt	Other Debt	Address	DebtIncomeRatio
1	41	2	6	19	0.124	1.073	NBA001	6.3
2	47	1	26	100	4.582	8.218	NBA021	12.8
3	33	2	10	57	6.111	5.802	NBA013	20.9
4	29	2	4	19	0.681	0.516	NBA009	6.3
5	47	1	31	253	9.308	8.908	NBA008	7.2
6	40	1	23	81	0.998	7.831	NBA016	10.9
7	38	2	4	56	0.442	0.454	NBA013	1.6
8	42	3	0	64	0.279	3.945	NBA009	6.6
9	26	1	5	18	0.575	2.215	NBA006	15.5
10	47	3	23	115	0.653	3.947	NBA011	4
11	44	3	8	88	0.285	5.083	NBA010	6.1
12	34	2	9	40	0.374	0.266	NBA003	1.6

Unsupervised learning techniques:

- Dimension reduction
- Density estimation



ALL OF THIS DATA
IS UNLABELED

The model works on its own
to discover information.

What is unsupervised learning?

Customer Id	Age	Edu	Years Employed	Income	Card Debt	Other Debt	Address	DebtIncomeRatio
1	41	2	6	19	0.124	1.073	NBA001	6.3
2	47	1	26	100	4.582	8.218	NBA021	12.8
3	33	2	10	57	6.111	5.802	NBA013	20.9
4	29	2	4	19	0.681	0.516	NBA009	6.3
5	47	1	31	253	9.308	8.908	NBA008	7.2
6	40	1	23	81	0.998	7.831	NBA016	10.9
7	38	2	4	56	0.442	0.454	NBA013	1.6
8	42	3	0	64	0.279	3.945	NBA009	6.6
9	26	1	5	18	0.575	2.215	NBA006	15.5
10	47	3	23	115	0.653	3.947	NBA011	4
11	44	3	8	88	0.285	5.083	NBA010	6.1
12	34	2	9	40	0.374	0.266	NBA003	1.6

Unsupervised learning techniques:

- Dimension reduction
- Density estimation
- Market basket analysis



ALL OF THIS DATA
IS UNLABELED

The model works on its own
to discover information.

What is unsupervised learning?

Customer Id	Age	Edu	Years Employed	Income	Card Debt	Other Debt	Address	DebtIncomeRatio
1	41	2	6	19	0.124	1.073	NBA001	6.3
2	47	1	26	100	4.582	8.218	NBA021	12.8
3	33	2	10	57	6.111	5.802	NBA013	20.9
4	29	2	4	19	0.681	0.516	NBA009	6.3
5	47	1	31	253	9.308	8.908	NBA008	7.2
6	40	1	23	81	0.998	7.831	NBA016	10.9
7	38	2	4	56	0.442	0.454	NBA013	1.6
8	42	3	0	64	0.279	3.945	NBA009	6.6
9	26	1	5	18	0.575	2.215	NBA006	15.5
10	47	3	23	115	0.653	3.947	NBA011	4
11	44	3	8	88	0.285	5.083	NBA010	6.1
12	34	2	9	40	0.374	0.266	NBA003	1.6

Unsupervised learning techniques:

- Dimension reduction
- Density estimation
- Market basket analysis
- Clustering

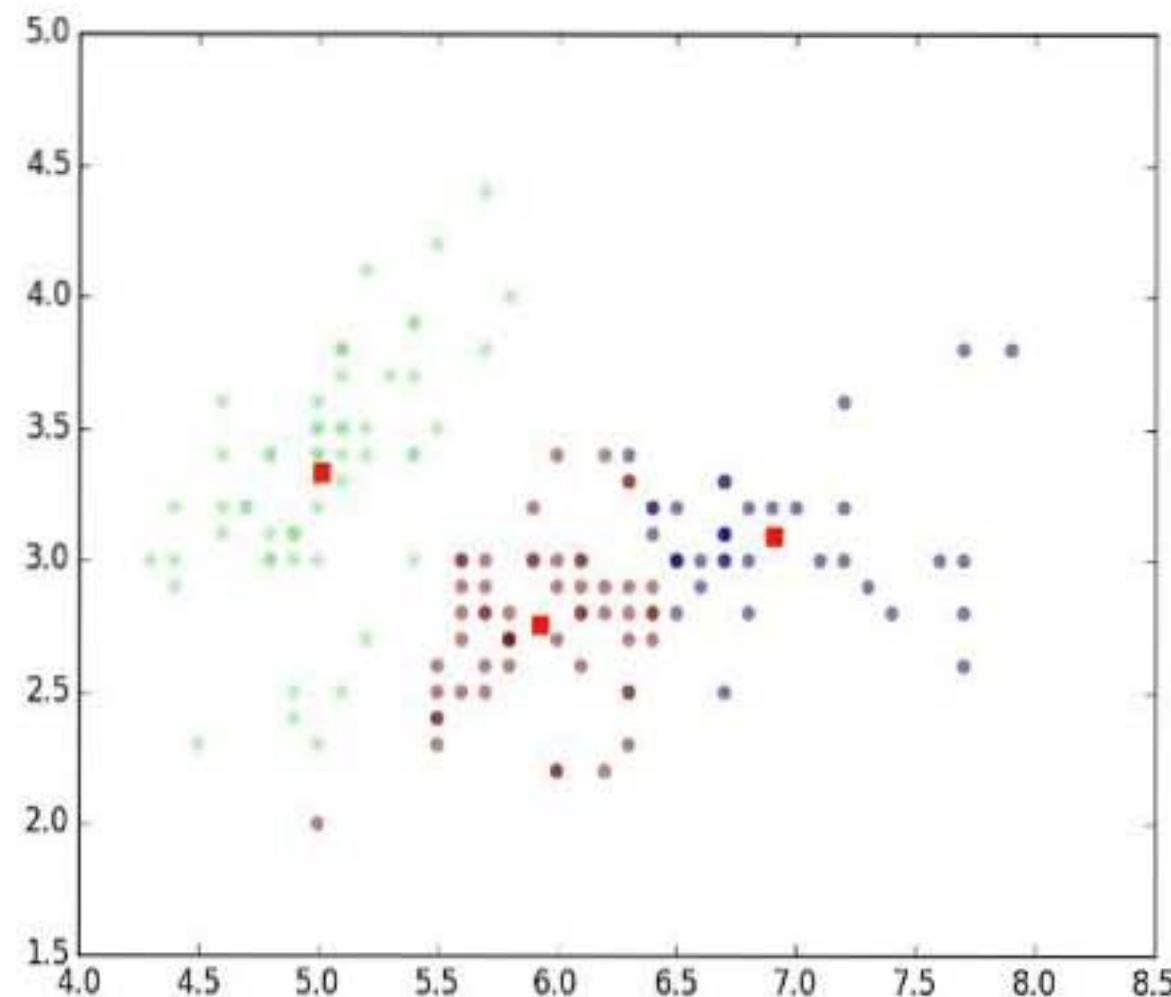
ALL OF THIS DATA
IS UNLABELED

The model works on its own
to discover information.

What is clustering?

Clustering is grouping of data points or objects that are somehow similar by:

- Discovering structure
- Summarization
- Anomaly detection



Supervised vs unsupervised learning

Supervised Learning

- **Classification:**
Classifies labeled data
- **Regression:**
Predicts trends using previous labeled data
- Has more evaluation methods than unsupervised learning
- Controlled environment

Unsupervised Learning

- **Clustering:**
Finds patterns and groupings from unlabeled data
- Has fewer evaluation methods than supervised learning
- Less controlled environment

Intro to Regression

Saeed Aghabozorgi



© IBM Corporation. All rights reserved.

1



What is regression?

	ENGINESIZE	CYLINDERS	FUELCONSUMPTION_COMB	CO2EMISSIONS
0	2.0	4	8.5	196
1	2.4	4	9.6	221
2	1.5	4	5.9	136
3	3.5	6	11.1	255
4	3.5	6	10.6	244
5	3.5	6	10.0	230
6	3.5	6	10.1	232
7	3.7	6	11.1	255
8	3.7	6	11.6	267
9	2.4	4	9.2	?

Regression is the process of predicting a continuous value



2

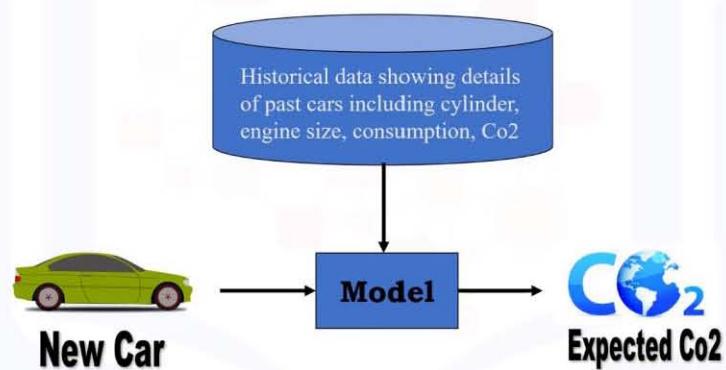


What is regression?

	X: Independent variable			Y: Dependent variable
	ENGINESIZE	CYLINDERS	FUELCONSUMPTION_COMB	CO2EMISSIONS
0	2.0	4	8.5	196
1	2.4	4	9.6	221
2	1.5	4	5.9	136
3	3.5	6	11.1	255
4	3.5	6	10.6	244
5	3.5	6	10.0	230
6	3.5	6	10.1	232
7	3.7	6	11.1	255
8	3.7	6	11.6	267
9	2.4	4	9.2	?

Regression is the process of predicting a continuous value

What is a regression model?



Types of regression models

- Simple Regression:

- Simple Linear Regression
- Simple Non-linear Regression

Predict `co2emission` vs `EngineSize` of all cars

- Multiple Regression:

- Multiple Linear Regression
- Multiple Non-linear Regression

Predict `co2emission` vs `EngineSize` and `Cylinders` of all cars

Regression algorithms

- Ordinal regression
- Poisson regression
- Fast forest quantile regression
- Linear, Polynomial, Lasso, Stepwise, Ridge regression
- Bayesian linear regression
- Neural network regression
- Decision forest regression
- Boosted decision tree regression
- KNN (K-nearest neighbors)

Simple Linear Regression

Saeed Aghabozorgi

Using linear regression to predict continuous values

X: Independent variable

Y: Dependent variable

	ENGINESIZE	CYLINDERS	FUELCONSUMPTION_COMB	CO2EMISSIONS
0	2.0	4	8.5	196
1	2.4	4	9.6	221
2	1.5	4	5.9	136
3	3.5	6	11.1	255
4	3.5	6	10.6	244
5	3.5	6	10.0	230
6	3.5	6	10.1	232
7	3.7	6	11.1	255
8	3.7	6	11.6	267
9	2.4	4	9.2	?

Linear regression topology

→ • Simple Linear Regression:

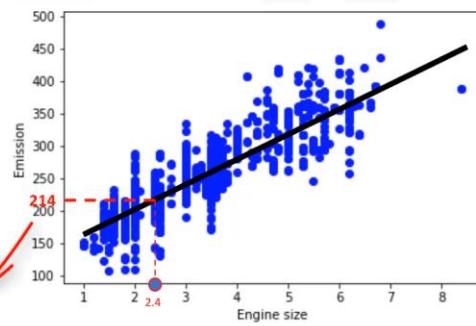
- Predict co2emission vs EngineSize of all cars
 - Independent variable (x): EngineSize
 - Dependent variable (y): co2emission

• Multiple Linear Regression:

- Predict co2emission vs EngineSize and Cylinders of all cars
 - Independent variable (x): EngineSize, Cylinders, etc
 - Dependent variable (y): co2emission

How does linear regression work?

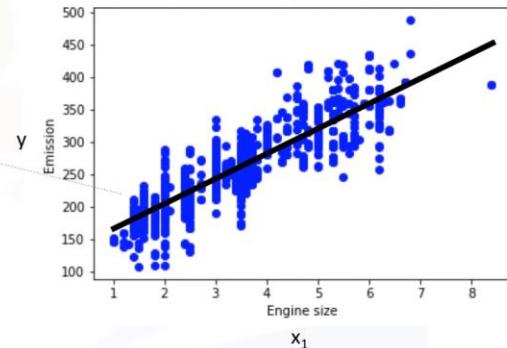
	ENGINESIZE	CYLINDERS	FUELCONSUMPTION_COMB	CO2EMISSIONS
0	2.0	4	8.5	196
1	2.4	4	9.6	221
2	1.5	4	5.9	136
3	3.5	6	11.1	255
4	3.5	6	10.6	244
5	3.5	6	10.0	230
6	3.5	6	10.1	232
7	3.7	6	11.1	255
8	3.7	6	11.6	267
9	2.4	4	9.2	?



Linear regression model representation

$$\hat{y} = \theta_0 + \theta_1 x_1$$

response variable
a single predictor



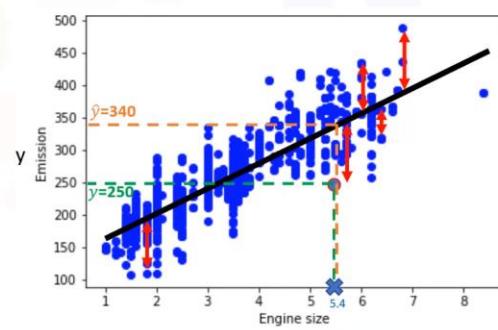
How to find the best fit?

$x_1 = 2.4$ independent variable
 $y = 250$ actual Co2 emission of x_1

$\hat{y} = \theta_0 + \theta_1 x_1$
 $\hat{y} = 340$ the predicted emission of x_1

$$\begin{aligned} \text{Error} &= y - \hat{y} \\ &= 250 - 340 \\ &= -90 \end{aligned}$$

$$MSE = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2$$



Estimating the parameters

	ENGINESIZE	CYLINDERS	FUELCONSUMPTION_COMB	CO2EMISSIONS
0	2.0	4	8.5	196
1	2.4	4	9.6	221
2	1.5	4	5.9	136
3	3.5	6	11.1	255
4	X ₁	3.5	6	10.6
5	3.5	6	10.0	244
6	3.5	6	10.1	230
7	3.7	6	11.1	232
8	3.7	6	11.6	255
9	3.7	6	11.6	267

$$\hat{y} = \theta_0 + \theta_1 x_1$$

$$\theta_1 = \frac{\sum_{i=1}^s (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^s (x_i - \bar{x})^2}$$

$$\bar{x} = (2.0 + 2.4 + 1.5 + \dots) / 9 = 3.34$$

$$\bar{y} = (196 + 221 + 136 + \dots) / 9 = 256$$

$$\theta_1 = \frac{(2.0 - 3.34)(196 - 256) + (2.4 - 3.34)(221 - 256) + \dots}{(2.0 - 3.34)^2 + (2.4 - 3.34)^2 + \dots}$$

$$\theta_1 = 39$$

$$\theta_0 = \bar{y} - \theta_1 \bar{x}$$

$$\theta_0 = 256 - 39 * 3.34$$

$$\theta_0 = 125.74$$

$$\hat{y} = 125.74 + 39x_1$$

Predictions with linear regression

	ENGINESIZE	CYLINDERS	FUELCONSUMPTION_COMB	CO2EMISSIONS
0	2.0	4	8.5	196
1	2.4	4	9.6	221
2	1.5	4	5.9	136
3	3.5	6	11.1	255
4	3.5	6	10.6	244
5	3.5	6	10.0	230
6	3.5	6	10.1	232
7	3.7	6	11.1	255
8	3.7	6	11.6	267
9	2.4	4	9.2	?

$$\hat{y} = \theta_0 + \theta_1 x_1$$

$$Co2Emission = \theta_0 + \theta_1 EngineSize$$

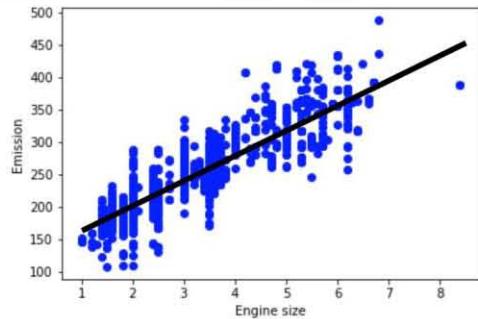
$$Co2Emission = 125 + 39 EngineSize$$

$$Co2Emission = 125 + 39 \times 2.4$$

$$Co2Emission = 218.6$$

Pros of linear regression

- Very fast
- No parameter tuning
- Easy to understand, and highly interpretable



Multiple Linear Regression

Saeed Aghabozorgi

Types of regression models

- Simple Linear Regression

- Predict Co2emission vs EngineSize of all cars
 - Independent variable (x): EngineSize
 - Dependent variable (y): Co2emission

- • Multiple Linear Regression

- Predict Co2emission vs EngineSize and Cylinders of all cars
 - Independent variable (x): EngineSize, Cylinders, etc.
 - Dependent variable (y): Co2emission

Examples of multiple linear regression

- Independent variables effectiveness on prediction
 - Does revision time, test anxiety, lecture attendance and gender have any effect on the exam performance of students?
- • Predicting impacts of changes
 - How much does blood pressure go up (or down) for every unit increase (or decrease) in the BMI of a patient?

Predicting continuous values with multiple linear regression

$$Co2 Em = \theta_0 + \theta_1 Engine\ size + \theta_2 Cylinders + \dots$$

$$\hat{y} = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \dots + \theta_n x_n$$

$$\hat{y} = \theta^T X$$

$$\theta^T = [\theta_0, \theta_1, \theta_2, \dots] \quad X = \begin{bmatrix} 1 \\ x_1 \\ x_2 \\ \dots \end{bmatrix}$$

X: Independent variable Y: Dependent variable

	ENGINESIZE	CYLINDERS	FUELCONSUMPTION_COMB	CO2EMISSIONS
0	2.0	4	8.5	196
1	2.4	4	9.6	221
2	1.5	4	5.9	136
3	3.5	6	11.1	255
4	3.5	6	10.6	244
5	3.5	6	10.0	230
6	3.5	6	10.1	232
7	3.7	6	11.1	255
8	3.7	6	11.6	267
9	2.4	4	9.2	?

Using MSE to expose the errors in the model

$$\hat{y} = \theta^T X$$

$\hat{y}_i = 140$ the predicted emission of x_i

$y_i = 196$ actual value of x_i

$y_i - \hat{y}_i = 196 - 140 = 56$ residual error

$$MSE = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

	ENGINESIZE	CYLINDERS	FUELCONSUMPTION_COMB	CO2EMISSIONS
0	2.0	4	8.5	196
1	2.4	4	9.6	221
2	1.5	4	5.9	136
3	3.5	6	11.1	255
4	3.5	6	10.6	244
5	3.5	6	10.0	230
6	3.5	6	10.1	232
7	3.7	6	11.1	255
8	3.7	6	11.6	267

Estimating multiple linear regression parameters

- How to estimate θ ?
 - Ordinary Least Squares
 - Linear algebra operations
 - Takes a long time for large datasets (10K+ rows)
 - An optimization algorithm
 - Gradient Descent
 - Proper approach if you have a very large dataset

Making predictions with multiple linear regression

	ENGINESIZE	CYLINDERS	FUELCONSUMPTION_COMB	CO2EMISSIONS	
0	2.0	4	8.5	196	$\hat{y} = \theta^T X$
1	2.4	4	9.6	221	$\theta^T = [125, 6.2, 14, \dots]$
2	1.5	4	5.9	136	$\hat{y} = 125 + 6.2x_1 + 14x_2 +$
3	3.5	6	11.1	255	$Co2Em = 125 + 6.2EngSize + 14 \text{ Cylinders} + \dots$
4	3.5	6	10.6	244	$Co2Em = 125 + 6.2 \times 2.4 + 14 \times 4 + \dots$
5	3.5	6	10.0	230	$Co2Em = 214.1$
6	3.5	6	10.1	232	
7	3.7	6	11.1	255	
8	3.7	6	11.6	267	
9	2.4	4	9.2	?	

Q&A – on multiple linear regression

- How to determine whether to use simple or multiple linear regression?
- How many independent variables should you use?
- Should the independent variable be continuous?
- What are the linear relationships between the dependent variable and the independent variables?

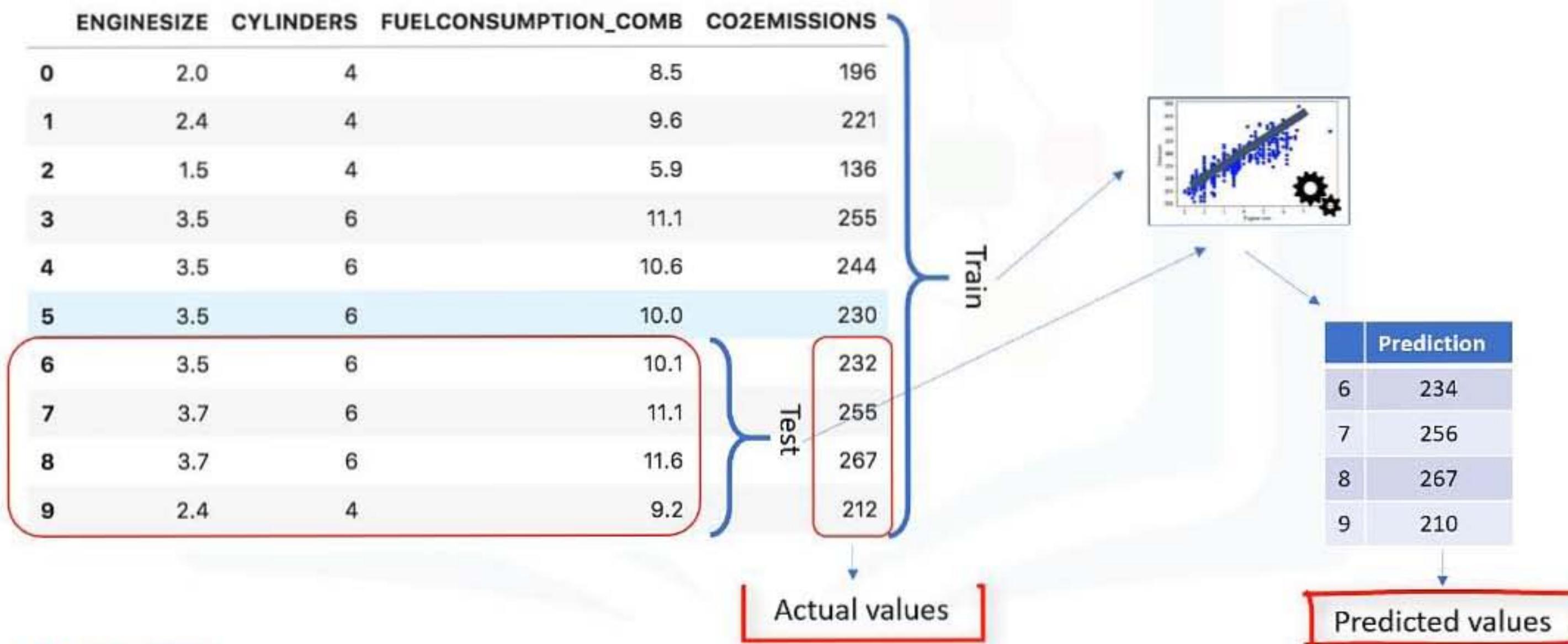
Model evaluation approaches

- Train and Test on the Same Dataset
- Train/Test Split

Regression Evaluation Metrics



Best approach for most accurate results?



Calculating the accuracy of a model

	ENGINESIZE	CYLINDERS	FUELCONSUMPTION_COMB	CO2EMISSIONS
0	2.0	4	8.5	196
1	2.4	4	9.6	221
2	1.5	4	5.9	136
3	3.5	6	11.1	255
4	3.5	6	10.6	244
5	3.5	6	10.0	230
6	3.5	6	10.1	
7	3.7	6	11.1	
8	3.7	6	11.6	
9	2.4	4	9.2	

Test

y

Actual values

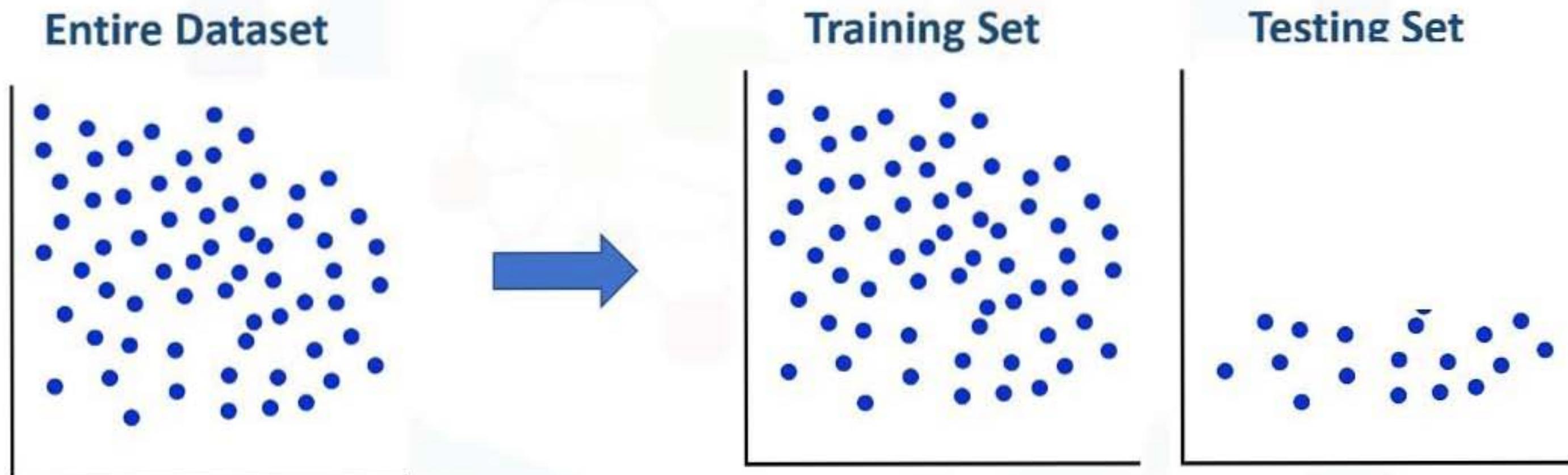
$$\text{Error} = \frac{1}{n} \sum_{j=1}^n |y_j - \hat{y}_j|$$

	Prediction
6	234
7	256
8	267
9	210

\hat{y}

Predicted values

Train and test on the same dataset



High “training accuracy”
Low “out-of-sample accuracy”

What is training & out-of-sample accuracy?

- **Training Accuracy**

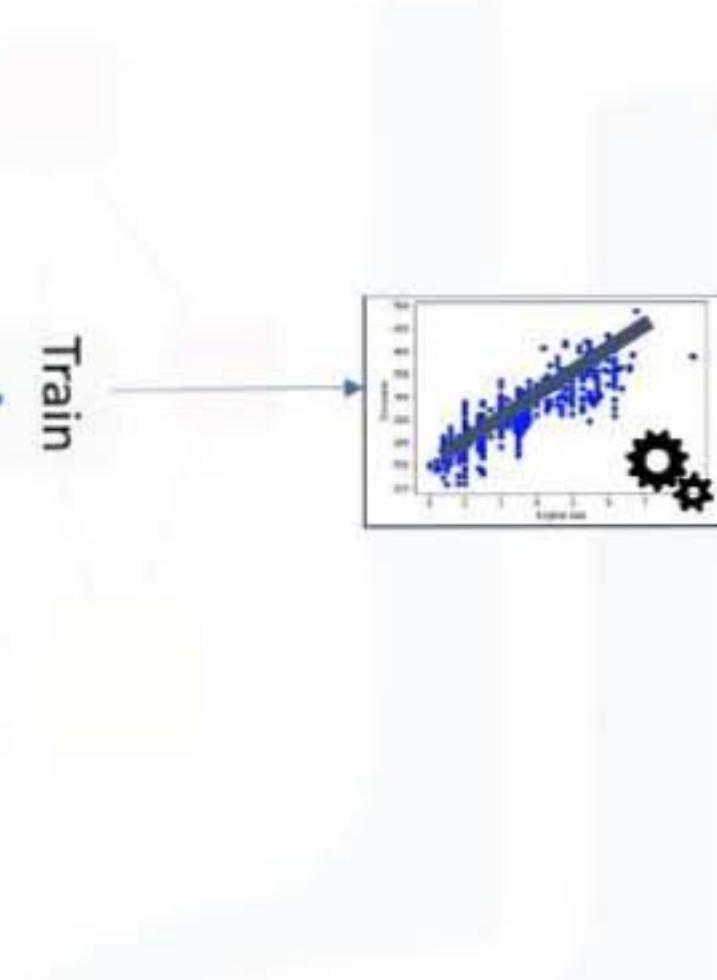
- High training accuracy isn't necessarily a good thing
- Result of over-fitting
 - **Over-fit:** the model is overly trained to the dataset, which may capture noise and produce a non-generalized model

- **Out-of-Sample Accuracy**

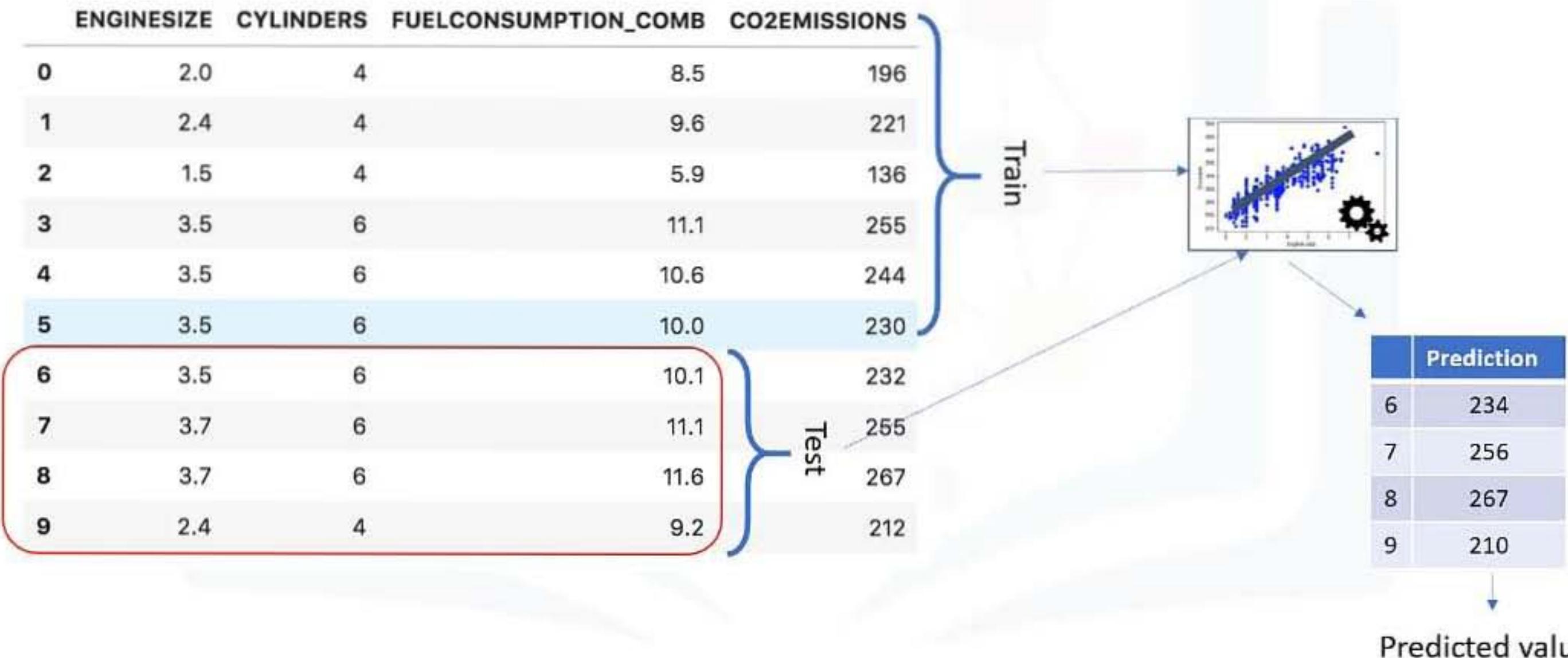
- It's important that our models have a high, out-of-sample accuracy
- How can we improve out-of-sample accuracy?

Train/Test split evaluation approach

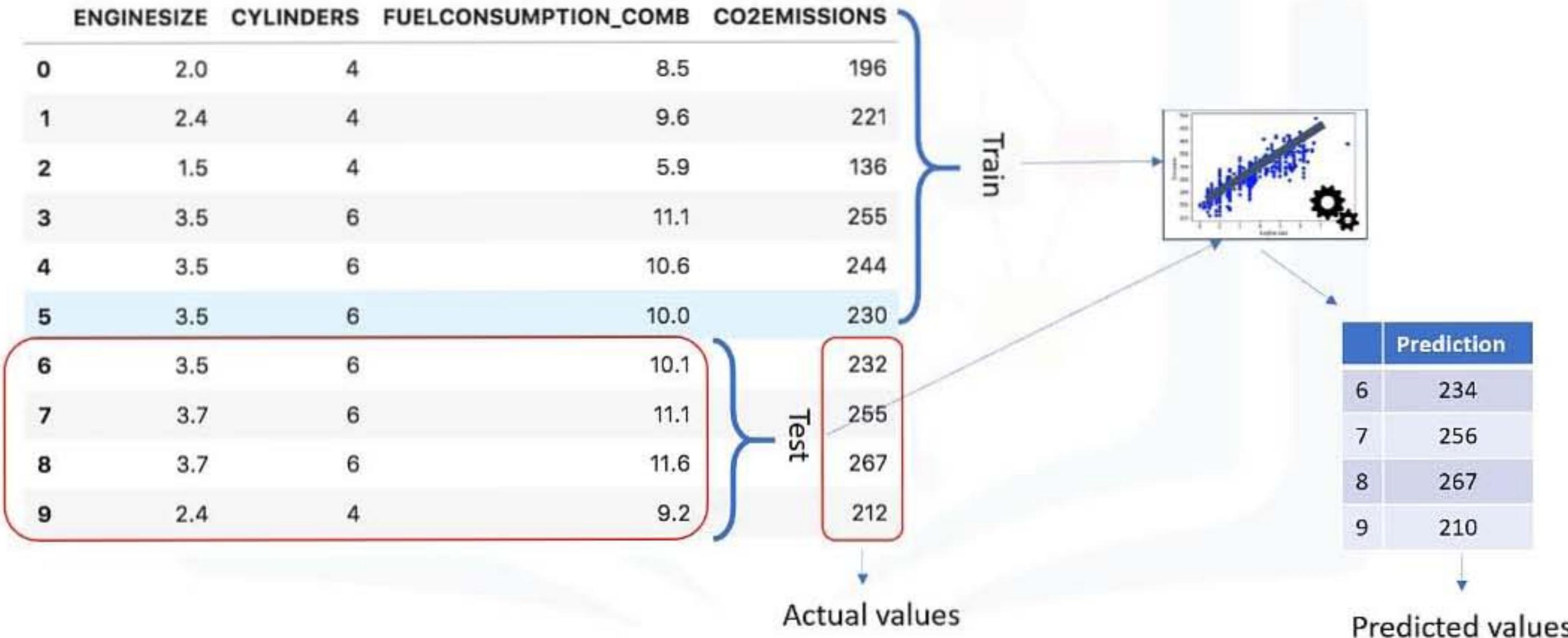
	ENGINESIZE	CYLINDERS	FUELCONSUMPTION_COMB	CO2EMISSIONS
0	2.0	4	8.5	196
1	2.4	4	9.6	221
2	1.5	4	5.9	136
3	3.5	6	11.1	255
4	3.5	6	10.6	244
5	3.5	6	10.0	230
6	3.5	6	10.1	232
7	3.7	6	11.1	255
8	3.7	6	11.6	267
9	2.4	4	9.2	212



Train/Test split evaluation approach

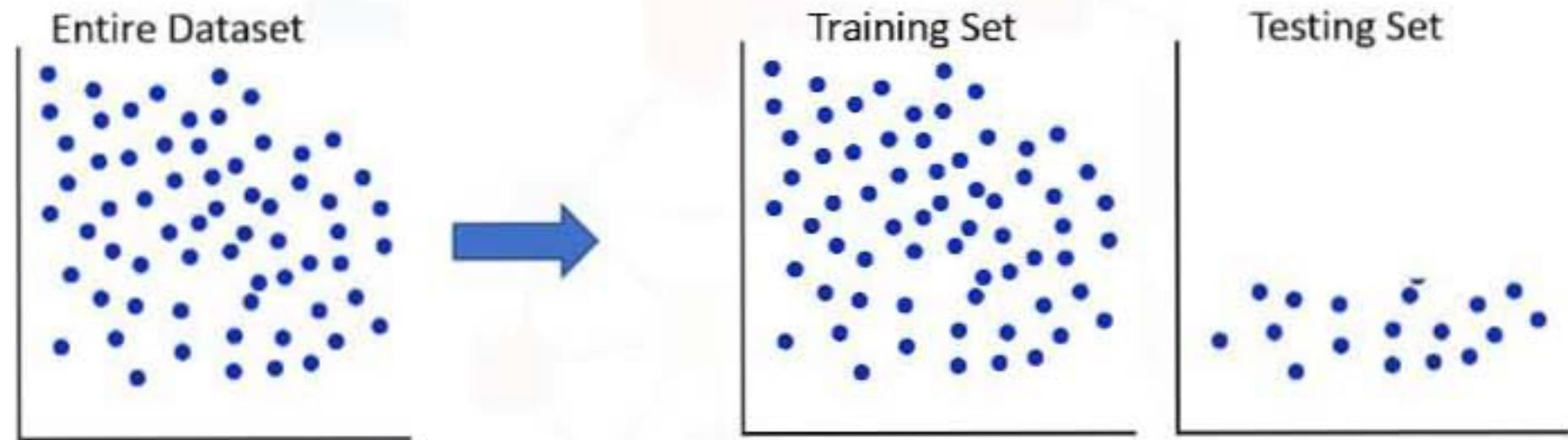


Train/Test split evaluation approach



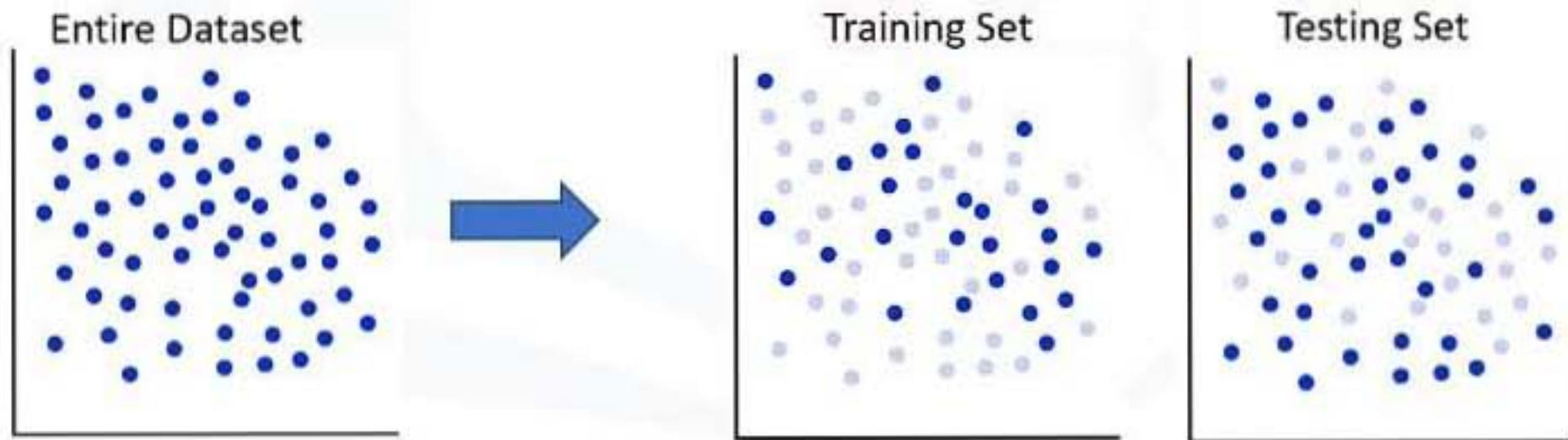
Train/Test split evaluation approach

Test on a portion of train set



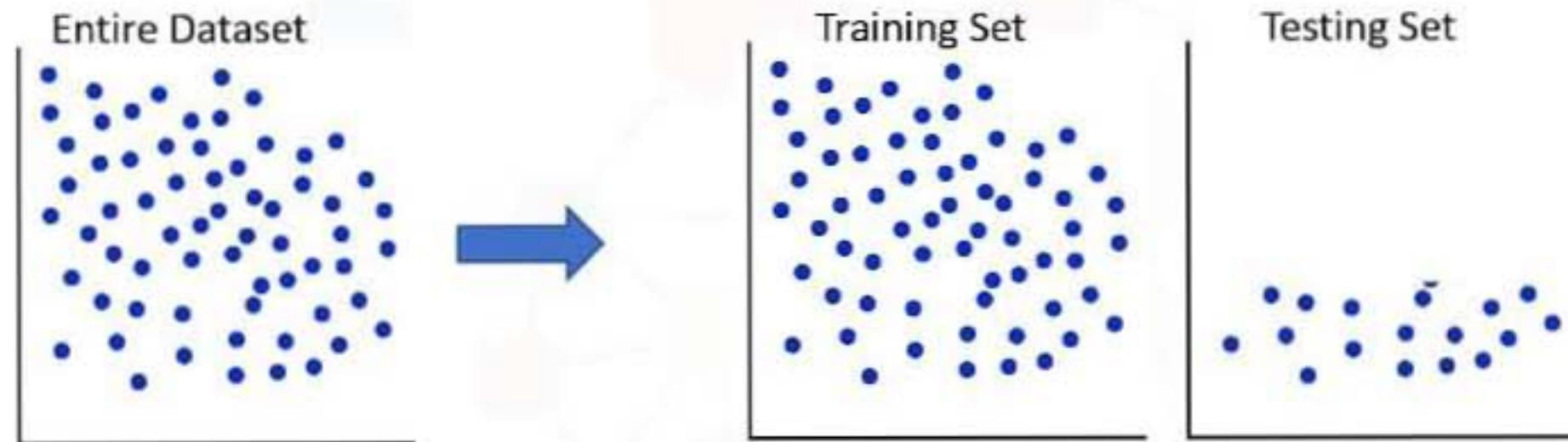
- Test-set is a portion of the train-set
- High “training accuracy”
- Low “out-of-sample accuracy”

Train/Test Split



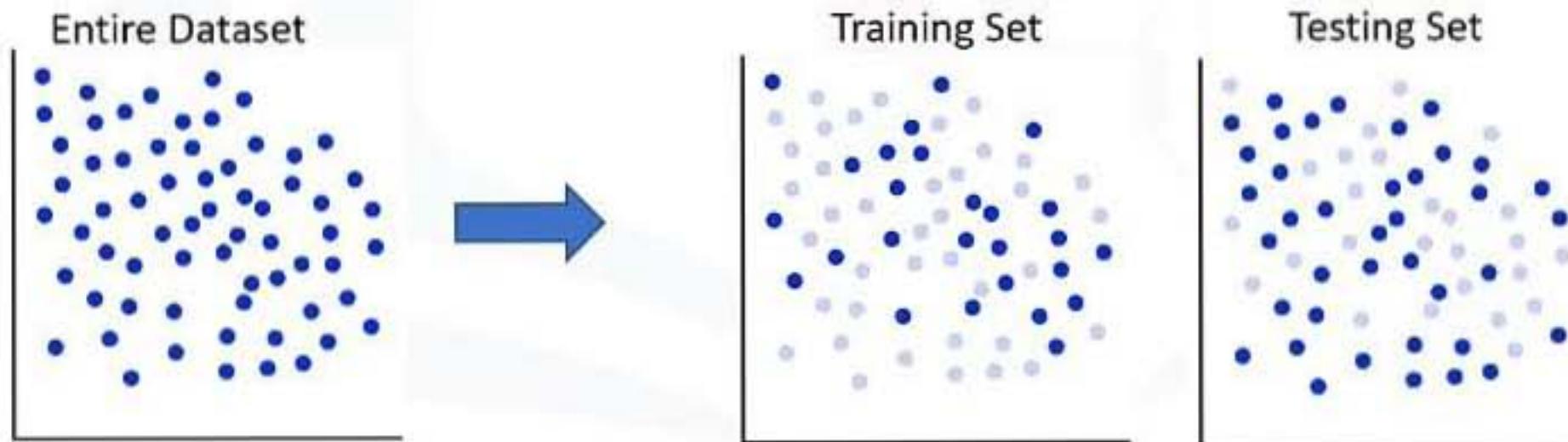
Train/Test split evaluation approach

Test on a portion of train set



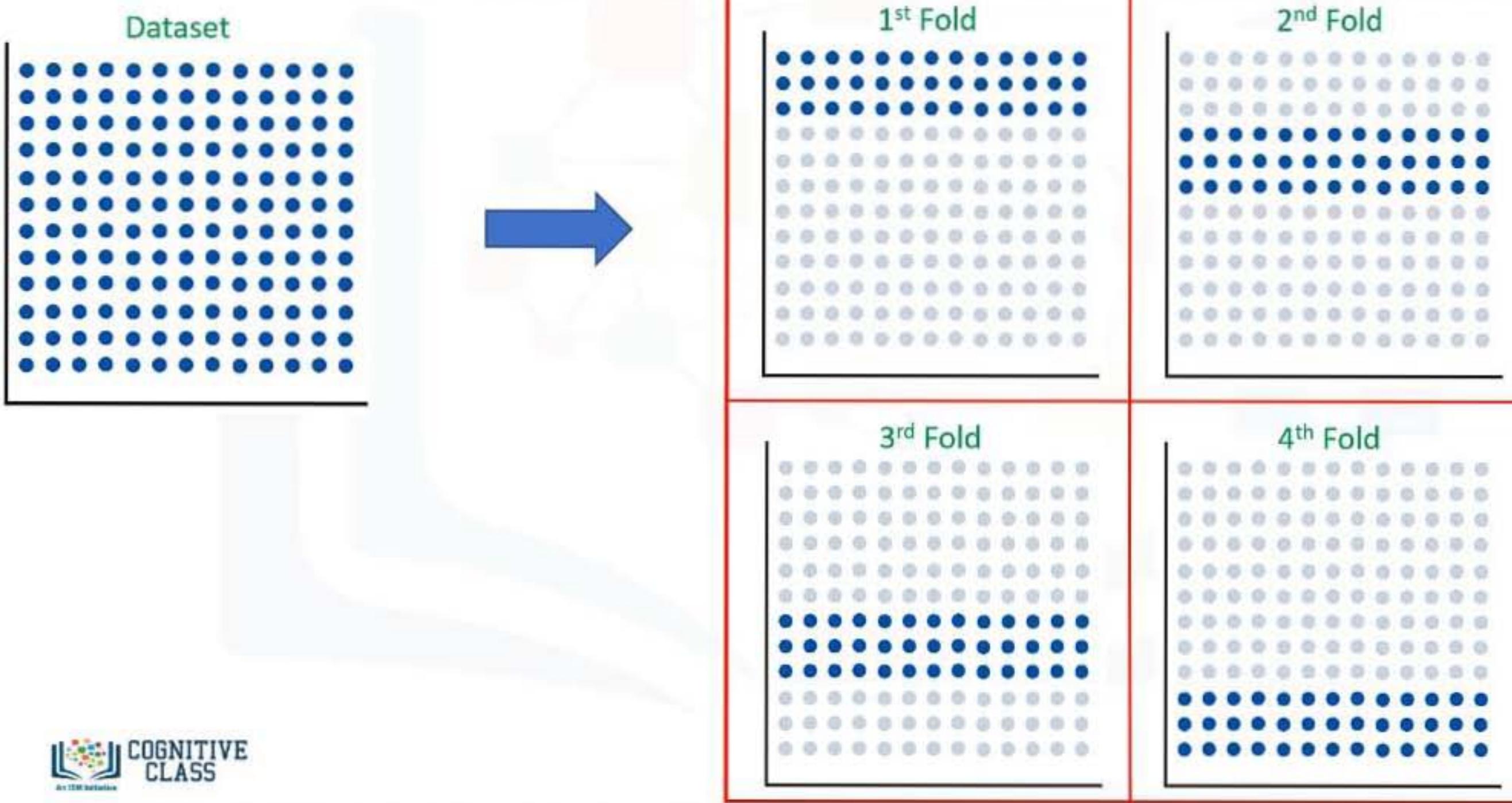
- Test-set is a portion of the train-set
- High “training accuracy”
- Low “out-of-sample accuracy”

Train/Test Split

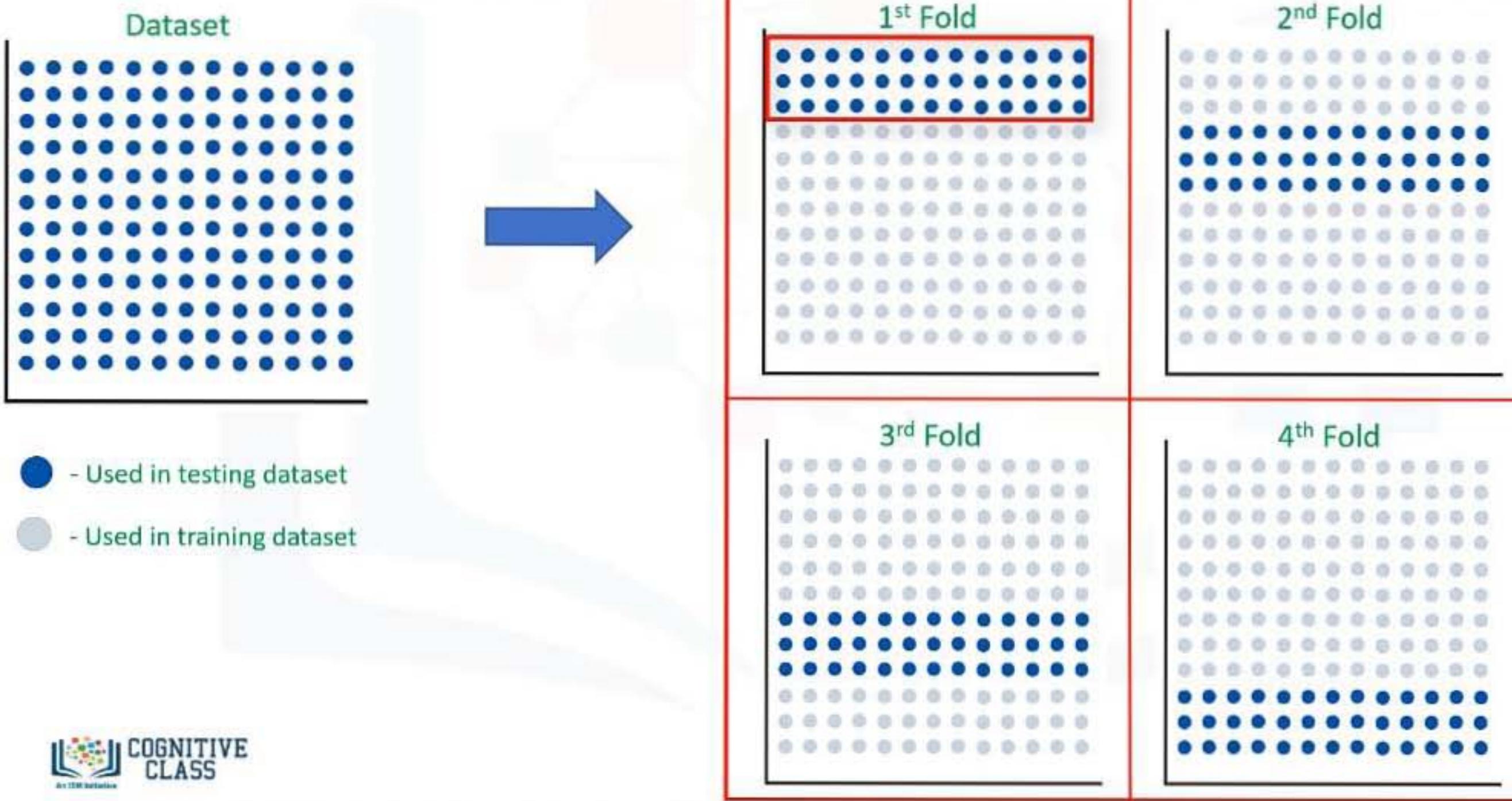


- Mutually exclusive
- More accurate evaluation on out-of-sample accuracy
- Highly dependent on which datasets the data is trained and tested

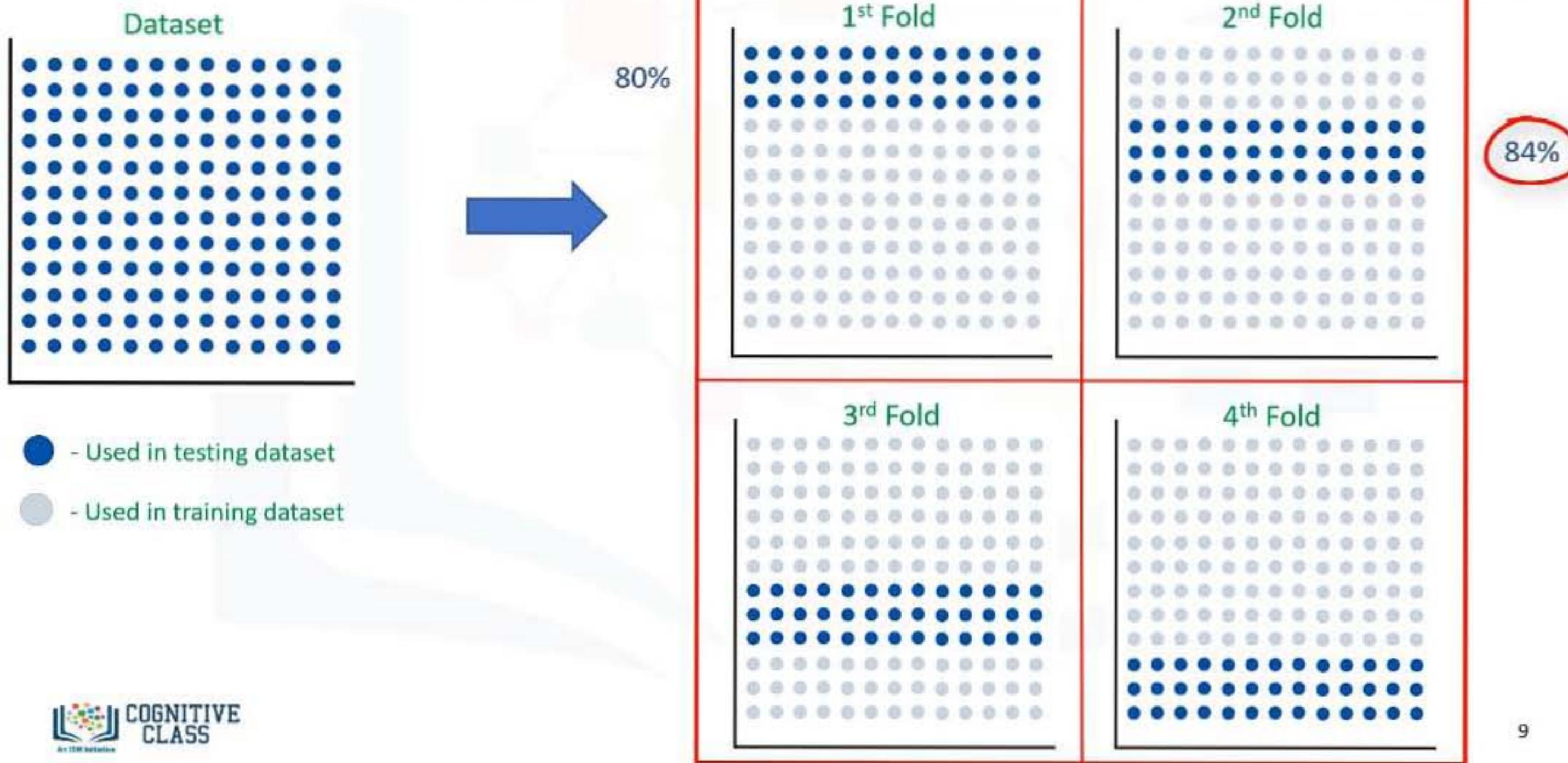
How to use K-fold cross-validation?



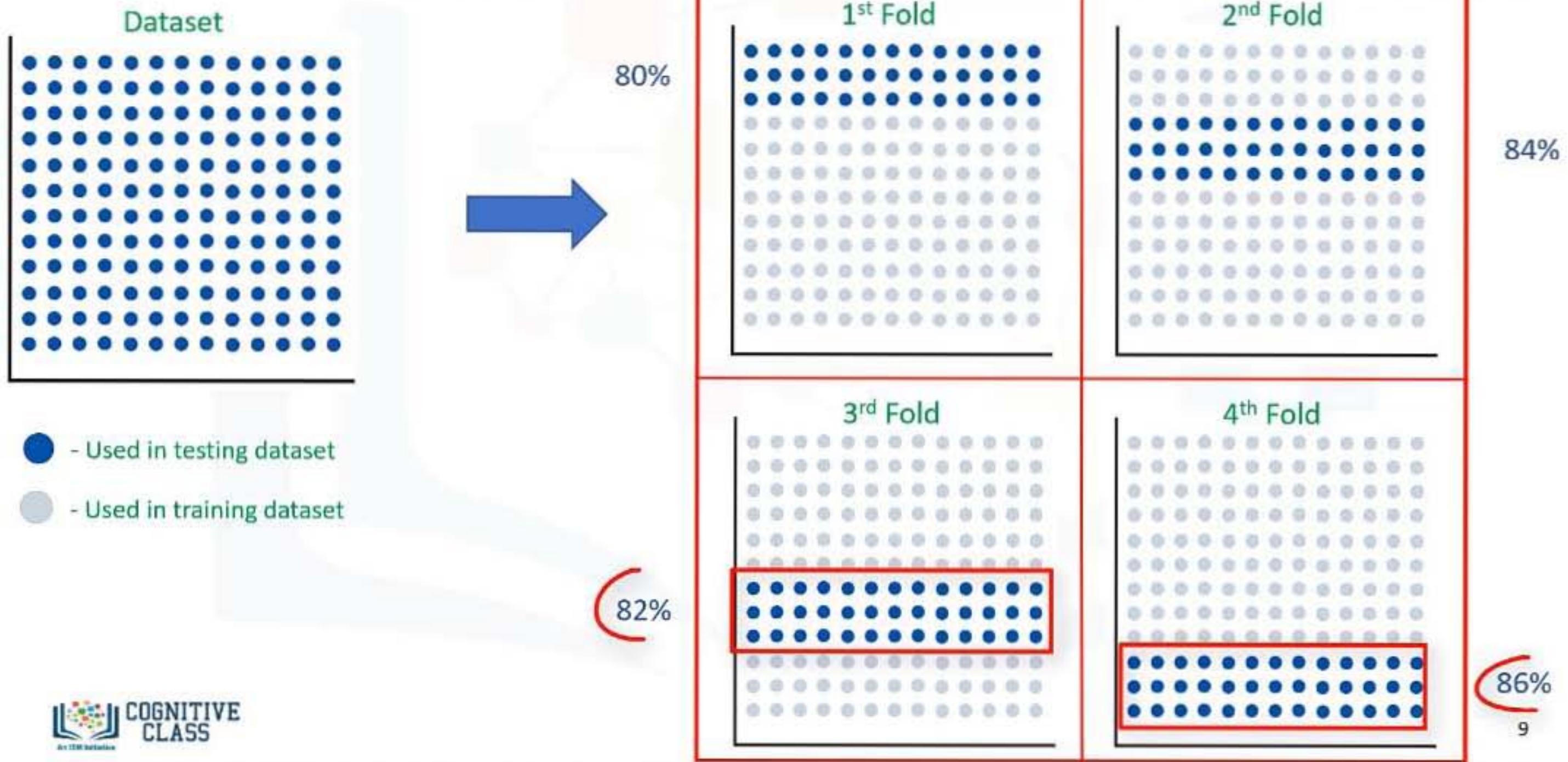
How to use K-fold cross-validation?



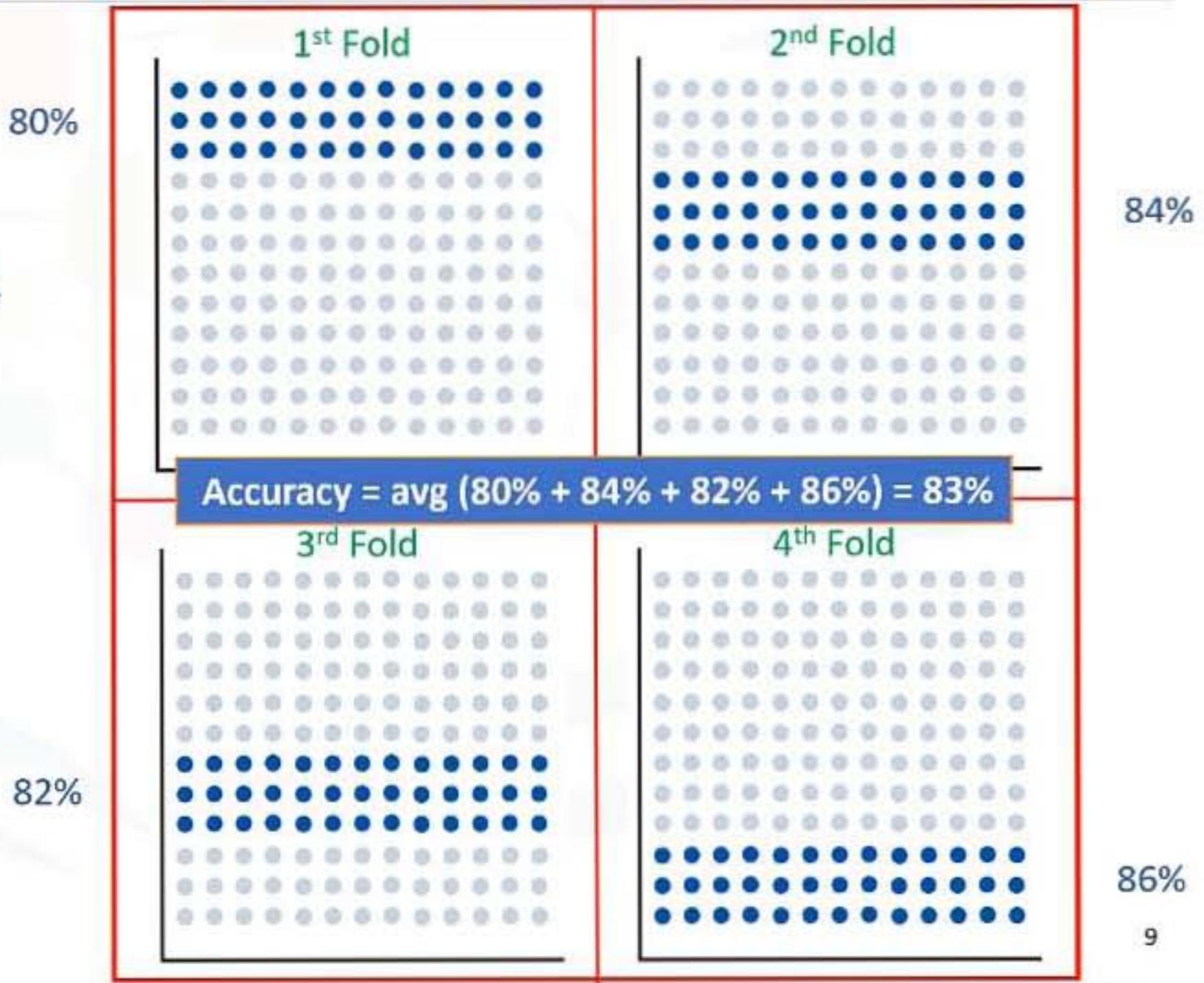
How to use K-fold cross-validation?



How to use K-fold cross-validation?



How to use K-fold cross-validation?



Regression accuracy

	ENGINESIZE	CYLINDERS	FUELCONSUMPTION_COMB	CO2EMISSIONS
0	2.0	4	8.5	196
1	2.4	4	9.6	221
2	1.5	4	5.9	136
3	3.5	6	11.1	255
4	3.5	6	10.6	244
5	3.5	6	10.0	230
6	3.5	6	10.1	232
7	3.7	6	11.1	255
8	3.7	6	11.6	267
9	2.4	4	9.2	212

Test

y

Actual values

$$\text{Error} = \frac{(232 - 234) + (255 - 256) + \dots}{4}$$

$$\text{Error} = \frac{1}{n} \sum_{j=1}^n |y_j - \hat{y}_j|$$

	Prediction
6	234
7	256
8	267
9	210

\hat{y}

Predicted values

Regression accuracy

	ENGINESIZE	CYLINDERS	FUELCONSUMPTION_COMB	CO2EMISSIONS
0	2.0	4	8.5	196
1	2.4	4	9.6	221
2	1.5	4	5.9	136
3	3.5	6	11.1	255
4	3.5	6	10.6	244
5	3.5	6	10.0	230
6	3.5	6	10.1	232
7	3.7	6	11.1	255
8	3.7	6	11.6	267
9	2.4	4	9.2	212

Actual values

- MAE
- MSE
- RMSE
- ...

\hat{y}

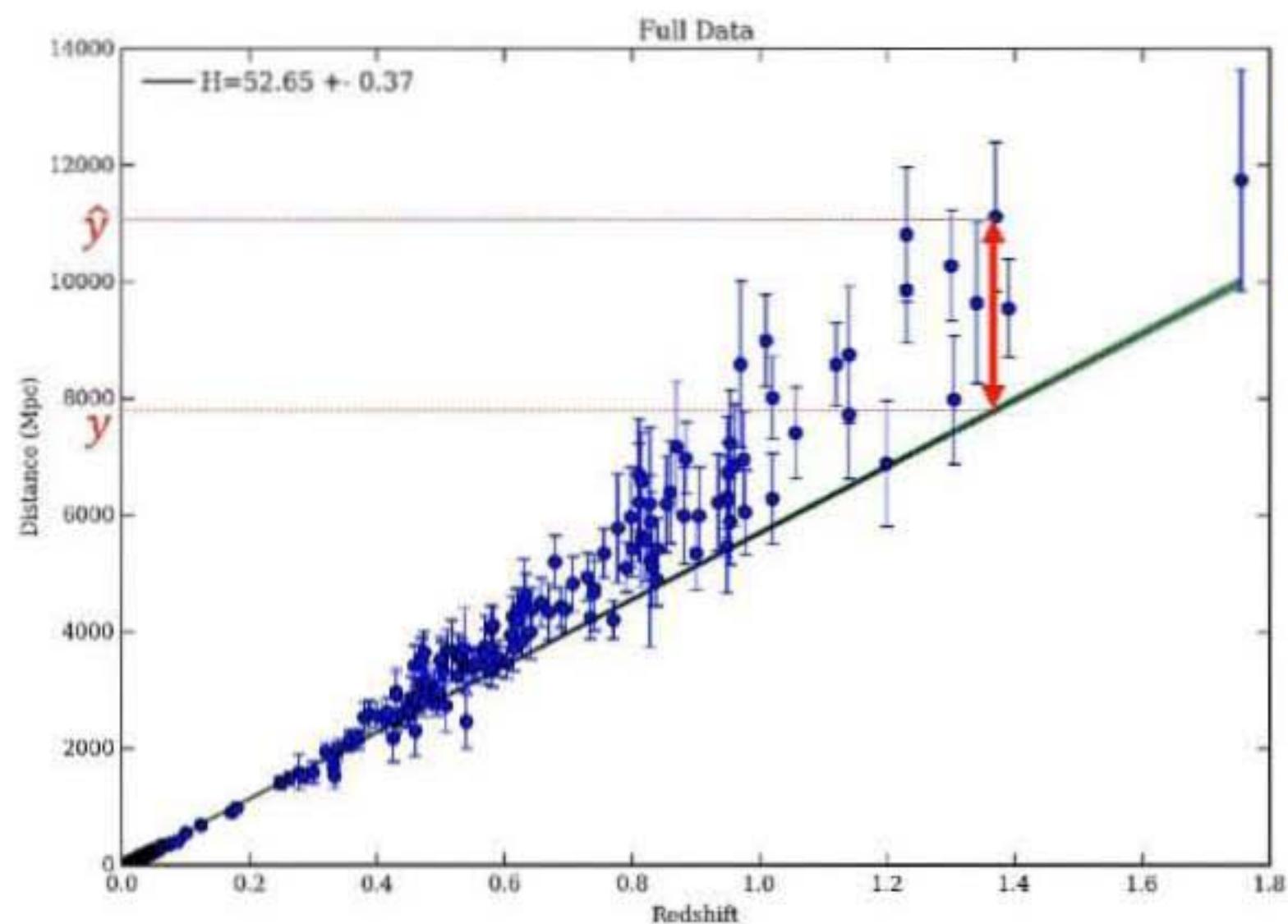
	Prediction
6	234
7	256
8	267
9	210

Predicted values

$$\text{Error} = \frac{1}{n} \sum_{j=1}^n |y_j - \hat{y}_j|$$

$$\text{Error} = \frac{(232 - 234) + (255 - 256) + \dots}{4}$$

What is an error of the model?

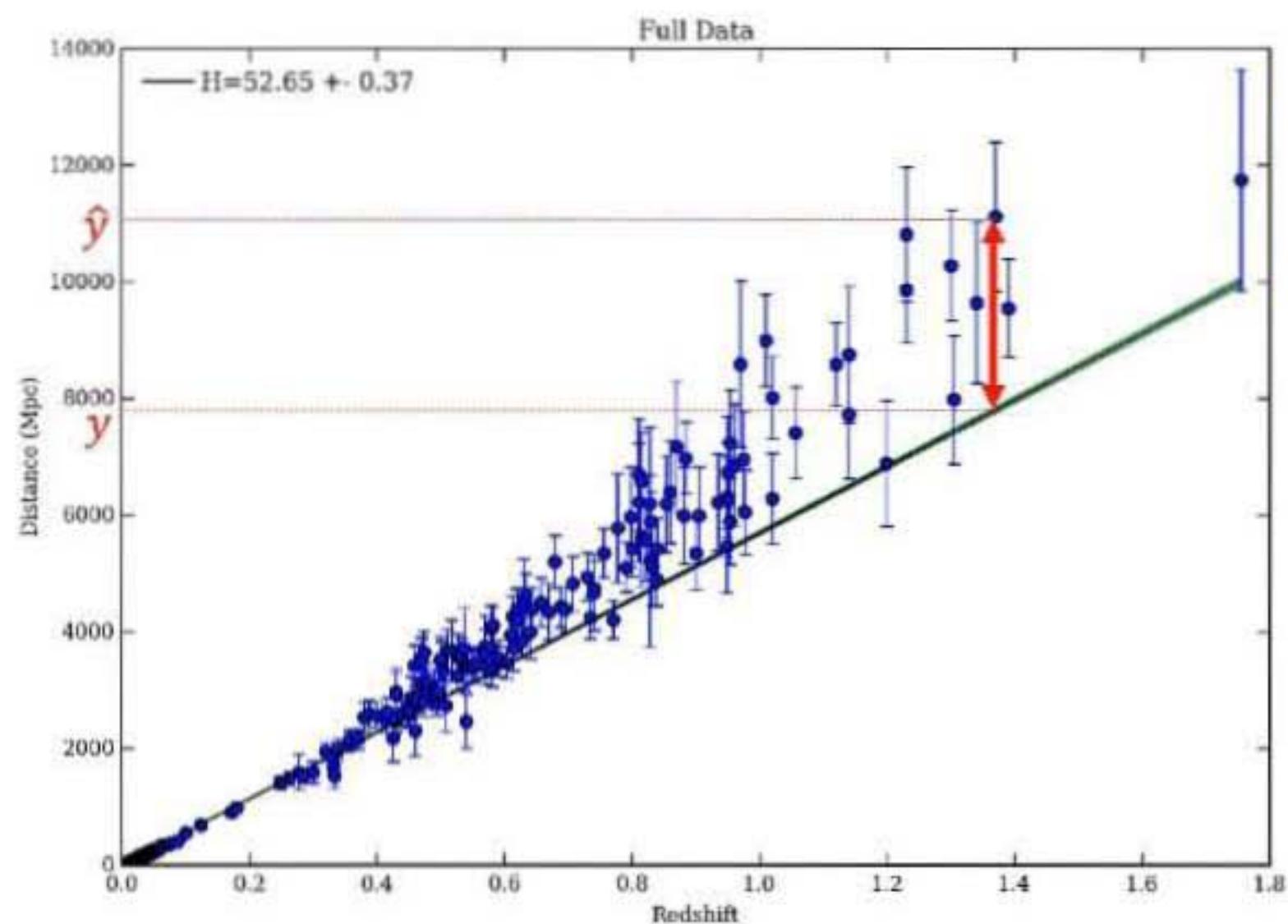


$$MAE = \frac{1}{n} \sum_{j=1}^n |y_j - \hat{y}_j|$$

$$MSE = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

$$RMSE = \sqrt{\frac{1}{n} \sum_{j=1}^n (y_j - \hat{y}_j)^2}$$

What is an error of the model?



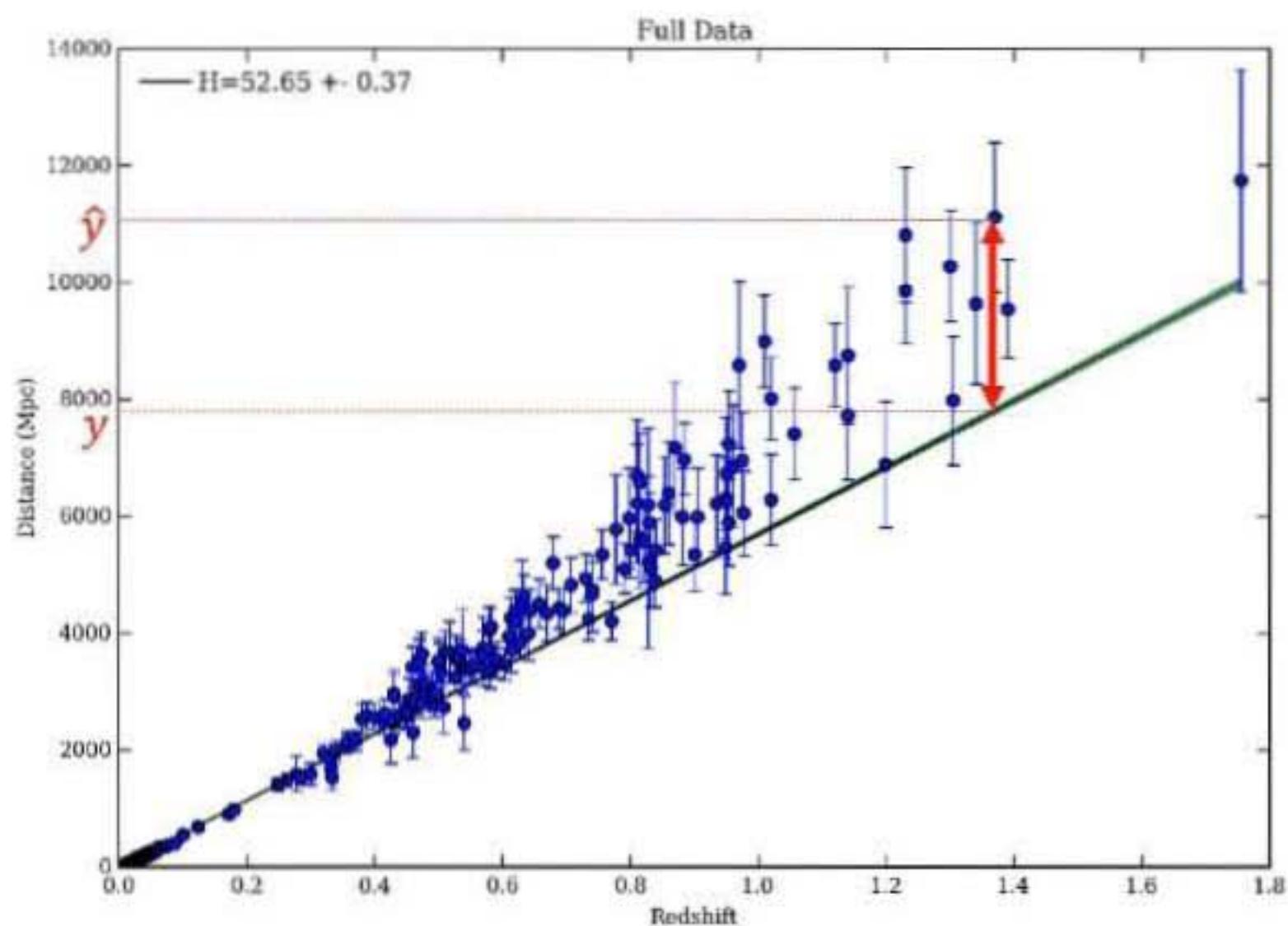
$$MAE = \frac{1}{n} \sum_{j=1}^n |y_j - \hat{y}_j|$$

$$MSE = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

$$RMSE = \sqrt{\frac{1}{n} \sum_{j=1}^n (y_j - \hat{y}_j)^2}$$

$$RAE = \frac{\sum_{j=1}^n |y_j - \hat{y}_j|}{\sum_{j=1}^n |y_j - \bar{y}|}$$

What is an error of the model?



$$MAE = \frac{1}{n} \sum_{j=1}^n |y_j - \hat{y}_j|$$

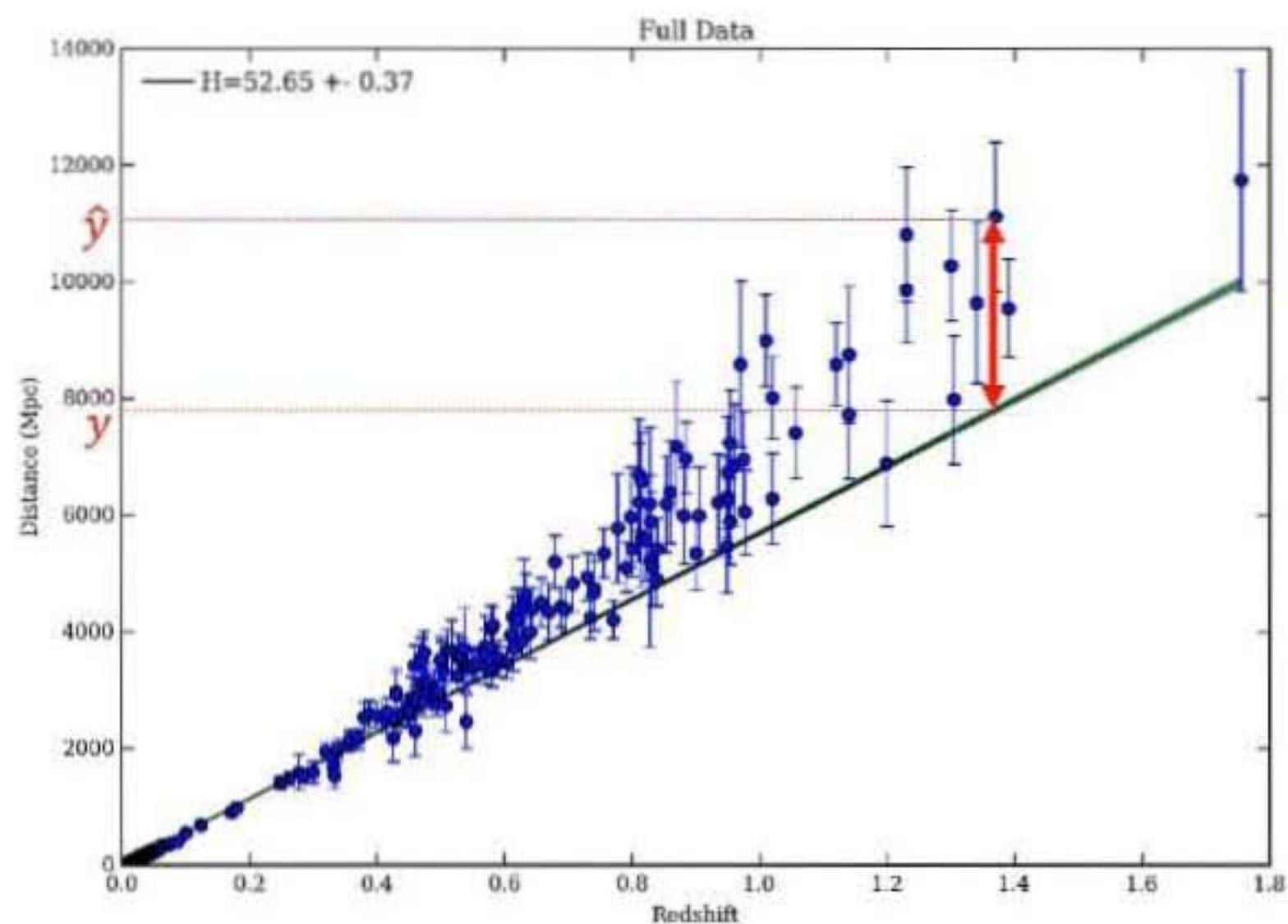
$$MSE = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

$$RMSE = \sqrt{\frac{1}{n} \sum_{j=1}^n (y_j - \hat{y}_j)^2}$$

$$RAE = \frac{\sum_{j=1}^n |y_j - \hat{y}_j|}{\sum_{j=1}^n |y_j - \bar{y}|}$$

$$RSE = \frac{\sum_{j=1}^n (y_j - \hat{y}_j)^2}{\sum_{j=1}^n (y_j - \bar{y})^2}$$

What is an error of the model?



$$MAE = \frac{1}{n} \sum_{j=1}^n |y_j - \hat{y}_j|$$

$$MSE = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

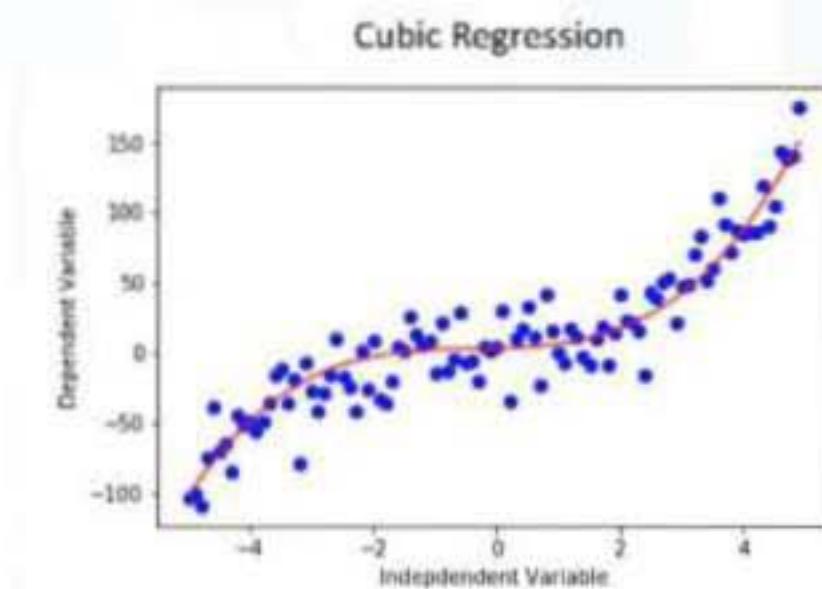
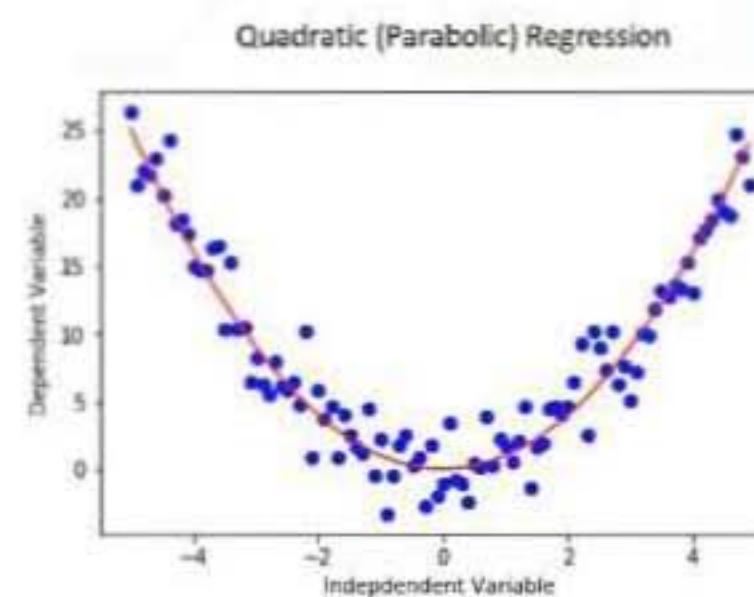
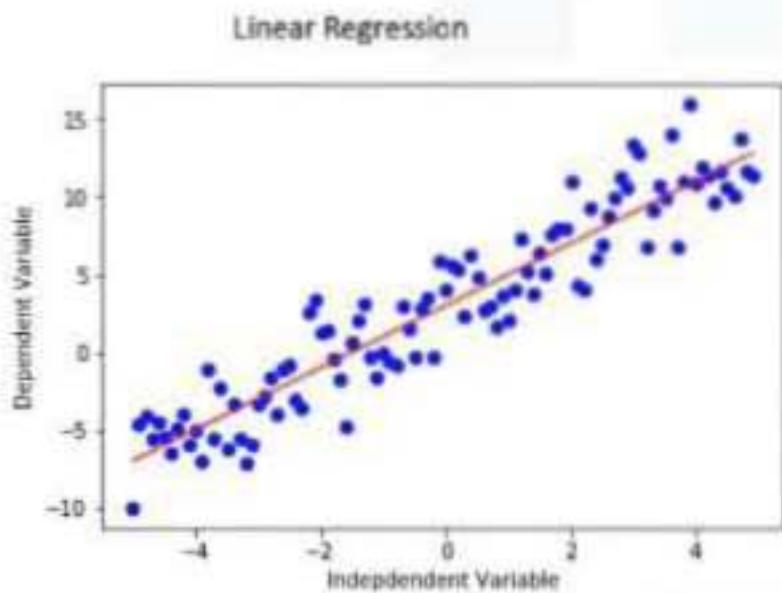
$$RMSE = \sqrt{\frac{1}{n} \sum_{j=1}^n (y_j - \hat{y}_j)^2}$$

$$RAE = \frac{\sum_{j=1}^n |y_j - \hat{y}_j|}{\sum_{j=1}^n |y_j - \bar{y}|}$$

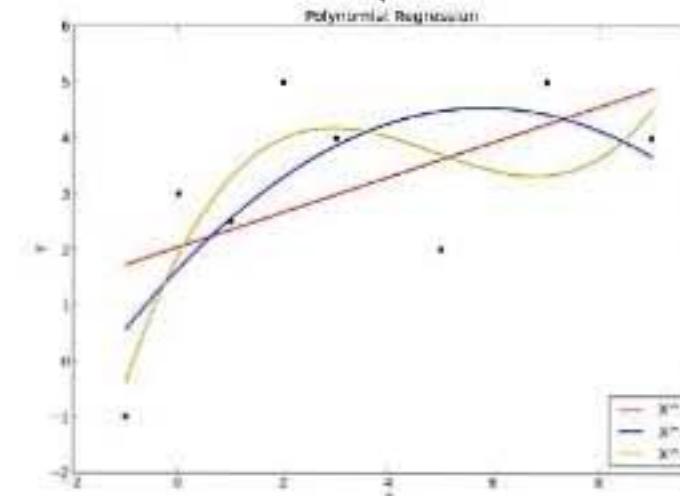
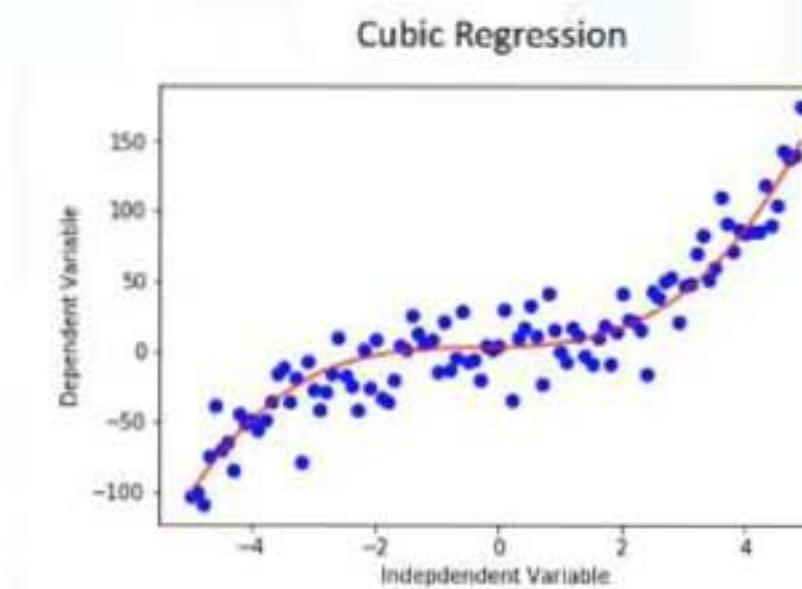
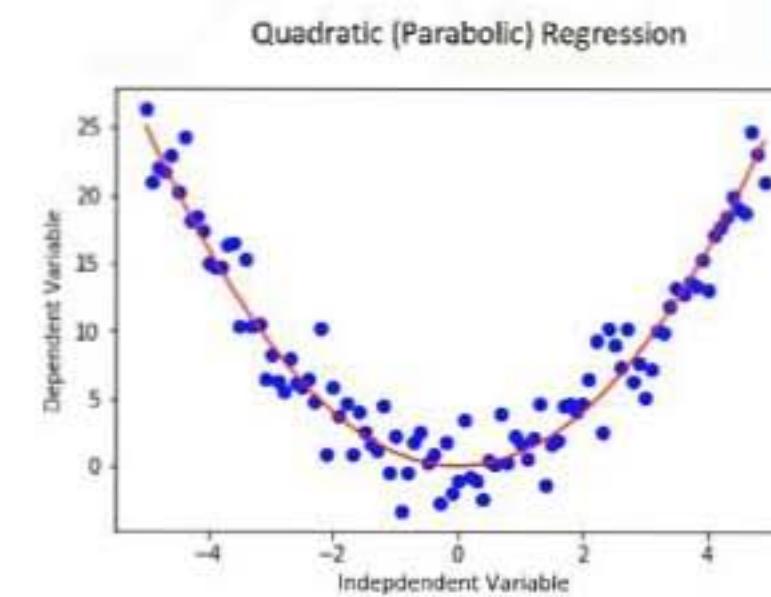
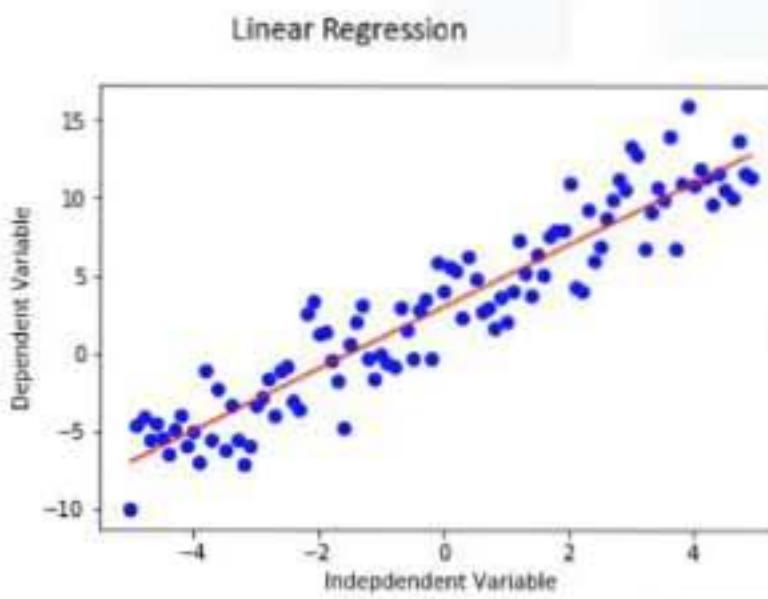
$$RSE = \frac{\sum_{j=1}^n (y_j - \hat{y}_j)^2}{\sum_{j=1}^n (y_j - \bar{y})^2}$$

$$R^2 = 1 - RSE$$

Different types of regression



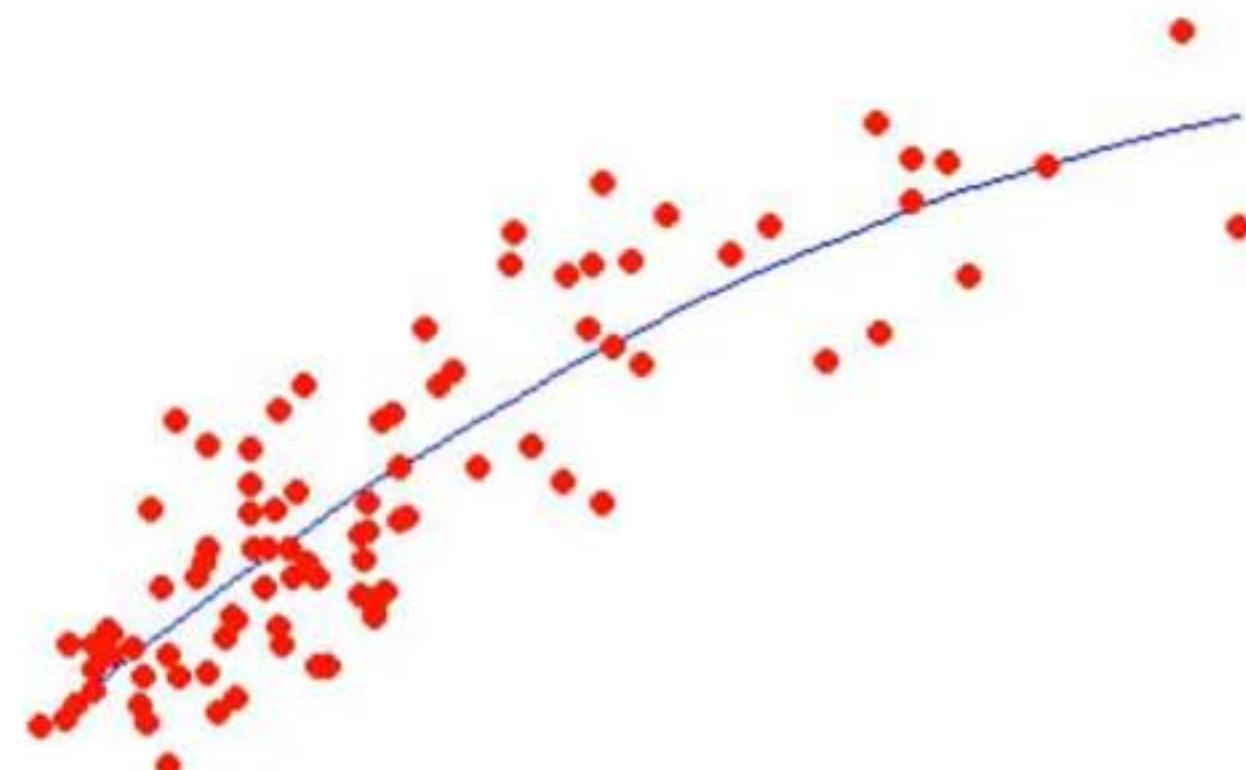
Different types of regression



What is polynomial regression?

- Some curvy data can be modeled by a **polynomial regression**
- For example:

$$\hat{y} = \theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3$$

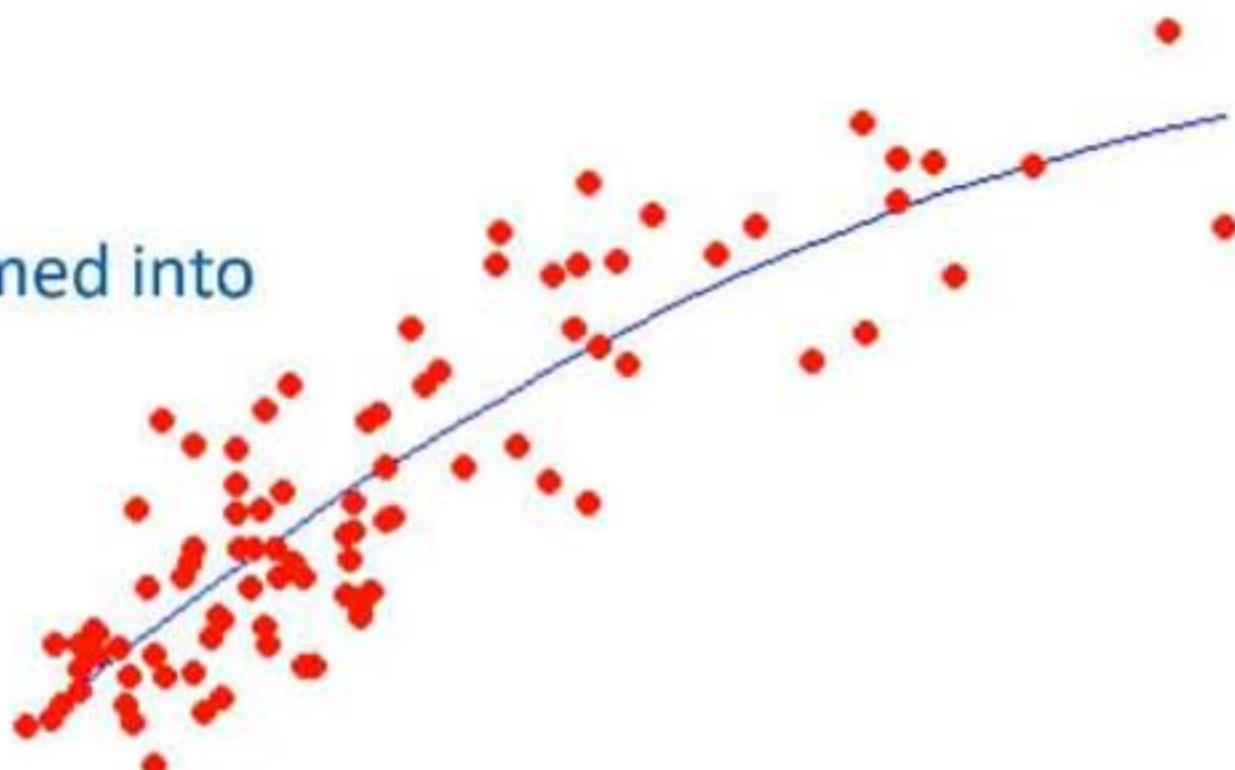


What is polynomial regression?

- Some curvy data can be modeled by a **polynomial regression**
- For example:

$$\hat{y} = \theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3$$

- A polynomial regression model can be transformed into linear regression model.



What is polynomial regression?

- Some curvy data can be modeled by a **polynomial regression**
- For example:

$$\hat{y} = \theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3$$

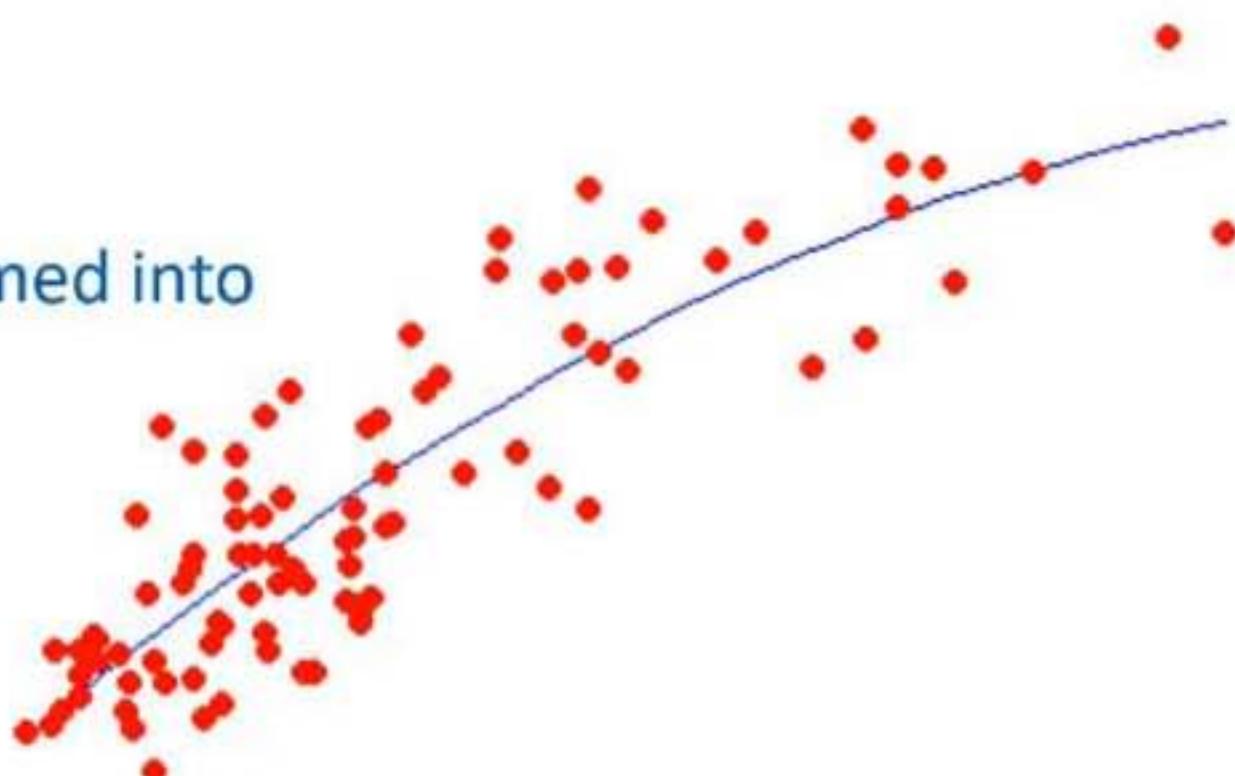
- A polynomial regression model can be transformed into linear regression model.

$$x_1 = x$$

$$x_2 = x^2$$

$$x_3 = x^3$$

$$\hat{y} = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_3$$



What is polynomial regression?

- Some curvy data can be modeled by a **polynomial regression**
- For example:

$$\hat{y} = \theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3$$

- A polynomial regression model can be transformed into linear regression model.

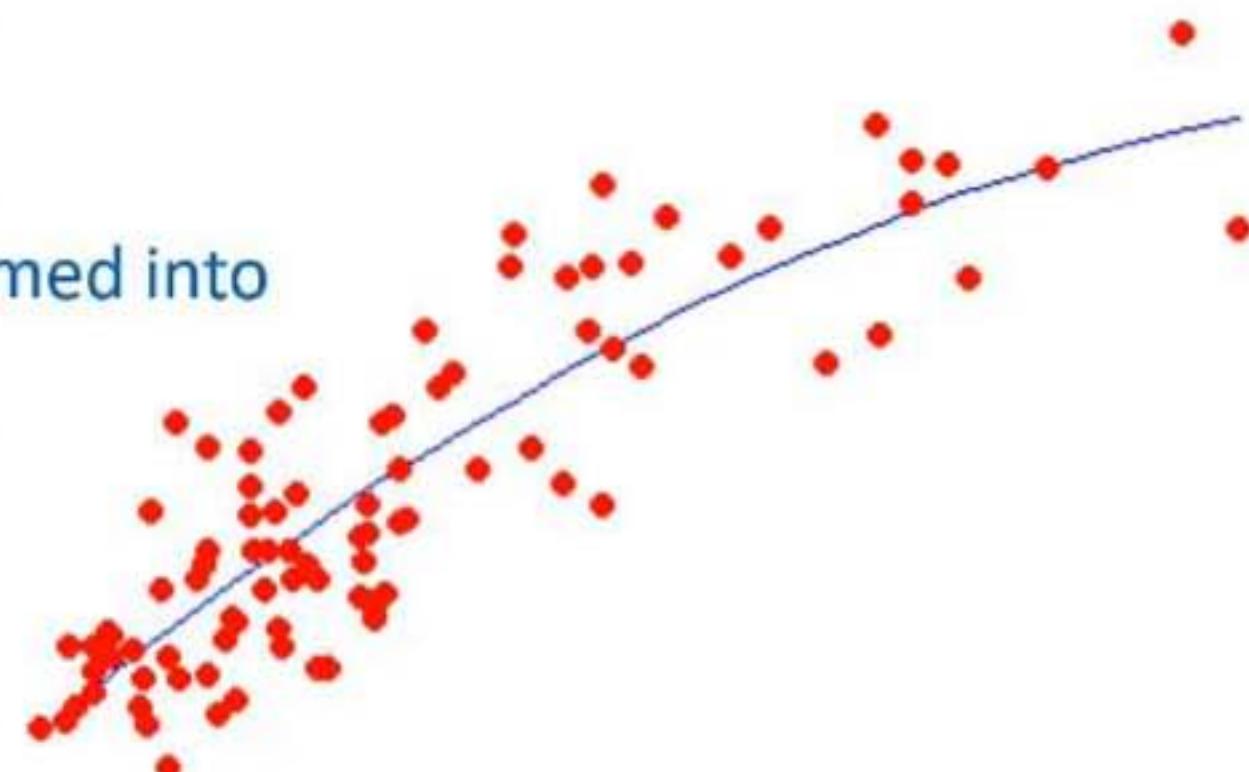
$$x_1 = x$$

$$x_2 = x^2$$

$$x_3 = x^3$$

$$\hat{y} = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_3$$

→ Multiple linear regression → Least Squares



What is non-linear regression?

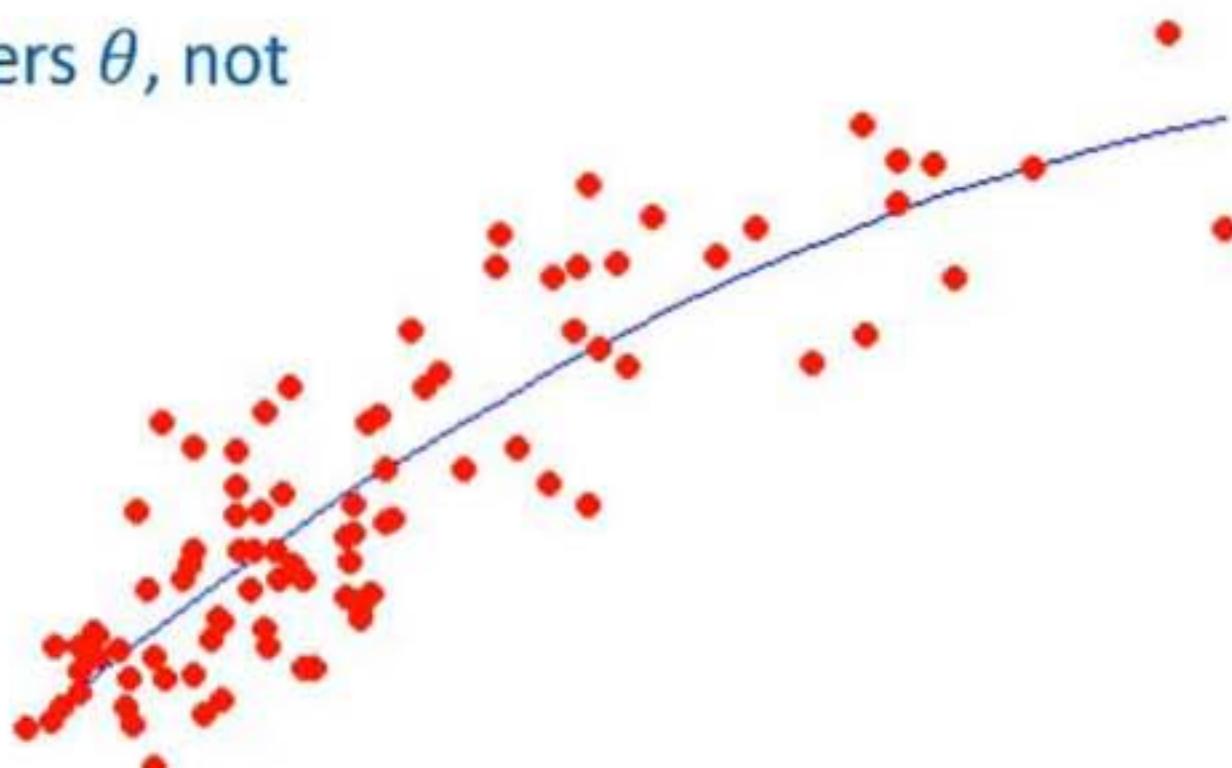
- To model non-linear relationship between the dependent variable and a set of independent variables
- \hat{y} must be a non-linear function of the parameters θ , not necessarily the features x

$$\hat{y} = \theta_0 + \theta_2 x^2$$

$$\hat{y} = \theta_0 + \theta_1 \theta_2 x$$

$$\hat{y} = \log(\theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3)$$

$$\hat{y} = \frac{\theta_0}{1 + \theta_1^{(x - \theta_2)}}$$



Linear vs non-linear regression

- How can I know if a problem is linear or non-linear in an easy way?
 - Inspect visually
 - Based on accuracy
- How should I model my data, if it displays non-linear on a scatter plot?

Linear vs non-linear regression

- How can I know if a problem is linear or non-linear in an easy way?
 - Inspect visually
 - Based on accuracy
- How should I model my data, if it displays non-linear on a scatter plot?
 - Polynomial regression
 - Non-linear regression model
 - Transform your data

Intro to Classification

Saeed Aghabozorgi



© IBM Corporation. All rights reserved.

1



What is classification?

- A supervised learning approach
- Categorizing some unknown items into a discrete set of categories or “classes”
- The target attribute is a categorical variable

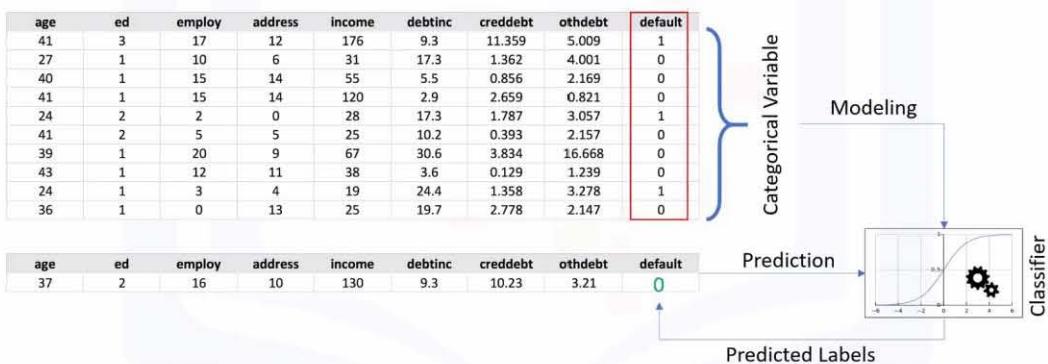


2

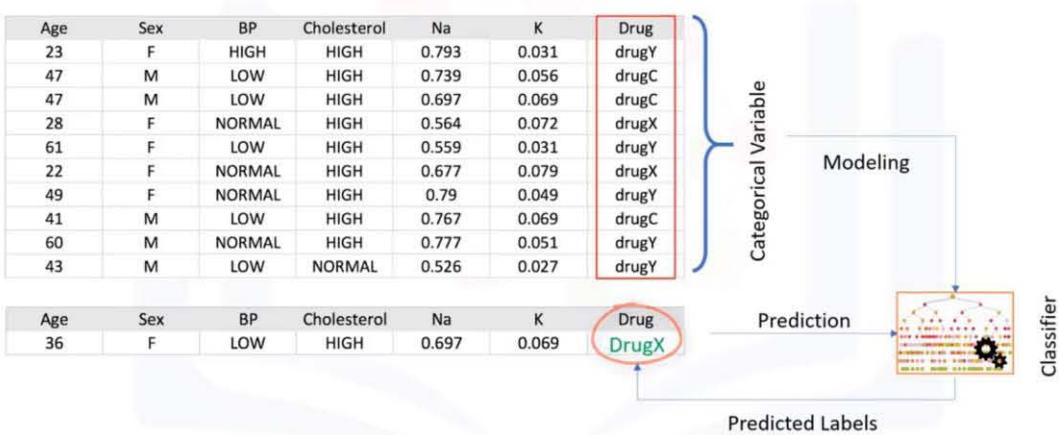


How does classification work?

Classification determines the class label for an unlabeled test case.



Example of multi-class classification

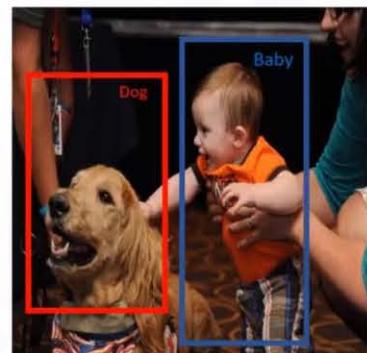


Classification use cases

	tenure	age	address	income	ed	employ	equip	callcard	wireless	churn
0	11.0	33.0	7.0	136.0	5.0	5.0	0.0	1.0	1.0	Yes
1	33.0	33.0	12.0	33.0	2.0	0.0	0.0	0.0	0.0	Yes
2	23.0	30.0	9.0	30.0	1.0	2.0	0.0	0.0	0.0	No
3	38.0	35.0	5.0	76.0	2.0	10.0	1.0	1.0	1.0	No
4	7.0	35.0	14.0	80.0	2.0	15.0	0.0	1.0	0.0	?

- Which category a customer belongs to?
- Whether a customer switches to another provider/brand?
- Whether a customer responds to a particular advertising campaign?

Classification applications



Classification algorithms in machine learning

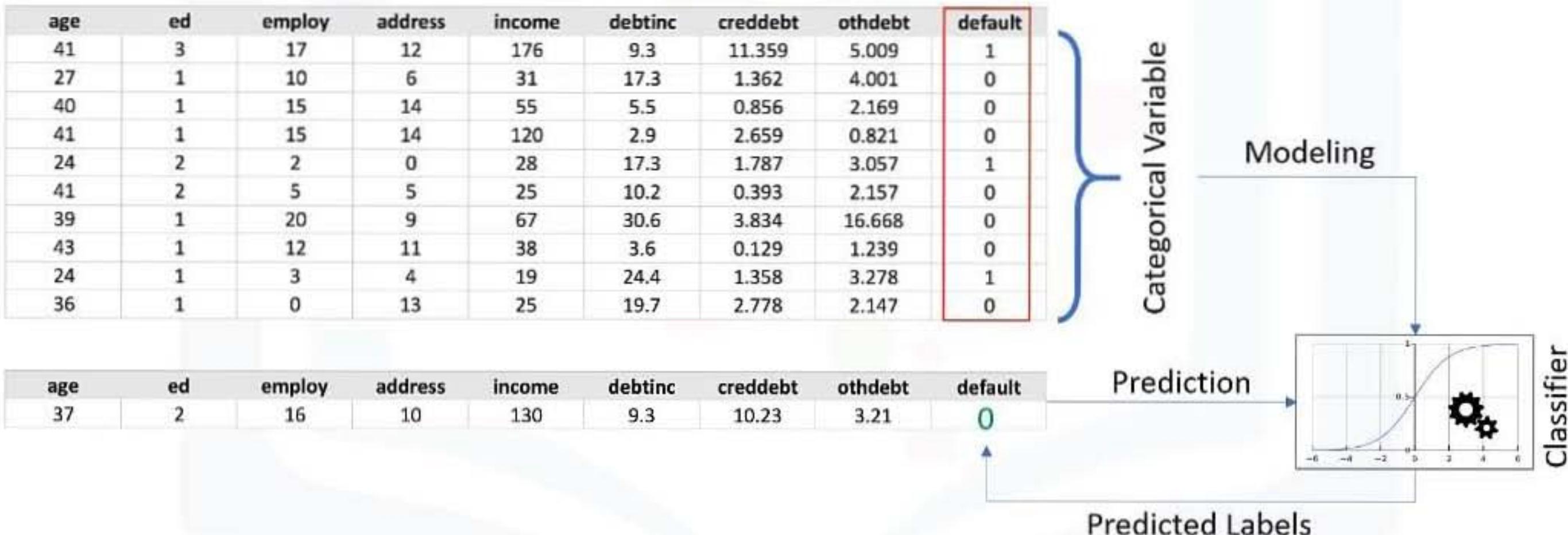
- Decision Trees (ID3, C4.5, C5.0)
- Naïve Bayes
- Linear Discriminant Analysis
- k -Nearest Neighbor
- Logistic Regression
- Neural Networks
- Support Vector Machines (SVM)

What is classification?

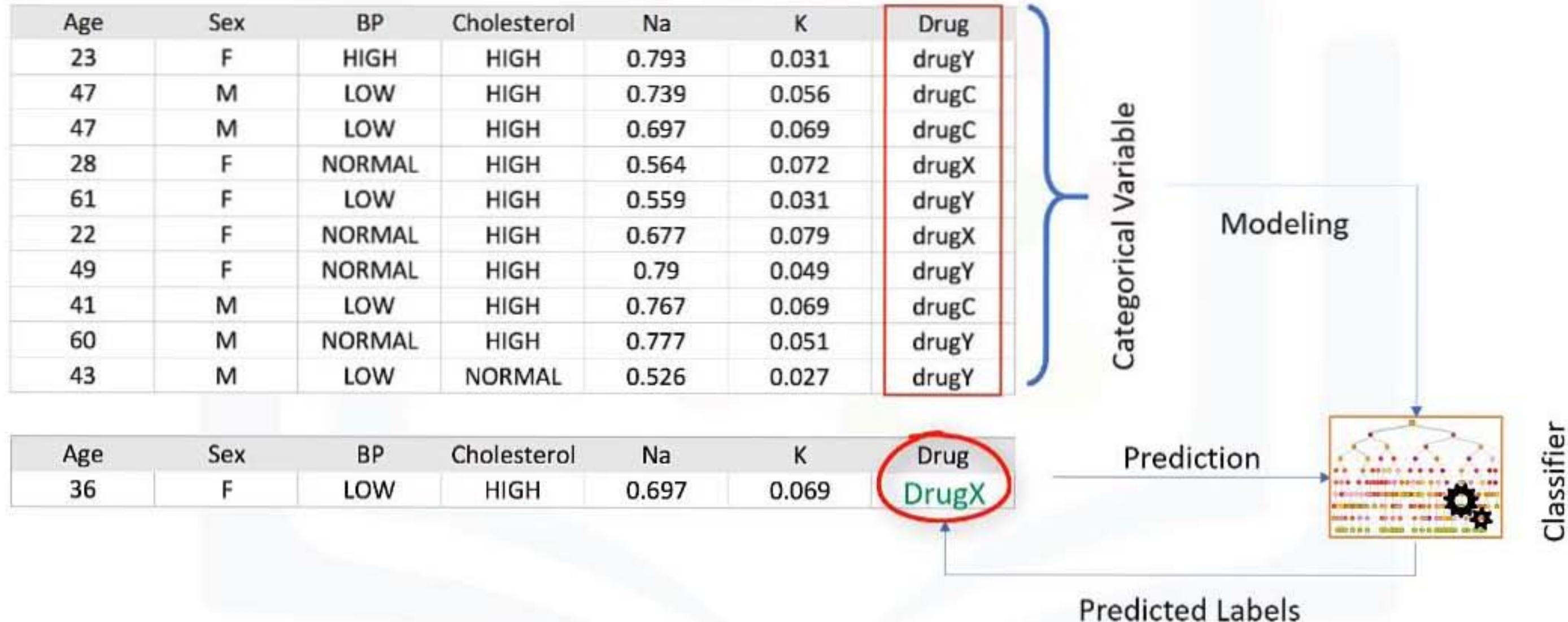
- A supervised learning approach
- Categorizing some unknown items into a discrete set of categories or “classes”
- The target attribute is a categorical variable

How does classification work?

Classification determines the class label for an unlabeled test case.



Example of multi-class classification



Classification use cases

	tenure	age	address	income	ed	employ	equip	callcard	wireless	churn
0	11.0	33.0	7.0	136.0	5.0	5.0	0.0	1.0	1.0	Yes
1	33.0	33.0	12.0	33.0	2.0	0.0	0.0	0.0	0.0	Yes
2	23.0	30.0	9.0	30.0	1.0	2.0	0.0	0.0	0.0	No
3	38.0	35.0	5.0	76.0	2.0	10.0	1.0	1.0	1.0	No
4	7.0	35.0	14.0	80.0	2.0	15.0	0.0	1.0	0.0	?

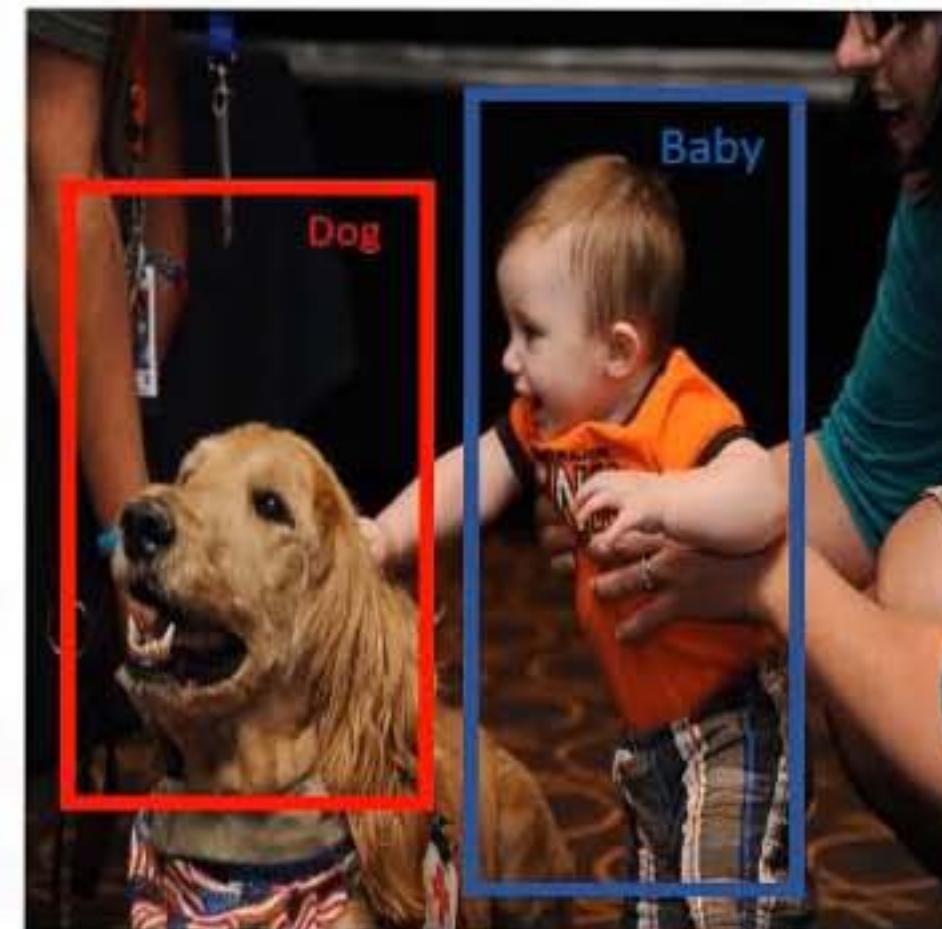
- Which category a customer belongs to?
- Whether a customer switches to another provider/brand?

Classification use cases

	tenure	age	address	income	ed	employ	equip	callcard	wireless	churn
0	11.0	33.0	7.0	136.0	5.0	5.0	0.0	1.0	1.0	Yes
1	33.0	33.0	12.0	33.0	2.0	0.0	0.0	0.0	0.0	Yes
2	23.0	30.0	9.0	30.0	1.0	2.0	0.0	0.0	0.0	No
3	38.0	35.0	5.0	76.0	2.0	10.0	1.0	1.0	1.0	No
4	7.0	35.0	14.0	80.0	2.0	15.0	0.0	1.0	0.0	?

- Which category a customer belongs to?
- Whether a customer switches to another provider/brand?
- Whether a customer responds to a particular advertising campaign?

Classification applications



Classification algorithms in machine learning

- Decision Trees (ID3, C4.5, C5.0)
- Naïve Bayes
- Linear Discriminant Analysis
- k -Nearest Neighbor
- Logistic Regression
- Neural Networks
- Support Vector Machines (SVM)

K-Nearest Neighbors

Saeed Aghabozorgi

Intro to KNN

X: Independent variable

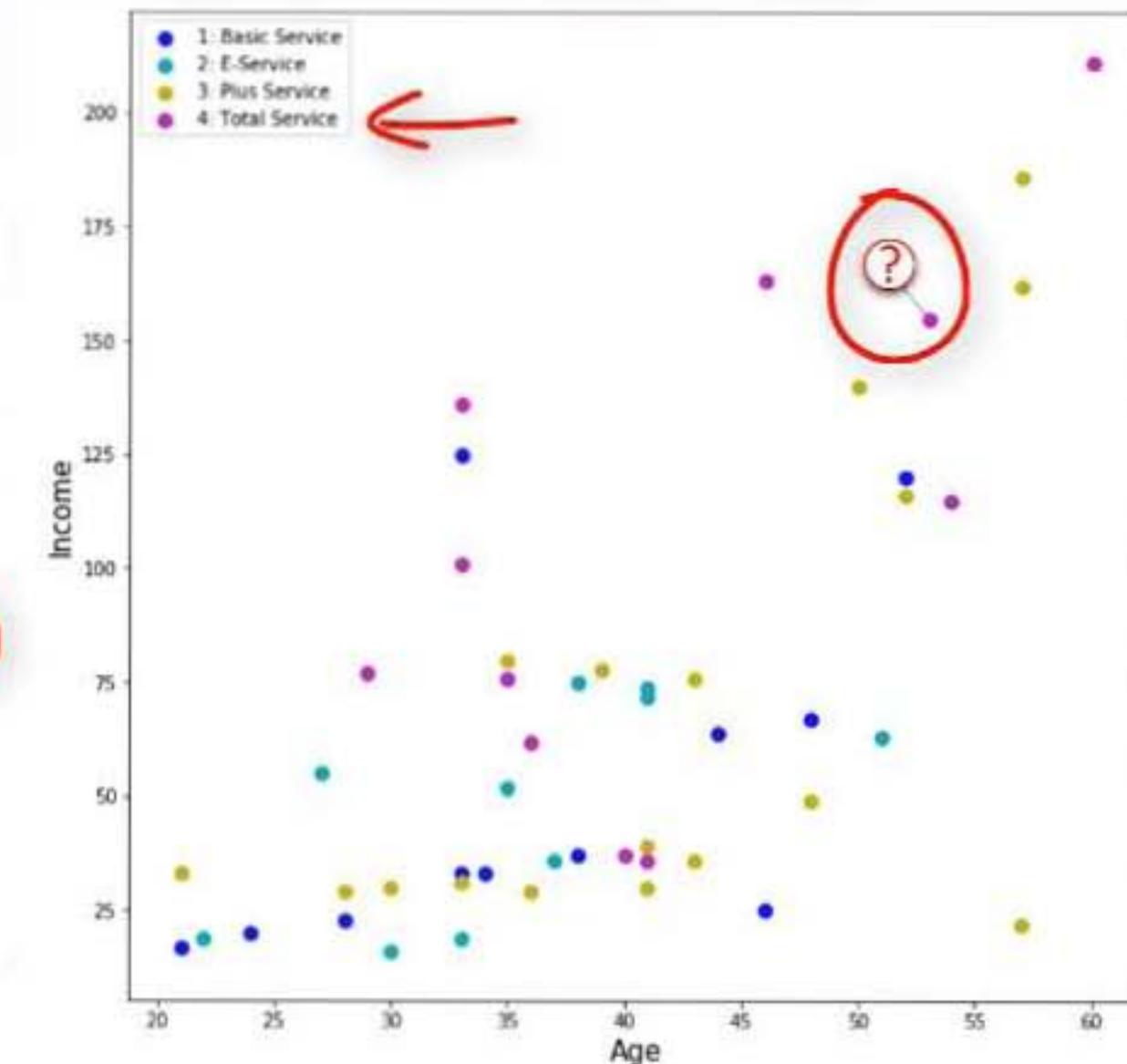
Y: Dependent variable

	region	age	marital	address	income	ed	employ	retire	gender	reside	custcat
0	2	44	1	9	64	4	5	0	0	2	1
1	3	33	1	7	136	5	5	0	0	6	4
2	3	52	1	24	116	1	29	0	1	2	3
3	2	33	0	12	33	2	0	0	1	1	1
4	2	30	1	9	30	1	2	0	0	4	3
5	2	39	0	17	78	2	16	0	1	1	3
6	3	22	1	2	19	2	4	0	1	5	2
7	2	35	0	5	76	2	10	0	0	3	4
8	3	50	1	7	166	4	31	0	0	5	?

Value	Label
1	Basic Service
2	E-Service
3	Plus Service
4	Total Service

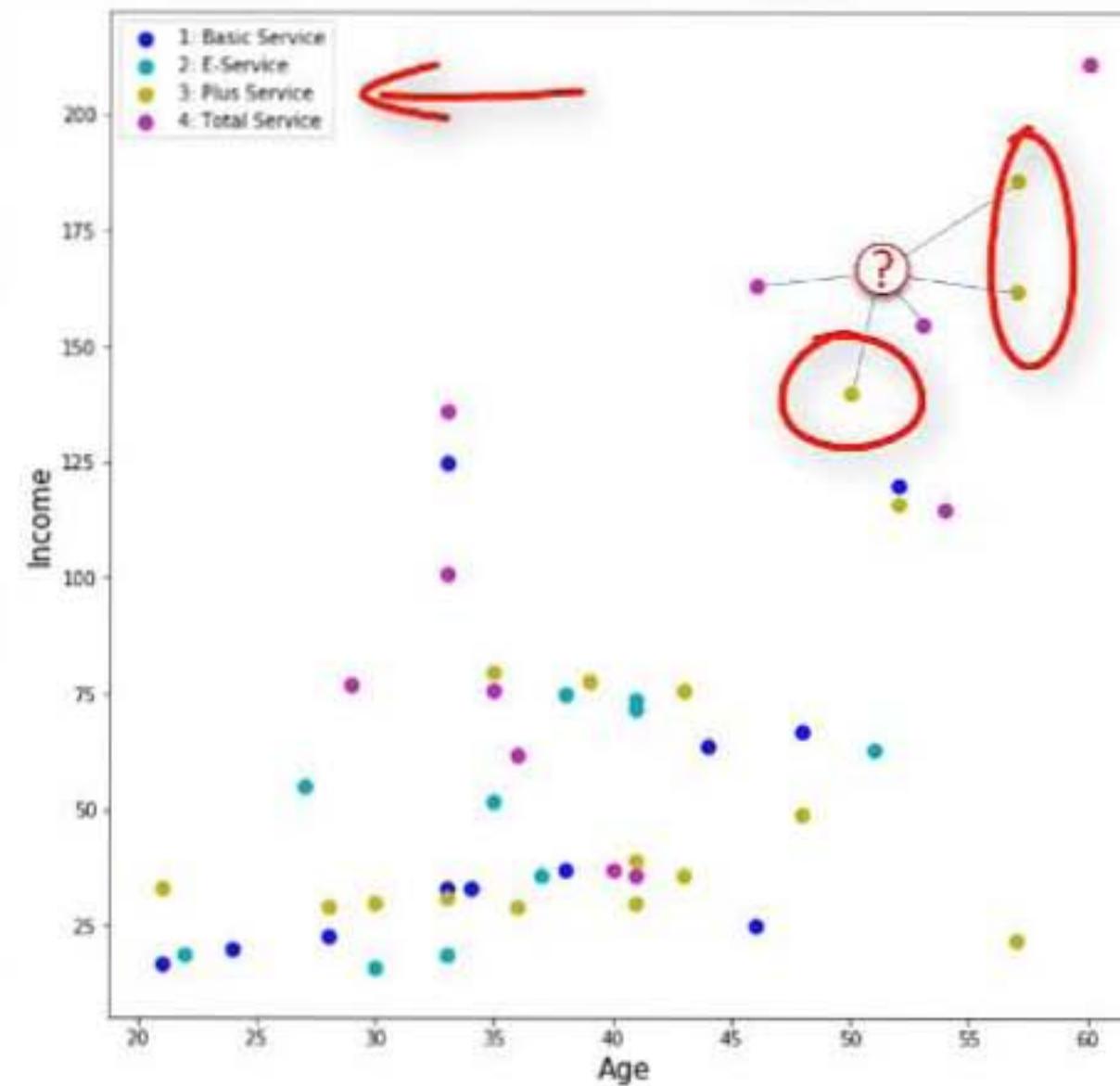
Determining the class using 1st KNN

	region	age	marital	address	income	ed	employ	retire	gender	reside	custcat
0	2	44	1	9	64	4	5	0	0	2	1
1	3	33	1	7	136	5	5	0	0	6	4
2	3	52	1	24	116	1	29	0	1	2	3
3	2	33	0	12	33	2	0	0	1	1	1
4	2	30	1	9	30	1	2	0	0	4	3
5	2	39	0	17	78	2	16	0	1	1	3
6	3	22	1	2	19	2	4	0	1	5	2
7	2	35	0	5	76	2	10	0	0	3	4
8	3	50	1	7	166	4	31	0	0	5	?



Determining the class using the 5 KNNs

	region	age	marital	address	income	ed	employ	retire	gender	reside	custcat
0	2	44	1	9	64	4	5	0	0	2	1
1	3	33	1	7	136	5	5	0	0	6	4
2	3	52	1	24	116	1	29	0	1	2	3
3	2	33	0	12	33	2	0	0	1	1	1
4	2	30	1	9	30	1	2	0	0	4	3
5	2	39	0	17	78	2	16	0	1	1	3
6	3	22	1	2	19	2	4	0	1	5	2
7	2	35	0	5	76	2	10	0	0	3	4
8	3	50	1	7	166	4	31	0	0	5	?

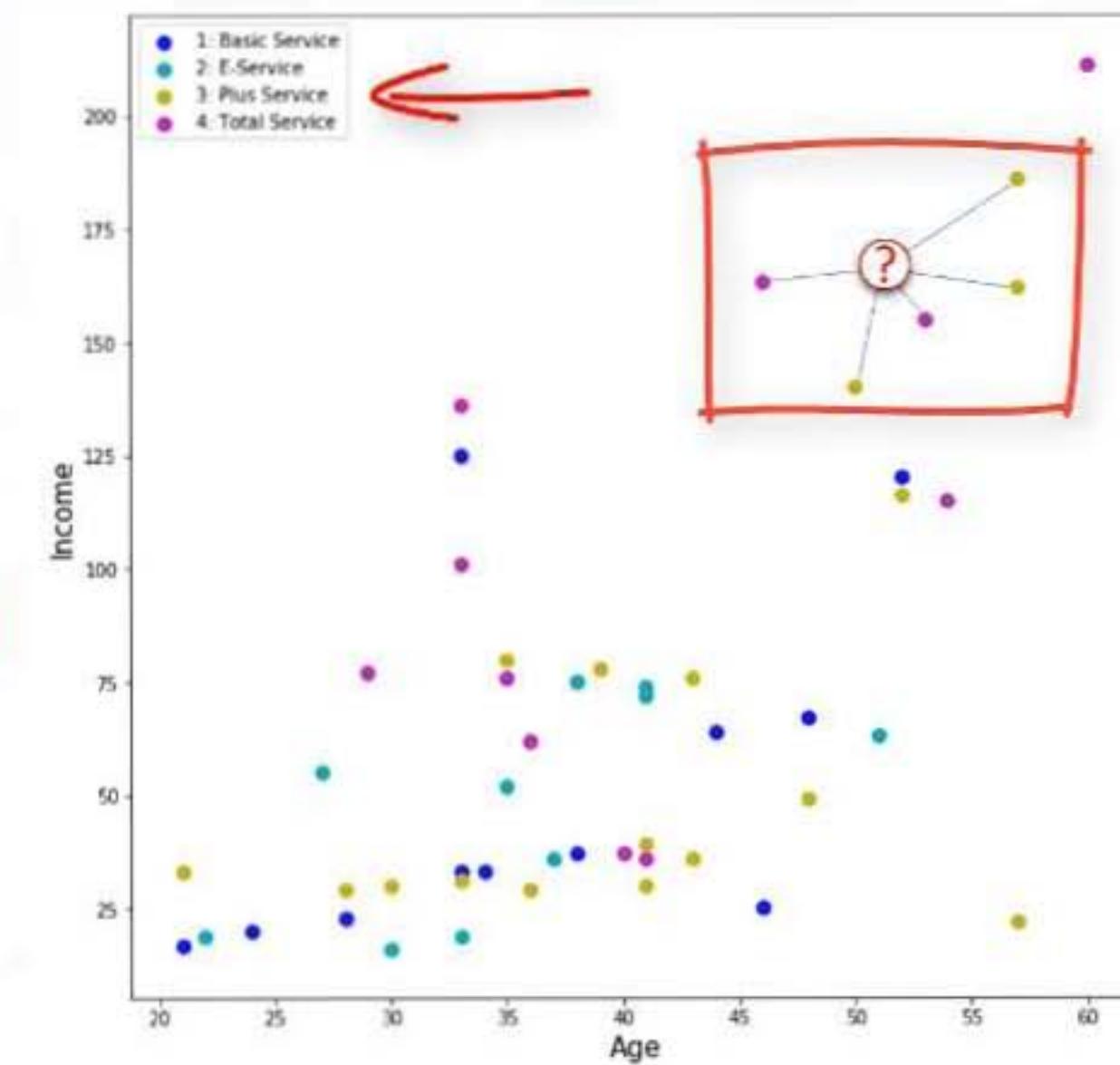


Determining the class using the 5 KNNs

	region	age	marital	address	income	ed	employ	retire	gender	reside	custcat
0	2	44	1	9	64	4	5	0	0	2	1
1	3	33	1	7	136	5	5	0	0	6	4
2	3	52	1	24	116	1	29	0	1	2	3
3	2	33	0	12	33	2	0	0	1	1	1
4	2	30	1	9	30	1	2	0	0	4	3
5	2	39	0	17	78	2	16	0	1	1	3
6	3	22	1	2	19	2	4	0	1	5	2
7	2	35	0	5	76	2	10	0	0	3	4
8	3	50	1	7	166	4	31	0	0	5	?

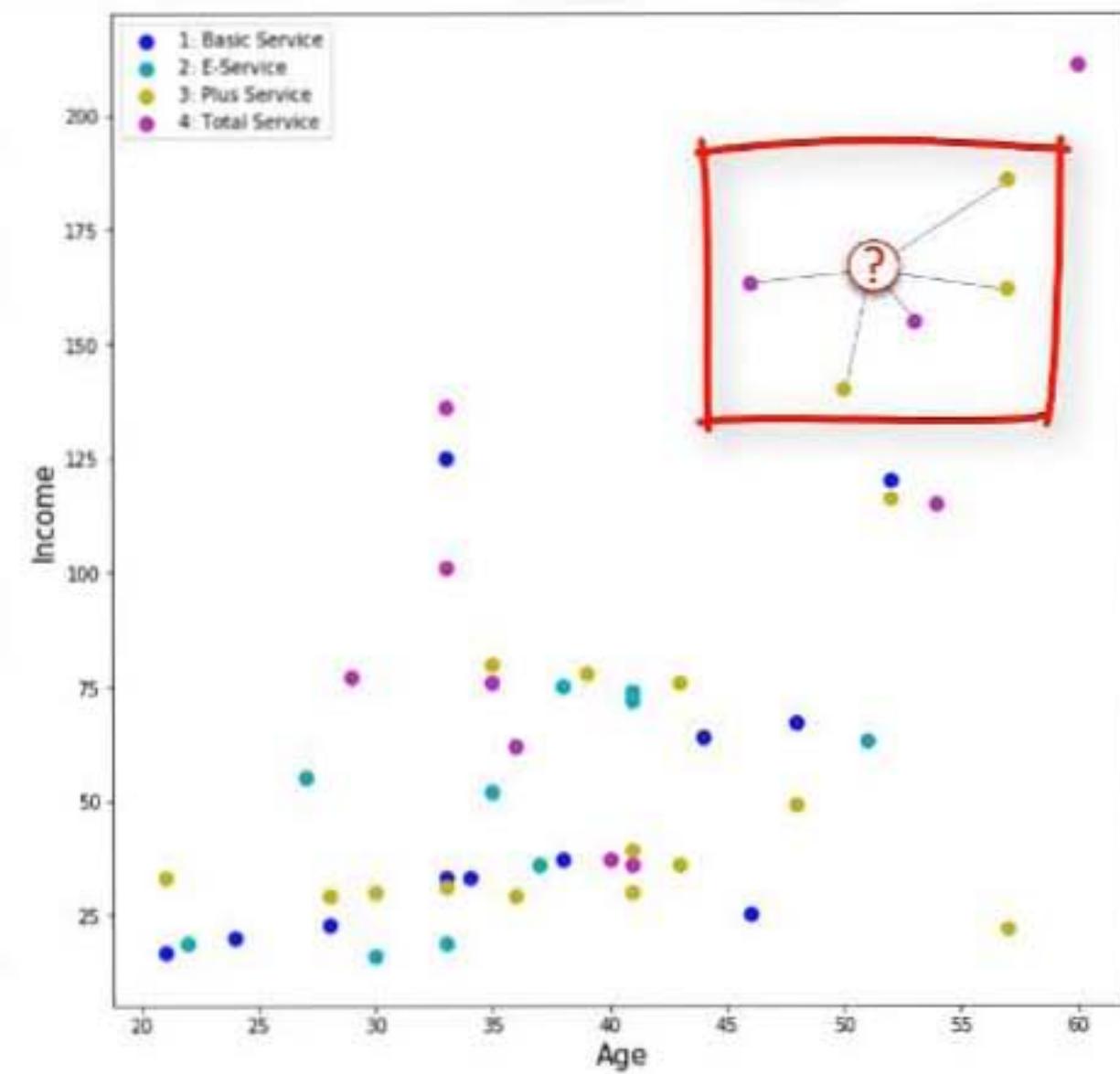
5-NN

→ 3: Plus Service



What is K-Nearest Neighbor (or KNN)?

- A method for **classifying** cases based on their similarity to other cases
- Cases that are near each other are said to be “**neighbors**”
- Based on **similar cases with same class labels** are near each other



The K-Nearest Neighbors algorithm

1. Pick a value for K.
2. Calculate the distance of unknown case from all cases.
3. Select the K-observations in the training data that are “nearest” to the unknown data point.
4. Predict the response of the unknown data point using the most popular response value from the K-nearest neighbors.

Calculating the similarity/distance in a 1-dimensional space



Customer 1

Age

54



Customer 2

Age

50

Calculating the similarity/distance in a 1-dimensional space



Customer 1
Age
54



Customer 2
Age
50

$$\text{Dis } (x_1, x_2) = \sqrt{\sum_{i=0}^n (x_{1i} - x_{2i})^2}$$

Calculating the similarity/distance in a 1-dimensional space



Customer 1

Age

54



Customer 2

Age

50

$$\text{Dis}(x_1, x_2) = \sqrt{\sum_{i=0}^n (x_{1i} - x_{2i})^2}$$

$$\text{Dis}(x_1, x_2) = \sqrt{(34 - 30)^2} = 4$$

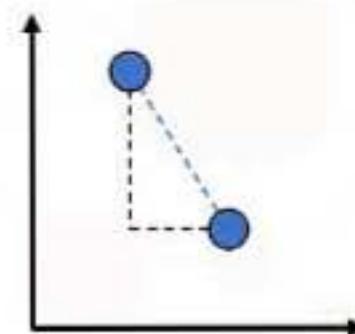
Calculating the similarity/distance in a 2-dimensional space



Customer 1	
Age	Income
54	190



Customer 2	
Age	Income
50	200



$$\begin{aligned}\text{Dis } (x_1, x_2) &= \sqrt{\sum_{i=0}^n (x_{1i} - x_{2i})^2} \\ &= \sqrt{(54 - 50)^2 + (190 - 200)^2} = 10.77\end{aligned}$$

Calculating the similarity/distance in a multi-dimensional space



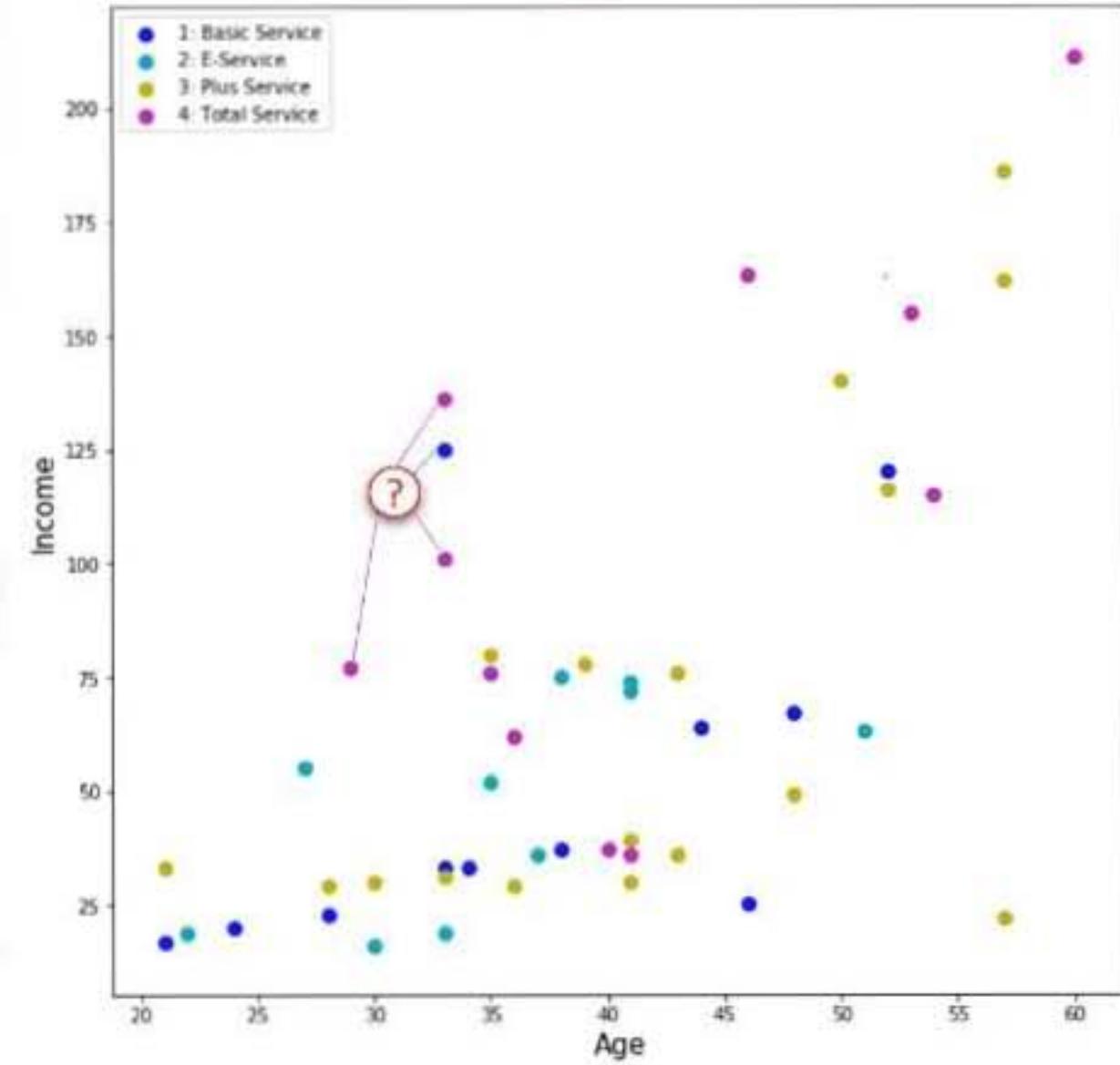
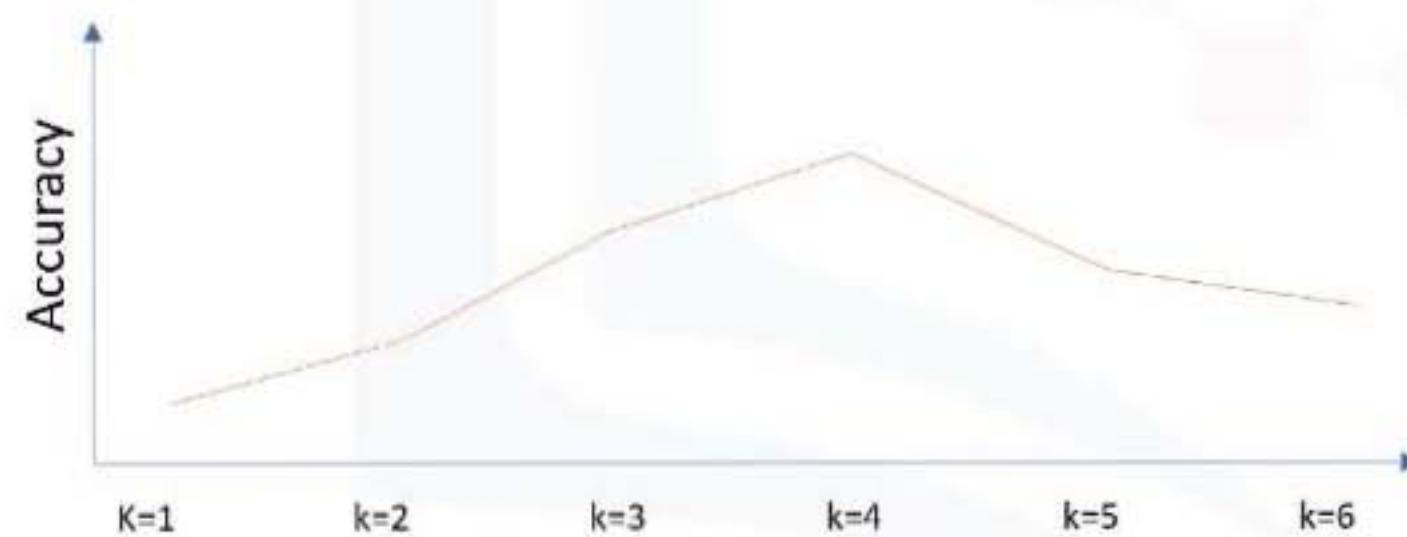
Customer 1		
Age	Income	Education
54	190	3

Customer 2		
Age	Income	Education
50	200	8

$$\begin{aligned}\text{Dis}(x_1, x_2) &= \sqrt{\sum_{i=0}^n (x_{1i} - x_{2i})^2} \\ &= \sqrt{(54 - 50)^2 + (190 - 200)^2 + (3 - 8)^2} = 11.87\end{aligned}$$

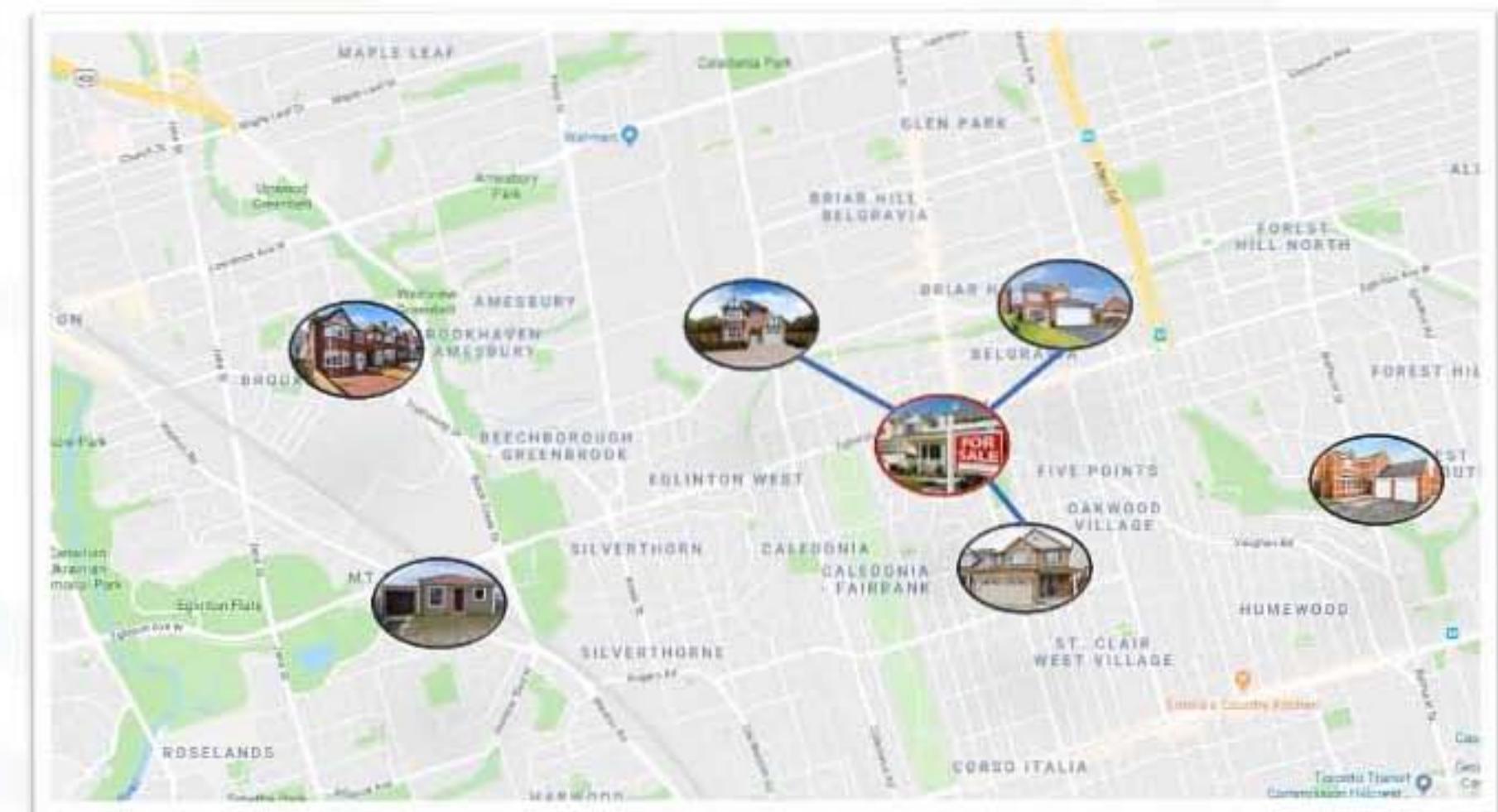
What is the best value of K for KNN?

- K =1 class 1
- K =20 ?



Computing continuous targets using KNN

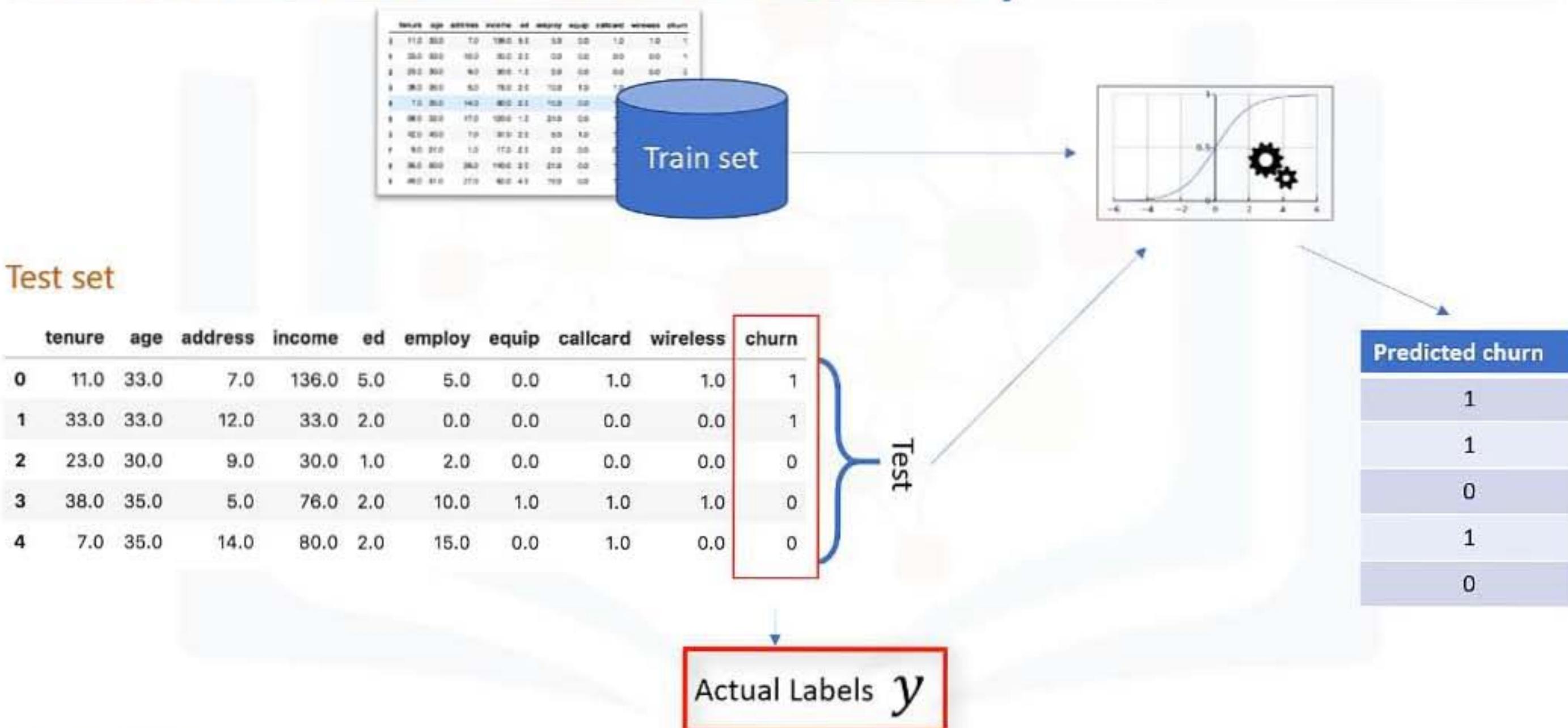
- KNN can also be used for regression



Evaluation Metrics in Classification

Saeed Aghabozorgi

Classification accuracy

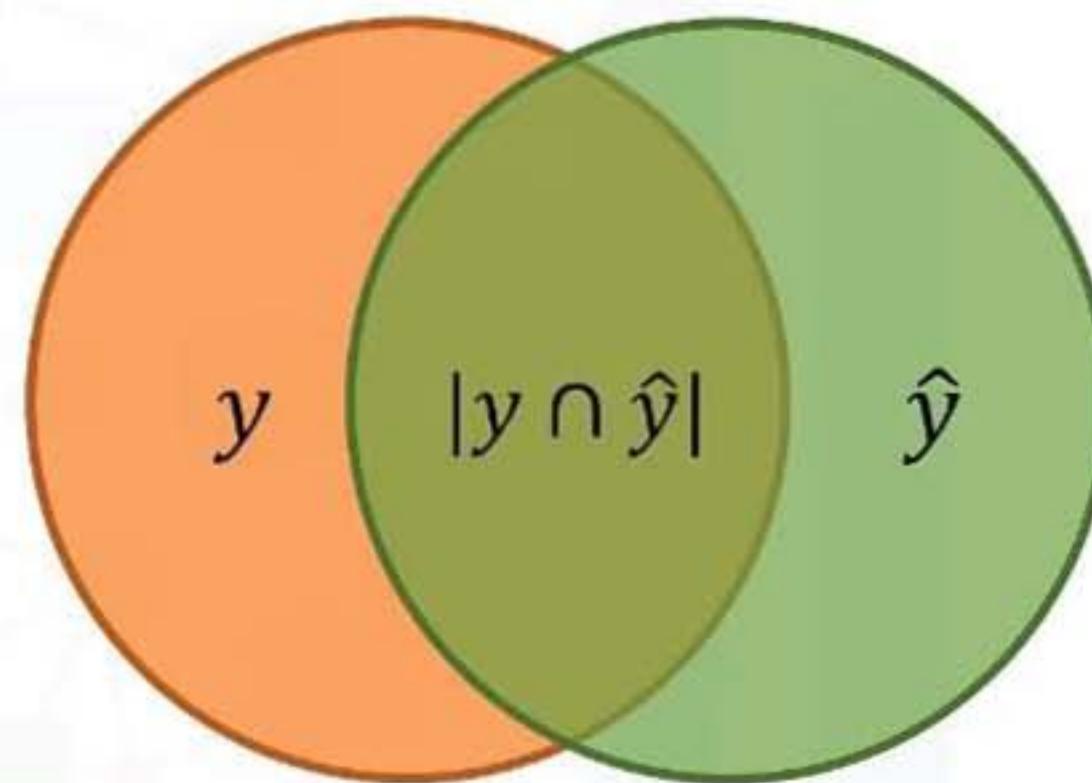


Jaccard index

y : Actual labels

\hat{y} : Predicted labels

$$J(y, \hat{y}) = \frac{|y \cap \hat{y}|}{|y \cup \hat{y}|} = \frac{|y \cap \hat{y}|}{|y| + |\hat{y}| - |y \cap \hat{y}|}$$



Jaccard index

y : Actual labels

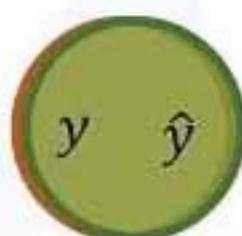
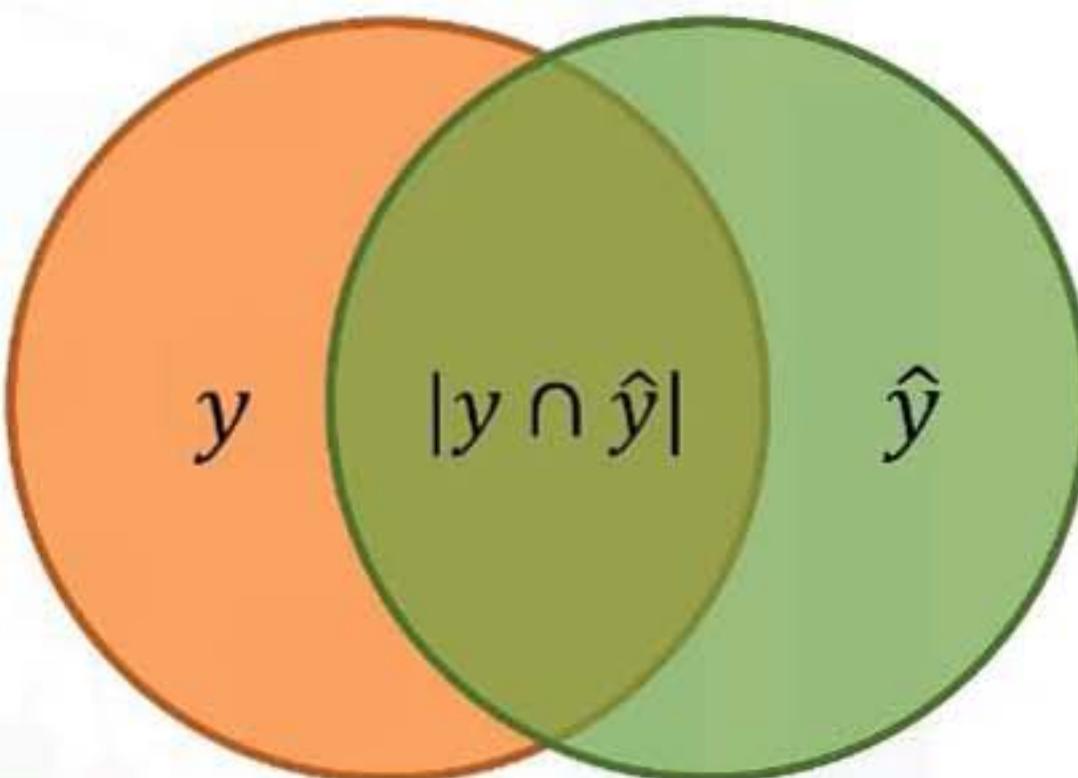
\hat{y} : Predicted labels

$$J(y, \hat{y}) = \frac{|y \cap \hat{y}|}{|y \cup \hat{y}|} = \frac{|y \cap \hat{y}|}{|y| + |\hat{y}| - |y \cap \hat{y}|}$$

y : [0, 0, 0, 0, 0, 1, 1, 1, 1, 1]

\hat{y} : [1, 1, 0, 0, 0, 1, 1, 1, 1, 1]

$$J(y, \hat{y}) = \frac{8}{10+10-8} = 0.66$$



$$J(y, \hat{y}) = 1.0$$

Higher Accuracy

Jaccard index

y : Actual labels

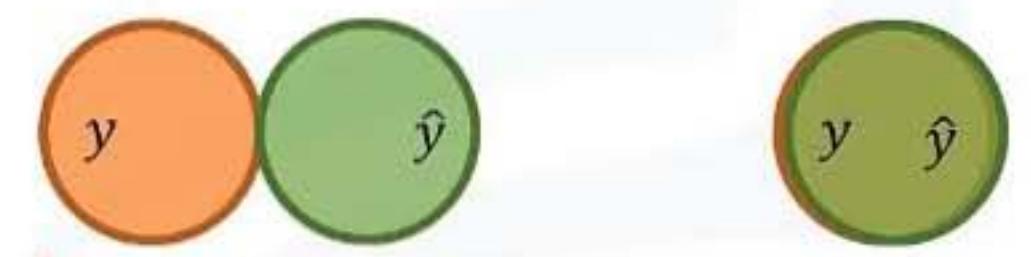
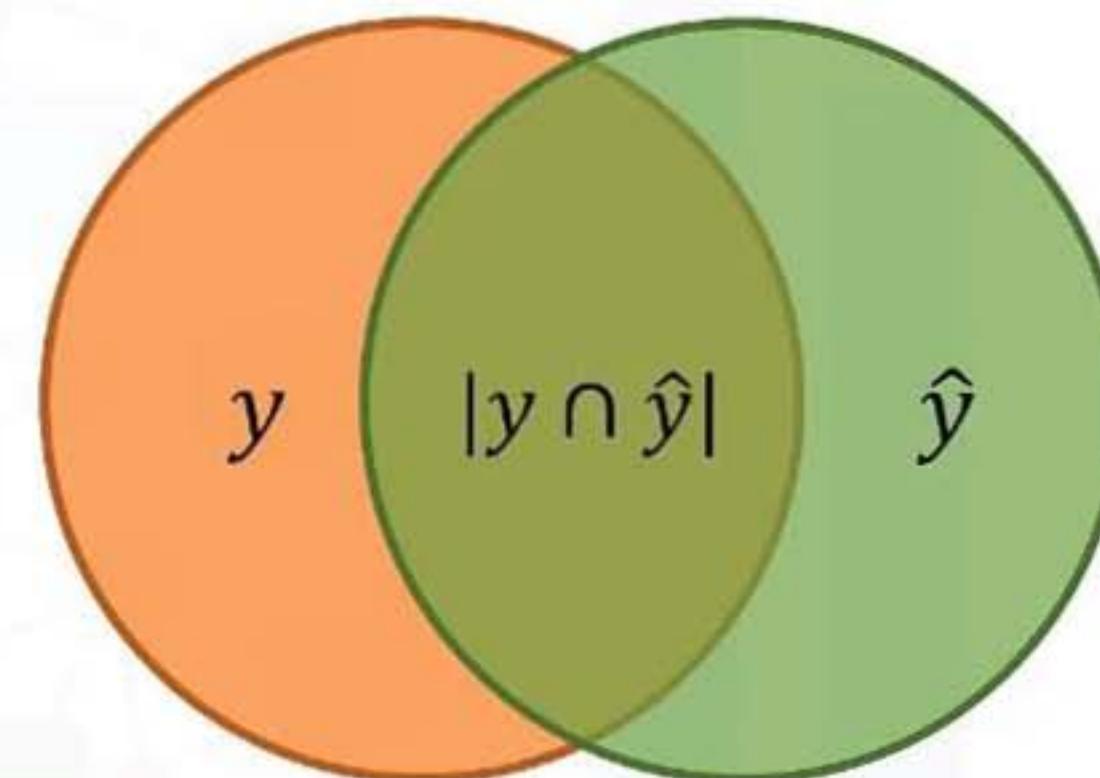
\hat{y} : Predicted labels

$$J(y, \hat{y}) = \frac{|y \cap \hat{y}|}{|y \cup \hat{y}|} = \frac{|y \cap \hat{y}|}{|y| + |\hat{y}| - |y \cap \hat{y}|}$$

y : [0, 0, 0, 0, 0, 1, 1, 1, 1, 1]

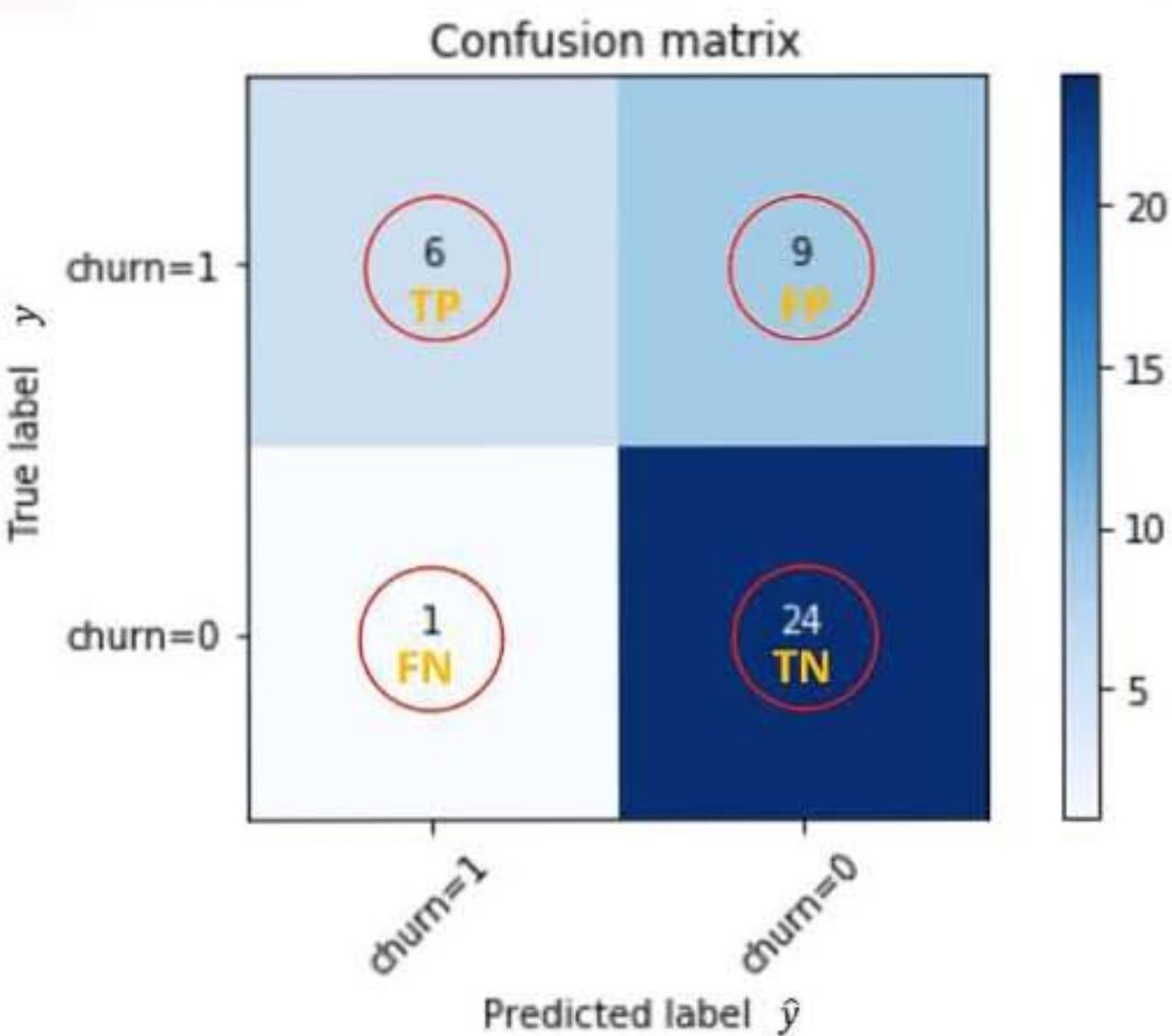
\hat{y} : [1, 1, 0, 0, 0, 1, 1, 1, 1, 1]

$$J(y, \hat{y}) = \frac{8}{10+10-8} = 0.66$$



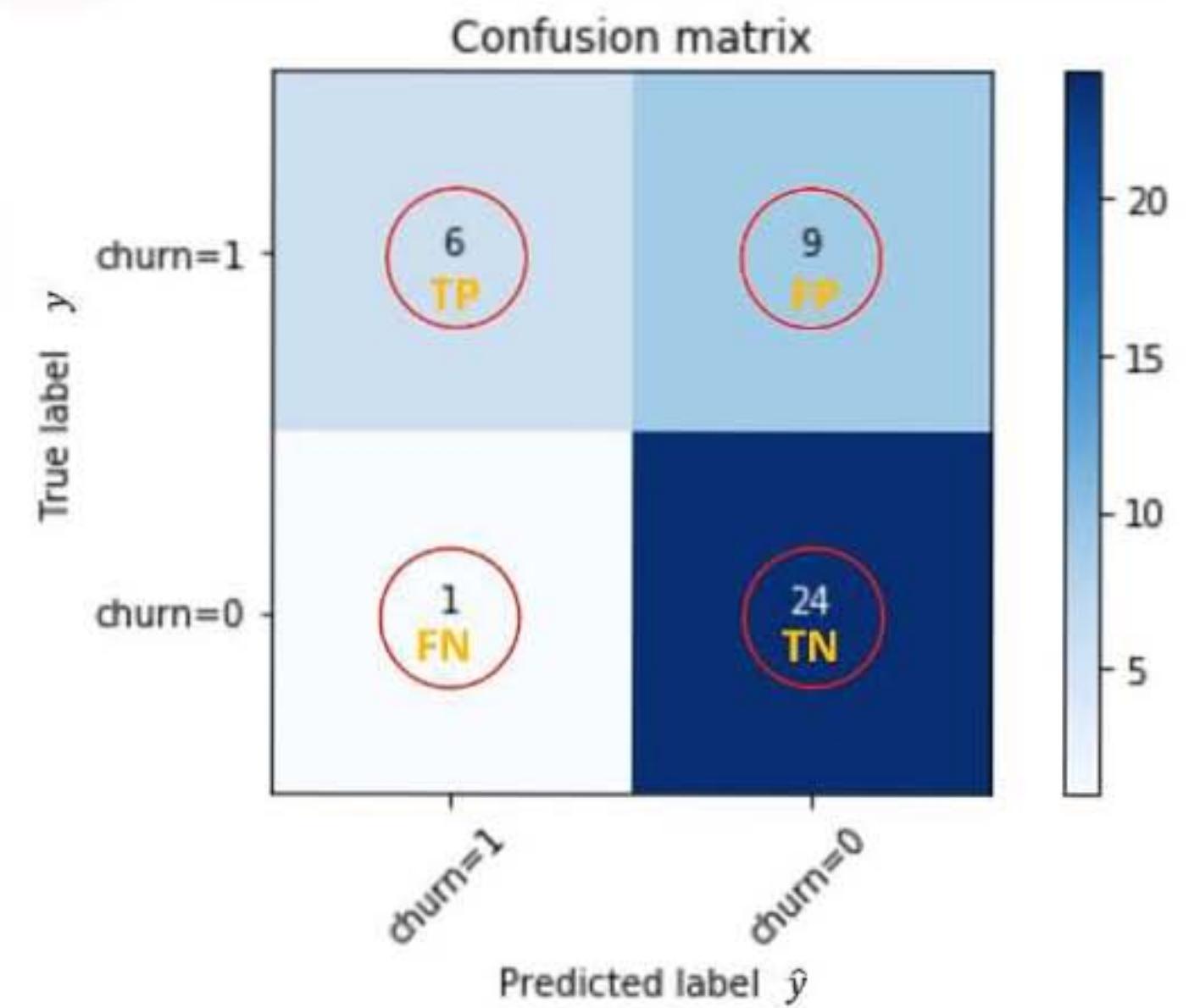
Higher Accuracy →

F1-score



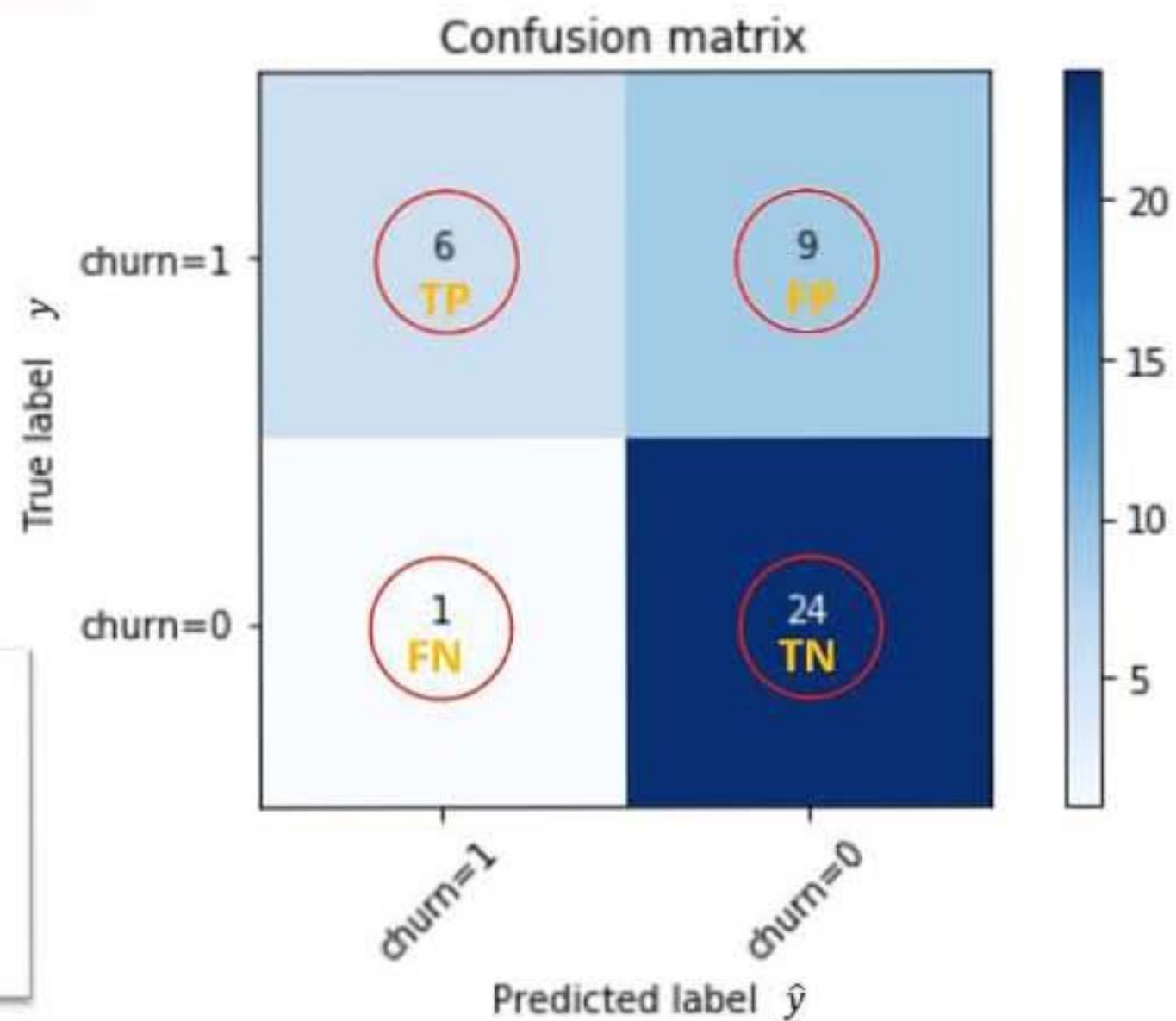
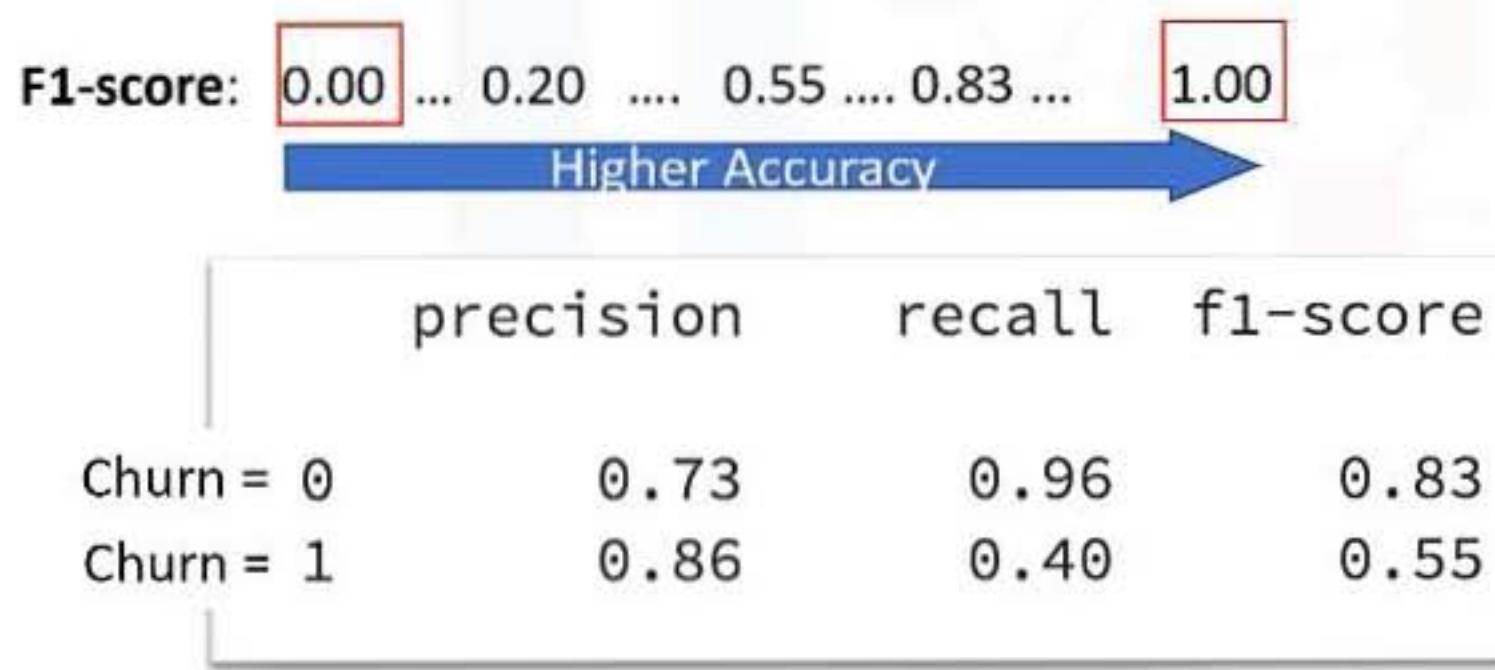
F1-score

- Precision = $TP / (TP + FP)$
- Recall = $TP / (TP + FN)$



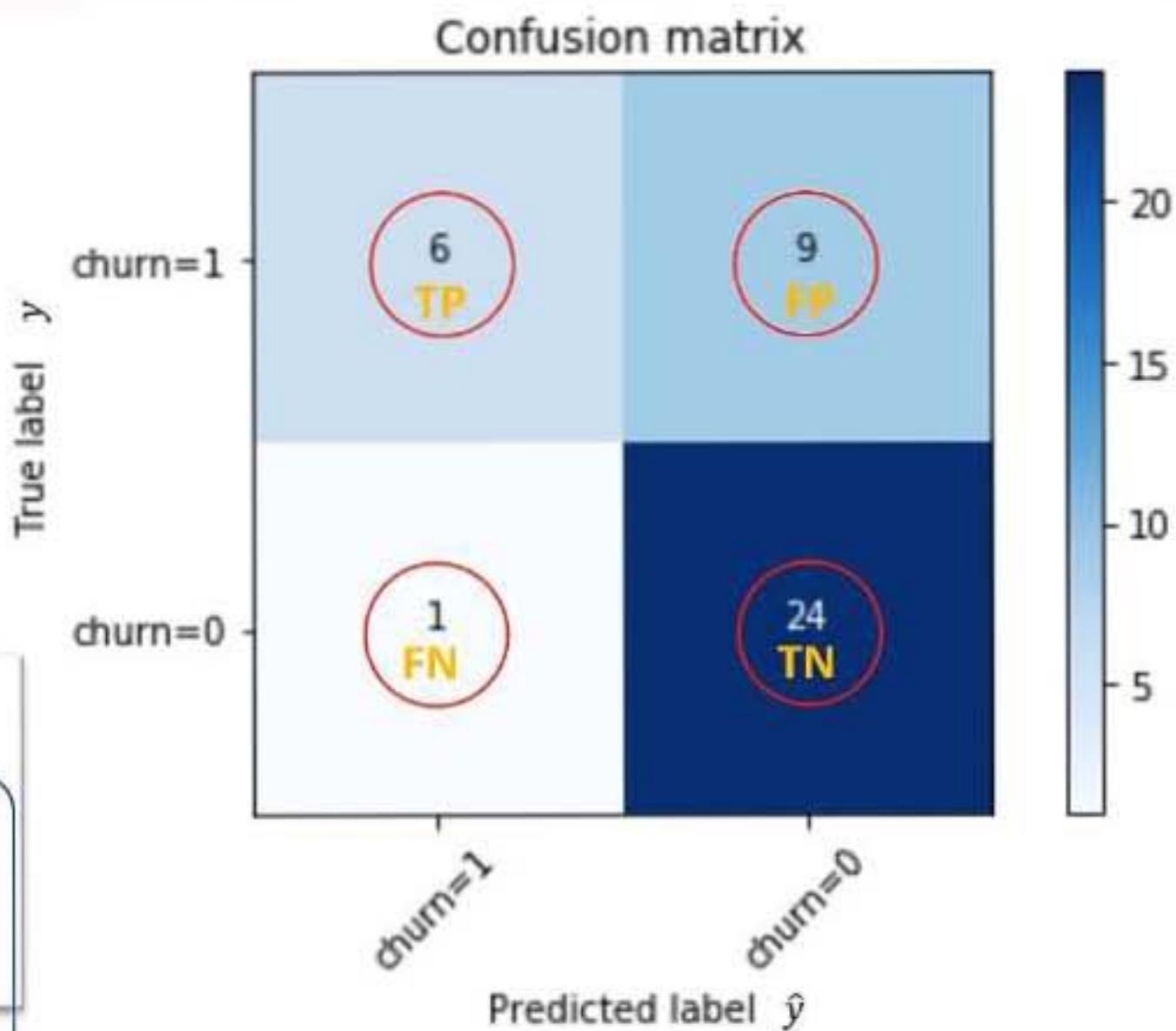
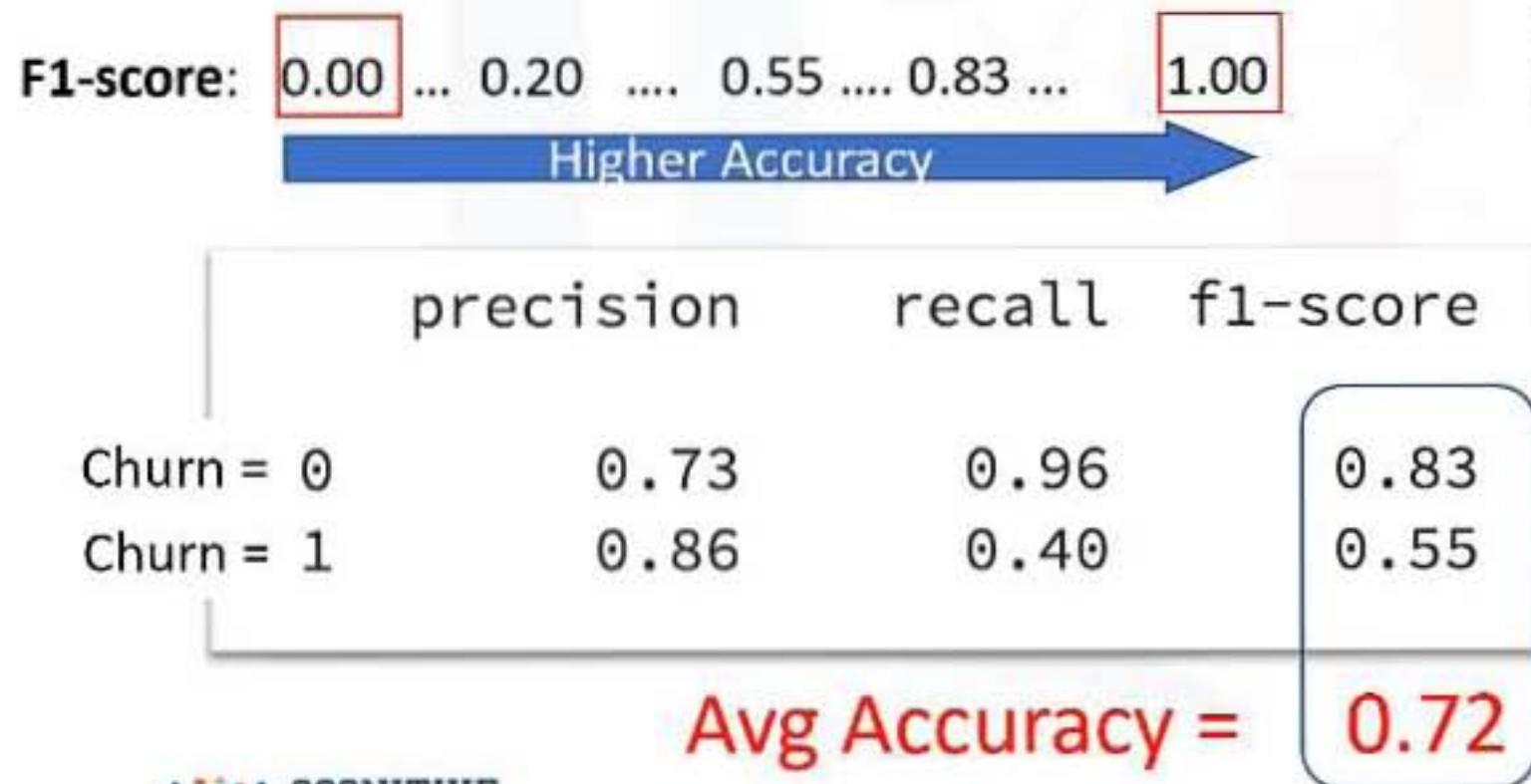
F1-score

- Precision = $TP / (TP + FP)$
- Recall = $TP / (TP + FN)$
- F1-score = $2x (prc \times rec) / (prc+rec)$



F1-score

- Precision = $TP / (TP + FP)$
- Recall = $TP / (TP + FN)$
- F1-score = $2x (prc \times rec) / (prc+rec)$

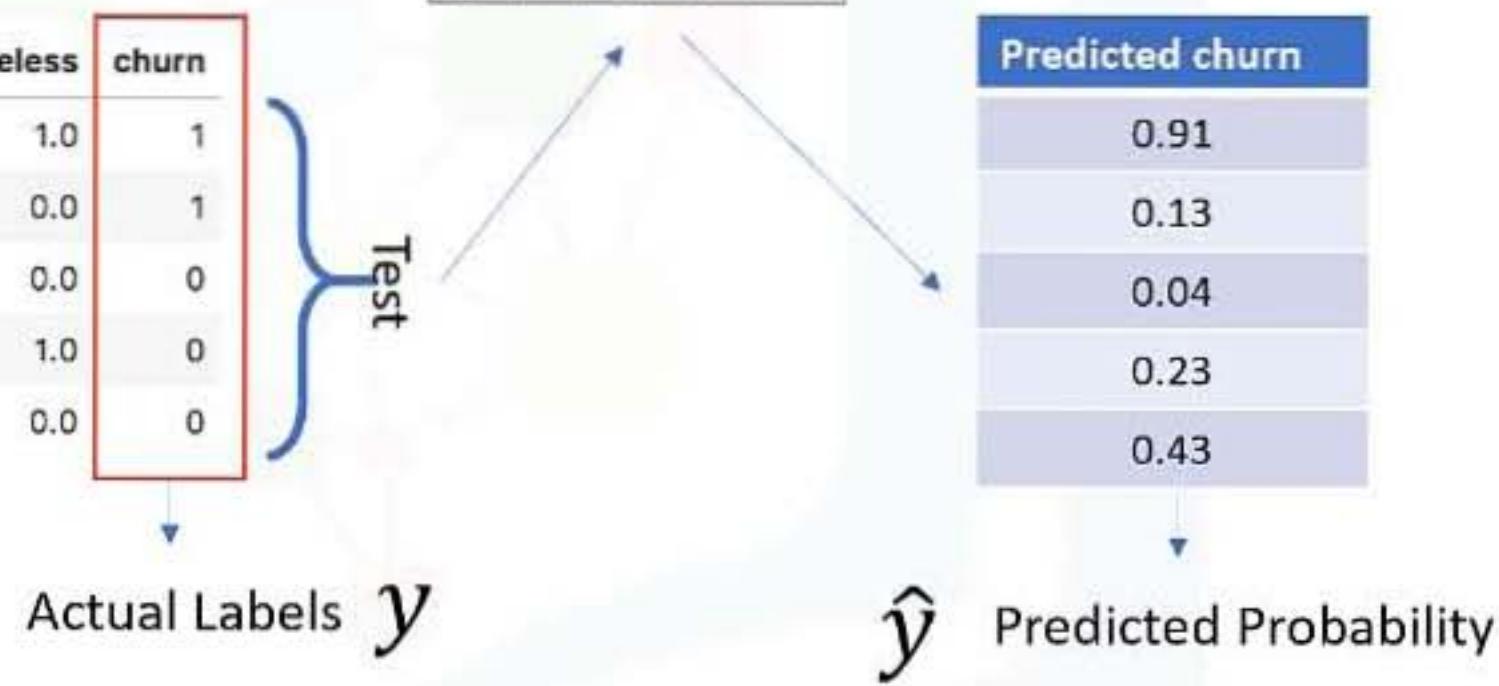
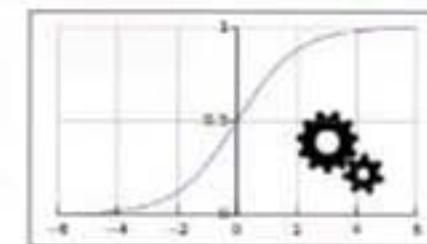


Log loss

Performance of a classifier where the predicted output is a probability value between 0 and 1.

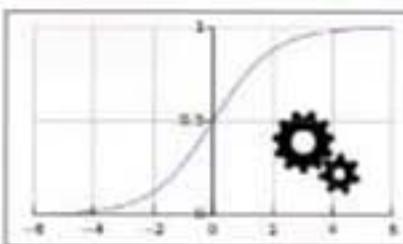
Test set

	tenure	age	address	income	ed	employ	equip	callcard	wireless	churn
0	11.0	33.0	7.0	136.0	5.0	5.0	0.0	1.0	1.0	1
1	33.0	33.0	12.0	33.0	2.0	0.0	0.0	0.0	0.0	1
2	23.0	30.0	9.0	30.0	1.0	2.0	0.0	0.0	0.0	0
3	38.0	35.0	5.0	76.0	2.0	10.0	1.0	1.0	1.0	0
4	7.0	35.0	14.0	80.0	2.0	15.0	0.0	1.0	0.0	0



Log loss

Performance of a classifier where the predicted output is a probability value between 0 and 1.



Test set

	tenure	age	address	income	ed	employ	equip	callcard	wireless	churn
0	11.0	33.0	7.0	136.0	5.0	5.0	0.0	1.0	1.0	1
1	33.0	33.0	12.0	33.0	2.0	0.0	0.0	0.0	0.0	1
2	23.0	30.0	9.0	30.0	1.0	2.0	0.0	0.0	0.0	0
3	38.0	35.0	5.0	76.0	2.0	10.0	1.0	1.0	1.0	0
4	7.0	35.0	14.0	80.0	2.0	15.0	0.0	1.0	0.0	0

Actual Labels y

$$(y \times \log(\hat{y}) + (1 - y) \times \log(1 - \hat{y}))$$

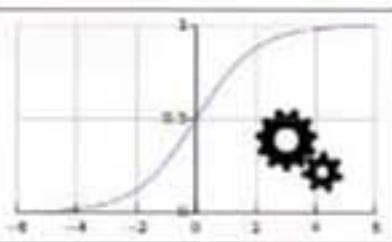
Test

Predicted churn	LogLoss
0.91	0.11
0.13	2.04
0.04	0.04
0.23	0.26
0.43	0.56

\hat{y} Predicted Probability

Log loss

Performance of a classifier where the predicted output is a probability value between 0 and 1.



Test set

	tenure	age	address	income	ed	employ	equip	callcard	wireless	churn
0	11.0	33.0	7.0	136.0	5.0	5.0	0.0	1.0	1.0	1
1	33.0	33.0	12.0	33.0	2.0	0.0	0.0	0.0	0.0	1
2	23.0	30.0	9.0	30.0	1.0	2.0	0.0	0.0	0.0	0
3	38.0	35.0	5.0	76.0	2.0	10.0	1.0	1.0	1.0	0
4	7.0	35.0	14.0	80.0	2.0	15.0	0.0	1.0	0.0	0

Actual Labels y

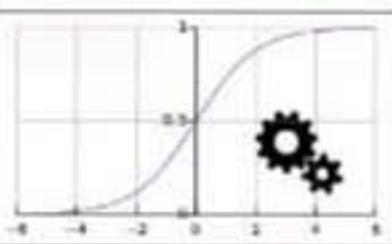
Predicted churn	LogLoss
0.91	0.11
0.13	2.04
0.04	0.04
0.23	0.26
0.43	0.56

\hat{y} Predicted Probability

$$\text{LogLoss} = -\frac{1}{n} \sum (y \times \log(\hat{y}) + (1 - y) \times \log(1 - \hat{y}))$$

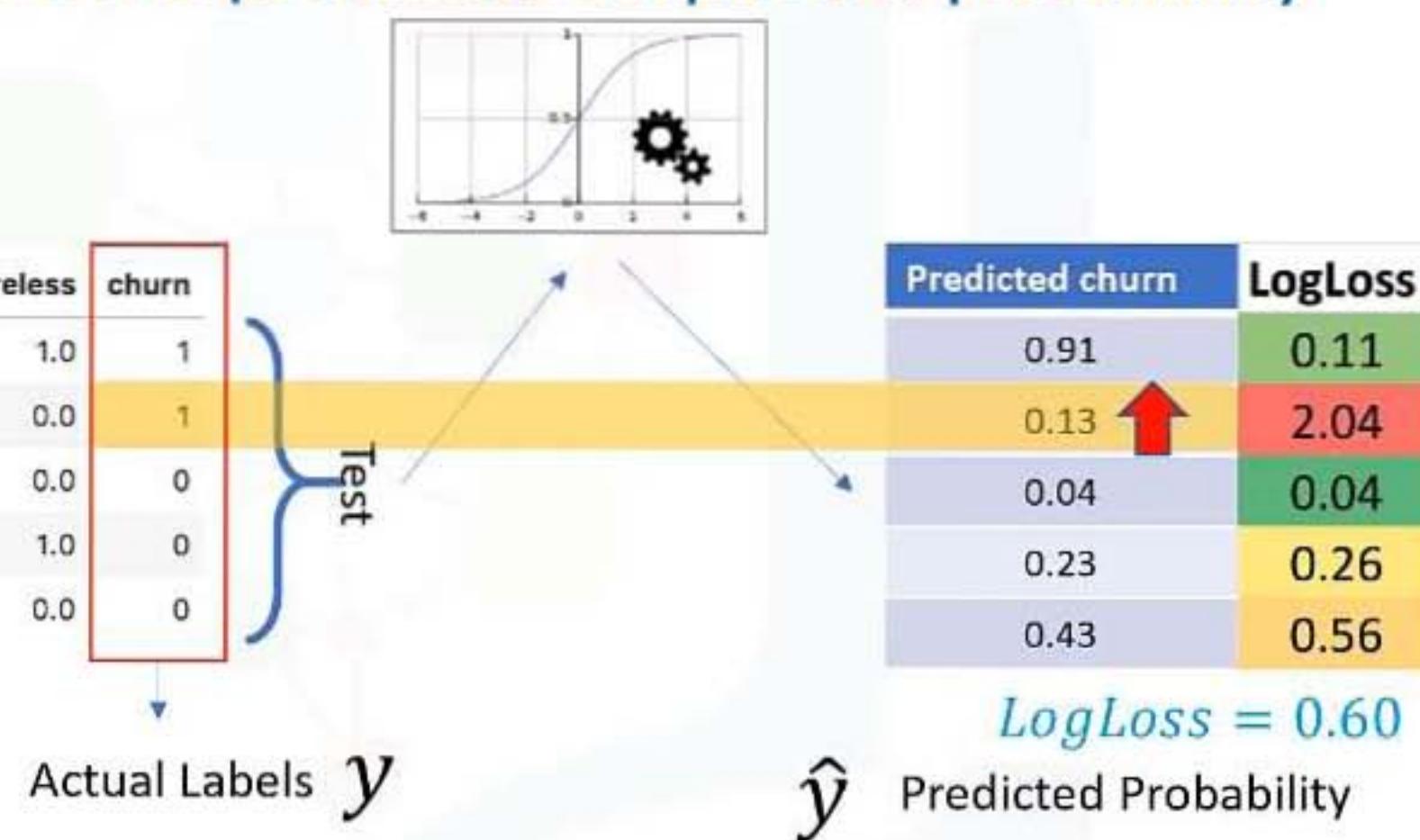
Log loss

Performance of a classifier where the predicted output is a probability value between 0 and 1.



Test set

	tenure	age	address	income	ed	employ	equip	callcard	wireless	churn
0	11.0	33.0	7.0	136.0	5.0	5.0	0.0	1.0	1.0	1
1	33.0	33.0	12.0	33.0	2.0	0.0	0.0	0.0	0.0	1
2	23.0	30.0	9.0	30.0	1.0	2.0	0.0	0.0	0.0	0
3	38.0	35.0	5.0	76.0	2.0	10.0	1.0	1.0	1.0	0
4	7.0	35.0	14.0	80.0	2.0	15.0	0.0	1.0	0.0	0



$$\text{LogLoss} = -\frac{1}{n} \sum (y \times \log(\hat{y}) + (1 - y) \times \log(1 - \hat{y}))$$

LogLoss: 0.00 ... 0.35 0.60 ... 1.00

← Higher Accuracy

Intro to Decision Trees

Saeed Aghabozorgi



© IBM Corporation. All rights reserved.

1

What is a decision tree?



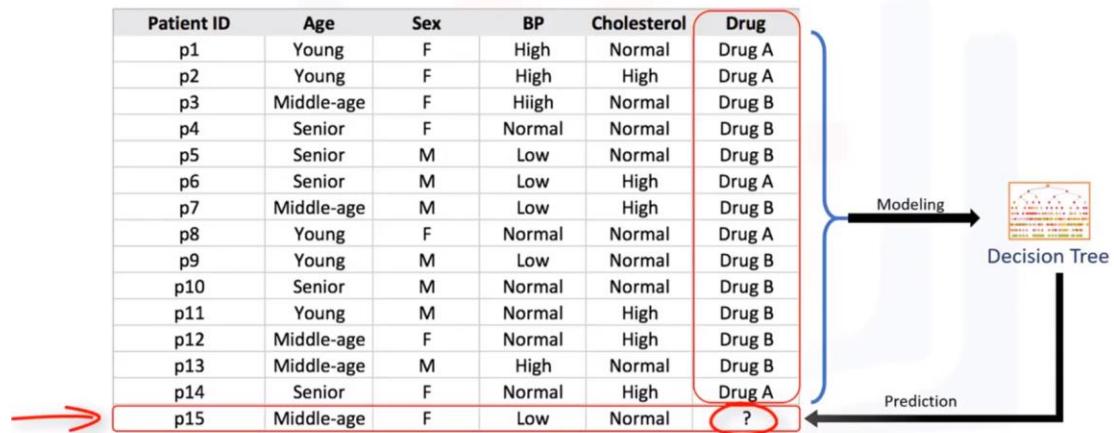
The basic intuition behind a decision tree is to map out all possible decision paths in the form of a tree.

Narendra Nath Joshi

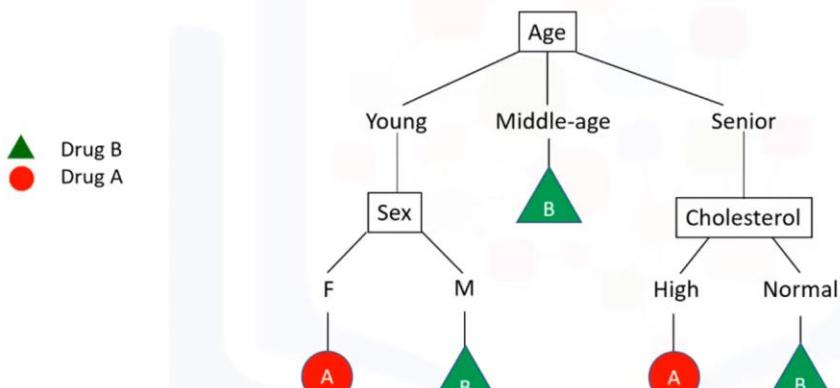


2

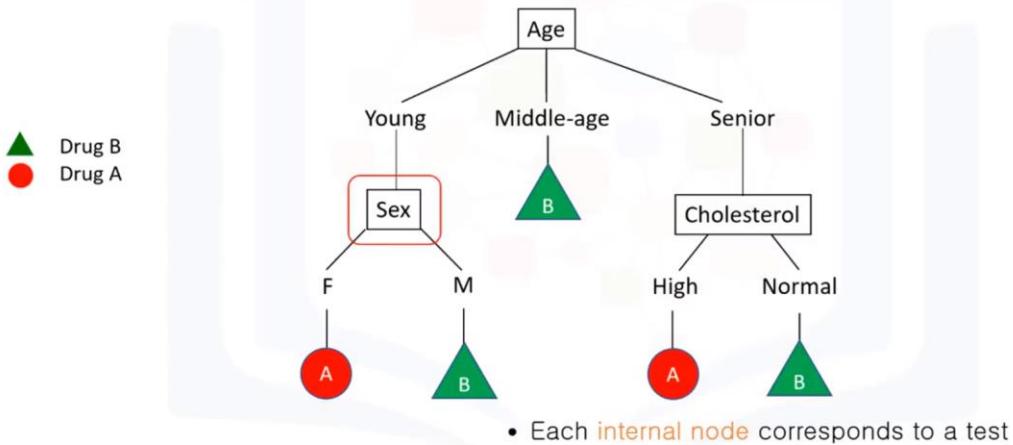
How to build a decision tree?



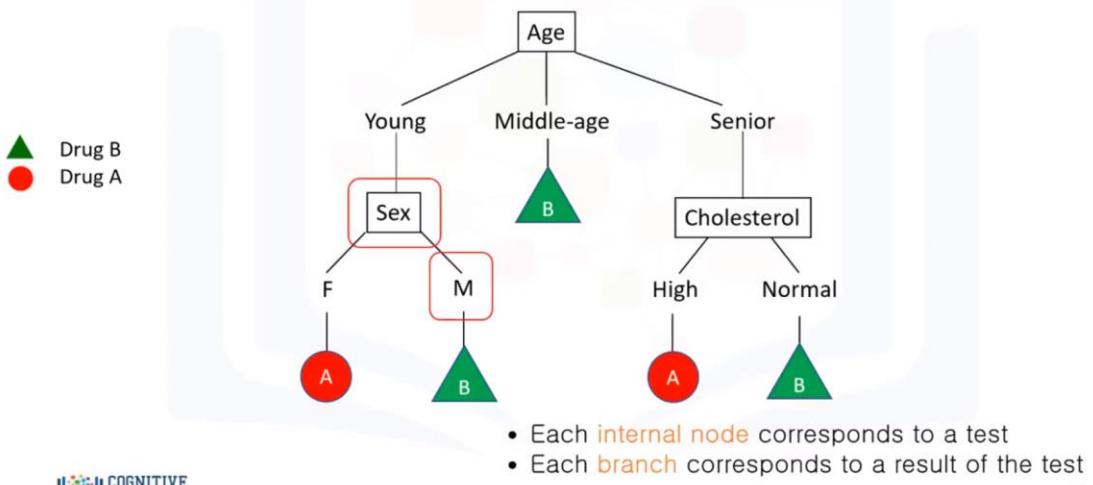
Building a decision tree with the training set



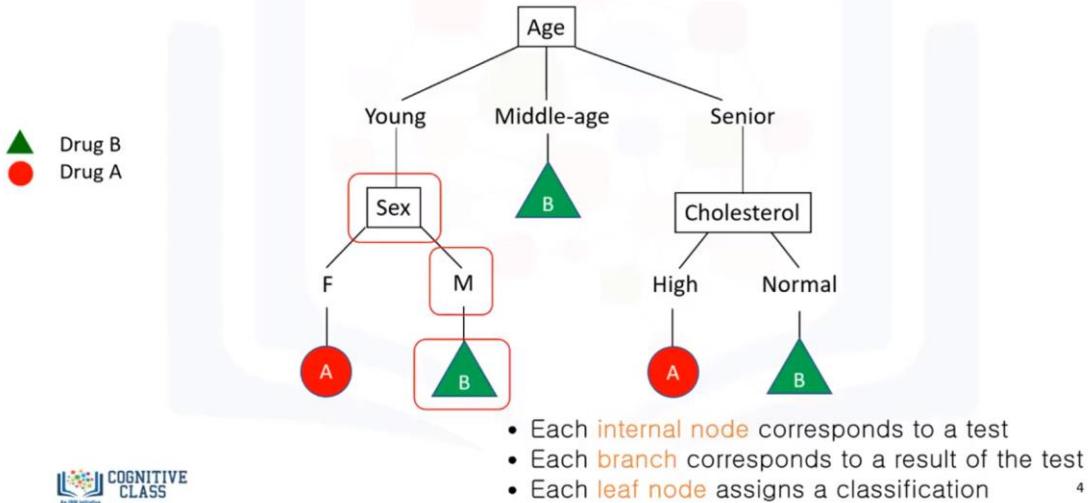
Building a decision tree with the training set



Building a decision tree with the training set

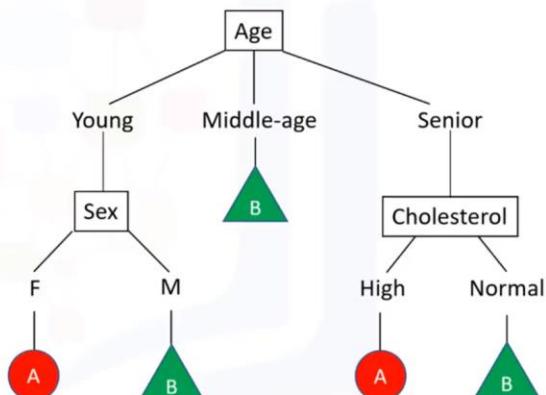


Building a decision tree with the training set



Decision tree learning algorithm

1. Choose an attribute from your dataset.
2. Calculate the significance of attribute in splitting of data.
3. Split data based on the value of the best attribute.
4. Go to step 1.

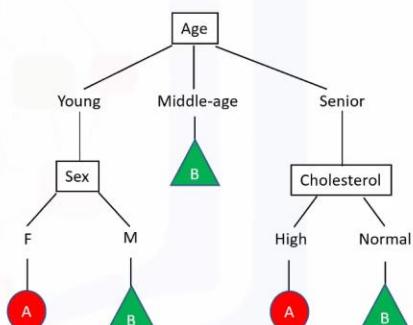


Building Decision Trees

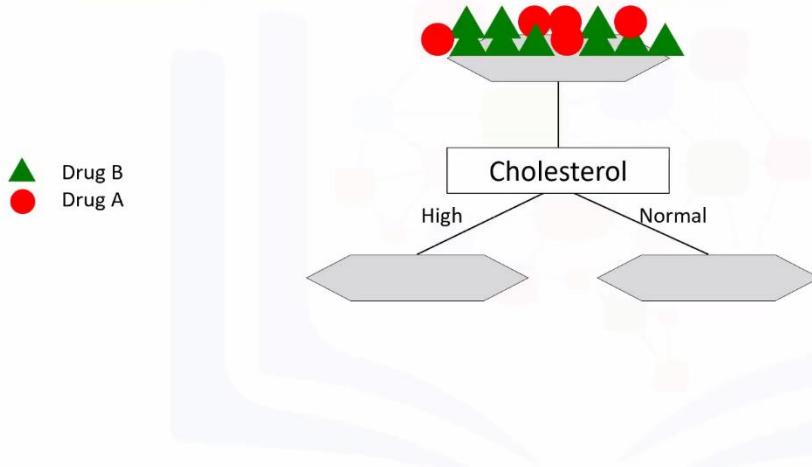
Saeed Aghabozorgi

How do you build a decision tree?

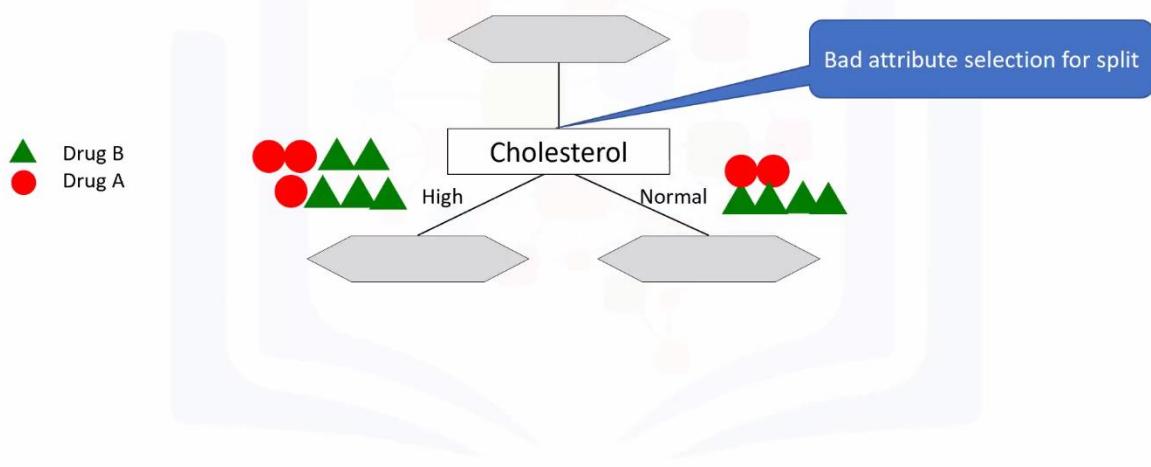
Patient ID	Age	Sex	BP	Cholesterol	Drug
p1	Young	F	High	Normal	Drug A
p2	Young	F	High	High	Drug A
p3	Middle-age	F	Hiigh	Normal	Drug B
p4	Senior	F	Normal	Normal	Drug B
p5	Senior	M	Low	Normal	Drug B
p6	Senior	M	Low	High	Drug A
p7	Middle-age	M	Low	High	Drug B
p8	Young	F	Normal	Normal	Drug A
p9	Young	M	Low	Normal	Drug B
p10	Senior	M	Normal	Normal	Drug B
p11	Young	M	Normal	High	Drug B
p12	Middle-age	F	Normal	High	Drug B
p13	Middle-age	M	High	Normal	Drug B
p14	Senior	F	Normal	High	Drug A
p15	Middle-age	F	Low	Normal	?



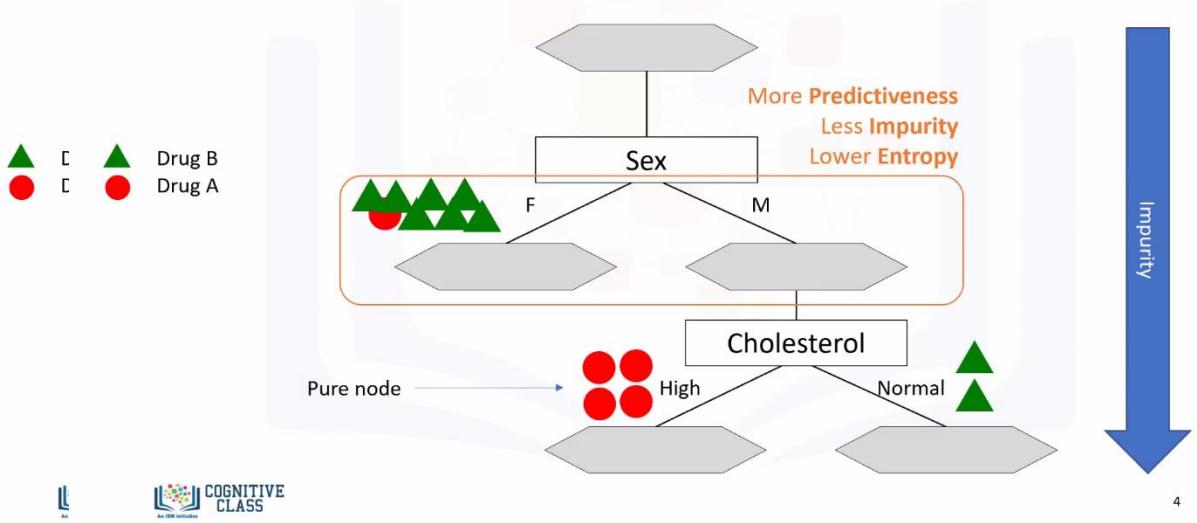
Which attribute is the best ?



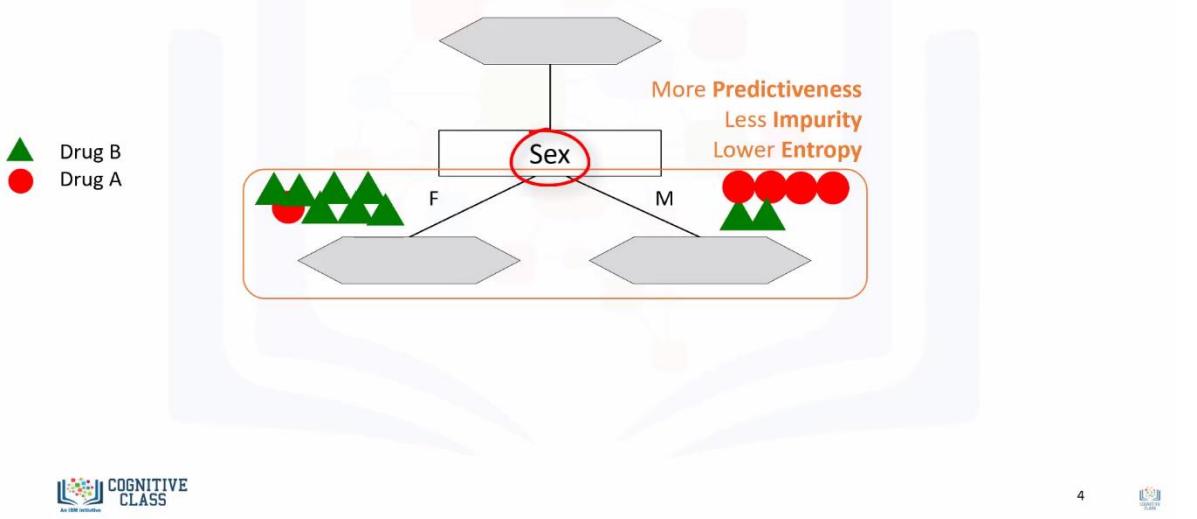
Which attribute is the best ?



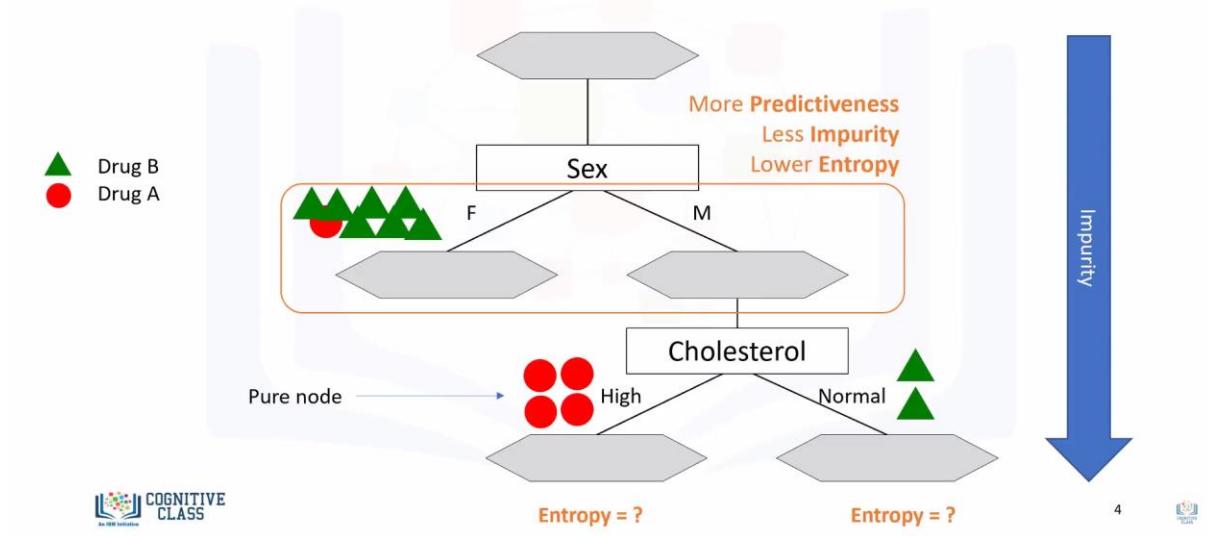
Which attribute is the best ?



Which attribute is the best ?



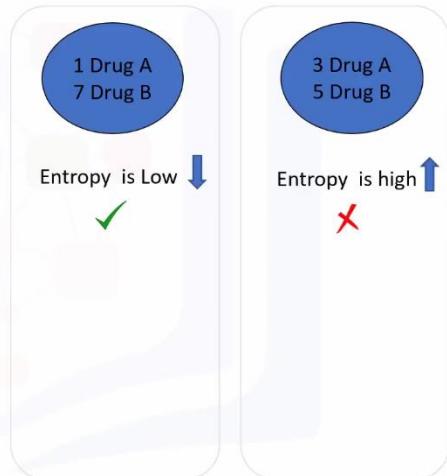
Which attribute is the best ?



Entropy

- Measure of randomness or uncertainty

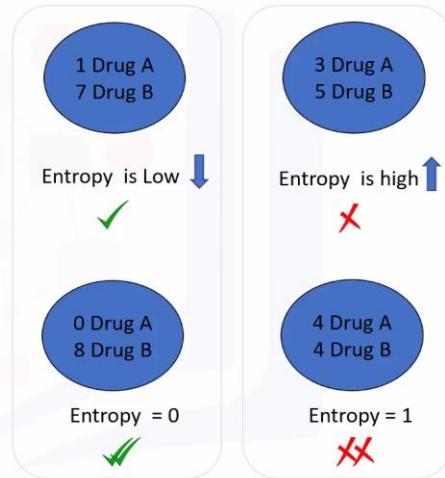
The lower the Entropy, the less uniform the distribution, the purer the node.



Entropy

- Measure of randomness or uncertainty

The lower the Entropy, the less uniform the distribution, the purer the node.

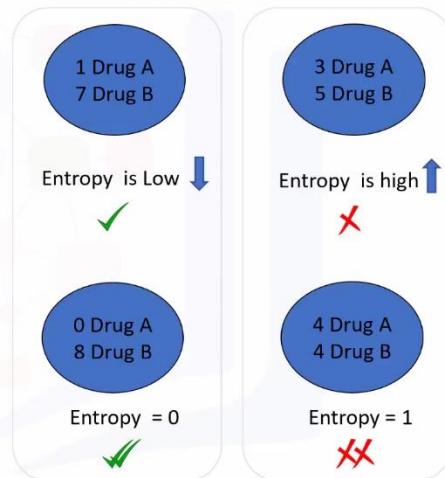


Entropy

- Measure of randomness or uncertainty

$$\text{Entropy} = -p(A)\log(p(A)) - p(B)\log(p(B))$$

The lower the Entropy, the less uniform the distribution, the purer the node.



Which attribute is the best one to use?

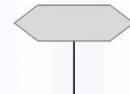
Patient ID	Age	Sex	BP	Cholesterol	Drug
p1	Young	F	High	Normal	Drug A
p2	Young	F	High	High	Drug A
p3	Middle-age	F	High	Normal	Drug B
p4	Senior	F	Normal	Normal	Drug B
p5	Senior	M	Low	Normal	Drug B
p6	Senior	M	Low	High	Drug A
p7	Middle-age	M	Low	High	Drug B
p8	Young	F	Normal	Normal	Drug A
p9	Young	M	Low	Normal	Drug B
p10	Senior	M	Normal	Normal	Drug B
p11	Young	M	Normal	High	Drug B
p12	Middle-age	F	Normal	High	Drug B
p13	Middle-age	M	High	Normal	Drug B
p14	Senior	F	Normal	High	Drug A

S: [9 B, 5 A]

$$E = - p(B) \log(p(B)) - p(A) \log(p(A))$$

$$E = - (9/14) \log(9/14) - (5/14) \log(5/14)$$

$$E = 0.940$$

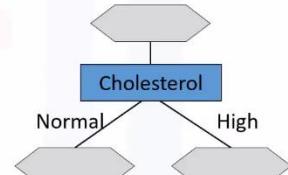


Is 'Cholesterol' the best attribute?

Patient ID	Age	Sex	BP	Cholesterol	Drug
p1	Young	F	High	Normal	Drug A
p2	Young	F	High	High	Drug A
p3	Middle-age	F	High	Normal	Drug B
p4	Senior	F	Normal	Normal	Drug B
p5	Senior	M	Low	Normal	Drug B
p6	Senior	M	Low	High	Drug A
p7	Middle-age	M	Low	High	Drug B
p8	Young	F	Normal	Normal	Drug A
p9	Young	M	Low	Normal	Drug B
p10	Senior	M	Normal	Normal	Drug B
p11	Young	M	Normal	High	Drug B
p12	Middle-age	F	Normal	High	Drug B
p13	Middle-age	M	High	Normal	Drug B
p14	Senior	F	Normal	High	Drug A

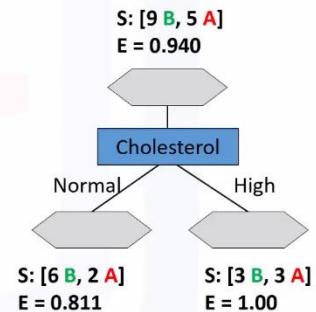
S: [9 B, 5 A]

$$E = 0.940$$



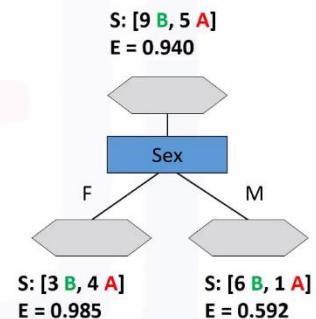
Is 'Cholesterol' the best attribute?

Patient ID	Age	Sex	BP	Cholesterol	Drug
p1	Young	F	High	Normal	Drug A
p2	Young	F	High	High	Drug A
p3	Middle-age	F	Hiigh	Normal	Drug B
p4	Senior	F	Normal	Normal	Drug B
p5	Senior	M	Low	Normal	Drug B
p6	Senior	M	Low	High	Drug A
p7	Middle-age	M	Low	High	Drug B
p8	Young	F	Normal	Normal	Drug A
p9	Young	M	Low	Normal	Drug B
p10	Senior	M	Normal	Normal	Drug B
p11	Young	M	Normal	High	Drug B
p12	Middle-age	F	Normal	High	Drug B
p13	Middle-age	M	High	Normal	Drug B
p14	Senior	F	Normal	High	Drug A

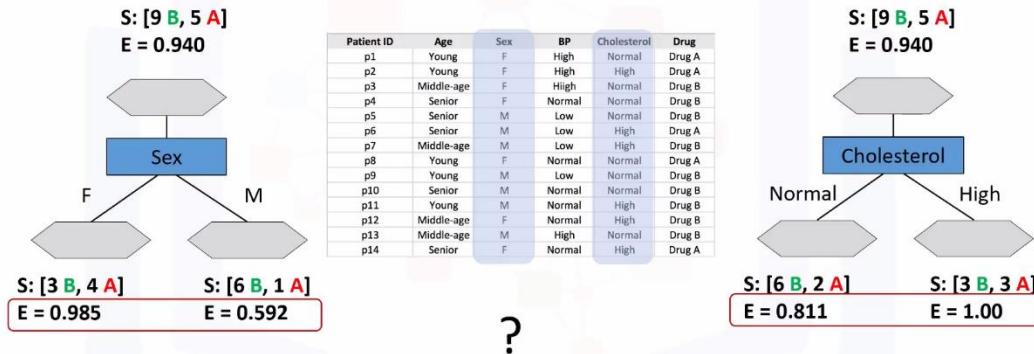


What about 'Sex'?

Patient ID	Age	Sex	BP	Cholesterol	Drug
p1	Young	F	High	Normal	Drug A
p2	Young	F	High	High	Drug A
p3	Middle-age	F	Hiigh	Normal	Drug B
p4	Senior	F	Normal	Normal	Drug B
p5	Senior	M	Low	Normal	Drug B
p6	Senior	M	Low	High	Drug A
p7	Middle-age	M	Low	High	Drug B
p8	Young	F	Normal	Normal	Drug A
p9	Young	M	Low	Normal	Drug B
p10	Senior	M	Normal	Normal	Drug B
p11	Young	M	Normal	High	Drug B
p12	Middle-age	F	Normal	High	Drug B
p13	Middle-age	M	High	Normal	Drug B
p14	Senior	F	Normal	High	Drug A



Which attribute is the best?

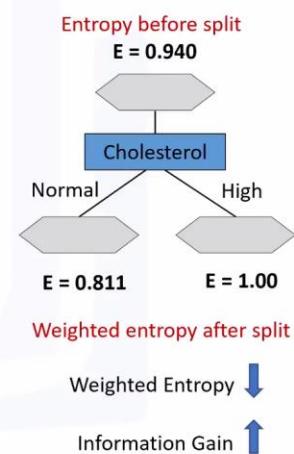


The tree with the higher Information Gain after splitting.

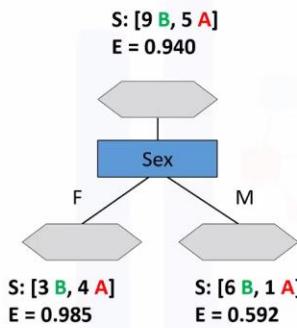
What is information gain?

Information gain is the information that can increase the level of certainty after splitting.

$$\text{Information Gain} = (\text{Entropy before split}) - (\text{weighted entropy after split})$$

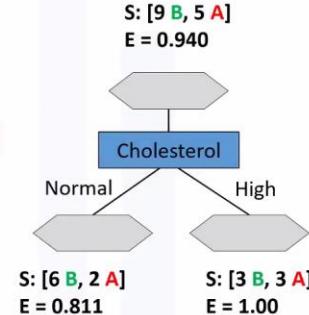


Which attribute is the best?



$$\text{Gain}(s, \text{Sex}) \\ = 0.940 - [(7/14)0.985 + (7/14)0.592] \\ = 0.151$$

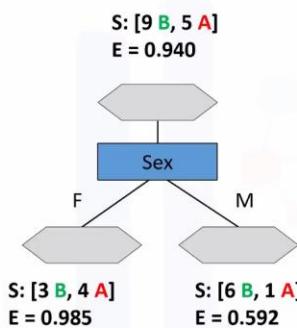
Patient ID	Age	Sex	BP	Cholesterol	Drug
p1	Young	F	High	Normal	Drug A
p2	Young	F	High	High	Drug A
p3	Middle-age	F	High	Normal	Drug B
p4	Senior	F	Normal	Normal	Drug B
p5	Senior	M	Low	Normal	Drug B
p6	Senior	M	Low	High	Drug B
p7	Middle-age	M	Normal	Normal	Drug A
p8	Young	F	Low	Normal	Drug A
p9	Young	M	Low	Normal	Drug B
p10	Senior	M	Normal	High	Drug B
p11	Young	M	Normal	High	Drug B
p12	Middle-age	F	Normal	High	Drug B
p13	Middle-age	M	High	Normal	Drug B
p14	Senior	F	Normal	High	Drug A



$$\text{Gain}(s, \text{Cholesterol}) \\ = 0.940 - [(8/14)0.811 + (6/14)1.0] \\ = 0.048$$

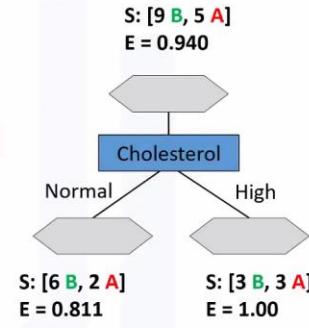
?

Which attribute is the best?



$$\text{Gain}(s, \text{Sex}) \\ = 0.940 - [(7/14)0.985 + (7/14)0.592] \\ = 0.151$$

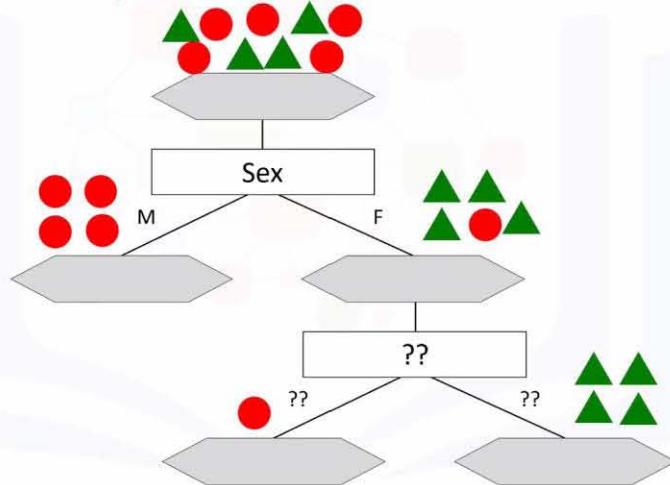
Patient ID	Age	Sex	BP	Cholesterol	Drug
p1	Young	F	High	Normal	Drug A
p2	Young	F	High	High	Drug A
p3	Middle-age	F	High	Normal	Drug B
p4	Senior	F	Normal	Normal	Drug B
p5	Senior	M	Low	Normal	Drug B
p6	Senior	M	Low	High	Drug A
p7	Middle-age	M	Low	Normal	Drug B
p8	Young	F	Normal	High	Drug A
p9	Young	M	Low	Normal	Drug B
p10	Senior	M	Normal	Normal	Drug B
p11	Young	M	Normal	High	Drug B
p12	Middle-age	F	Normal	High	Drug B
p13	Middle-age	M	High	Normal	Drug B
p14	Senior	F	Normal	High	Drug A



$$\text{Gain}(s, \text{Cholesterol}) \\ = 0.940 - [(8/14)0.811 + (6/14)1.0] \\ = 0.048$$

?

Correct way to build a decision tree



Intro to Logistic Regression

Saeed Aghabozorgi

What is logistic regression?

Logistic regression is a classification algorithm for categorical variables.

	tenure	age	address	income	ed	employ	equip	callcard	wireless	churn
0	11.0	33.0	7.0	136.0	5.0	5.0	0.0	1.0	1.0	Yes
1	33.0	33.0	12.0	33.0	2.0	0.0	0.0	0.0	0.0	Yes
2	23.0	30.0	9.0	30.0	1.0	2.0	0.0	0.0	0.0	No
3	38.0	35.0	5.0	76.0	2.0	10.0	1.0	1.0	1.0	No
4	7.0	35.0	14.0	80.0	2.0	15.0	0.0	1.0	0.0	?

What is logistic regression?

Logistic regression is a classification algorithm for categorical variables.

	Independent variables										Dependent variable
	tenure	age	address	income	ed	employ	equip	callcard	wireless	churn	
0	11.0	33.0	7.0	136.0	5.0	5.0	0.0	1.0	1.0	Yes	
1	33.0	33.0	12.0	33.0	2.0	0.0	0.0	0.0	0.0	Yes	
2	23.0	30.0	9.0	30.0	1.0	2.0	0.0	0.0	0.0	No	
3	38.0	35.0	5.0	76.0	2.0	10.0	1.0	1.0	1.0	No	
4	7.0	35.0	14.0	80.0	2.0	15.0	0.0	1.0	0.0	?	

What is logistic regression?

Logistic regression is a classification algorithm for categorical variables.

	Independent variables										Dependent variable
	tenure	age	address	income	ed	employ	equip	callcard	wireless	churn	
0	11.0	33.0	7.0	136.0	5.0	5.0	0.0	1.0	1.0	Yes	
1	33.0	33.0	12.0	33.0	2.0	0.0	0.0	0.0	0.0	Yes	
2	23.0	30.0	9.0	30.0	1.0	2.0	0.0	0.0	0.0	No	
3	38.0	35.0	5.0	76.0	2.0	10.0	1.0	1.0	1.0	No	
4	7.0	35.0	14.0	80.0	2.0	15.0	0.0	1.0	0.0	?	

Continuous/Categorical variables

Categorical Variable

What is logistic regression?

Logistic regression is a classification algorithm for categorical variables.

	Independent variables										Dependent variable
	tenure	age	address	income	ed	employ	equip	callcard	wireless	churn	
0	11.0	33.0	7.0	136.0	5.0	5.0	0.0	1.0	1.0	Yes	
1	33.0	33.0	12.0	33.0	2.0	0.0	0.0	0.0	0.0	Yes	
2	23.0	30.0	9.0	30.0	1.0	2.0	0.0	0.0	0.0	No	
3	38.0	35.0	5.0	76.0	2.0	10.0	1.0	1.0	1.0	No	
4	7.0	35.0	14.0	80.0	2.0	15.0	0.0	1.0	0.0	?	

Continuous/Categorical variables

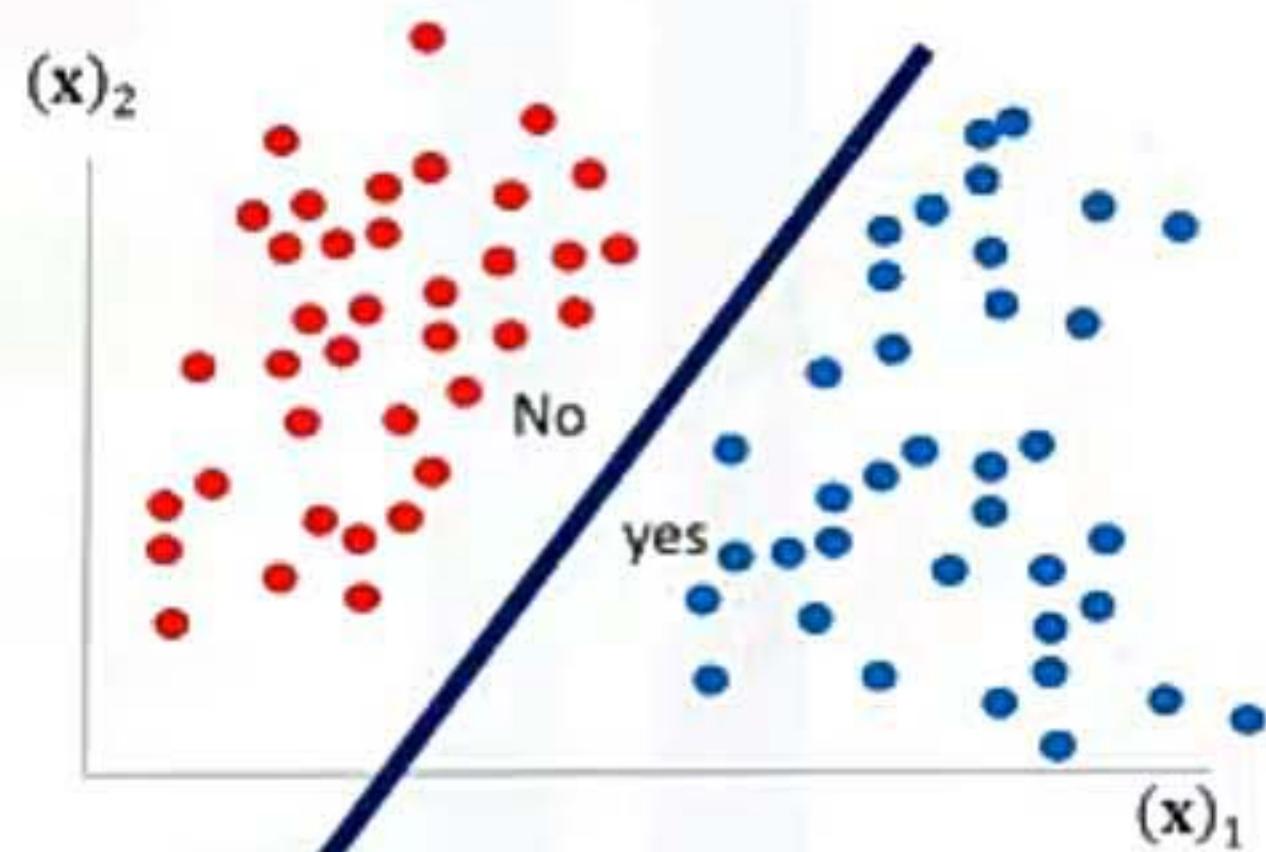
Categorical Variable

Logistic regression applications

- Predicting the probability of a person having a heart attack
- Predicting the mortality in injured patients
- Predicting a customer's propensity to purchase a product or halt a subscription
- Predicting the probability of failure of a given process or product
- Predicting the likelihood of a homeowner defaulting on a mortgage

When is logistic regression suitable?

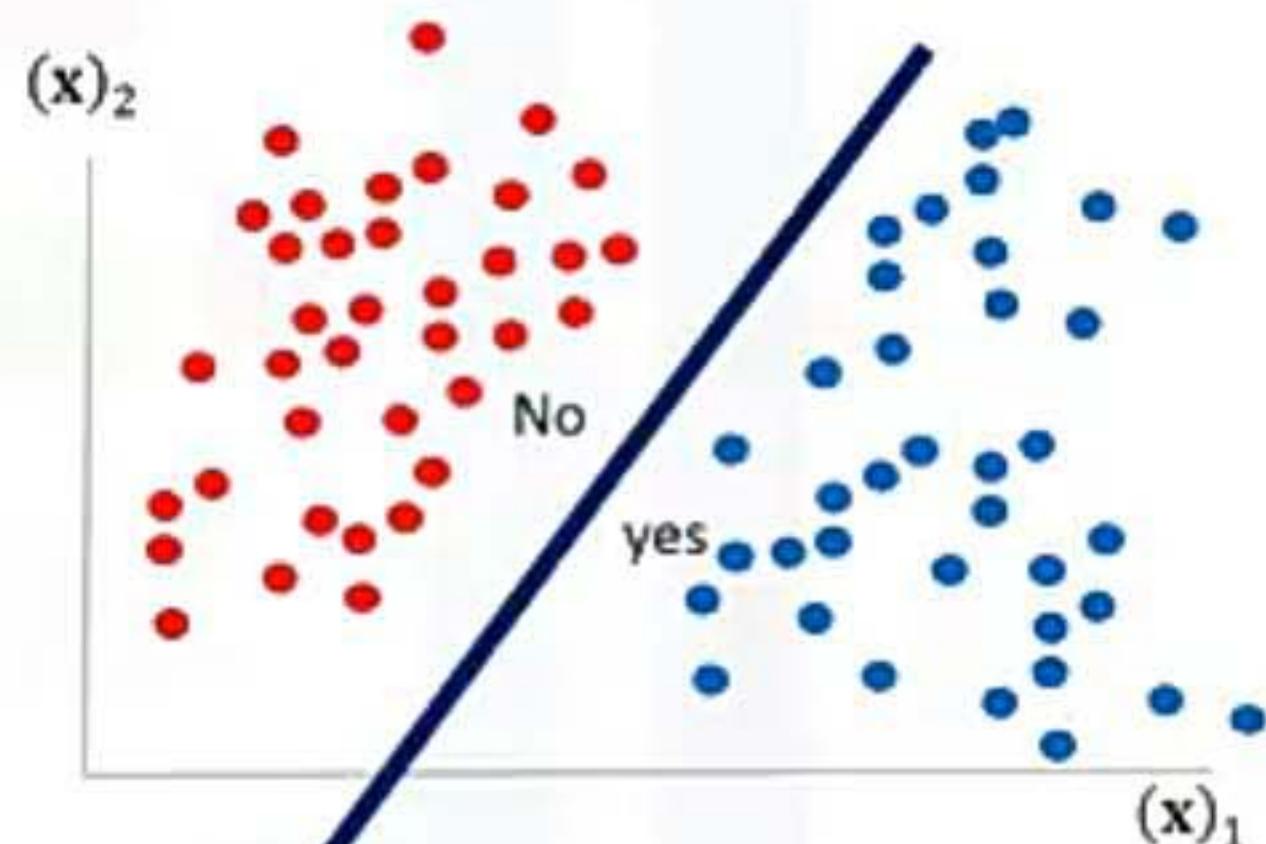
- If your data is binary
 - 0/1, YES/NO, True/False
- If you need probabilistic results
- When you need a linear decision boundary



$$\theta_0 + \theta_1 x_1 + \theta_2 x_2 > 0$$

When is logistic regression suitable?

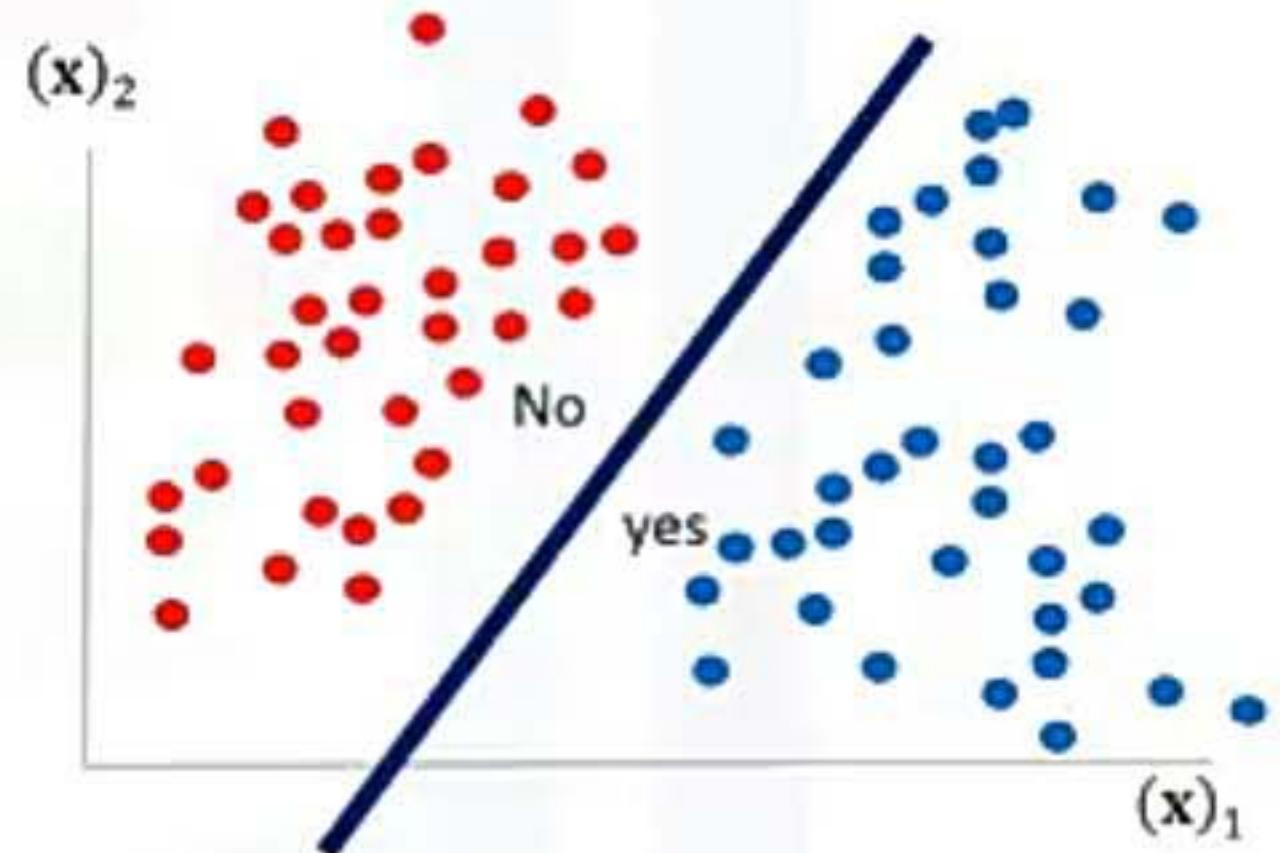
- If your data is binary
 - 0/1, YES/NO, True/False
- If you need probabilistic results
- When you need a linear decision boundary



$$\theta_0 + \theta_1 x_1 + \theta_2 x_2 > 0$$

When is logistic regression suitable?

- If your data is binary
 - 0/1, YES/NO, True/False
- If you need probabilistic results
- When you need a linear decision boundary
- If you need to understand the impact of a feature



$$\theta_0 + \theta_1 x_1 + \theta_2 x_2 > 0$$

Building a model for customer churn

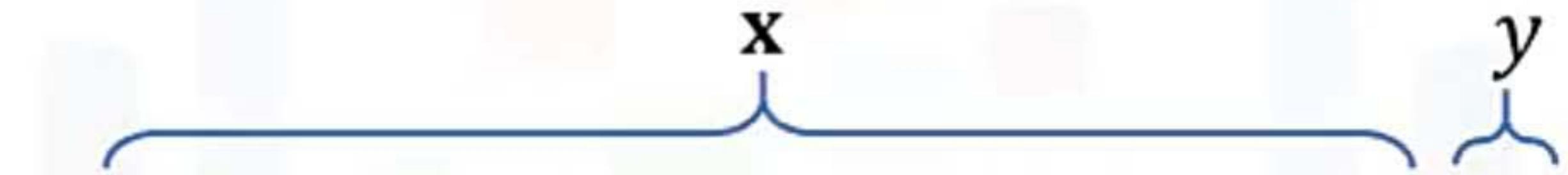
The diagram illustrates a machine learning model architecture. At the top, a red-outlined circle contains the letter 'X'. A blue bracket below it spans across the first eleven columns of the table, indicating the input feature matrix. To the right of the table, a blue bracket points upwards to a single vertical line ending in a blue-outlined circle containing the letter 'y', representing the output target variable.

	tenure	age	address	income	ed	employ	equip	callcard	wireless	churn
0	11.0	33.0	7.0	136.0	5.0	5.0	0.0	1.0	1.0	1.0
1	33.0	33.0	12.0	33.0	2.0	0.0	0.0	0.0	0.0	1.0
2	23.0	30.0	9.0	30.0	1.0	2.0	0.0	0.0	0.0	0.0
3	38.0	35.0	5.0	76.0	2.0	10.0	1.0	1.0	1.0	0.0

$$X \in \mathbb{R}^{m \times n}$$

$$y \in \{0,1\}$$

Building a model for customer churn



The diagram shows a horizontal bracket above the feature names labeled 'x' and a vertical bracket to the right of the target name labeled 'y'.

	tenure	age	address	income	ed	employ	equip	callcard	wireless	churn
0	11.0	33.0	7.0	136.0	5.0	5.0	0.0	1.0	1.0	1.0
1	33.0	33.0	12.0	33.0	2.0	0.0	0.0	0.0	0.0	1.0
2	23.0	30.0	9.0	30.0	1.0	2.0	0.0	0.0	0.0	0.0
3	38.0	35.0	5.0	76.0	2.0	10.0	1.0	1.0	1.0	0.0

$$X \in \mathbb{R}^{m \times n}$$
$$y \in \{0,1\}$$

$$\hat{y} = P(y=1|x)$$

Building a model for customer churn

	tenure	age	address	income	ed	employ	equip	callcard	wireless	churn
0	11.0	33.0	7.0	136.0	5.0	5.0	0.0	1.0	1.0	1.0
1	33.0	33.0	12.0	33.0	2.0	0.0	0.0	0.0	0.0	1.0
2	23.0	30.0	9.0	30.0	1.0	2.0	0.0	0.0	0.0	0.0
3	38.0	35.0	5.0	76.0	2.0	10.0	1.0	1.0	1.0	0.0

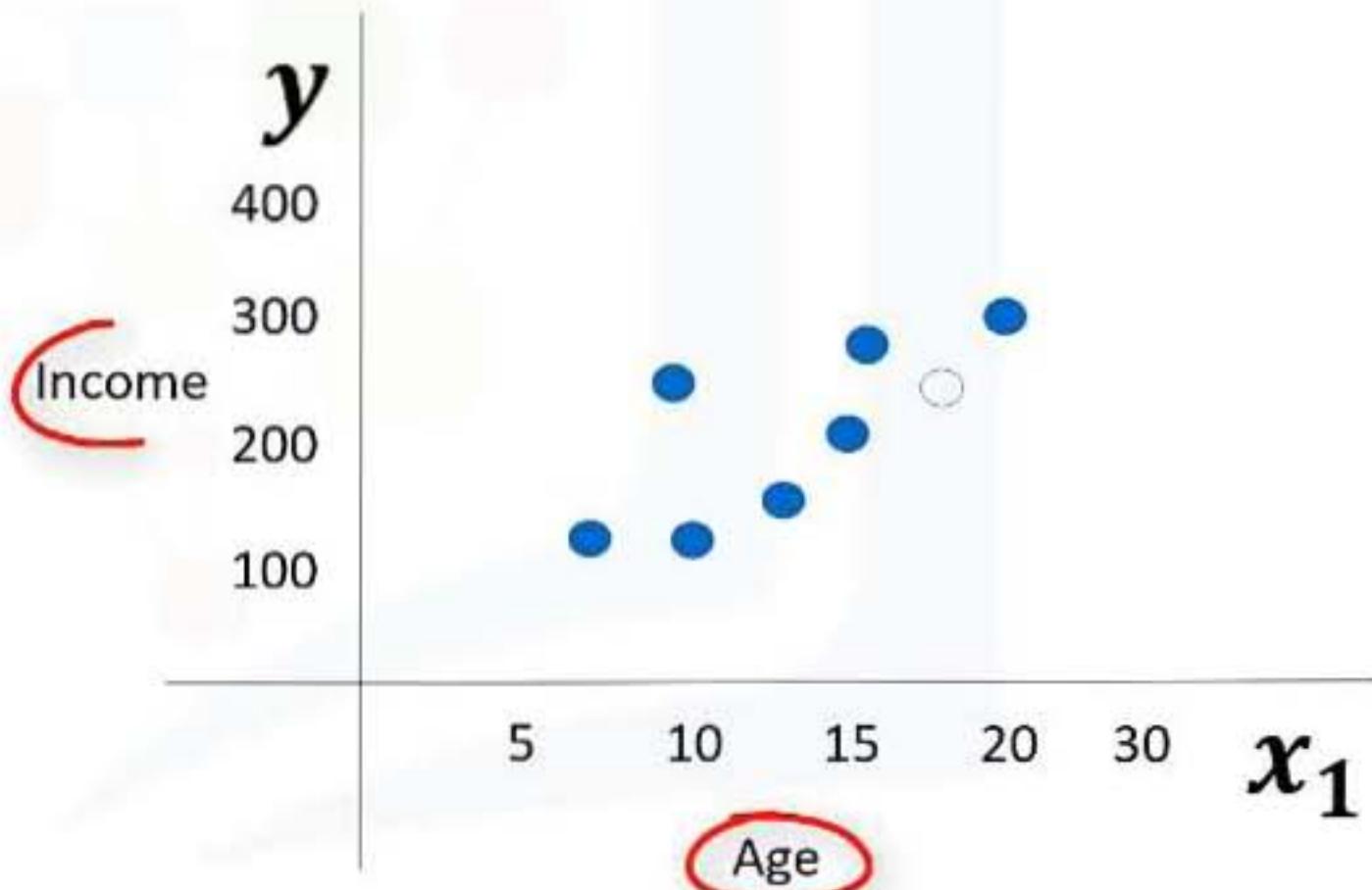
$$X \in \mathbb{R}^{m \times n}$$
$$y \in \{0,1\}$$

$$\hat{y} = P(y=1|x)$$

$$P(y=0|x) = 1 - P(y=1|x)$$

Predicting customer income

	tenure	age	address	income	ed	employ	equip	callcard	wireless	churn
0	11.0	33.0	7.0	136.0	5.0	5.0	0.0	1.0	1.0	1
1	33.0	33.0	12.0	33.0	2.0	0.0	0.0	0.0	0.0	1
2	23.0	30.0	9.0	30.0	1.0	2.0	0.0	0.0	0.0	0
3	38.0	35.0	5.0	76.0	2.0	10.0	1.0	1.0	1.0	0
4	7.0	35.0	14.0	80.0	2.0	15.0	0.0	1.0	0.0	0



Predicting churn using linear regression

	tenure	age	address	income	ed	employ	equip	callcard	wireless	churn
0	11.0	33.0	7.0	136.0	5.0	5.0	0.0	1.0	1.0	1
1	33.0	33.0	12.0	33.0	2.0	0.0	0.0	0.0	0.0	1
2	23.0	30.0	9.0	30.0	1.0	2.0	0.0	0.0	0.0	0
3	38.0	35.0	5.0	76.0	2.0	10.0	1.0	1.0	1.0	0
4	7.0	35.0	14.0	80.0	2.0	15.0	0.0	1.0	0.0	0

y

Churn

Yes (1)



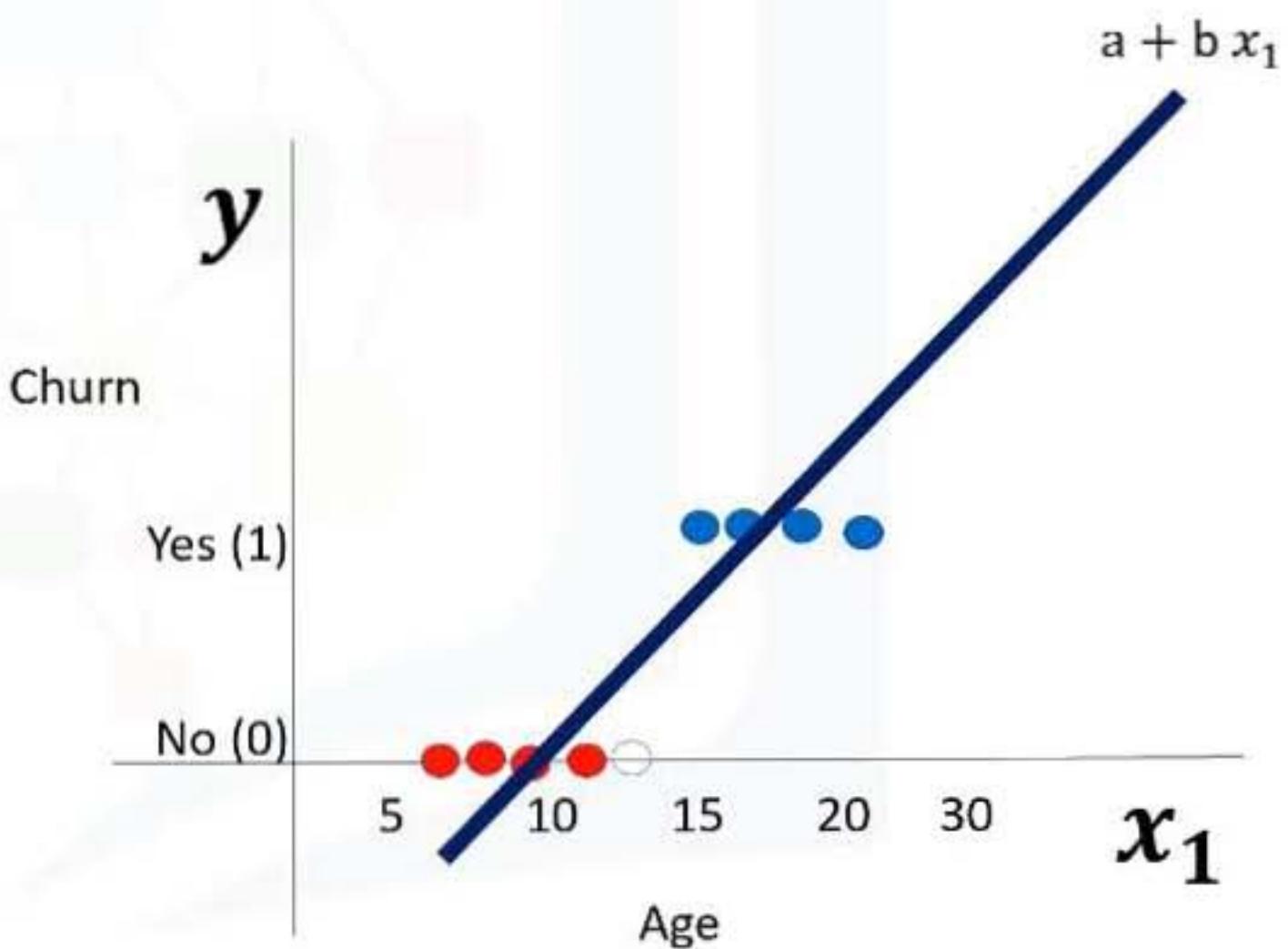
No (0)



Age

x_1

Predicting churn using linear regression

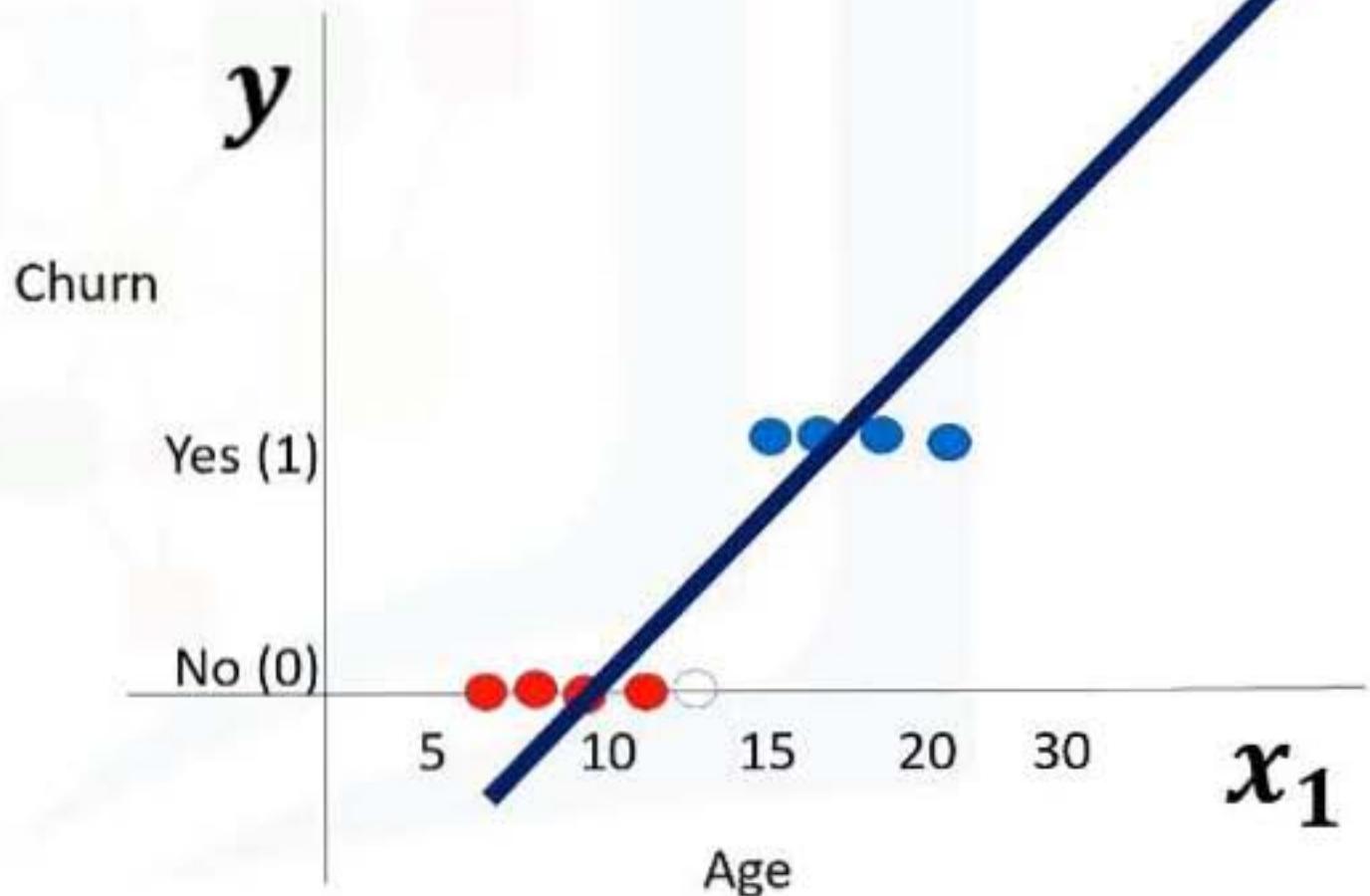


Predicting churn using linear regression

$$\theta^T = [\theta_0, \theta_1]$$

$$\theta_0 + \theta_1 x_1$$

$$a + b x_1$$



Predicting churn using linear regression

$$\theta^T X = \theta_0 + \theta_1 x_1$$

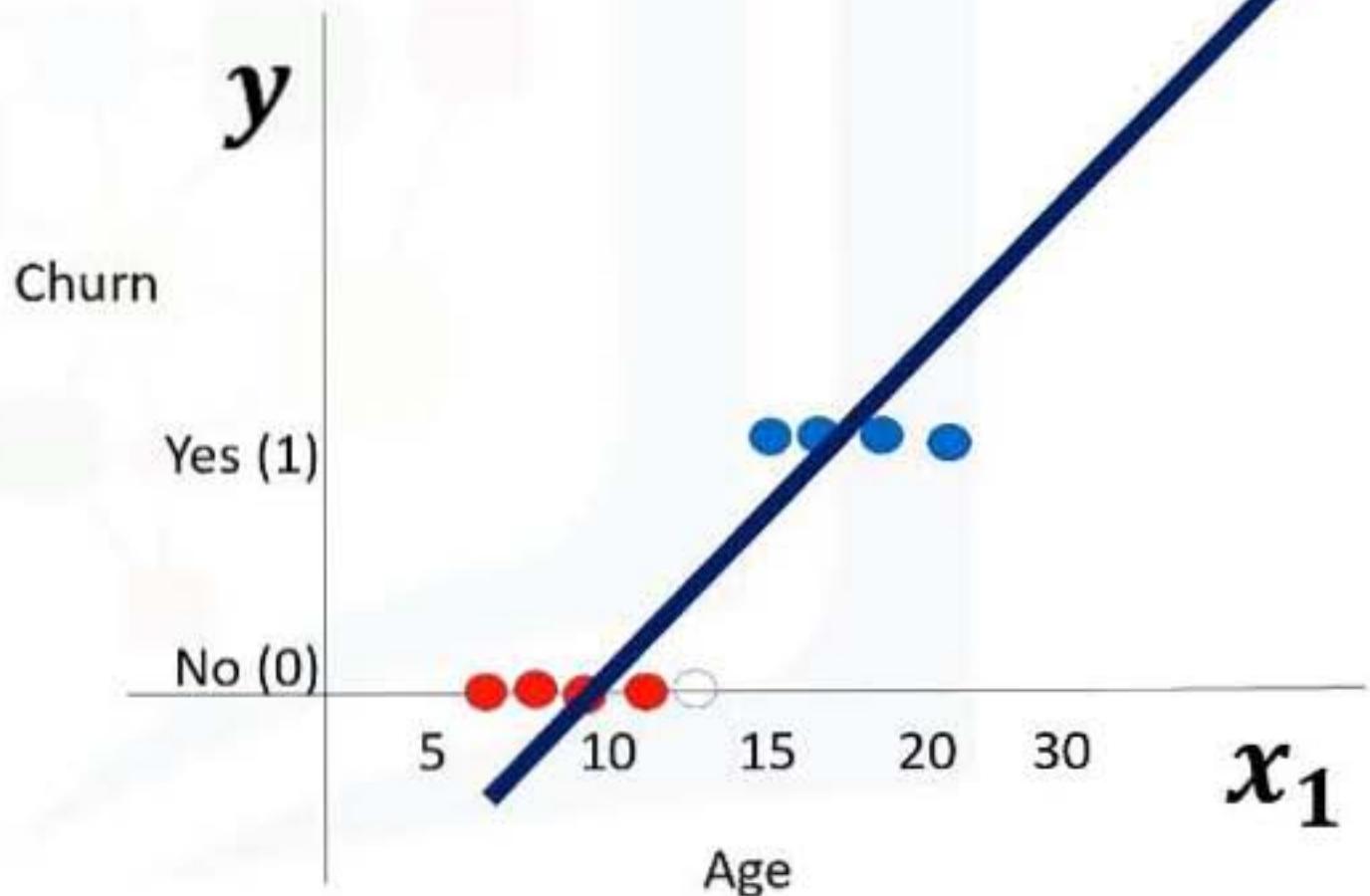
$$\theta^T X = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \dots$$

$$\theta^T = [\theta_0, \theta_1, \theta_2, \dots]$$

$$\theta^T = [\theta_0, \theta_1]$$

$$\theta_0 + \theta_1 x_1$$

$$a + b x_1$$



Predicting churn using linear regression

$$\theta^T X = \theta_0 + \theta_1 x_1$$

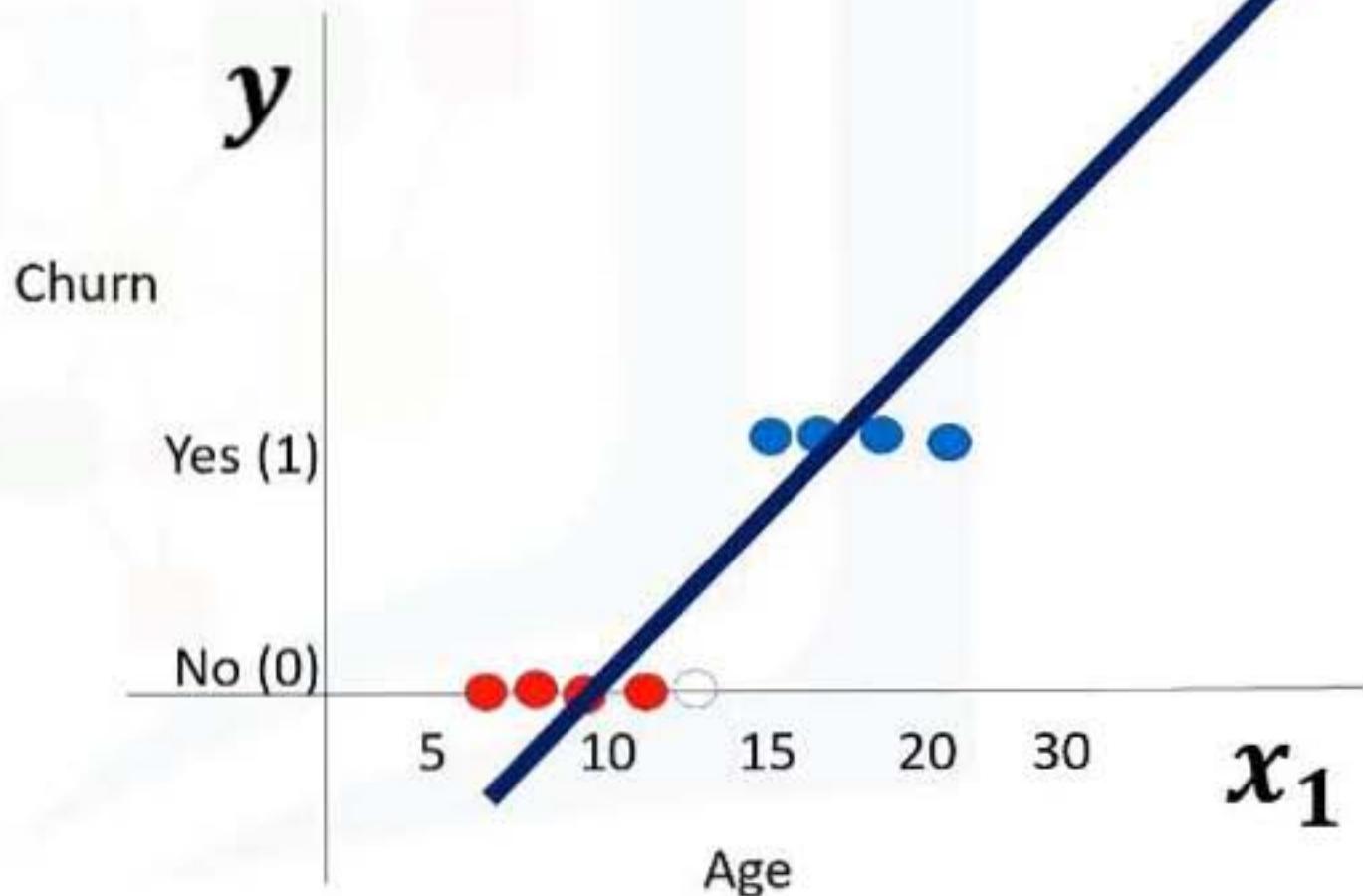
$$\theta^T X = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \dots$$

$$\theta^T = [\theta_0, \theta_1, \theta_2, \dots] \quad X = \begin{bmatrix} 1 \\ x_1 \\ x_2 \\ \dots \end{bmatrix}$$

$$\theta^T = [\theta_0, \theta_1]$$

$$\theta_0 + \theta_1 x_1$$

$$a + b x_1$$



Predicting churn using linear regression

$$\theta^T X = \theta_0 + \theta_1 x_1$$

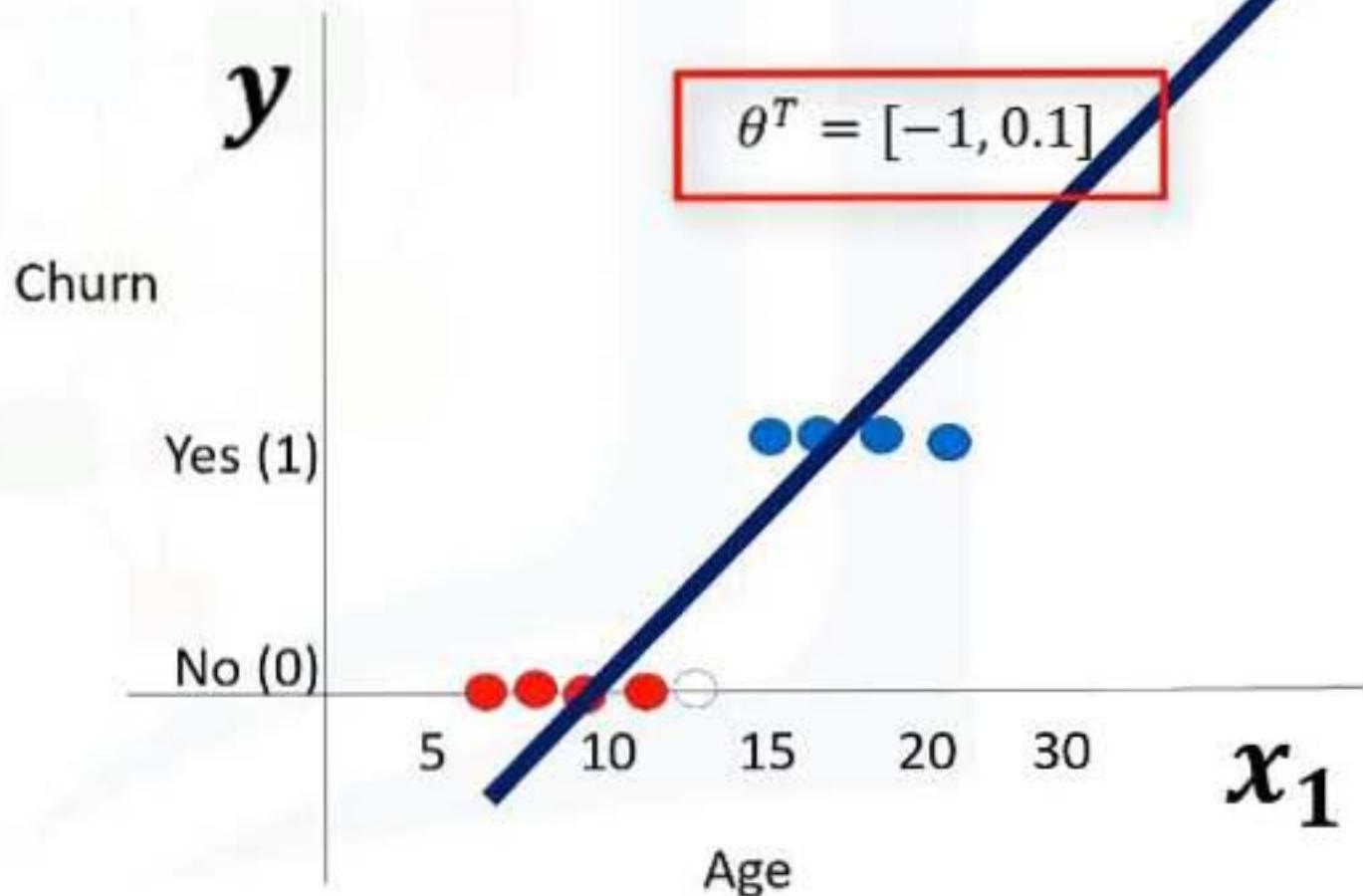
$$\theta^T X = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \dots$$

$$\theta^T = [\theta_0, \theta_1, \theta_2, \dots] \quad X = \begin{bmatrix} 1 \\ x_1 \\ x_2 \\ \dots \end{bmatrix}$$

$$\theta^T = [\theta_0, \theta_1]$$

$$\theta_0 + \theta_1 x_1$$

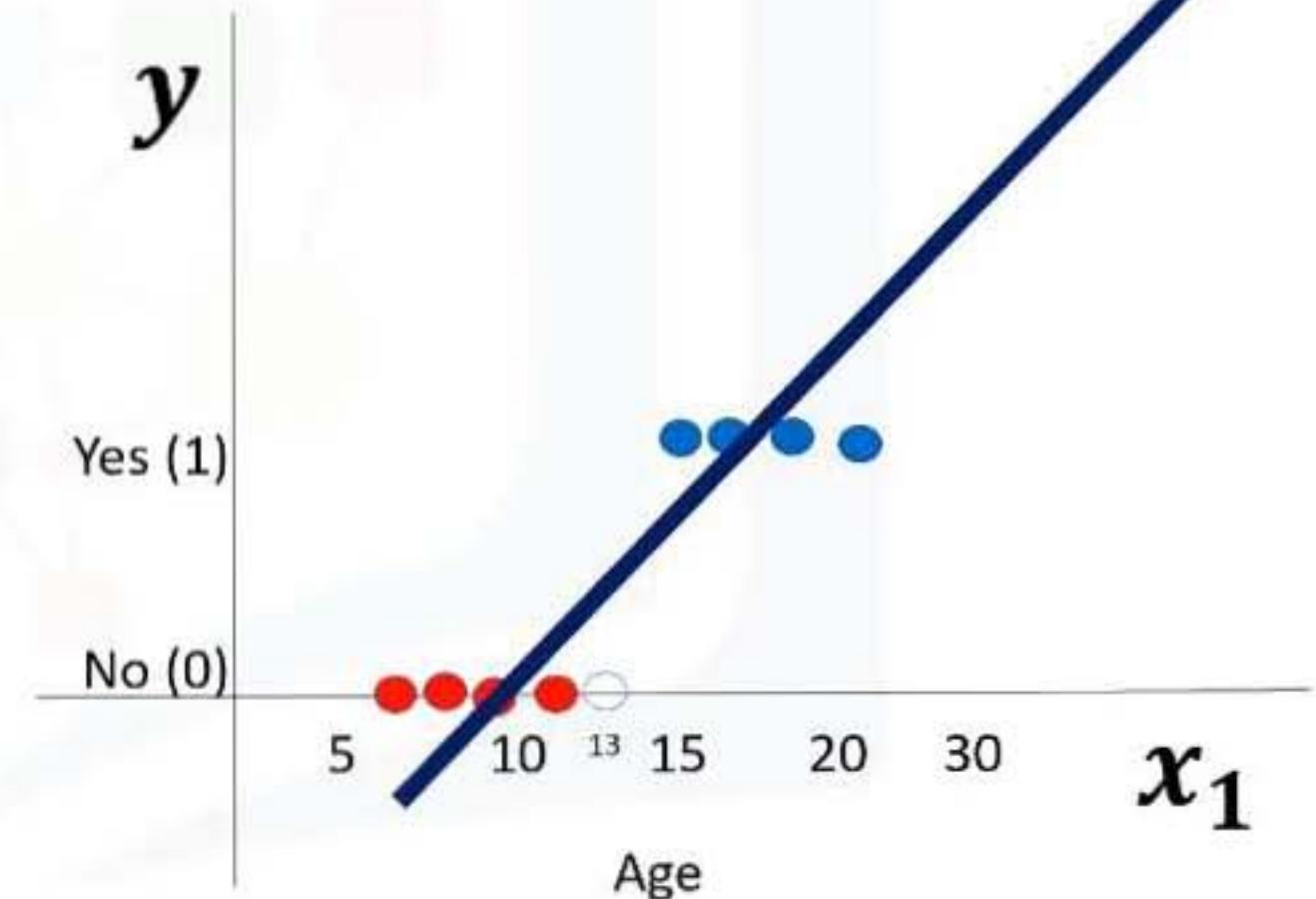
$$a + b x_1$$



Linear regression in classification problems?

$$\theta^T X = \theta_0 + \theta_1 x_1$$

$$\theta^T X = -1 + 0.1 \cdot x_1$$

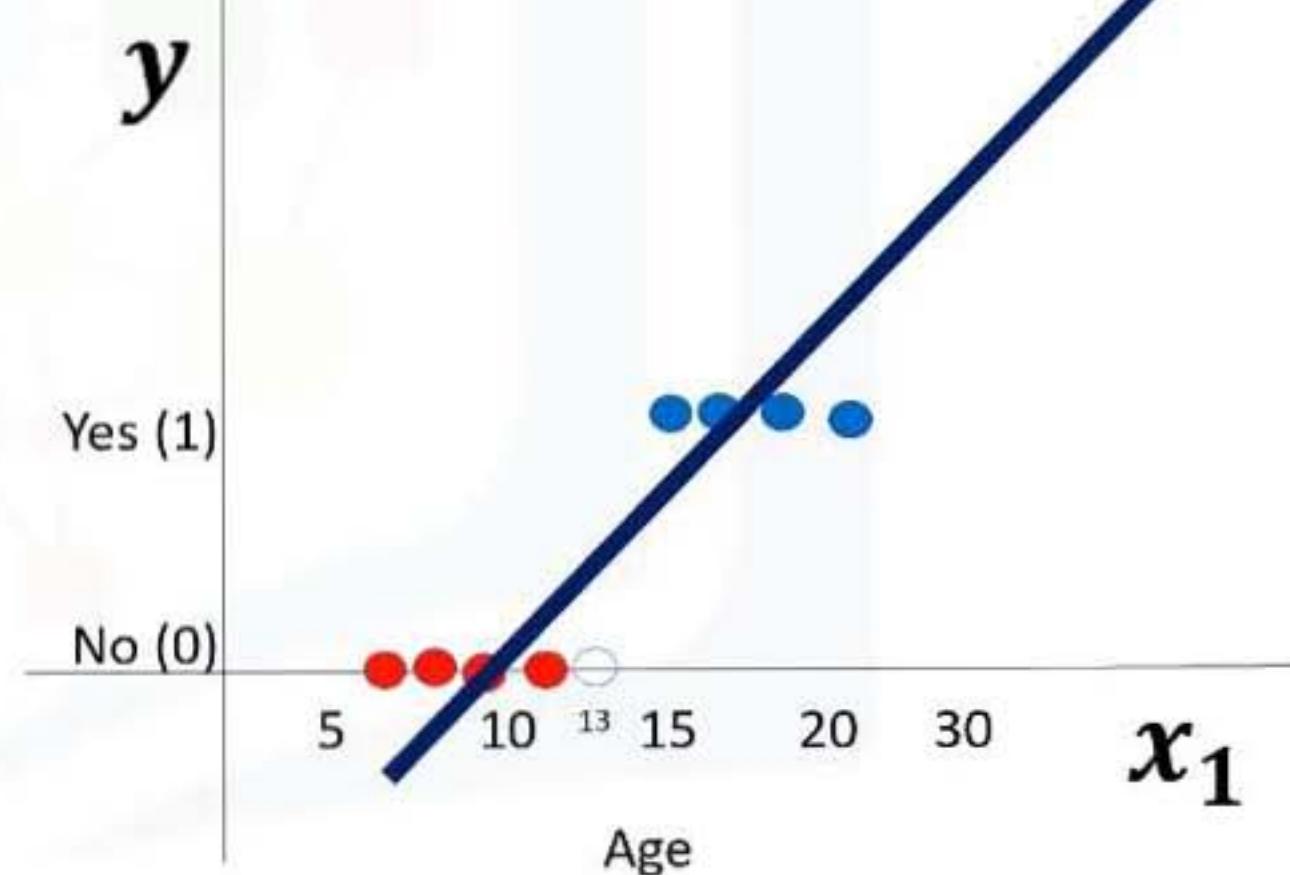


Linear regression in classification problems?

$$\theta^T X = \theta_0 + \theta_1 x_1$$

$$p_1 = [13] \rightarrow \theta^T X = -1 + 0.1 \cdot x_1 \\ = -1 + 0.1 \times 13 \\ = 0.3$$

$$\theta^T X = -1 + 0.1 \cdot x_1$$



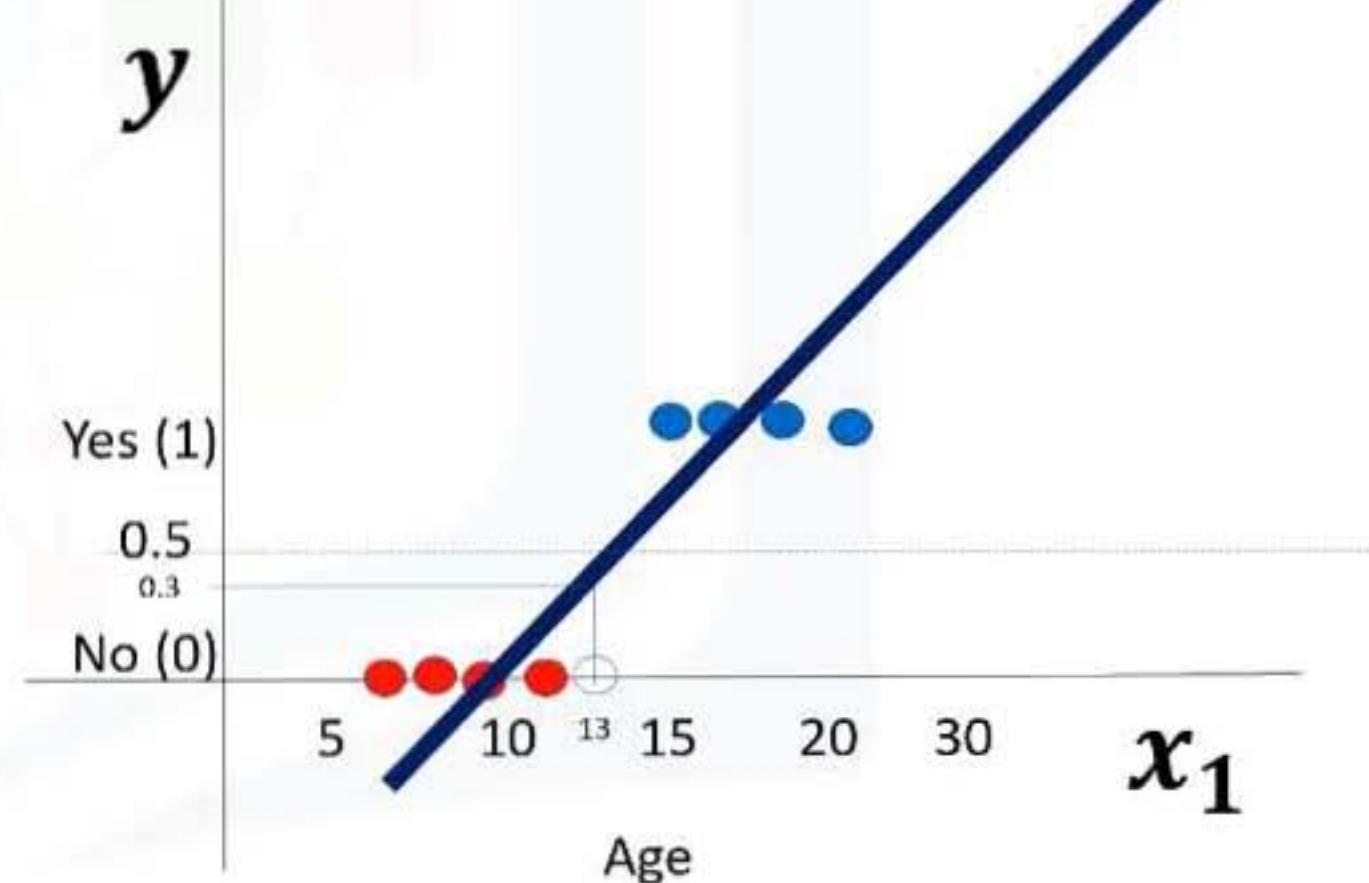
Linear regression in classification problems?

$$\theta^T X = \theta_0 + \theta_1 x_1$$

$$p_1 = [13] \rightarrow \theta^T X = -1 + 0.1 \cdot x_1 \\ = -1 + 0.1 \times 13 \\ = 0.3$$

$$\hat{y} = \begin{cases} 0 & \text{if } \theta^T X < 0.5 \\ 1 & \text{if } \theta^T X \geq 0.5 \end{cases}$$

$$\theta^T X = -1 + 0.1 \cdot x_1$$



Linear regression in classification problems?

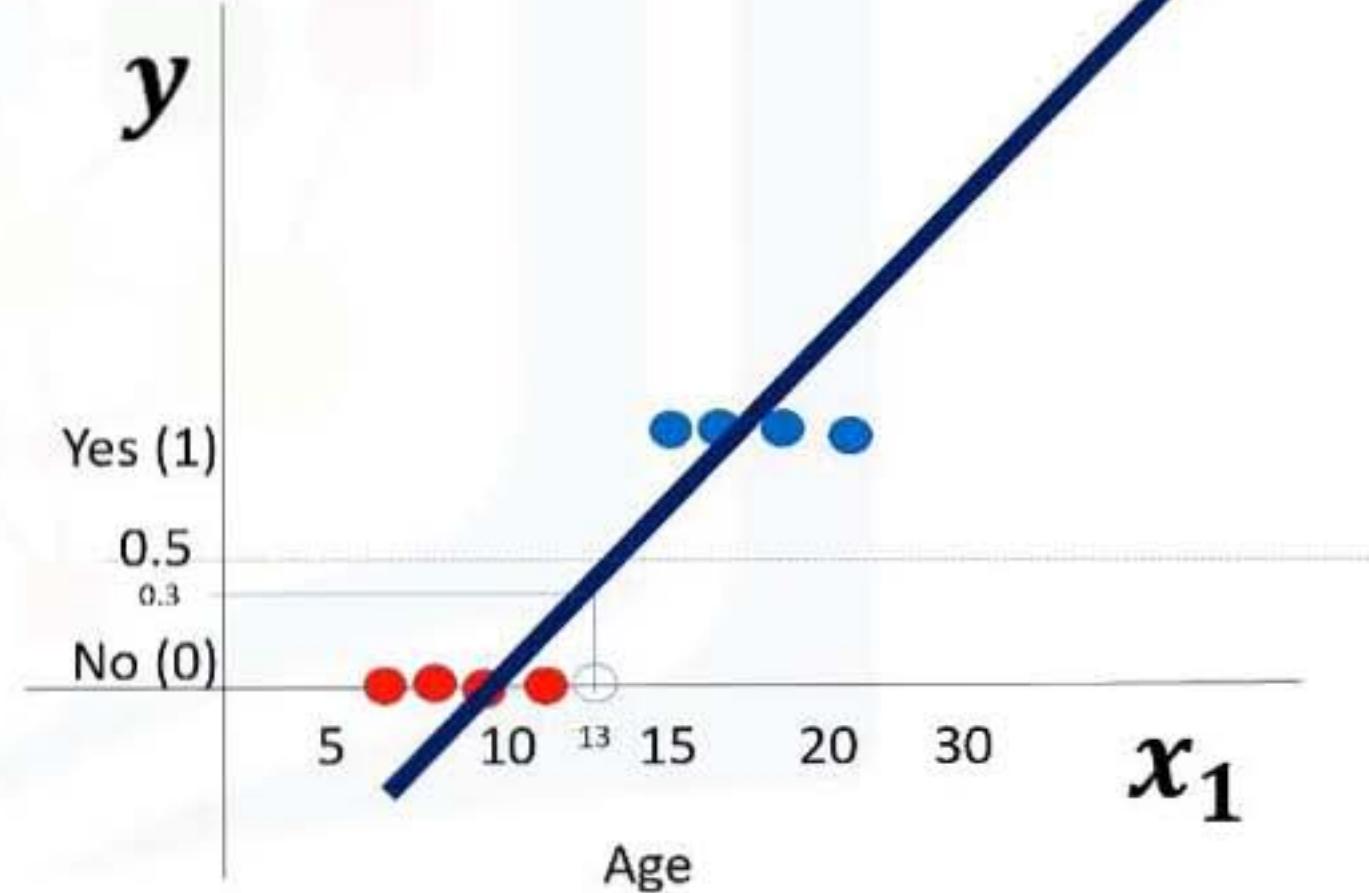
$$\theta^T X = \theta_0 + \theta_1 x_1$$

$$p_1 = [13] \rightarrow \theta^T X = -1 + 0.1 \cdot x_1 \\ = -1 + 0.1 \times 13 \\ = 0.3$$

$$\hat{y} = \begin{cases} 0 & \text{if } \theta^T X < 0.5 \\ 1 & \text{if } \theta^T X \geq 0.5 \end{cases}$$

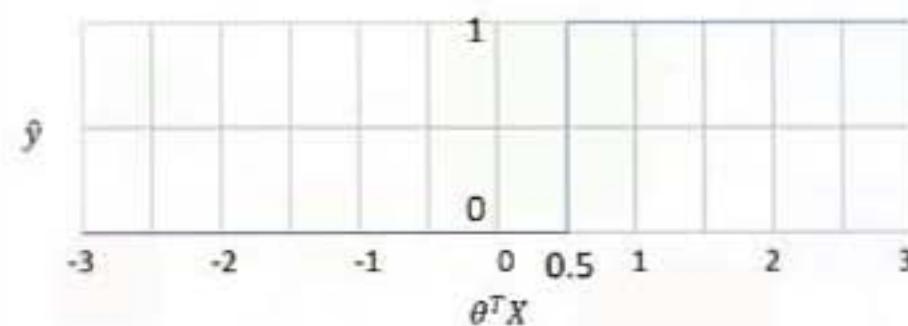
$$\theta^T X = 0.3 \\ \theta^T X < 0.5 \rightarrow \text{Class 0}$$

$$\theta^T X = -1 + 0.1 \cdot x_1$$



The problem with using linear regression

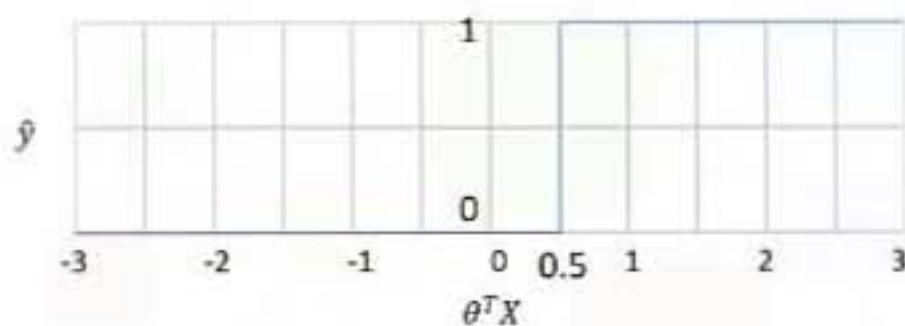
$$\theta^T X = \theta_0 + \theta_1 x_1 + \dots$$



$$\hat{y} = \begin{cases} 0 & \text{if } \theta^T X < 0.5 \\ 1 & \text{if } \theta^T X \geq 0.5 \end{cases}$$

The problem with using linear regression

$$\theta^T X = \theta_0 + \theta_1 x_1 + \dots$$

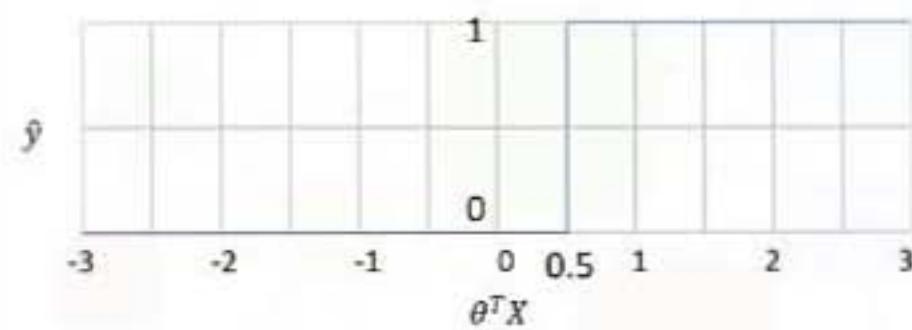


$$\hat{y} = \begin{cases} 0 & \text{if } \theta^T X < 0.5 \\ 1 & \text{if } \theta^T X \geq 0.5 \end{cases}$$

$$\sigma(\theta^T X) = \sigma(\theta_0 + \theta_1 x_1 + \dots)$$

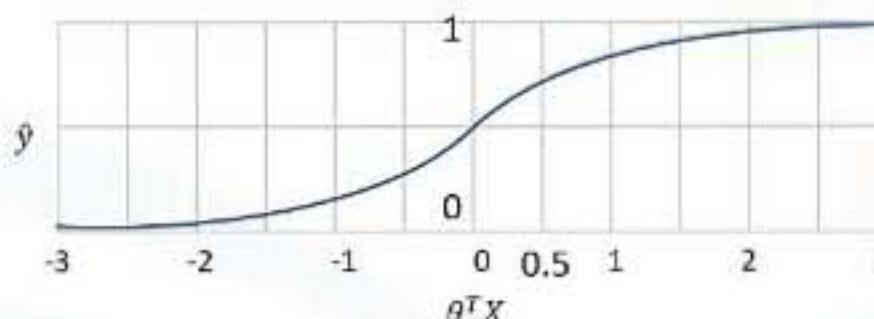
The problem with using linear regression

$$\theta^T X = \theta_0 + \theta_1 x_1 + \dots$$



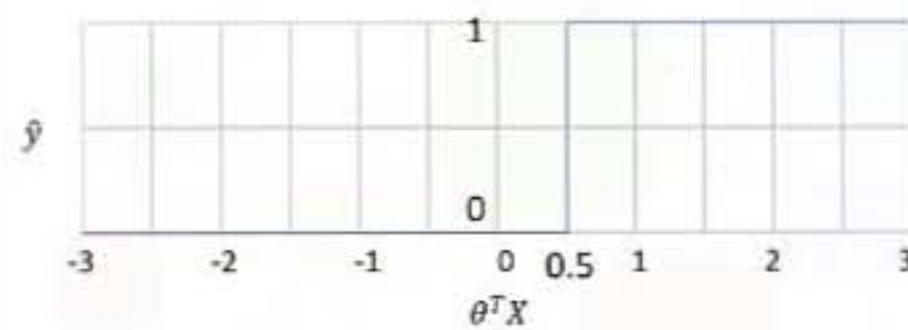
$$\hat{y} = \begin{cases} 0 & \text{if } \theta^T X < 0.5 \\ 1 & \text{if } \theta^T X \geq 0.5 \end{cases}$$

$$\sigma(\theta^T X) = \sigma(\theta_0 + \theta_1 x_1 + \dots)$$

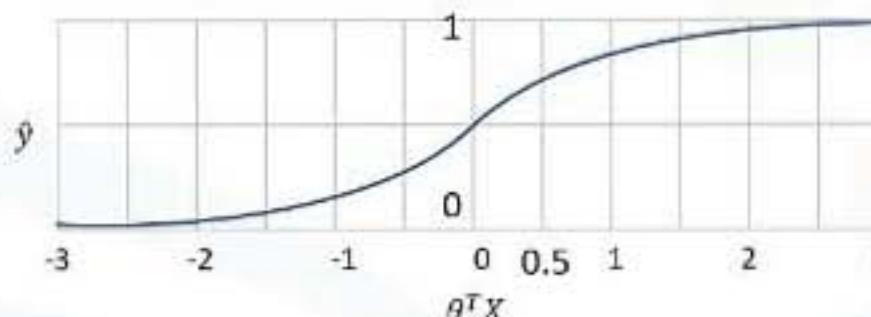


The problem with using linear regression

$$\theta^T X = \theta_0 + \theta_1 x_1 + \dots$$



$$\sigma(\theta^T X) = \sigma(\theta_0 + \theta_1 x_1 + \dots)$$



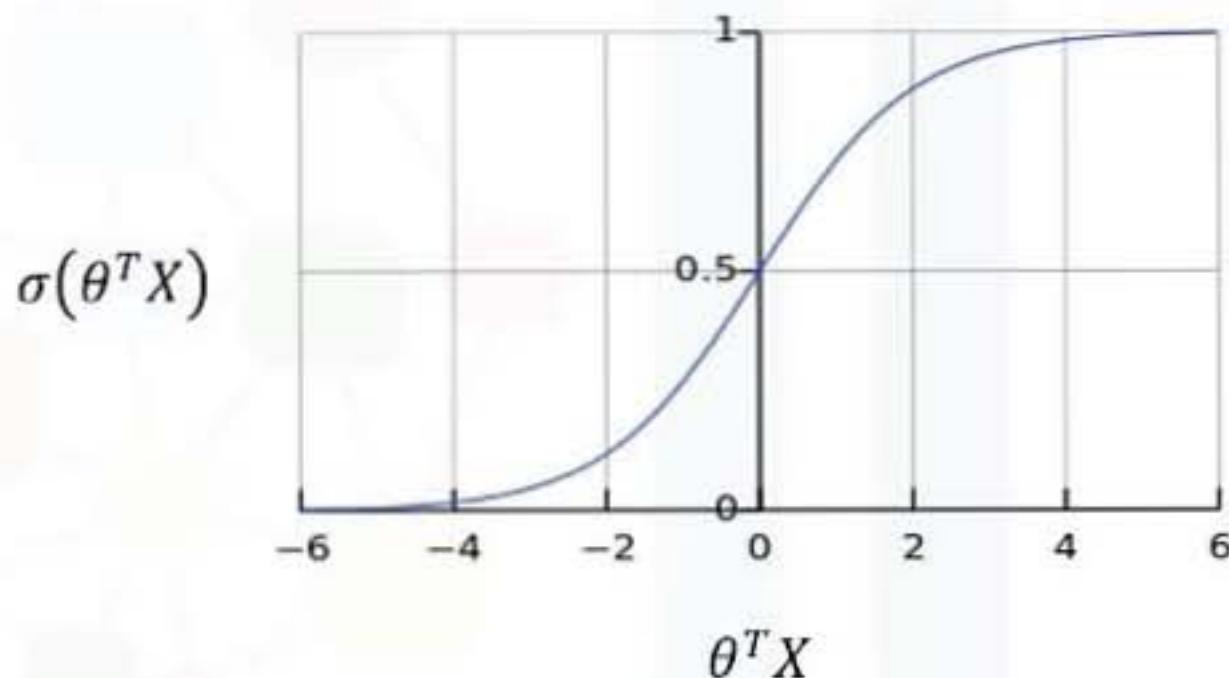
$$\hat{y} = \begin{cases} 0 & \text{if } \theta^T X < 0.5 \\ 1 & \text{if } \theta^T X \geq 0.5 \end{cases}$$

$$\hat{y} = \sigma(\theta^T X)$$

Sigmoid function in logistic regression

- Logistic Function

$$\sigma(\theta^T X) = \frac{1}{1 + e^{-\theta^T X}}$$



Sigmoid function in logistic regression

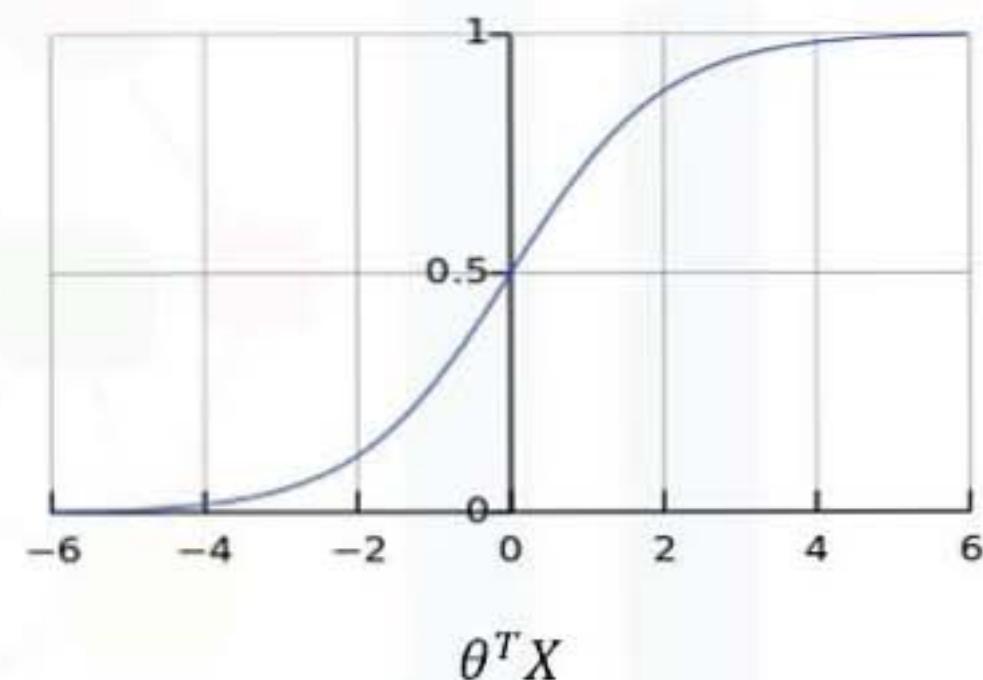
- Logistic Function

$$\sigma(\theta^T X) = \frac{1}{1 + e^{-\theta^T X}}$$

$$\sigma(\theta^T X) = 1$$

$$\sigma(\theta^T X) = 0$$

$$\sigma(\theta^T X)$$



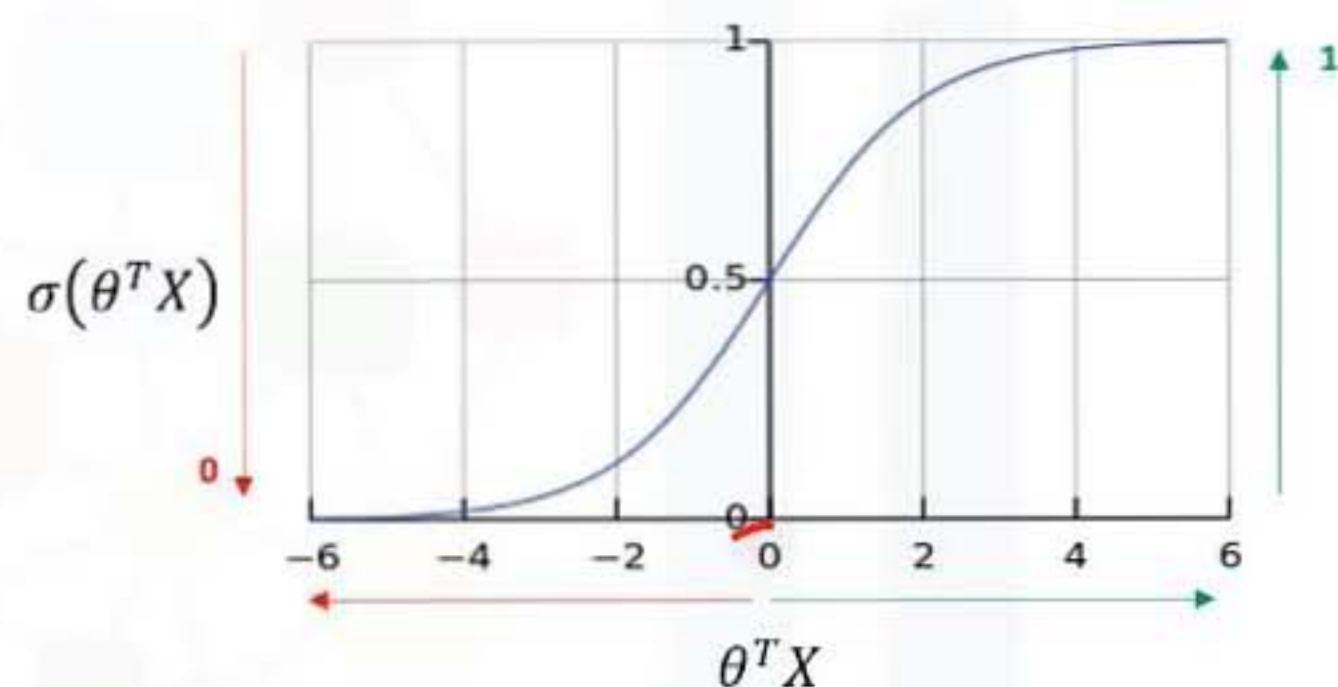
Sigmoid function in logistic regression

- Logistic Function

$$\sigma(\theta^T X) = \frac{1}{1 + e^{-\theta^T X}}$$

$$\sigma(\theta^T X) = 1$$

$$\sigma(\theta^T X) = 0$$



Sigmoid function in logistic regression

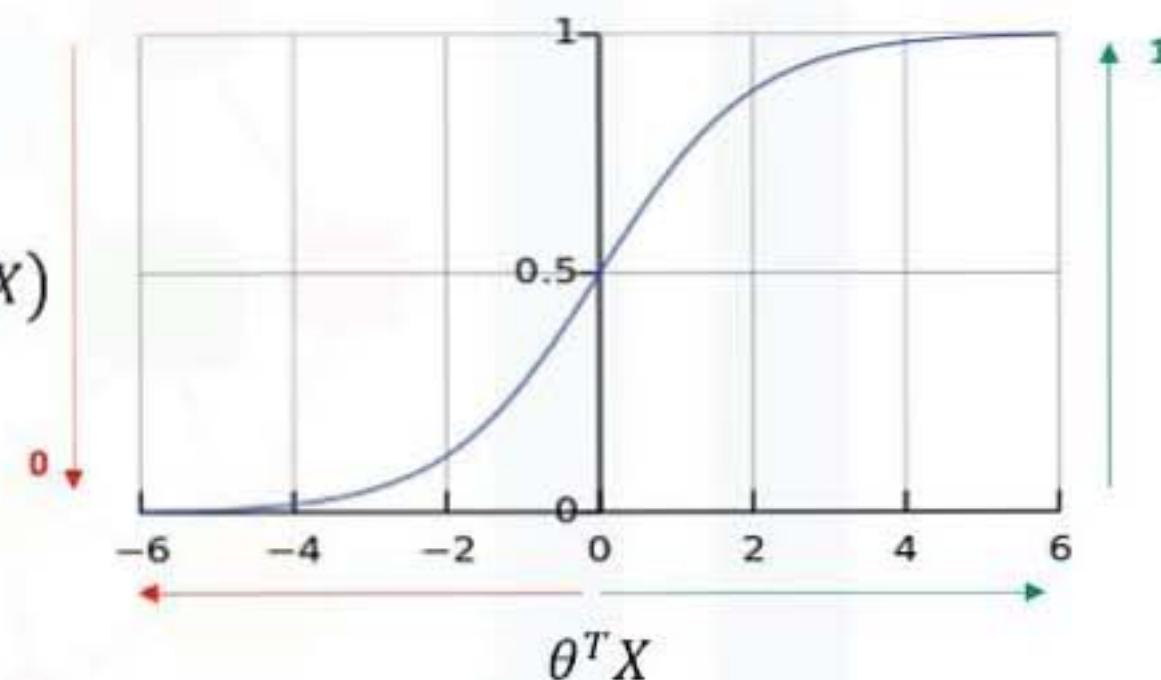
- Logistic Function

$$\sigma(\theta^T X) = \frac{1}{1 + e^{-\theta^T X}}$$

$$\sigma(\theta^T X) = 1$$

$$\sigma(\theta^T X) = 0$$

[0, 1]



Sigmoid function in logistic regression

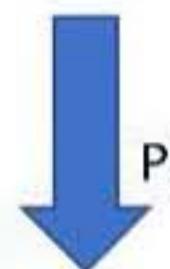
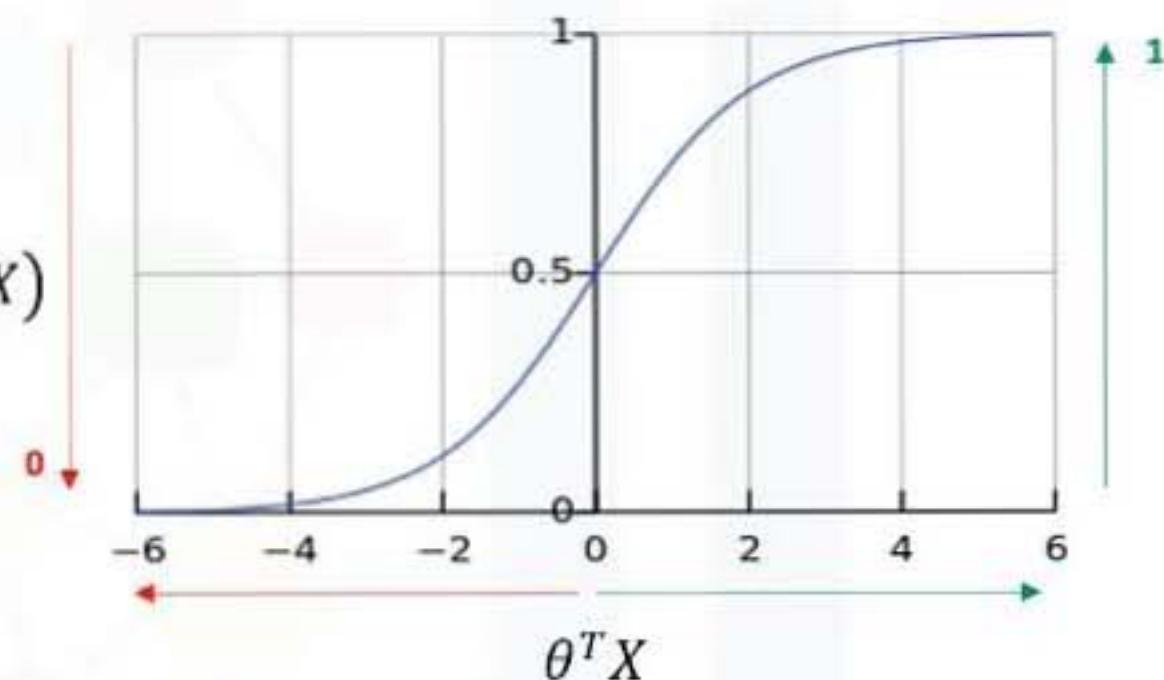
- Logistic Function

$$\sigma(\theta^T X) = \frac{1}{1 + e^{-\theta^T X}}$$

$$\sigma(\theta^T X) = 1$$

$$\sigma(\theta^T X) = 0$$

[0, 1]



$P(y=1|x)$



$P(y=1|x)$

Clarification of the customer churn model

What is the output of our model?

- $P(Y=1|X)$
 - $P(y=0|X) = 1 - P(y=1|x)$
-
- $P(\text{Churn}=1|\text{income,age}) = 0.8$
 - $P(\text{Churn}=0|\text{income,age}) = 1 - 0.8 = 0.2$

$$\sigma(\theta^T X) \longrightarrow P(y=1|x)$$

Clarification of the customer churn model

What is the output of our model?

- $P(Y=1|X)$
 - $P(y=0|X) = 1 - P(y=1|x)$
-
- $P(\text{Churn}=1|\text{income,age}) = 0.8$
 - $P(\text{Churn}=0|\text{income,age}) = 1 - 0.8 = 0.2$

$$\sigma(\theta^T X) \longrightarrow P(y=1|x)$$

$$1 - \sigma(\theta^T X) \longrightarrow P(y=0|x)$$

The training process

$$\sigma(\theta^T X) \longrightarrow P(y=1|x)$$

1. Initialize θ .
2. Calculate $\hat{y} = \sigma(\theta^T X)$ for a customer.

$$\theta = [-1, 2]$$

$$\hat{y} = \sigma([-1, 2] \times [2, 5]) = 0.7$$

The training process

$$\sigma(\theta^T X) \longrightarrow P(y=1|x)$$

1. Initialize θ .
2. Calculate $\hat{y} = \sigma(\theta^T X)$ for a customer.
3. Compare the output of \hat{y} with actual output of customer, y , and record it as error.

$$\theta = [-1, 2]$$

$$\hat{y} = \sigma([-1, 2] \times [2, 5]) = 0.7$$

$$\text{Error} = 1 - 0.7 = 0.3$$

The training process

$$\sigma(\theta^T X) \longrightarrow P(y=1|x)$$

1. Initialize θ .
2. Calculate $\hat{y} = \sigma(\theta^T X)$ for a customer.
3. Compare the output of \hat{y} with actual output of customer, y , and record it as error.
4. Calculate the error for all customers.

$$\theta = [-1, 2]$$

$$\hat{y} = \sigma([-1, 2] \times [2, 5]) = 0.7$$

$$\text{Error} = 1 - 0.7 = 0.3$$

$$Cost = J(\theta)$$

The training process

$$\sigma(\theta^T X) \longrightarrow P(y=1|x)$$

1. Initialize θ .
2. Calculate $\hat{y} = \sigma(\theta^T X)$ for a customer.
3. Compare the output of \hat{y} with actual output of customer, y , and record it as error.
4. Calculate the error for all customers.

$$\theta = [-1, 2]$$

$$\hat{y} = \sigma([-1, 2] \times [2, 5]) = 0.7$$

$$\text{Error} = 1 - 0.7 = 0.3$$

$$Cost = J(\theta)$$

The training process

$$\sigma(\theta^T X) \longrightarrow P(y=1|x)$$

1. Initialize θ .
2. Calculate $\hat{y} = \sigma(\theta^T X)$ for a customer.
3. Compare the output of \hat{y} with actual output of customer, y , and record it as error.
4. Calculate the error for all customers.

$$\theta = [-1, 2]$$

$$\hat{y} = \sigma([-1, 2] \times [2, 5]) = 0.7$$

$$\text{Error} = 1 - 0.7 = 0.3$$

$$Cost = J(\theta)$$

The training process

$$\sigma(\theta^T X) \longrightarrow P(y=1|x)$$

1. Initialize θ . $\theta = [-1, 2]$
2. Calculate $\hat{y} = \sigma(\theta^T X)$ for a customer. $\hat{y} = \sigma([-1, 2] \times [2, 5]) = 0.7$
3. Compare the output of \hat{y} with actual output of customer, y , and record it as error. Error = $1 - 0.7 = 0.3$
4. Calculate the error for all customers. Cost = $J(\theta)$
5. Change the θ to reduce the cost. θ_{new}

The training process

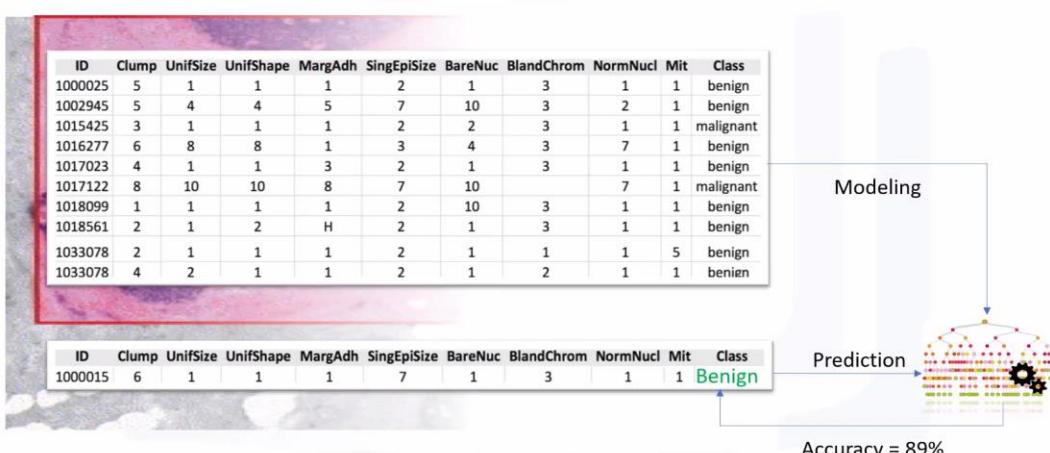
$$\sigma(\theta^T X) \longrightarrow P(y=1|x)$$

1. Initialize θ . $\theta = [-1, 2]$
2. Calculate $\hat{y} = \sigma(\theta^T X)$ for a customer. $\hat{y} = \sigma([-1, 2] \times [2, 5]) = 0.7$
3. Compare the output of \hat{y} with actual output of customer, y , and record it as error. Error = $1 - 0.7 = 0.3$
4. Calculate the error for all customers. Cost = $J(\theta)$
5. Change the θ to reduce the cost. θ_{new}
6. Go back to step 2.

Support Vector Machine (SVM)

Saeed Aghabozorgi

Classification with SVM



What is SVM?

SVM is a supervised algorithm that classifies cases by finding a separator.

1. Mapping data to a **high-dimensional** feature space
2. Finding a **separator**

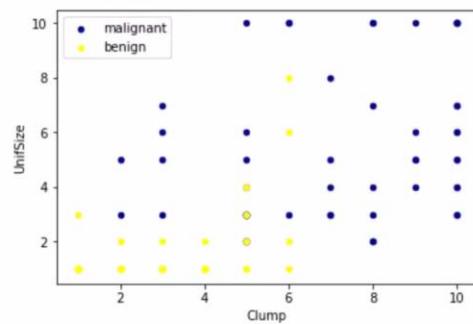
Clump	UnifSize	UnifShape	MargAdh	SingEpiSize	BareNuc	BlandChrom	NormNucl	Mit	Class
5	1	1	1	2	1	3	1	1	benign
5	4	4	5	7	10	3	2	1	benign
3	1	1	1	2	2	3	1	1	malignant
6	8	8	1	3	4	3	7	1	benign
4	1	1	3	2	1	3	1	1	benign
8	10	10	8	7	10		7	1	malignant
1	1	1	1	2	10	3	1	1	benign
2	1	2	H	2	1	3	1	1	benign
2	1	1	1	2	1	1	1	5	benign
4	2	1	1	2	1	2	1	1	benign

What is SVM?

SVM is a supervised algorithm that classifies cases by finding a separator.

1. Mapping data to a **high-dimensional** feature space
2. Finding a **separator**

Clump	UnifSize	UnifShape	MargAdh	SingEpiSize	BareNuc	BlandChrom	NormNucl	Mit	Class
5	1	1	1	2	1	3	1	1	benign
5	4	4	5	7	10	3	2	1	benign
3	1	1	1	2	2	3	1	1	malignant
6	8	8	1	3	4	3	7	1	benign
4	1	1	3	2	1	3	1	1	benign
8	10	10	8	7	10		7	1	malignant
1	1	1	1	2	10	3	1	1	benign
2	1	2	H	2	1	3	1	1	benign
2	1	1	1	2	1	1	1	5	benign
4	2	1	1	2	1	2	1	1	benign

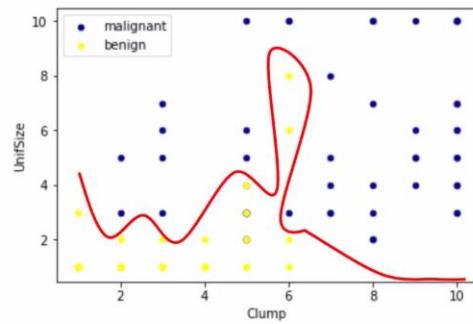


What is SVM?

SVM is a supervised algorithm that classifies cases by finding a separator.

1. Mapping data to a **high-dimensional** feature space
2. Finding a **separator**

Clump	UnifSize	UnifShape	MargAdh	SingEpiSize	BareNuc	BlandChrom	NormNucl	Mit	Class
5	1	1	1	2	1	3	1	1	benign
5	4	4	5	7	10	3	2	1	benign
3	1	1	1	2	2	3	1	1	malignant
6	8	8	1	3	4	3	7	1	benign
4	1	1	3	2	1	3	1	1	benign
8	10	10	8	7	10		7	1	malignant
1	1	1	1	2	10	3	1	1	benign
2	1	2	H	2	1	3	1	1	benign
2	1	1	1	2	1	1	1	5	benign
4	2	1	1	2	1	2	1	1	benign

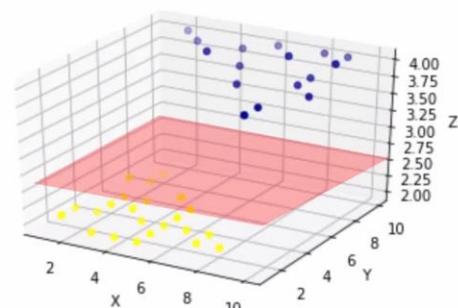


What is SVM?

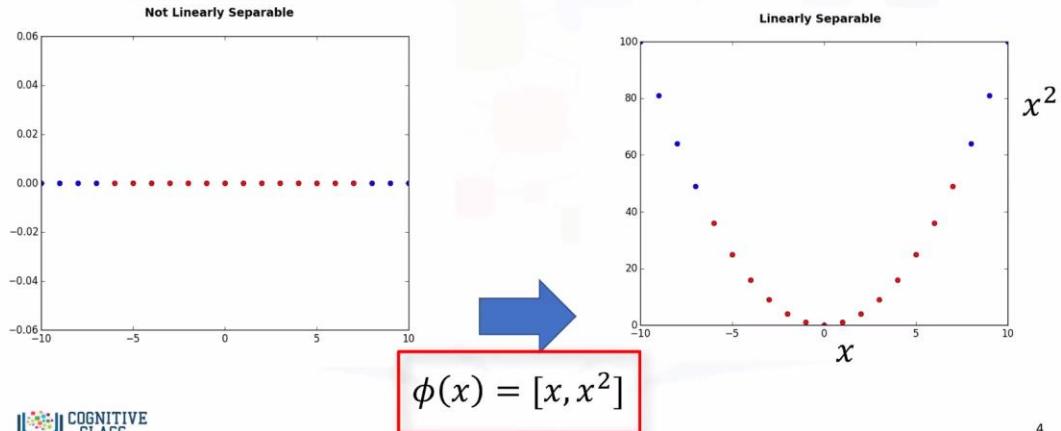
SVM is a supervised algorithm that classifies cases by finding a separator.

1. Mapping data to a **high-dimensional** feature space
2. Finding a **separator**

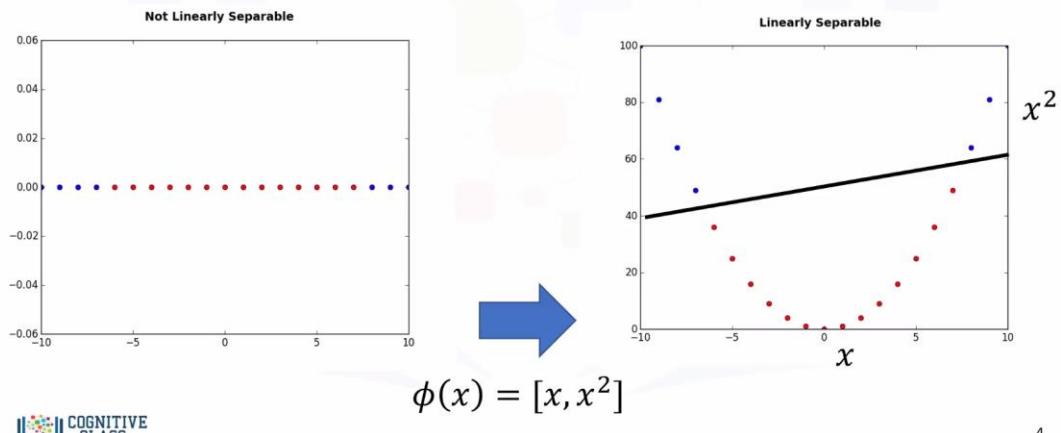
Clump	UnifSize	UnifShape	MargAdh	SingEpiSize	BareNuc	BlandChrom	NormNucl	Mit	Class
5	1	1	1	2	1	3	1	1	benign
5	4	4	5	7	10	3	2	1	benign
3	1	1	1	2	2	3	1	1	malignant
6	8	8	1	3	4	3	7	1	benign
4	1	1	3	2	1	3	1	1	benign
8	10	10	8	7	10		7	1	malignant
1	1	1	1	2	10	3	1	1	benign
2	1	2	H	2	1	3	1	1	benign
2	1	1	1	2	1	1	1	5	benign
4	2	1	1	2	1	2	1	1	benign



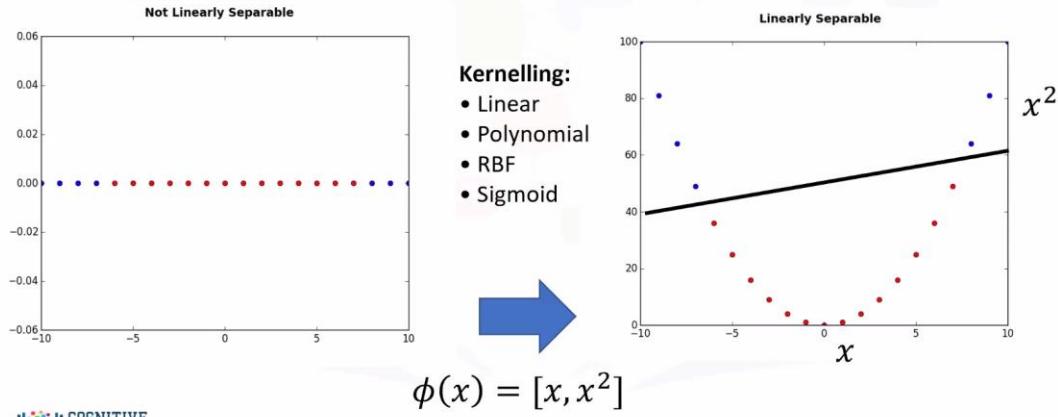
Data transformation



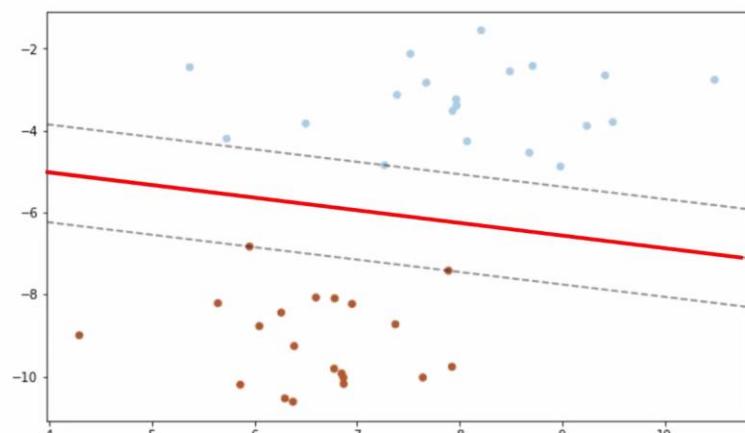
Data transformation



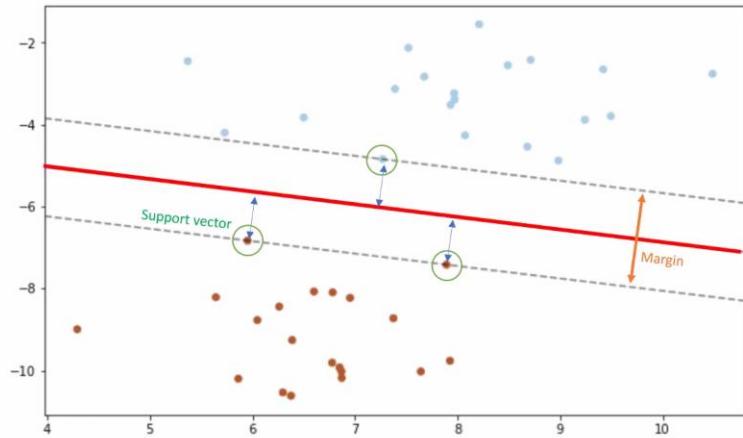
Data transformation



Using SVM to find the hyperplane

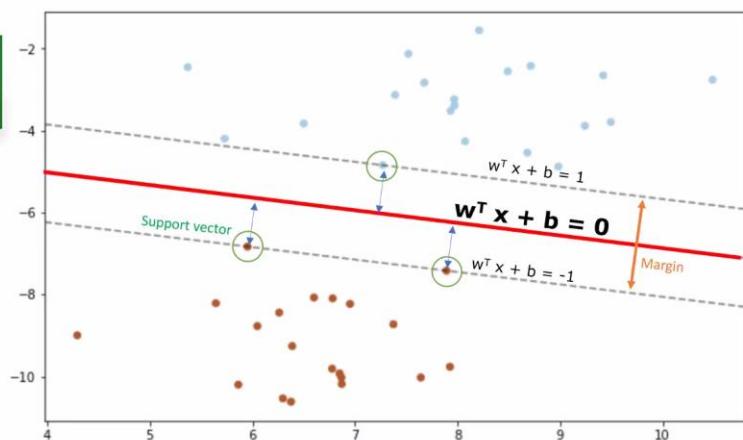


Using SVM to find the hyperplane



Using SVM to find the hyperplane

Find \mathbf{w} and b such that
 $\Phi(\mathbf{w}) = \frac{1}{2} \mathbf{w}^T \mathbf{w}$ is minimized;
and for all $\{(\mathbf{x}_i, y_i)\}$: $y_i (\mathbf{w}^T \mathbf{x}_i + b) \geq 1$



Pros and cons of SVM

- Advantages:
 - Accurate in high-dimensional spaces
 - Memory efficient
- Disadvantages:
 - Prone to over-fitting
 - No probability estimation
 - Small datasets

SVM applications

- Image recognition
- Text category assignment
- Detecting spam
- Sentiment analysis
- Gene Expression Classification
- Regression, outlier detection and clustering

Intro to Clustering

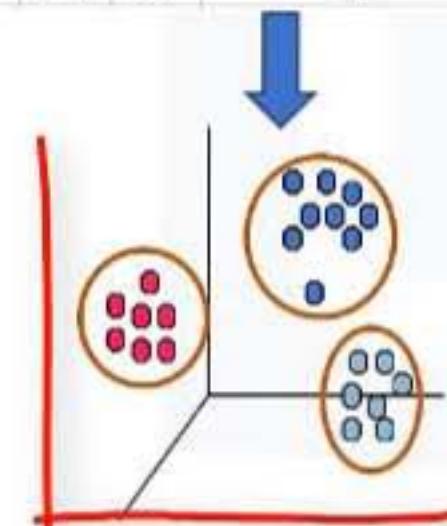
Saeed Aghabozorgi

Clustering for segmentation

Customer Id	Age	Edu	Years Employed	Income	Card Debt	Other Debt	Address	DebtIncomeRatio	Defaulted
1	41	2	6	19	0.124	1.073	NBA001	6.3	0
2	47	1	26	100	4.582	8.218	NBA021	12.8	0
3	33	2	10	57	6.111	5.802	NBA013	20.9	1
4	29	2	4	19	0.681	0.516	NBA009	6.3	0
5	47	1	31	253	9.308	8.908	NBA008	7.2	0
6	40	1	23	81	0.998	7.831	NBA016	10.9	1
7	38	2	4	56	0.442	0.454	NBA013	1.6	0
8	42	3	0	64	0.279	3.945	NBA009	6.6	0
9	26	1	5	18	0.575	2.215	NBA006	15.5	1

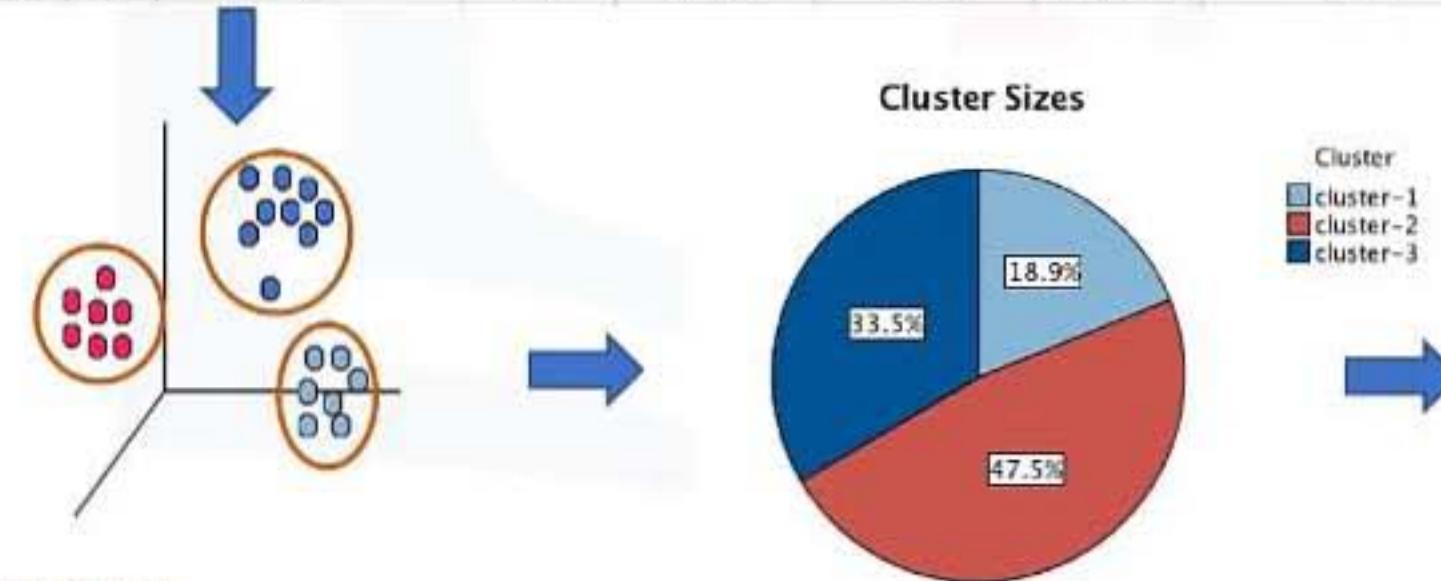
Clustering for segmentation

Customer Id	Age	Edu	Years Employed	Income	Card Debt	Other Debt	Address	DebtIncomeRatio	Defaulted
1	41	2	6	19	0.124	1.073	NBA001	6.3	0
2	47	1	26	100	4.582	8.218	NBA021	12.8	0
3	33	2	10	57	6.111	5.802	NBA013	20.9	1
4	29	2	4	19	0.681	0.516	NBA009	6.3	0
5	47	1	31	253	9.308	8.908	NBA008	7.2	0
6	40	1	23	81	0.998	7.831	NBA016	10.9	1
7	38	2	4	56	0.442	0.454	NBA013	1.6	0
8	42	3	0	64	0.279	3.945	NBA009	6.6	0
9	26	1	5	18	0.575	2.215	NBA006	15.5	1



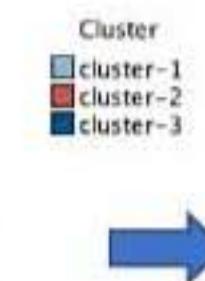
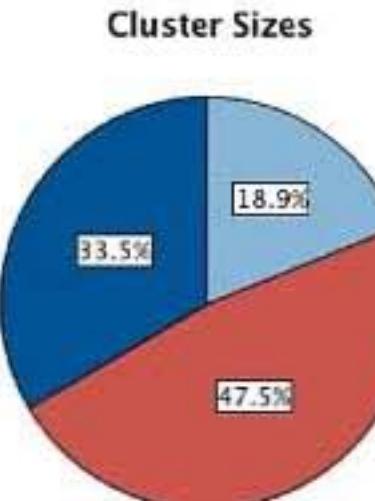
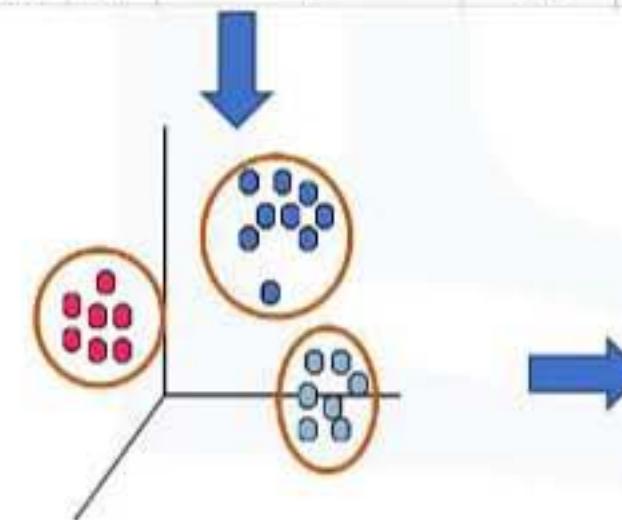
Clustering for segmentation

Customer Id	Age	Edu	Years Employed	Income	Card Debt	Other Debt	Address	DebtIncomeRatio	Defaulted
1	41	2	6	19	0.124	1.073	NBA001	6.3	0
2	47	1	26	100	4.582	8.218	NBA021	12.8	0
3	33	2	10	57	6.111	5.802	NBA013	20.9	1
4	29	2	4	19	0.681	0.516	NBA009	6.3	0
5	47	1	31	253	9.308	8.908	NBA008	7.2	0
6	40	1	23	81	0.998	7.831	NBA016	10.9	1
7	38	2	4	56	0.442	0.454	NBA013	1.6	0
8	42	3	0	64	0.279	3.945	NBA009	6.6	0
9	26	1	5	18	0.575	2.215	NBA006	15.5	1



Clustering for segmentation

Customer Id	Age	Edu	Years Employed	Income	Card Debt	Other Debt	Address	DebtIncomeRatio	Defaulted
1	41	2	6	19	0.124	1.073	NBA001	6.3	0
2	47	1	26	100	4.582	8.218	NBA021	12.8	0
3	33	2	10	57	6.111	5.802	NBA013	20.9	1
4	29	2	4	19	0.681	0.516	NBA009	6.3	0
5	47	1	31	253	9.308	8.908	NBA008	7.2	0
6	40	1	23	81	0.998	7.831	NBA016	10.9	1
7	38	2	4	56	0.442	0.454	NBA013	1.6	0
8	42	3	0	64	0.279	3.945	NBA009	6.6	0
9	26	1	5	18	0.575	2.215	NBA006	15.5	1

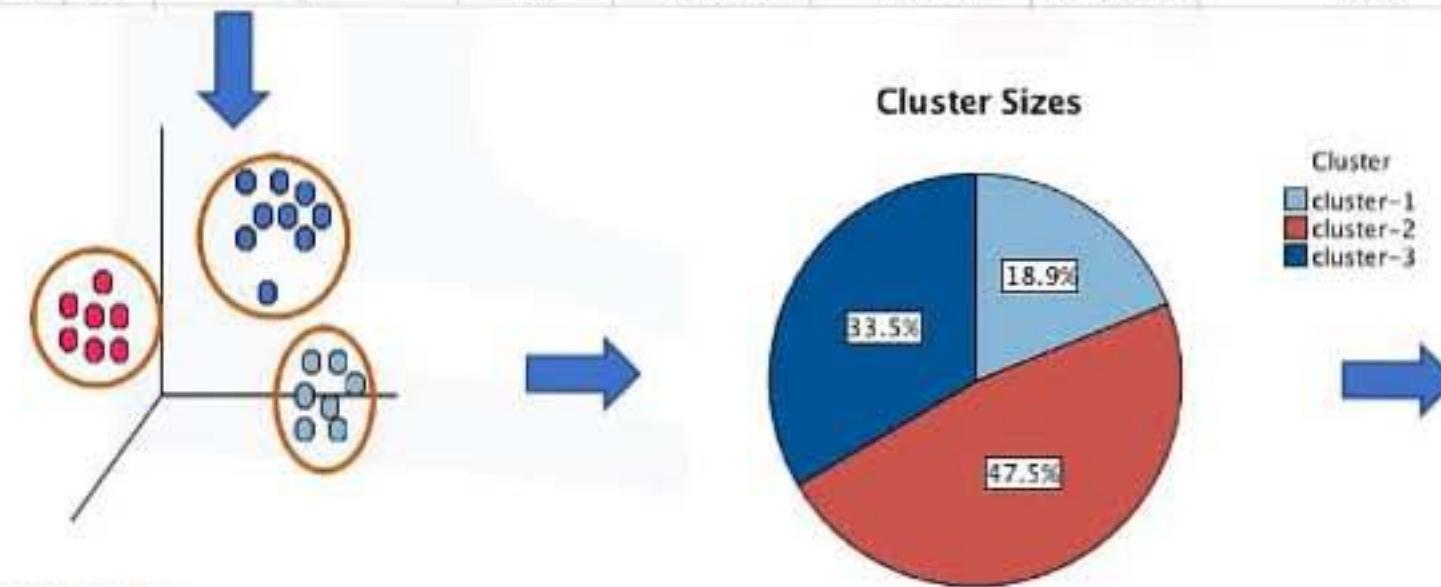


Cluster	Segment Name
cluster-1	AFFLUENT AND MIDDLE AGED
cluster-2	YOUNG EDUCATED AND MIDDLE INCOME
cluster-3	YOUNG AND LOW INCOME

Clustering for segmentation

Customer Id	Age	Edu	Years Employed	Income	Card Debt	Other Debt	Address	DebtIncomeRatio	Defaulted
1	41	2	6	19	0.124	1.073	NBA001	6.3	0
2	47	1	26	100	4.582	8.218	NBA021	12.8	0
3	33	2	10	57	6.111	5.802	NBA013	20.9	1
4	29	2	4	19	0.681	0.516	NBA009	6.3	0
5	47	1	31	253	9.308	8.908	NBA008	7.2	0
6	40	1	23	81	0.998	7.831	NBA016	10.9	1
7	38	2	4	56	0.442	0.454	NBA013	1.6	0
8	42	3	0	64	0.279	3.945	NBA009	6.6	0
9	26	1	5	18	0.575	2.215	NBA006	15.5	1

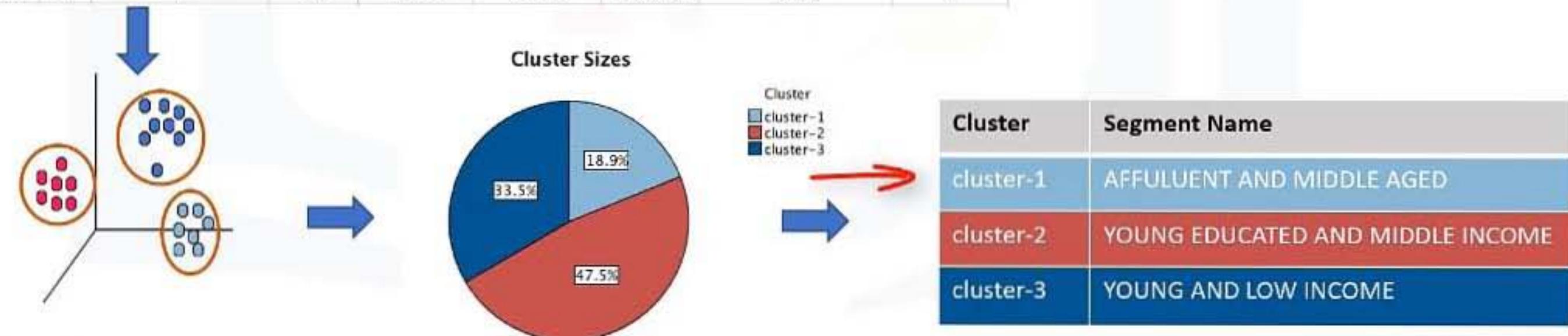
Customer ID	Segment
1	YOUNG AND LOW INCOME
2	AFFLUENT AND MIDDLE AGED
3	AFFLUENT AND MIDDLE AGED
4	YOUNG AND LOW INCOME
5	AFFLUENT AND MIDDLE AGED
6	AFFLUENT AND MIDDLE AGED
7	YOUNG AND LOW INCOME
8	YOUNG AND LOW INCOME
9	AFFLUENT AND MIDDLE AGED



Cluster	Segment Name
cluster-1	AFFLUENT AND MIDDLE AGED
cluster-2	YOUNG EDUCATED AND MIDDLE INCOME
cluster-3	YOUNG AND LOW INCOME

Clustering for segmentation

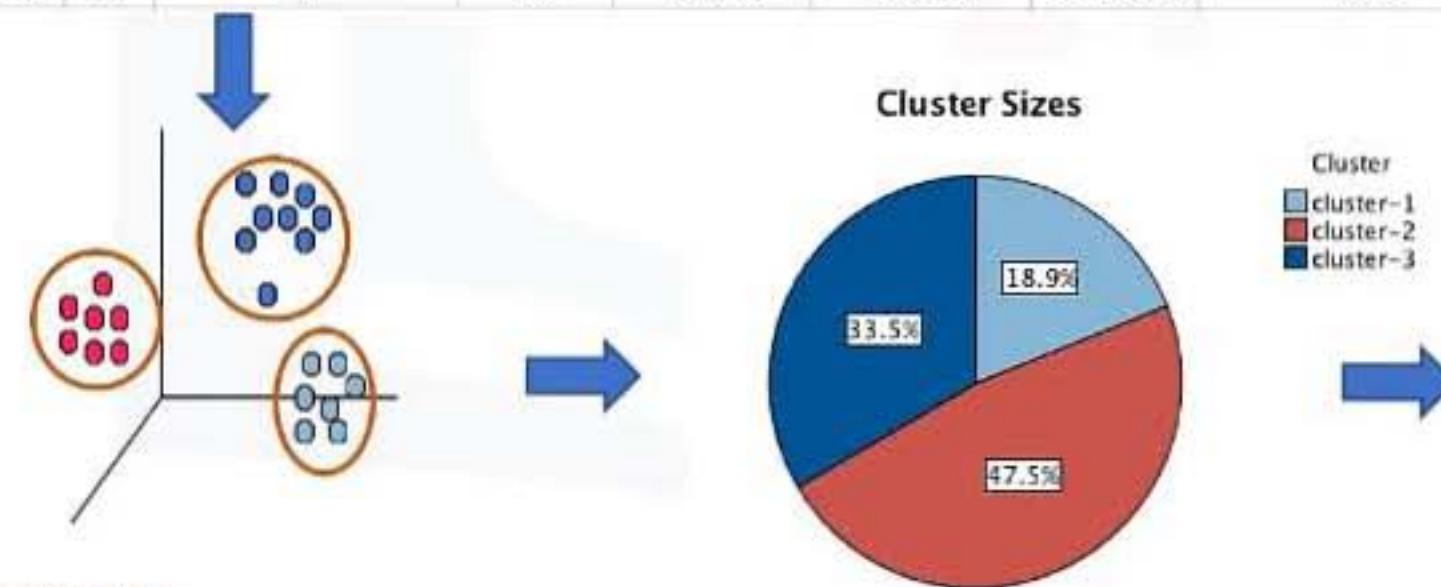
Customer Id	Age	Edu	Years Employed	Income	Card Debt	Other Debt	Address	DebtIncomeRatio	Defaulted
1	41	2	6	19	0.124	1.073	NBA001	6.3	0
2	47	1	26	100	4.582	8.218	NBA021	12.8	0
3	33	2	10	57	6.111	5.802	NBA013	20.9	1
4	29	2	4	19	0.681	0.516	NBA009	6.3	0
5	47	1	31	253	9.308	8.908	NBA008	7.2	0
6	40	1	23	81	0.998	7.831	NBA016	10.9	1
7	38	2	4	56	0.442	0.454	NBA013	1.6	0
8	42	3	0	64	0.279	3.945	NBA009	6.6	0
9	26	1	5	18	0.575	2.215	NBA006	15.5	1



Clustering for segmentation

Customer Id	Age	Edu	Years Employed	Income	Card Debt	Other Debt	Address	DebtIncomeRatio	Defaulted
1	41	2	6	19	0.124	1.073	NBA001	6.3	0
2	47	1	26	100	4.582	8.218	NBA021	12.8	0
3	33	2	10	57	6.111	5.802	NBA013	20.9	1
4	29	2	4	19	0.681	0.516	NBA009	6.3	0
5	47	1	31	253	9.308	8.908	NBA008	7.2	0
6	40	1	23	81	0.998	7.831	NBA016	10.9	1
7	38	2	4	56	0.442	0.454	NBA013	1.6	0
8	42	3	0	64	0.279	3.945	NBA009	6.6	0
9	26	1	5	18	0.575	2.215	NBA006	15.5	1

Customer ID	Segment
1	YOUNG AND LOW INCOME
2	AFFLUENT AND MIDDLE AGED
3	AFFLUENT AND MIDDLE AGED
4	YOUNG AND LOW INCOME
5	AFFLUENT AND MIDDLE AGED
6	AFFLUENT AND MIDDLE AGED
7	YOUNG AND LOW INCOME
8	YOUNG AND LOW INCOME
9	AFFLUENT AND MIDDLE AGED

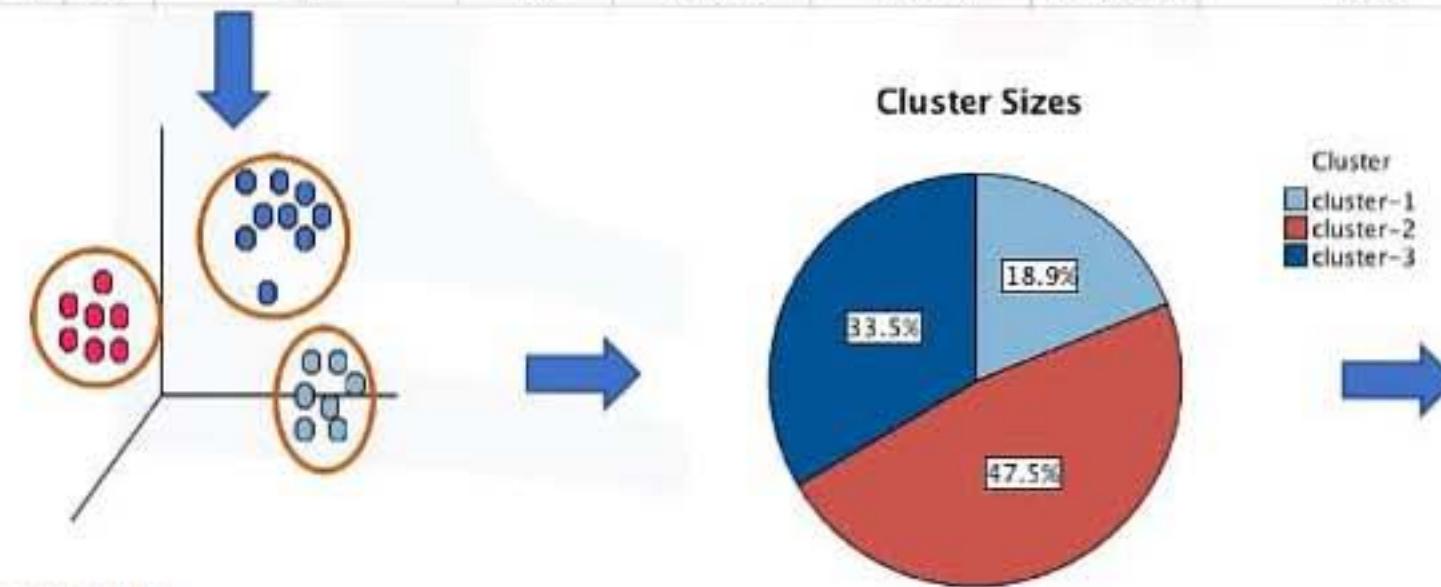


Cluster	Segment Name
cluster-1	AFFLUENT AND MIDDLE AGED
cluster-2	YOUNG EDUCATED AND MIDDLE INCOME
cluster-3	YOUNG AND LOW INCOME

Clustering for segmentation

Customer Id	Age	Edu	Years Employed	Income	Card Debt	Other Debt	Address	DebtIncomeRatio	Defaulted
1	41	2	6	19	0.124	1.073	NBA001	6.3	0
2	47	1	26	100	4.582	8.218	NBA021	12.8	0
3	33	2	10	57	6.111	5.802	NBA013	20.9	1
4	29	2	4	19	0.681	0.516	NBA009	6.3	0
5	47	1	31	253	9.308	8.908	NBA008	7.2	0
6	40	1	23	81	0.998	7.831	NBA016	10.9	1
7	38	2	4	56	0.442	0.454	NBA013	1.6	0
8	42	3	0	64	0.279	3.945	NBA009	6.6	0
9	26	1	5	18	0.575	2.215	NBA006	15.5	1

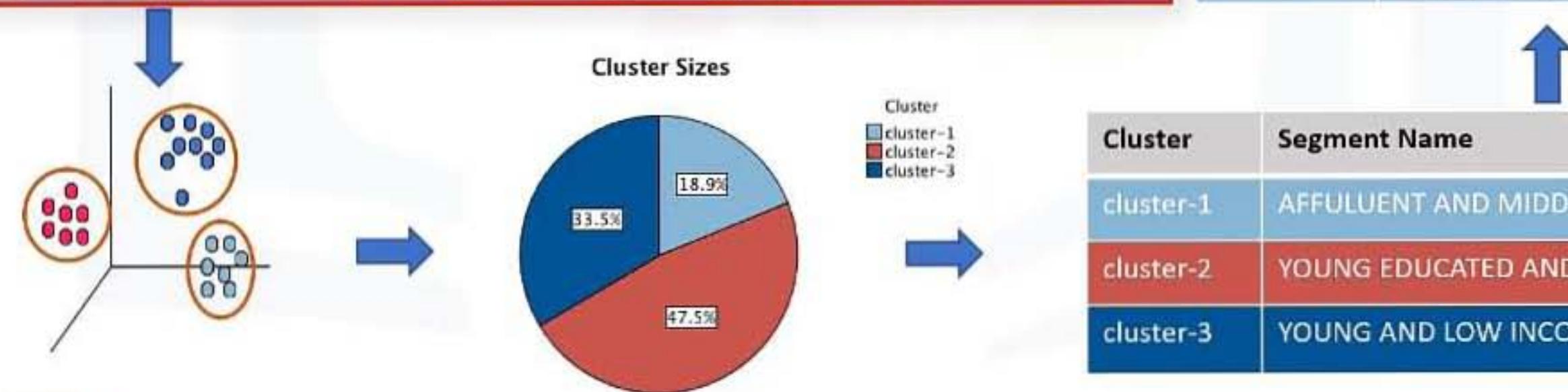
Customer ID	Segment
1	YOUNG AND LOW INCOME
2	AFFLUENT AND MIDDLE AGED
3	AFFLUENT AND MIDDLE AGED
4	YOUNG AND LOW INCOME
5	AFFLUENT AND MIDDLE AGED
6	AFFLUENT AND MIDDLE AGED
7	YOUNG AND LOW INCOME
8	YOUNG AND LOW INCOME
9	AFFLUENT AND MIDDLE AGED



Clustering for segmentation

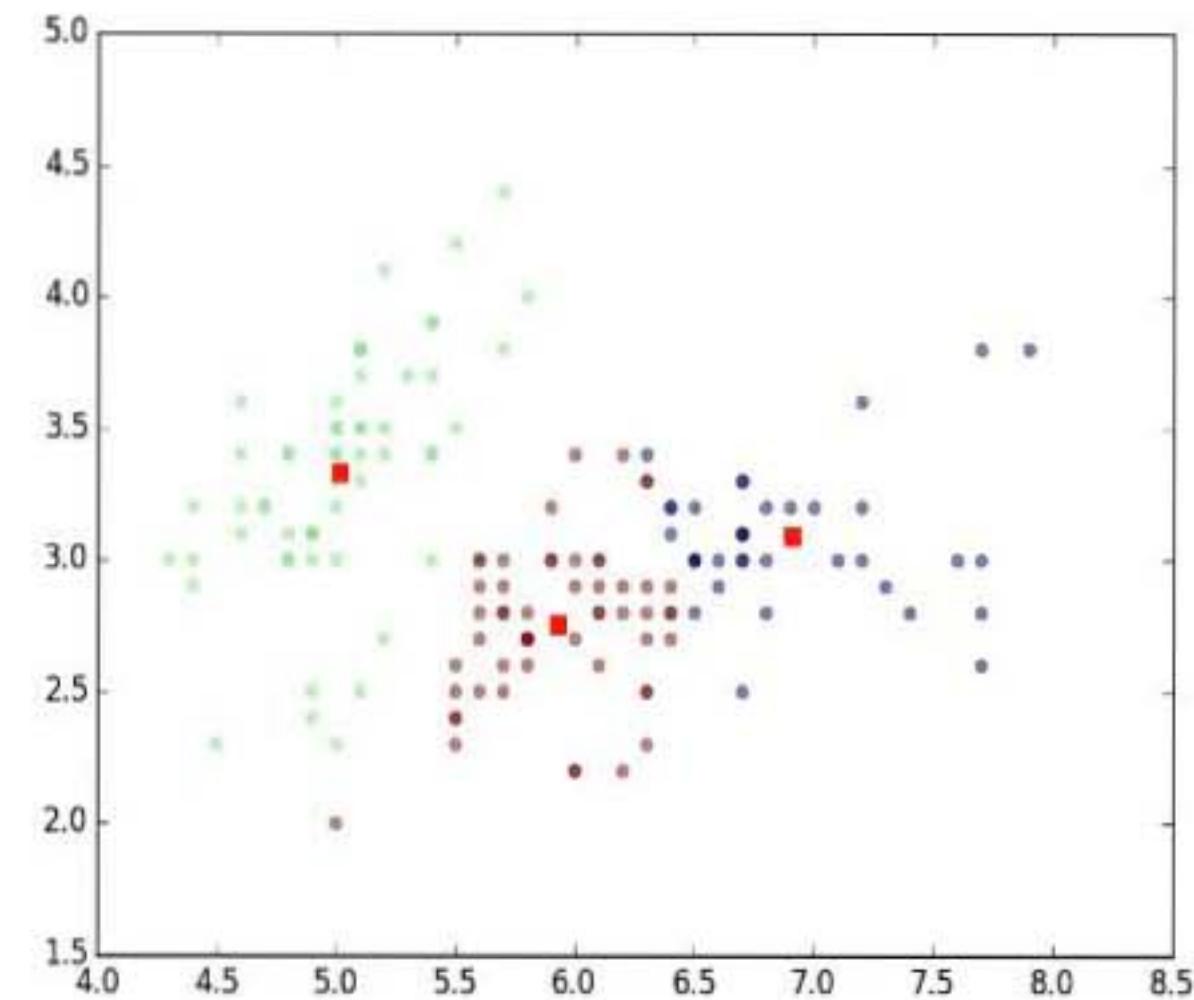
Customer Id	Age	Edu	Years Employed	Income	Card Debt	Other Debt	Address	DebtIncomeRatio	Defaulted
1	41	2	6	19	0.124	1.073	NBA001	6.3	0
2	47	1	26	100	4.582	8.218	NBA021	12.8	0
3	33	2	10	57	6.111	5.802	NBA013	20.9	1
4	29	2	4	19	0.681	0.516	NBA009	6.3	0
5	47	1	31	253	9.308	8.908	NBA008	7.2	0
6	40	1	23	81	0.998	7.831	NBA016	10.9	1
7	38	2	4	56	0.442	0.454	NBA013	1.6	0
8	42	3	0	64	0.279	3.945	NBA009	6.6	0
9	26	1	5	18	0.575	2.215	NBA006	15.5	1

Customer ID	Segment
1	YOUNG AND LOW INCOME
2	AFFLUENT AND MIDDLE AGED
3	AFFLUENT AND MIDDLE AGED
4	YOUNG AND LOW INCOME
5	AFFLUENT AND MIDDLE AGED
6	AFFLUENT AND MIDDLE AGED
7	YOUNG AND LOW INCOME
8	YOUNG AND LOW INCOME
9	AFFLUENT AND MIDDLE AGED



What is clustering?

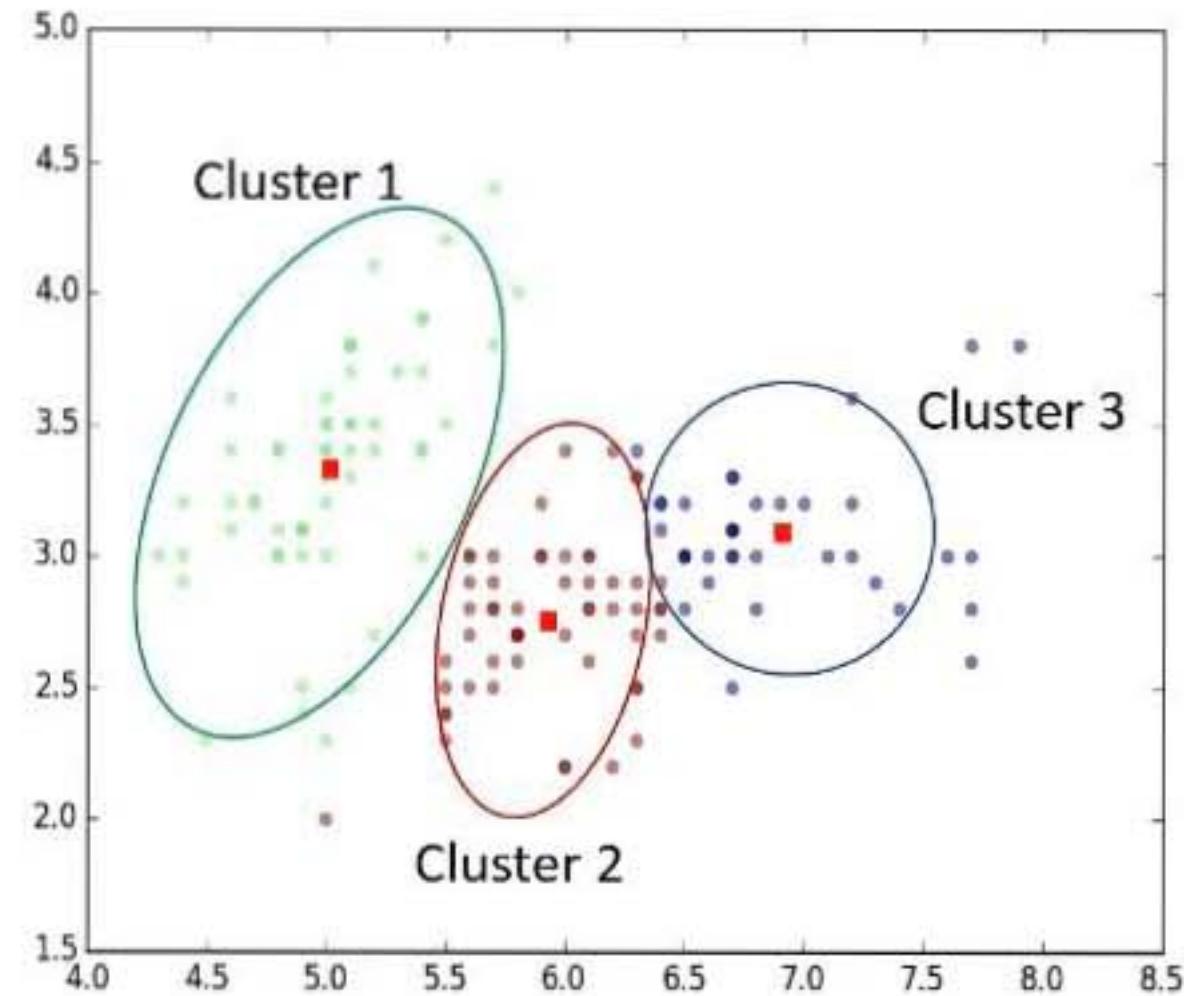
What is a cluster?



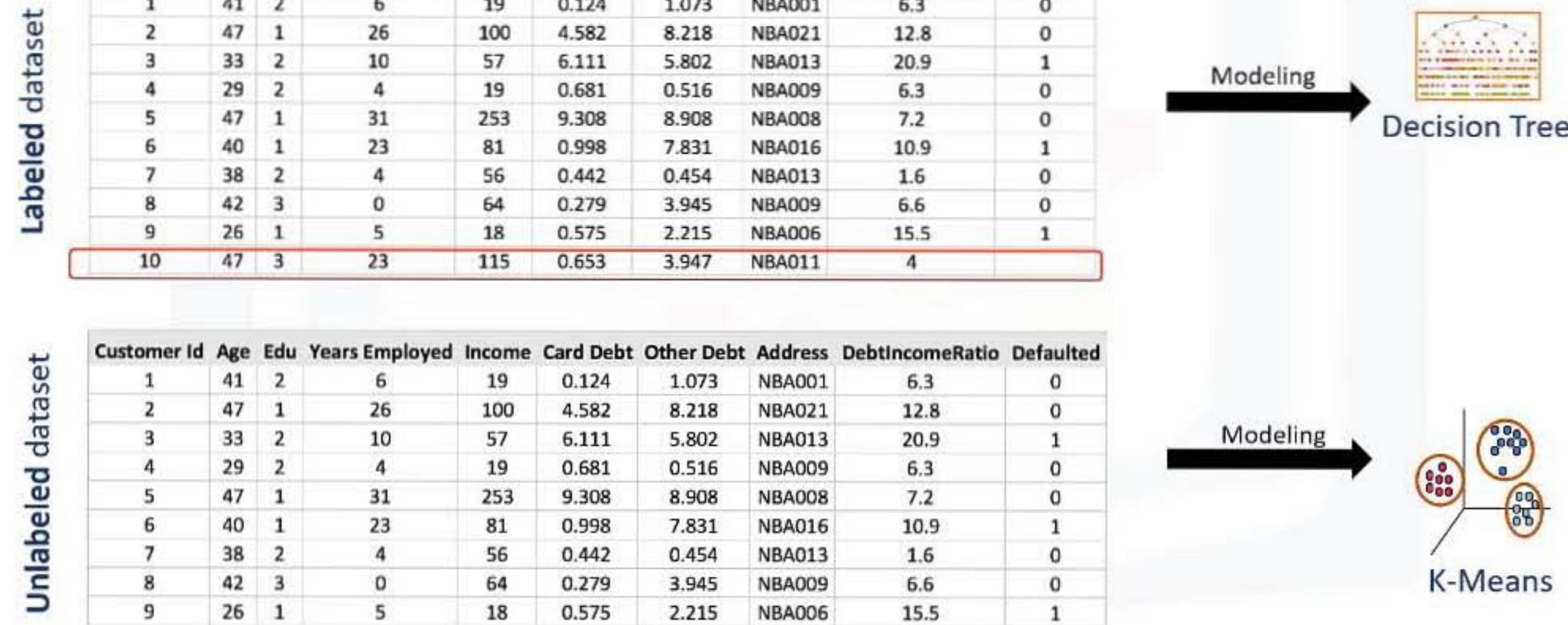
What is clustering?

What is a cluster?

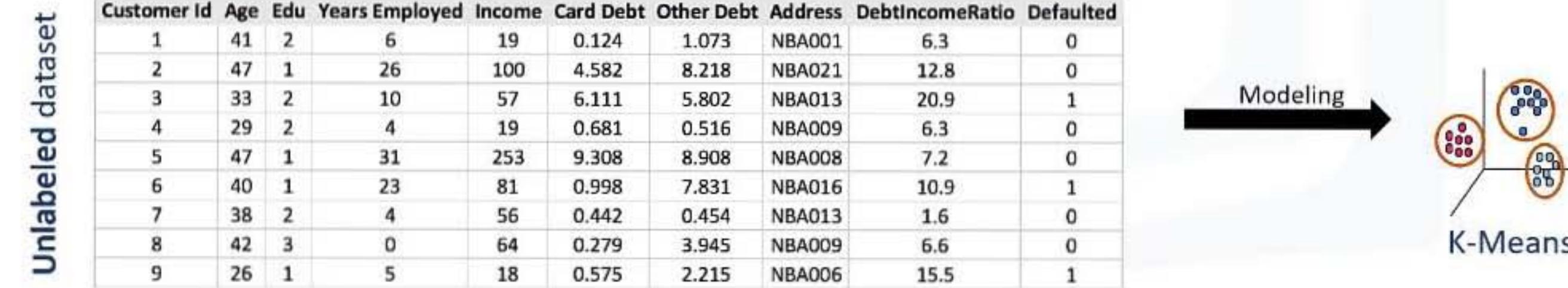
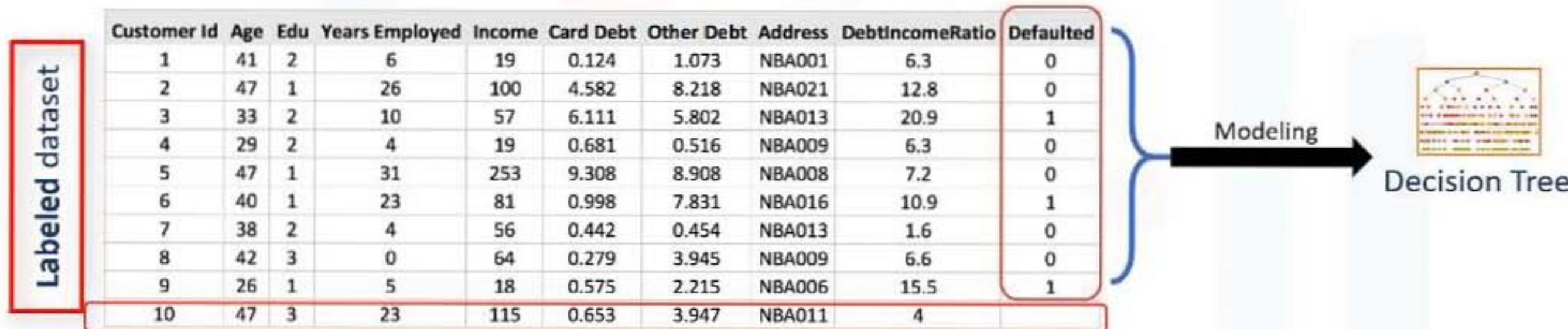
A group of objects that are **similar to other objects** in the cluster, and **dissimilar to data points** in other clusters.



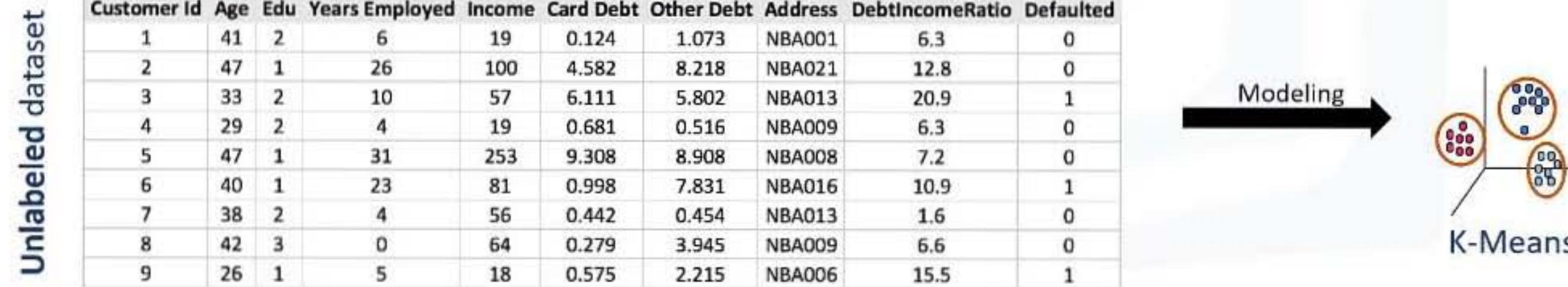
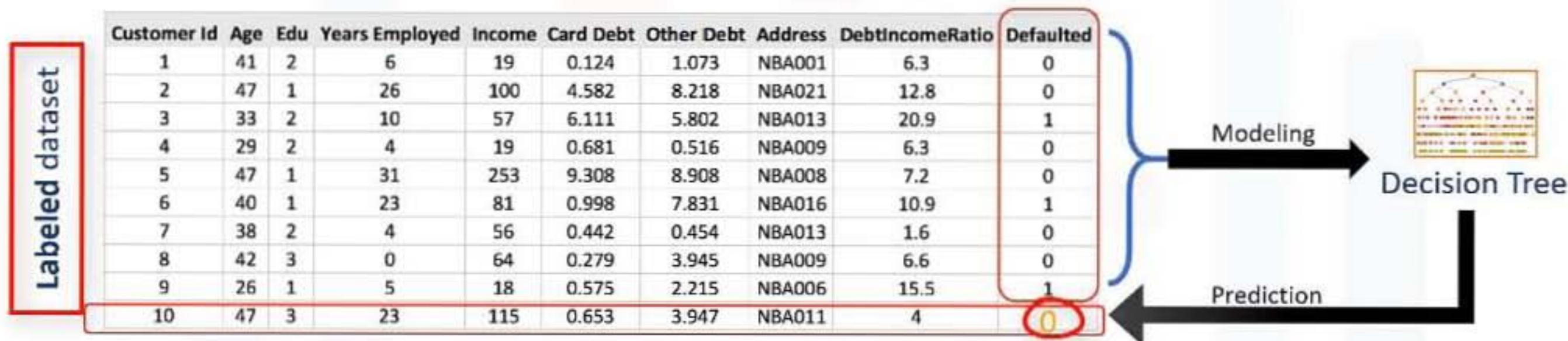
Clustering Vs. classification



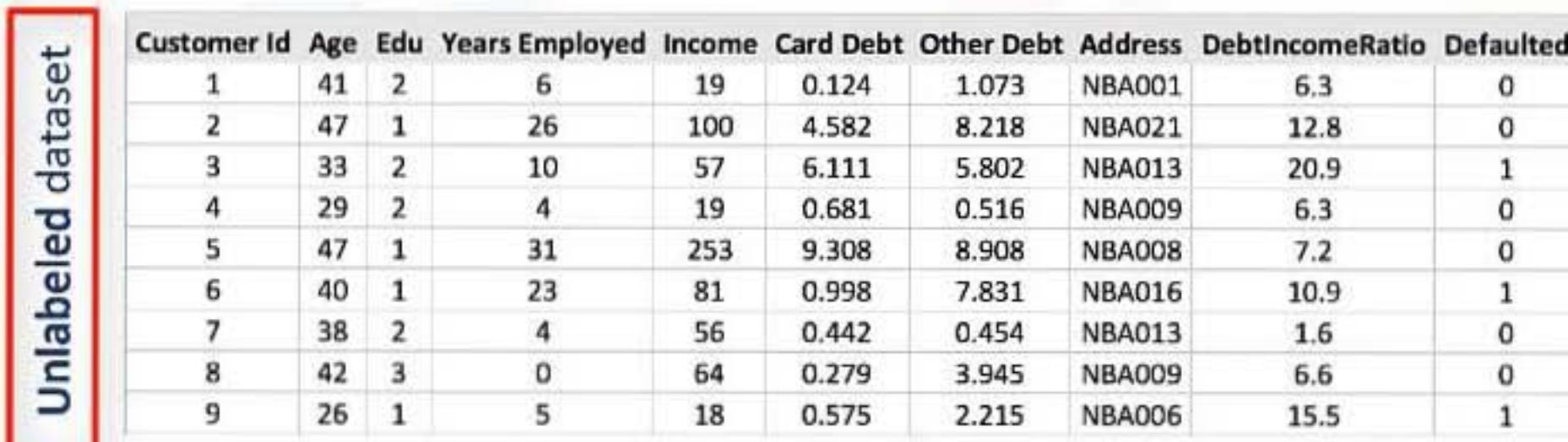
Clustering Vs. classification



Clustering Vs. classification



Clustering Vs. classification



Clustering Vs. classification



Clustering applications

- RETAIL/MARKETING:
 - Identifying buying patterns of customers
- BANKING:
- INSURANCE:

Clustering applications

- RETAIL/MARKETING:
 - Identifying buying patterns of customers
 - Recommending new books or movies to new customers
- BANKING:
- INSURANCE:

Clustering applications

- RETAIL/MARKETING:
 - Identifying buying patterns of customers
 - Recommending new books or movies to new customers
- BANKING:
 - Fraud detection in credit card use
 - Identifying clusters of customers (e.g., loyal)
- INSURANCE:

Clustering applications

- RETAIL/MARKETING:
 - Identifying buying patterns of customers
 - Recommending new books or movies to new customers
- BANKING:
 - Fraud detection in credit card use
 - Identifying clusters of customers (e.g., loyal)
- INSURANCE:
 - Fraud detection in claims analysis
 - Insurance risk of customers

Clustering applications

- **PUBLICATION:**

- Auto-categorizing news based on their content
- Recommending similar news articles

- **MEDICINE:**

- Characterizing patient behavior

- **BIOLOGY:**

Clustering applications

- **PUBLICATION:**

- Auto-categorizing news based on their content
- Recommending similar news articles

- **MEDICINE:**

- Characterizing patient behavior

- **BIOLOGY:**

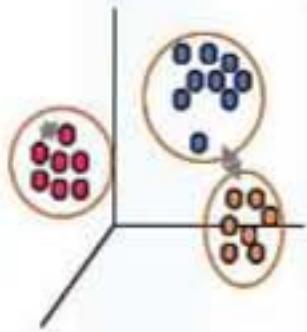
- Clustering genetic markers to identify family ties

Why clustering?

- Exploratory data analysis
- Summary generation
- Outlier detection
- Finding duplicates
- Pre-processing step

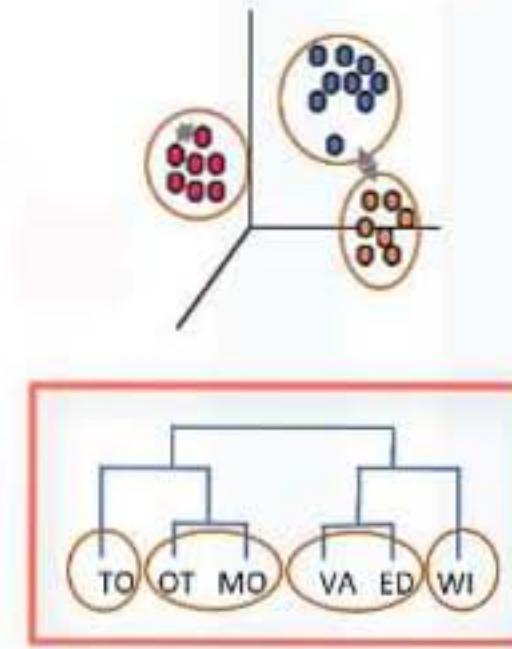
Clustering algorithms

- Partitioned-based Clustering ,
 - Relatively efficient
 - E.g. k-Means, k-Median, Fuzzy c-Means
- Hierarchical Clustering
- Density-based Clustering



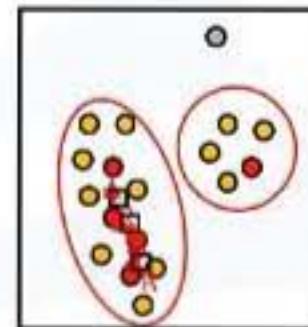
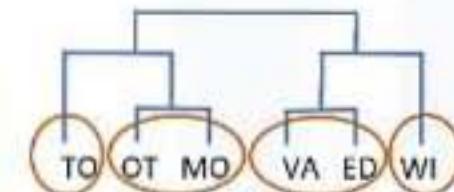
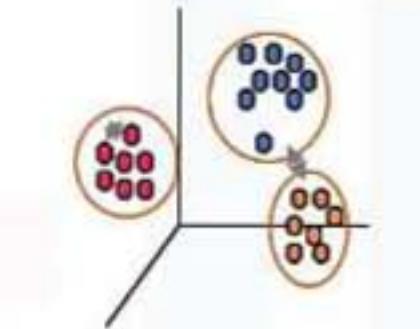
Clustering algorithms

- Partitioned-based Clustering
 - Relatively efficient
 - E.g. k-Means, k-Median, Fuzzy c-Means
- Hierarchical Clustering 
 - Produces trees of clusters
 - E.g. Agglomerative, Divisive
- Density-based Clustering



Clustering algorithms

- Partitioned-based Clustering
 - Relatively efficient
 - E.g. k-Means, k-Median, Fuzzy c-Means
- Hierarchical Clustering
 - Produces trees of clusters
 - E.g. Agglomerative, Divisive
- Density-based Clustering
 - Produces arbitrary shaped clusters
 - E.g. DBSCAN



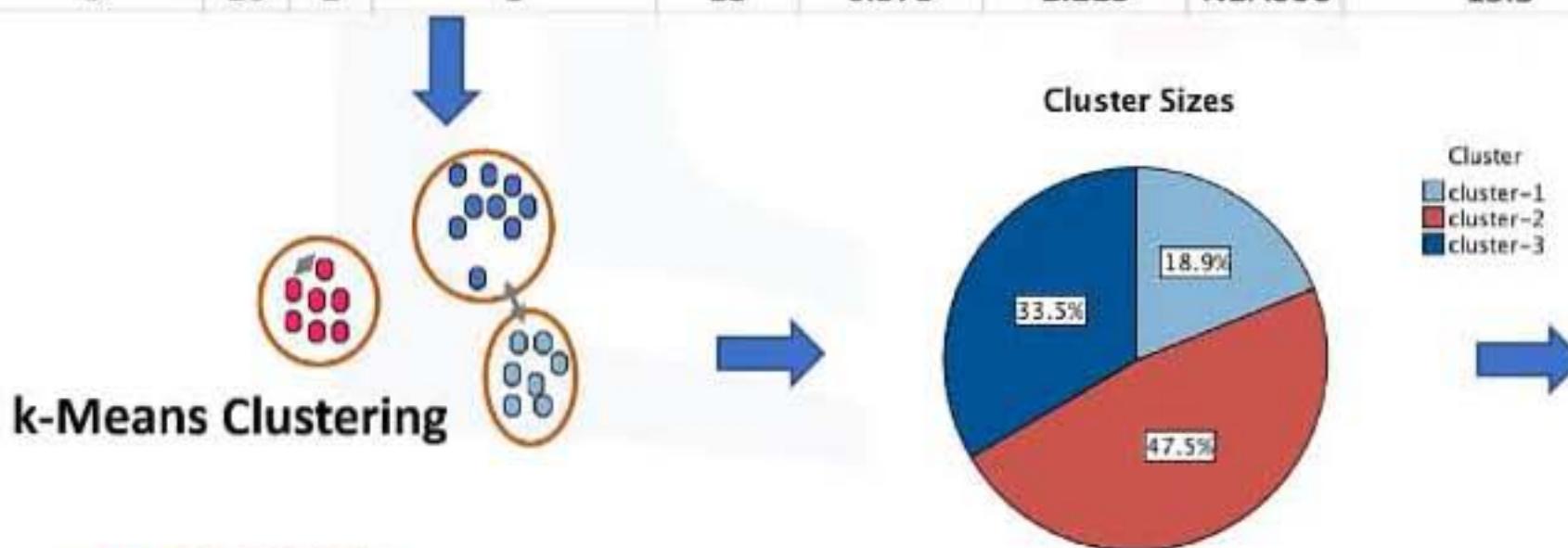
k-Means Clustering

Saeed Aghabozorgi

What is k-Means clustering?

Customer ID	Age	Edu	Years Employed	Income	Card Debt	Other Debt	Address	DebtIncomeRatio	Defaulted
1	41	2	6	19	0.124	1.073	NBA001	6.3	0
2	47	1	26	100	4.582	8.218	NBA021	12.8	0
3	33	2	10	57	6.111	5.802	NBA013	20.9	1
4	29	2	4	19	0.681	0.516	NBA009	6.3	0
5	47	1	31	253	9.308	8.908	NBA008	7.2	0
6	40	1	23	81	0.998	7.831	NBA016	10.9	1
7	38	2	4	56	0.442	0.454	NBA013	1.6	0
8	42	3	0	64	0.279	3.945	NBA009	6.6	0
9	26	1	5	18	0.575	2.215	NBA006	15.5	1

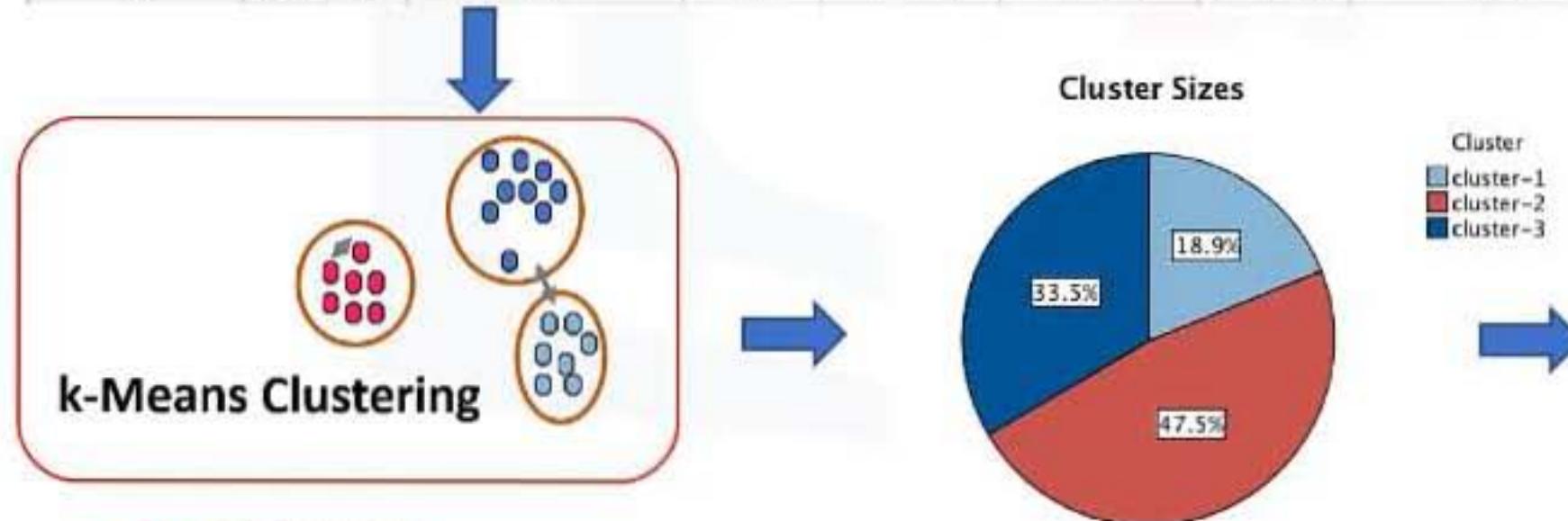
Customer ID	Segment
1	YOUNG AND LOW INCOME
2	AFFLUENT AND MIDDLE AGED
3	AFFLUENT AND MIDDLE AGED
4	YOUNG AND LOW INCOME
5	AFFLUENT AND MIDDLE AGED
6	AFFLUENT AND MIDDLE AGED
7	YOUNG AND LOW INCOME
8	YOUNG AND LOW INCOME
9	AFFLUENT AND MIDDLE AGED



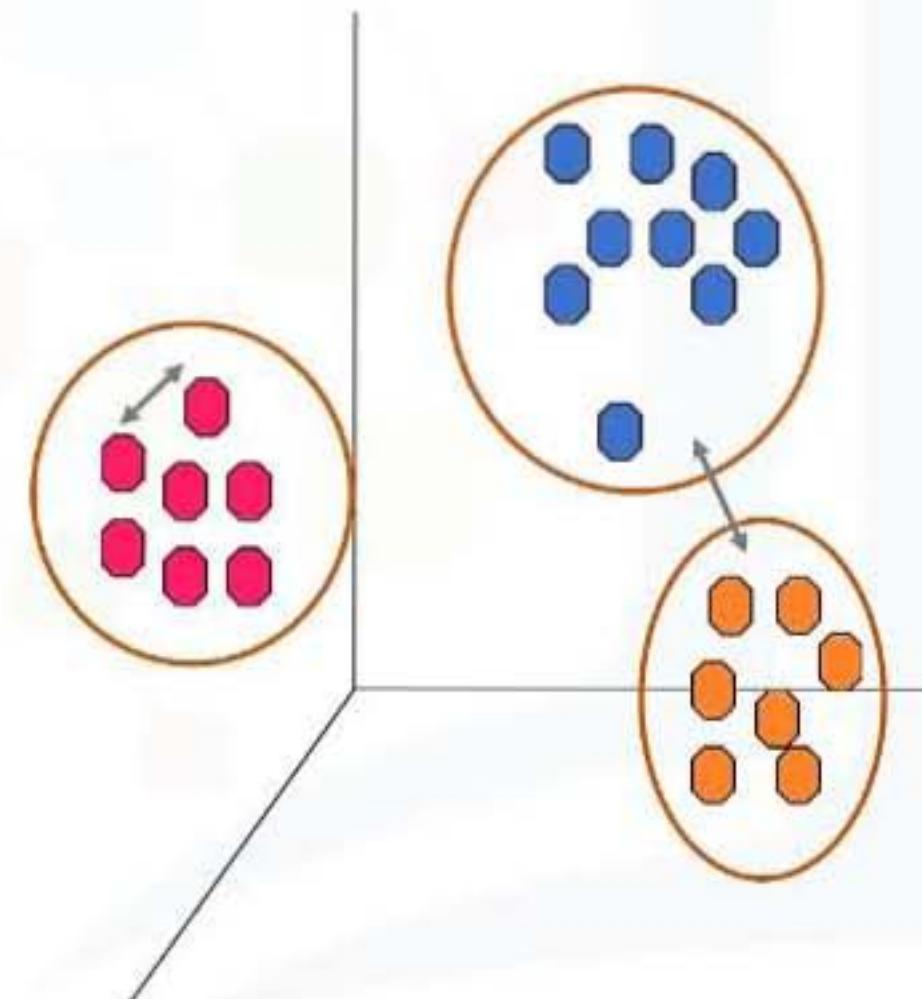
What is k-Means clustering?

Customer ID	Age	Edu	Years Employed	Income	Card Debt	Other Debt	Address	DebtIncomeRatio	Defaulted
1	41	2	6	19	0.124	1.073	NBA001	6.3	0
2	47	1	26	100	4.582	8.218	NBA021	12.8	0
3	33	2	10	57	6.111	5.802	NBA013	20.9	1
4	29	2	4	19	0.681	0.516	NBA009	6.3	0
5	47	1	31	253	9.308	8.908	NBA008	7.2	0
6	40	1	23	81	0.998	7.831	NBA016	10.9	1
7	38	2	4	56	0.442	0.454	NBA013	1.6	0
8	42	3	0	64	0.279	3.945	NBA009	6.6	0
9	26	1	5	18	0.575	2.215	NBA006	15.5	1

Customer ID	Segment
1	YOUNG AND LOW INCOME
2	AFFLUENT AND MIDDLE AGED
3	AFFLUENT AND MIDDLE AGED
4	YOUNG AND LOW INCOME
5	AFFLUENT AND MIDDLE AGED
6	AFFLUENT AND MIDDLE AGED
7	YOUNG AND LOW INCOME
8	YOUNG AND LOW INCOME
9	AFFLUENT AND MIDDLE AGED

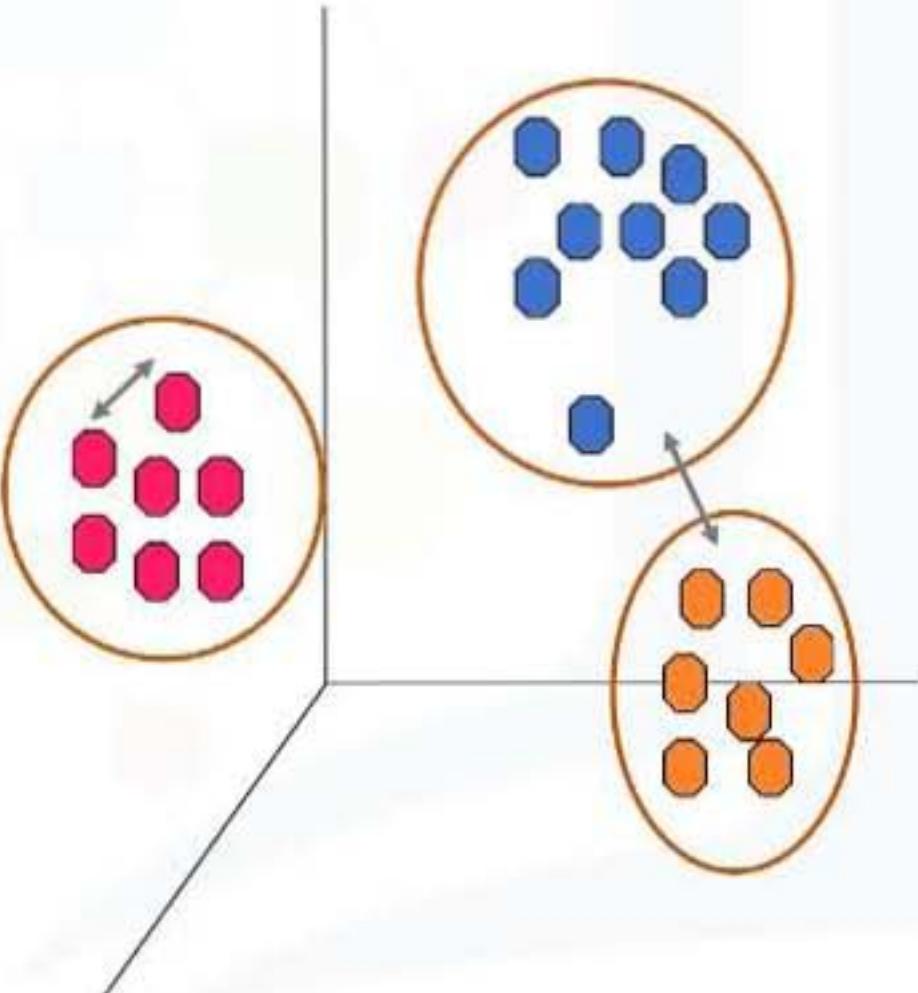


k-Means algorithms



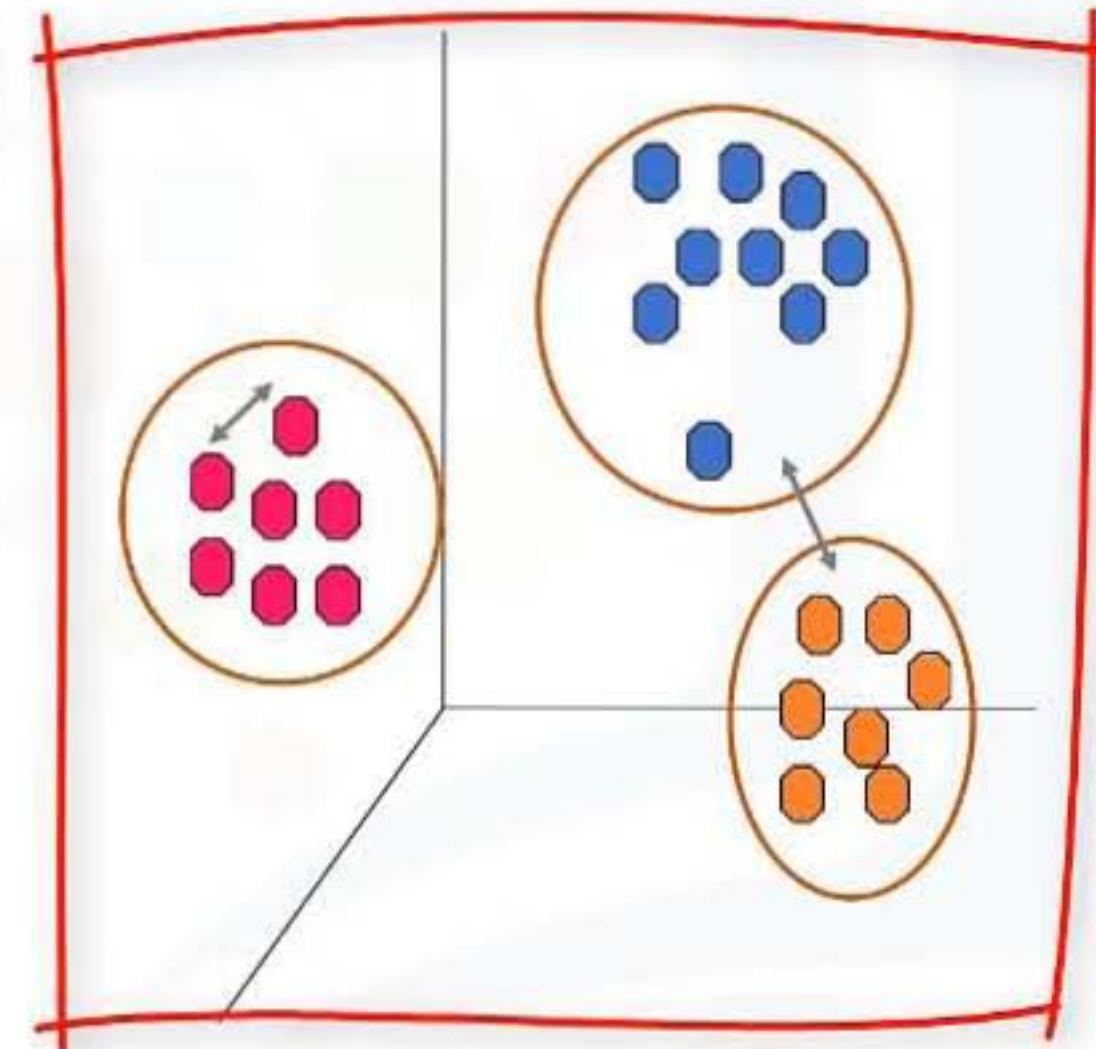
k-Means algorithms

- Partitioning Clustering
- K-means divides the data into non-overlapping subsets (clusters) without any cluster-internal structure



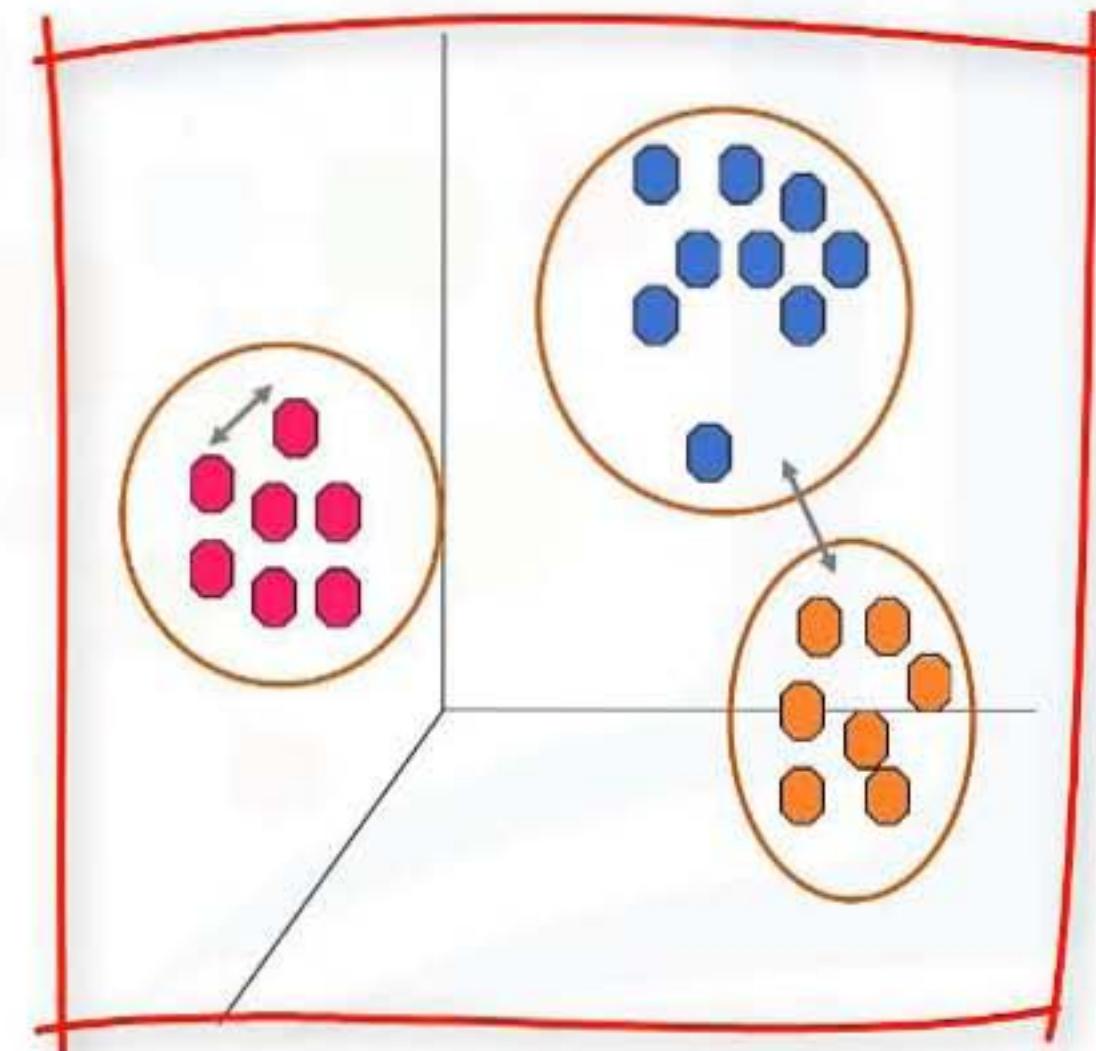
k-Means algorithms

- Partitioning Clustering
- K-means divides the data into non-overlapping subsets (clusters) without any cluster-internal structure
- Examples within a cluster are very similar



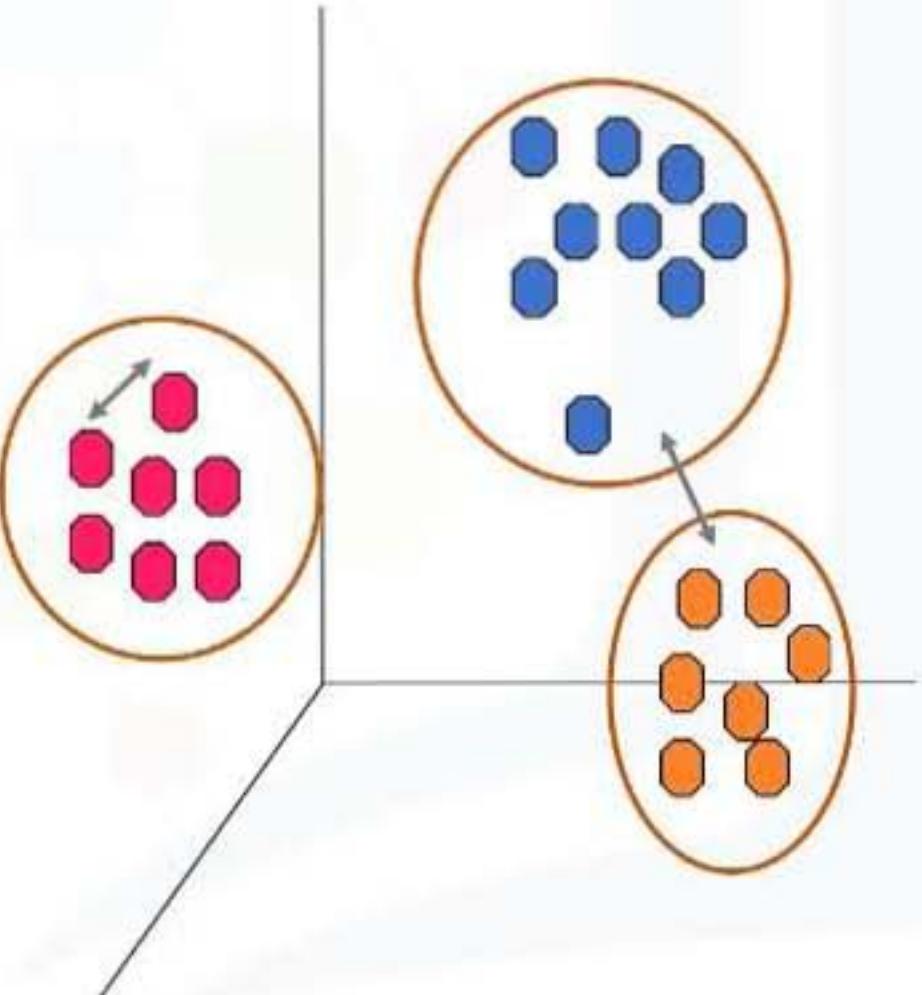
k-Means algorithms

- Partitioning Clustering
- K-means divides the data into non-overlapping subsets (clusters) without any cluster-internal structure
- Examples within a cluster are very similar
- Examples across different clusters are very different

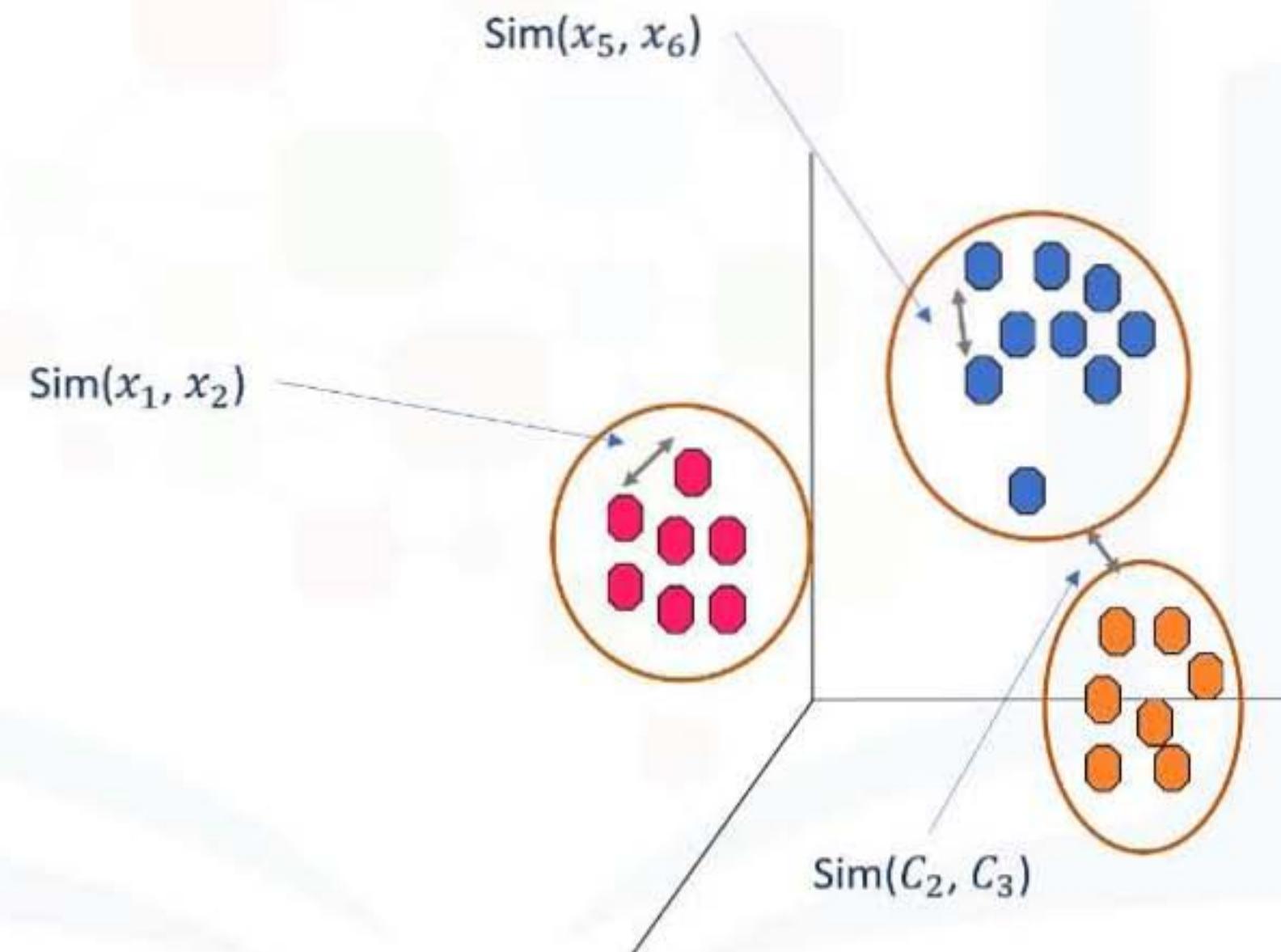


k-Means algorithms

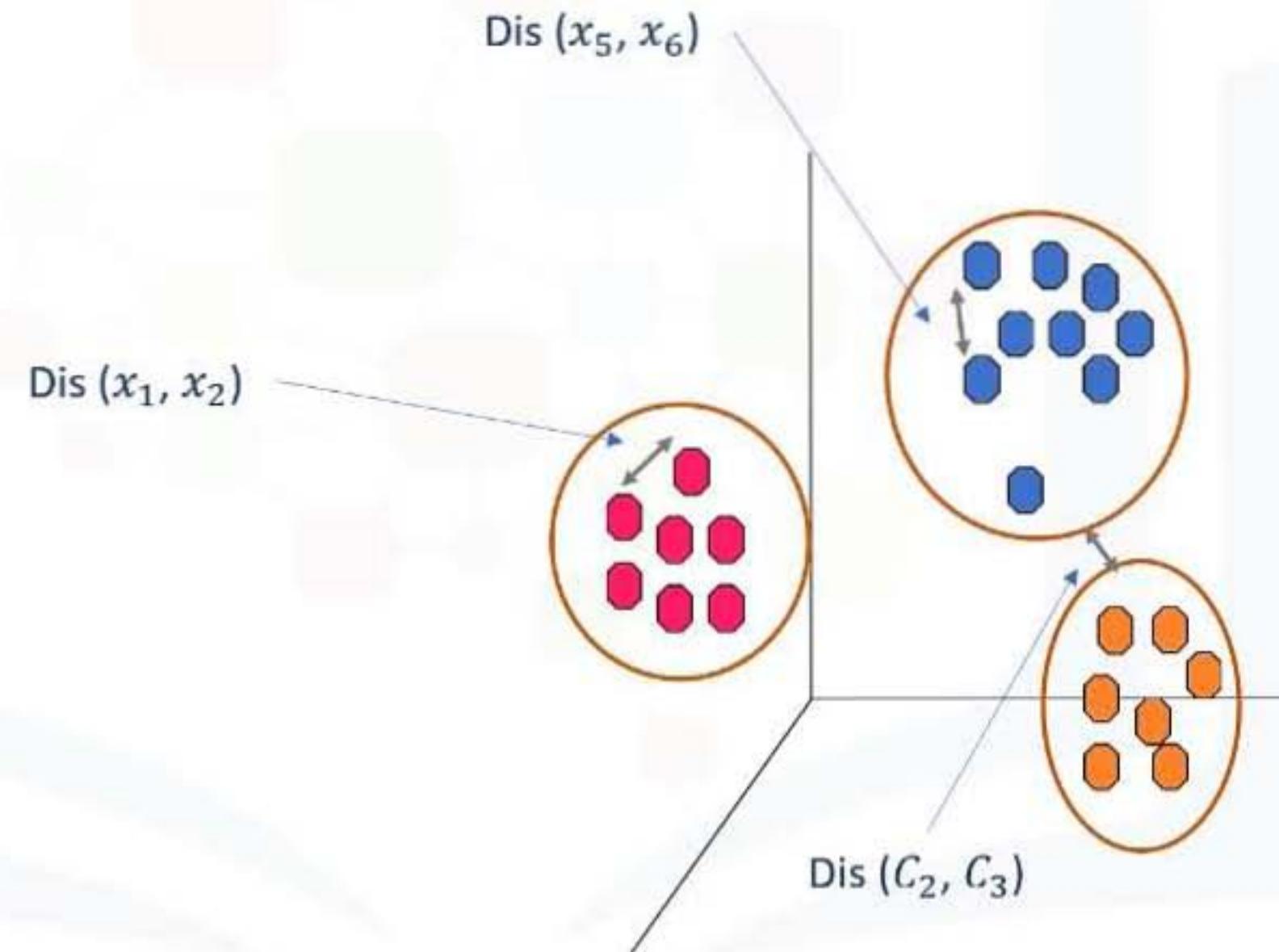
- Partitioning Clustering
- K-means divides the data into non-overlapping subsets (clusters) without any cluster-internal structure
- Examples within a cluster are very similar
- Examples across different clusters are very different



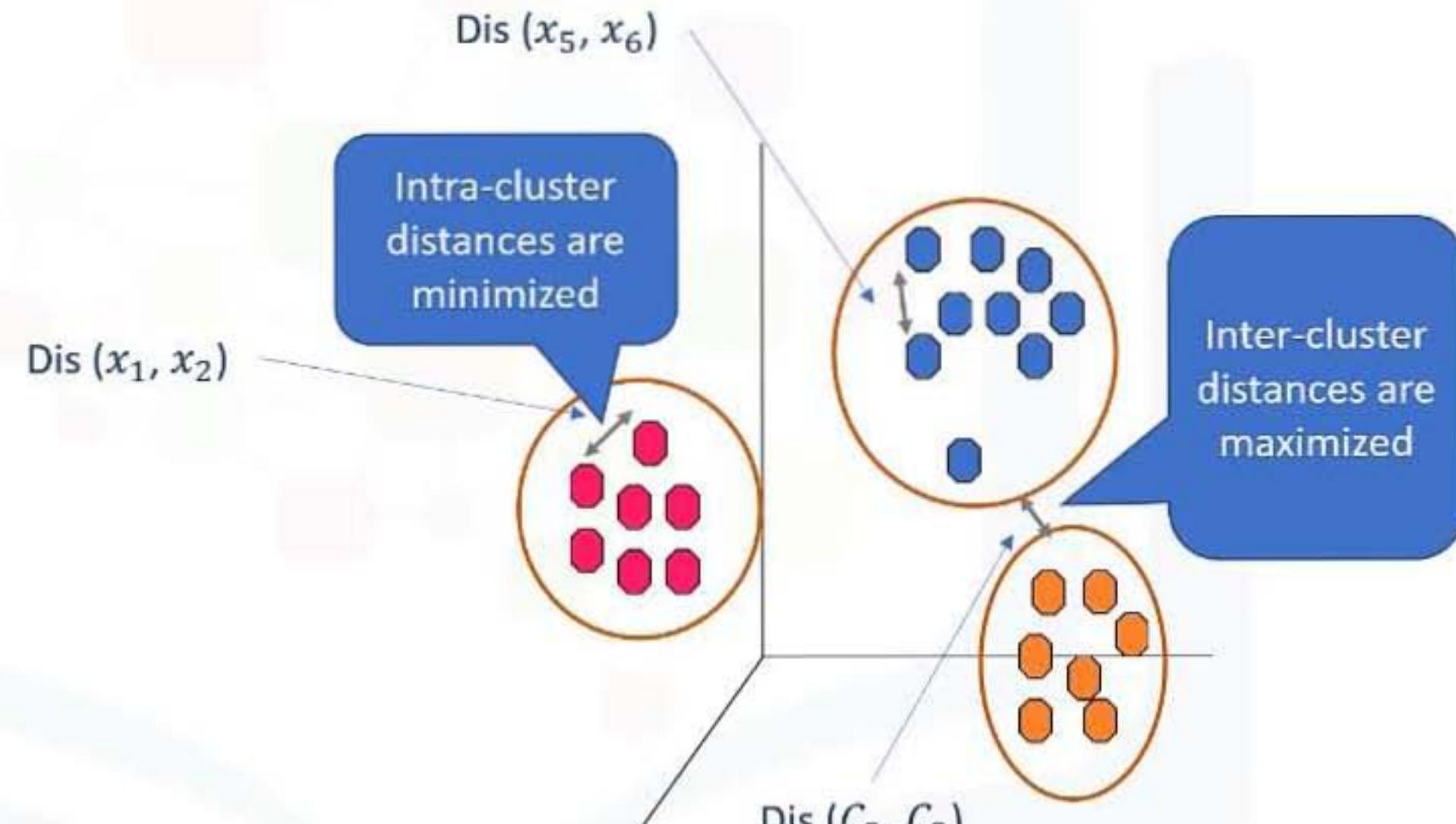
Determine the similarity or dissimilarity



Determine the similarity or dissimilarity



Determine the similarity or dissimilarity



1-dimensional similarity/distance



Customer 1

Age

54



Customer 2

Age

50

1-dimensional similarity/distance



Customer 1

Age

54



Customer 2

Age

50

$$\text{Dis } (x_1, x_2) = \sqrt{\sum_{i=0}^n (x_{1i} - x_{2i})^2}$$

2-dimensional similarity/distance



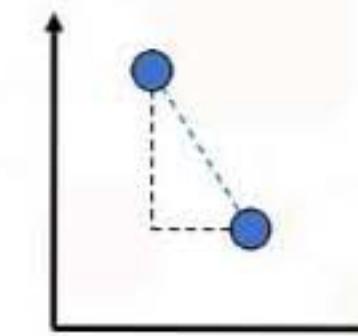
Customer 1

Age	Income
54	190



Customer 2

Age	Income
50	200



$$\begin{aligned} \text{Dis}(x_1, x_2) &= \sqrt{\sum_{i=0}^n (x_{1i} - x_{2i})^2} \\ &= \sqrt{(54 - 50)^2 + (190 - 200)^2} = 10.77 \end{aligned}$$

2-dimensional similarity/distance



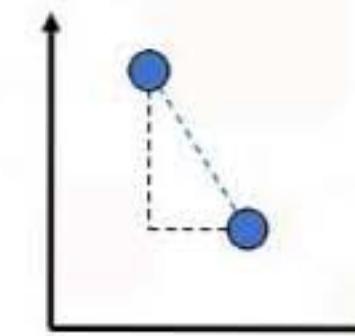
Customer 1

Age	Income
54	190



Customer 2

Age	Income
50	200



$$\begin{aligned}\text{Dis}(x_1, x_2) &= \sqrt{\sum_{i=0}^n (x_{1i} - x_{2i})^2} \\ &= \sqrt{(54 - 50)^2 + (190 - 200)^2} = 10.77\end{aligned}$$

Multi-dimensional similarity/distance



Customer 1		
Age	Income	education
54	190	3

Customer 2		
Age	Income	education
50	200	8

$$\begin{aligned}\text{Dis}(x_1, x_2) &= \sqrt{\sum_{i=0}^n (x_{1i} - x_{2i})^2} \\ &= \sqrt{(54 - 50)^2 + (190 - 200)^2 + (3 - 8)^2} = 11.87\end{aligned}$$

Multi-dimensional similarity/distance



Customer 1

Age	Income	education
54	190	3



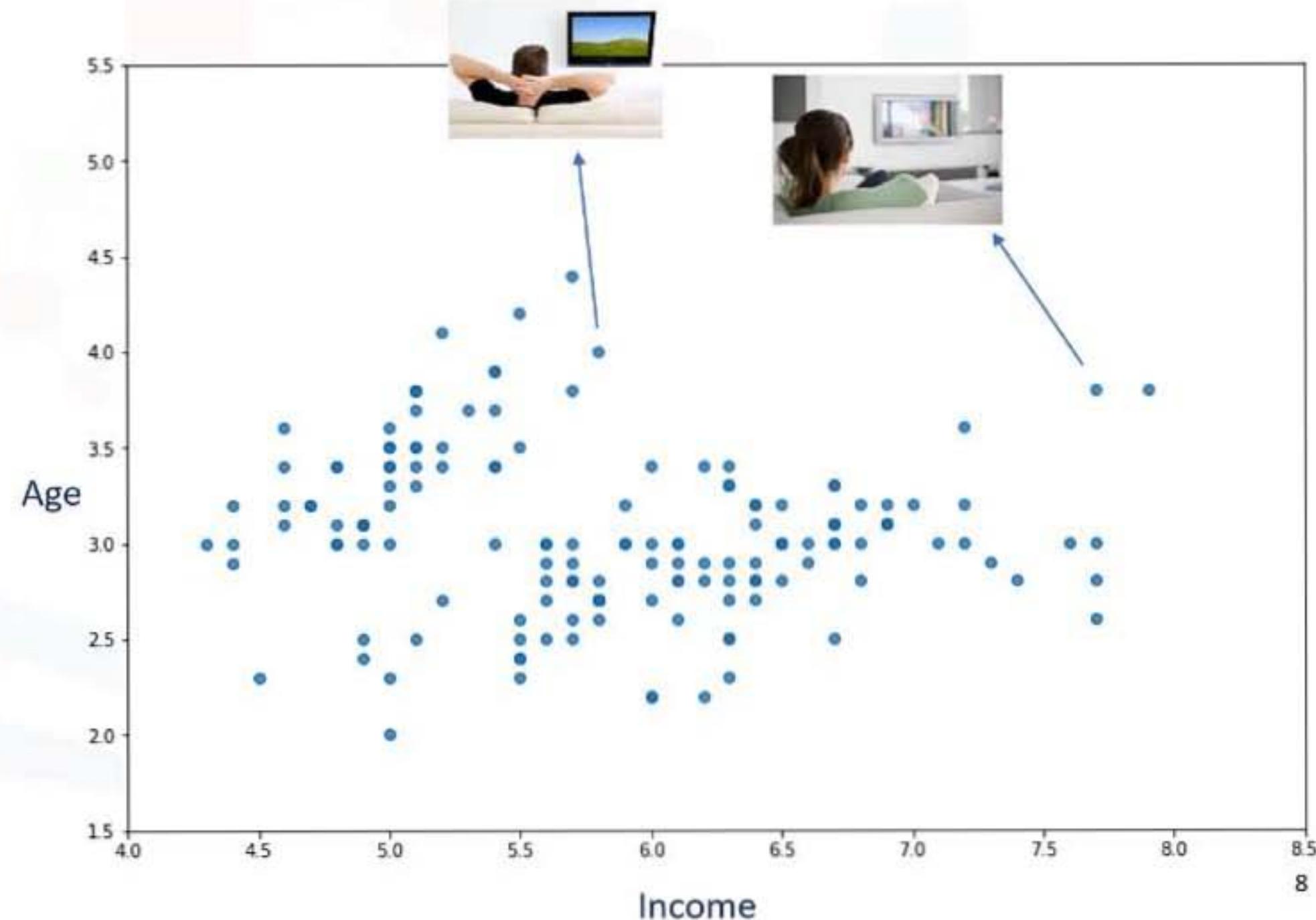
Customer 2

Age	Income	education
50	200	8

$$\begin{aligned}\text{Dis } (x_1, x_2) &= \sqrt{\sum_{i=0}^n (x_{1i} - x_{2i})^2} \\ &= \sqrt{(54 - 50)^2 + (190 - 200)^2 + (3 - 8)^2} = 11.87\end{aligned}$$

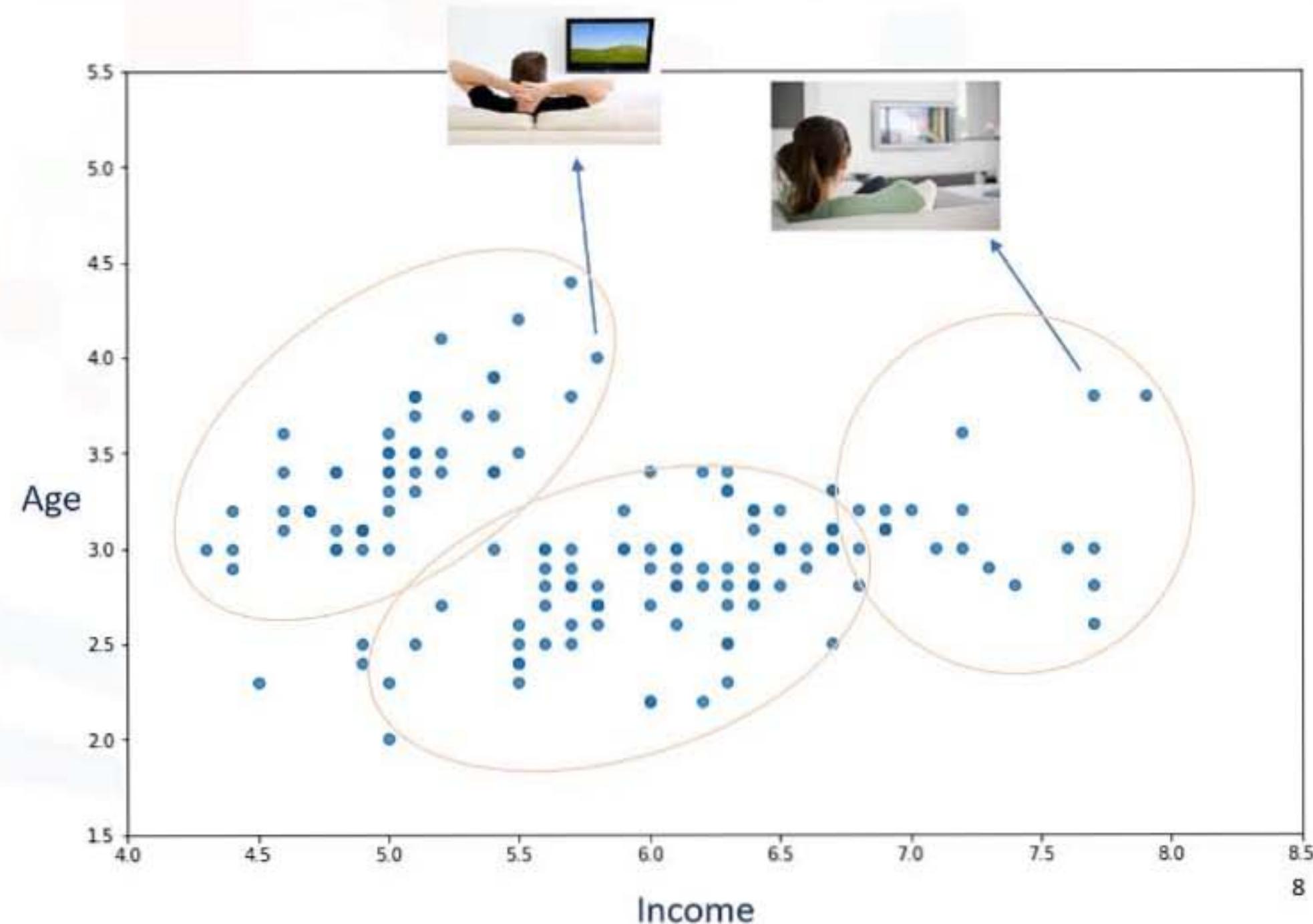
How does k-Means clustering work?

Customer ID	Age	Income
1	3	4
2	2	6
3	3.5	2
...

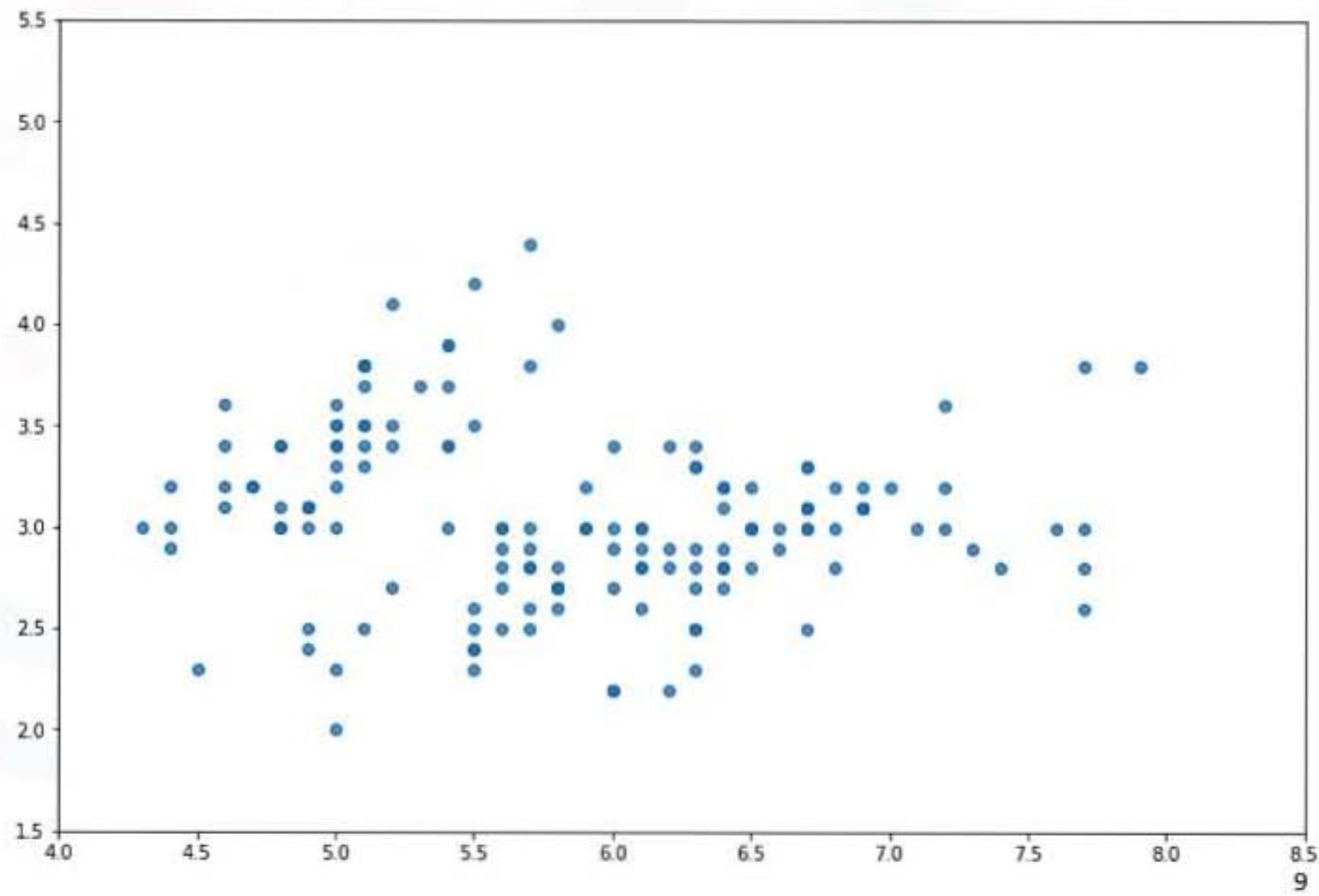


How does k-Means clustering work?

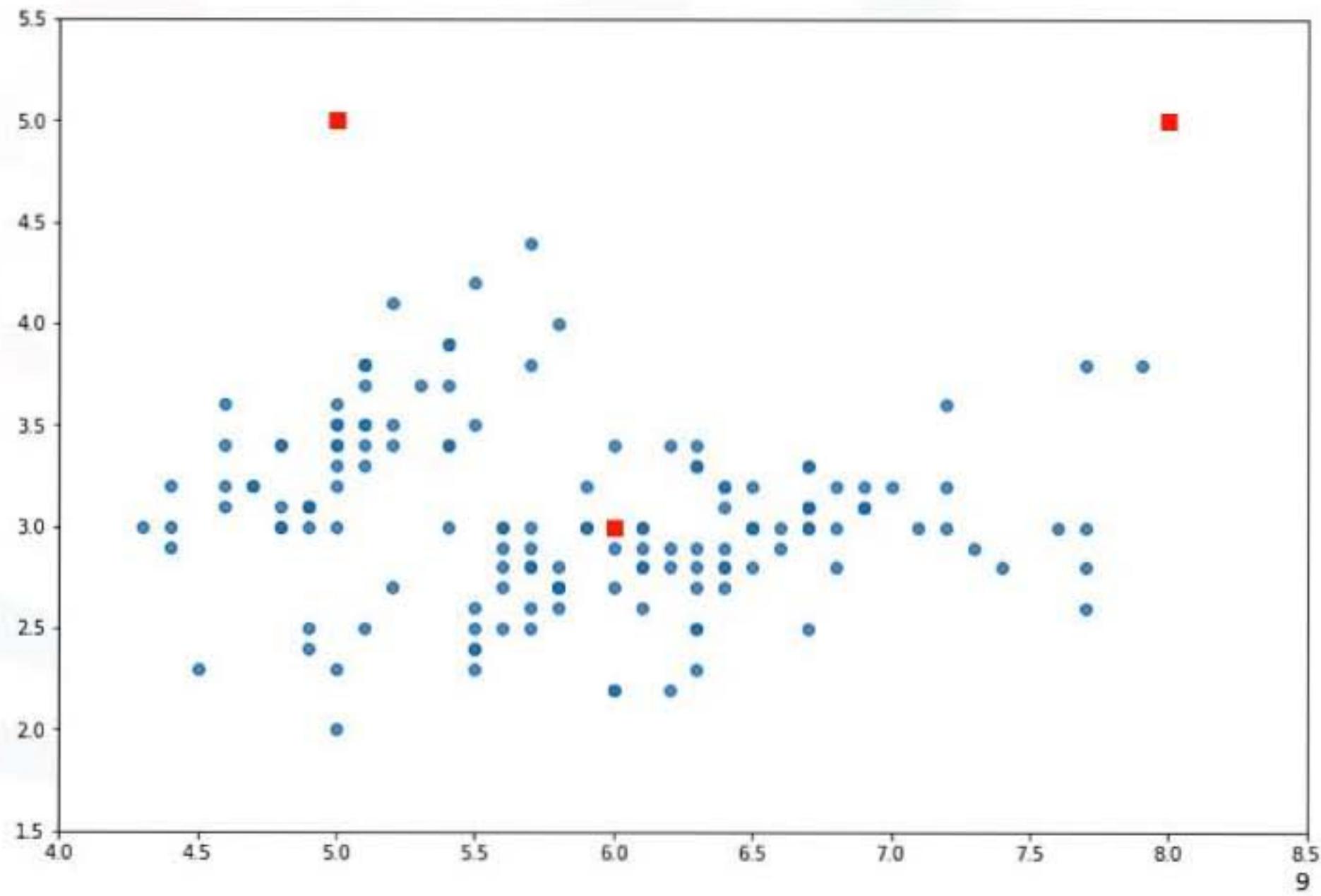
Customer ID	Age	Income
1	3	4
2	2	6
3	3.5	2
...



k-Means clustering – initialize k

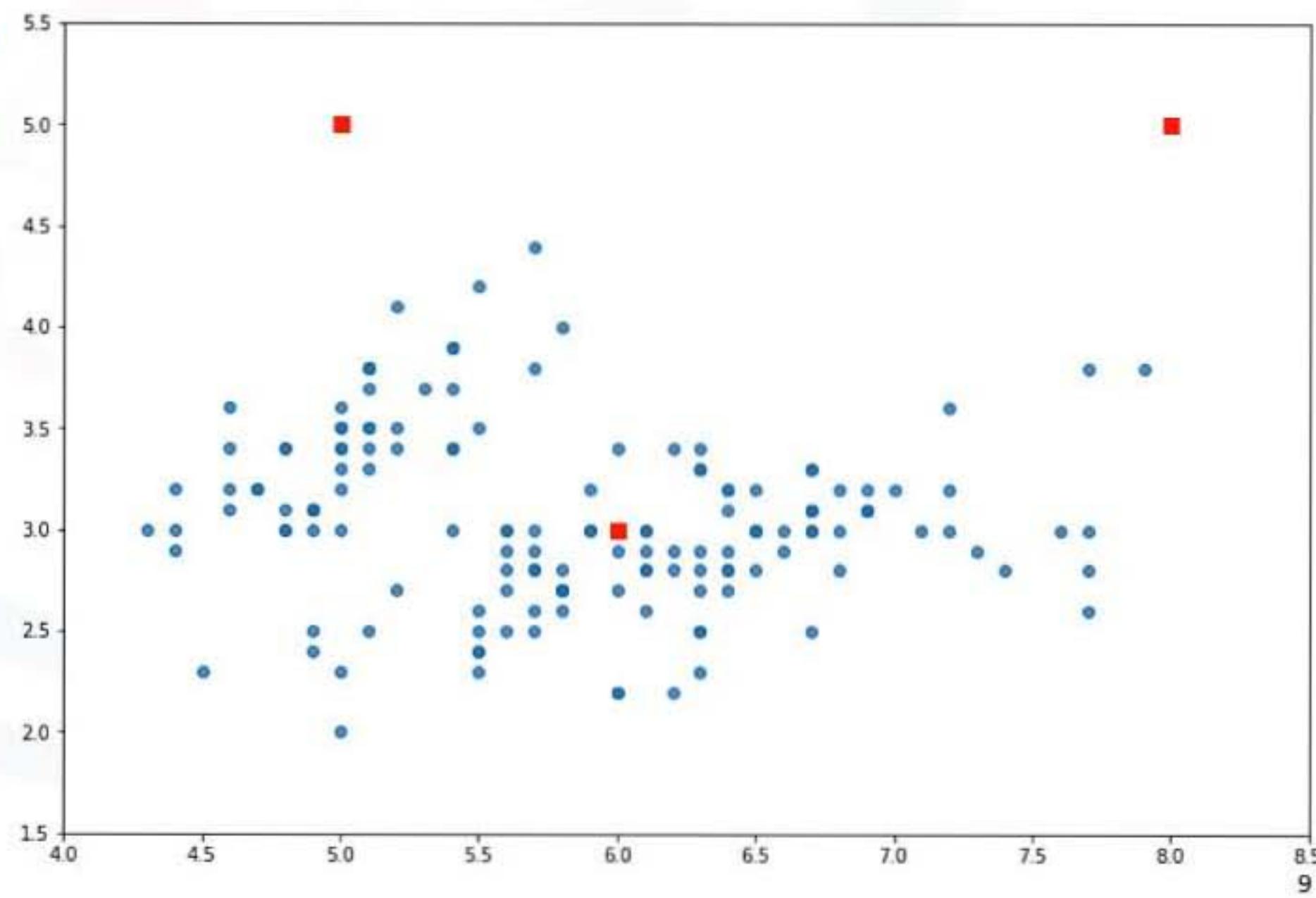


k-Means clustering – initialize k



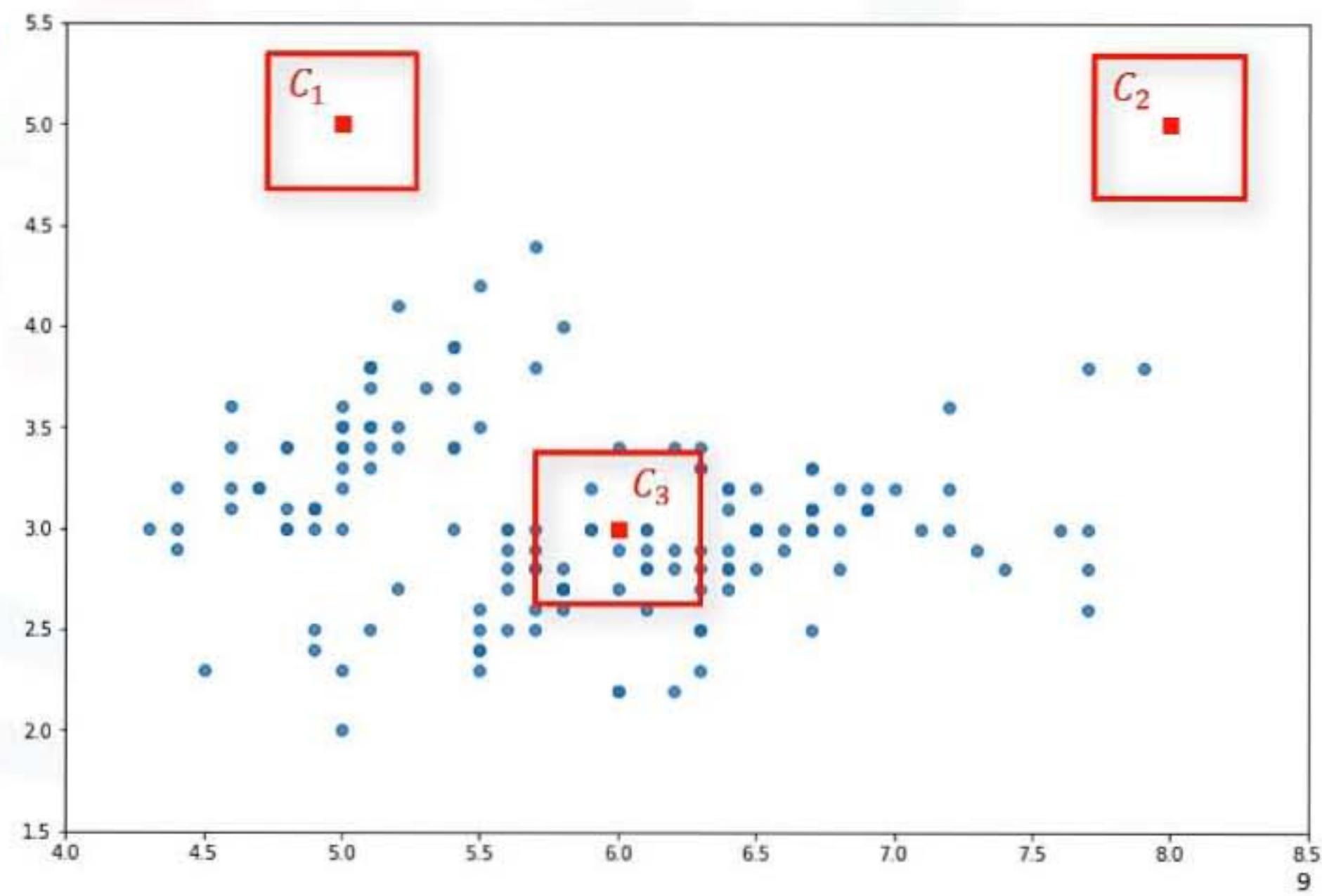
k-Means clustering – initialize k

1) Initialize $k=3$
centroids randomly



k-Means clustering – initialize k

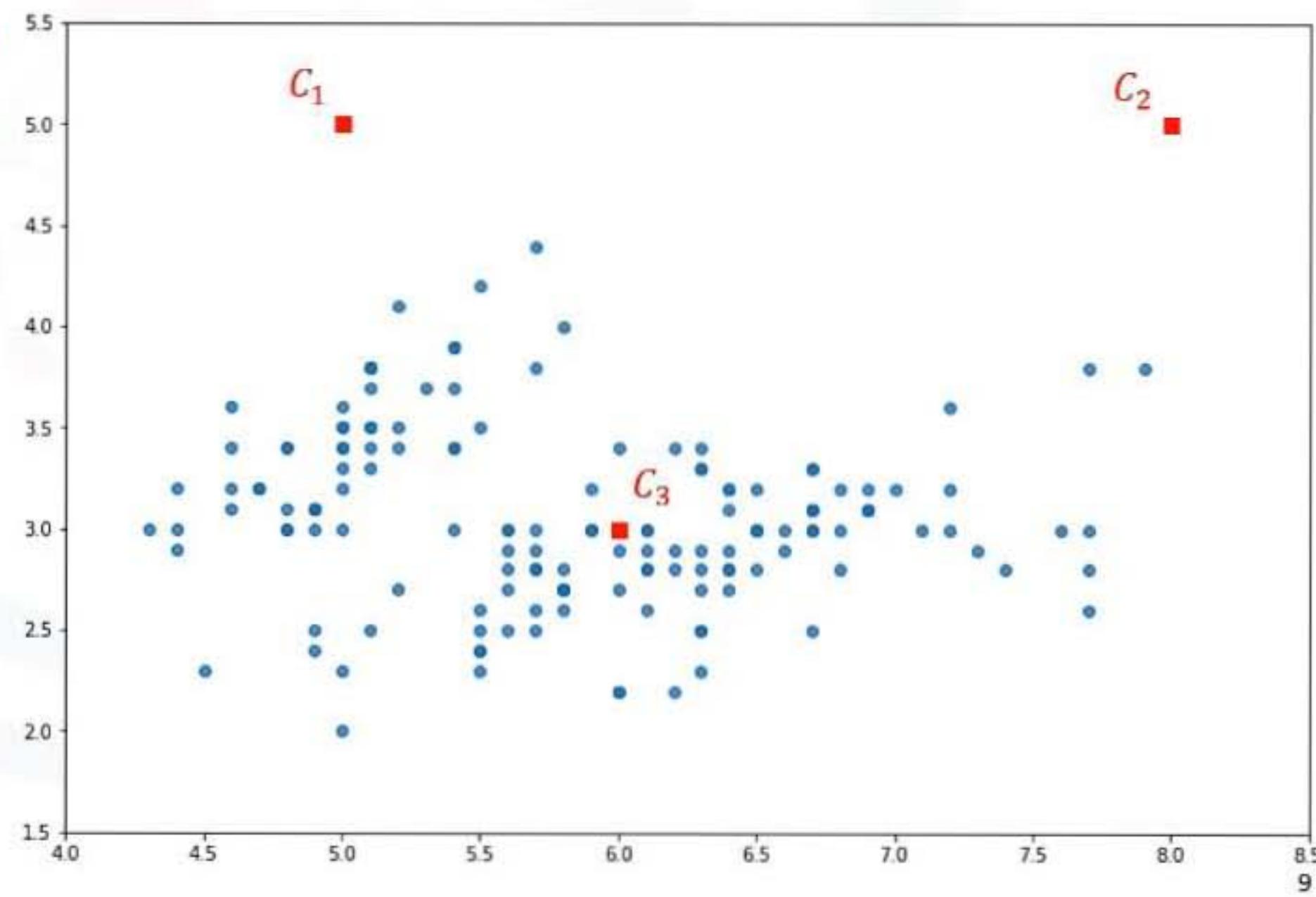
1) Initialize $k=3$
centroids randomly



k-Means clustering – initialize k

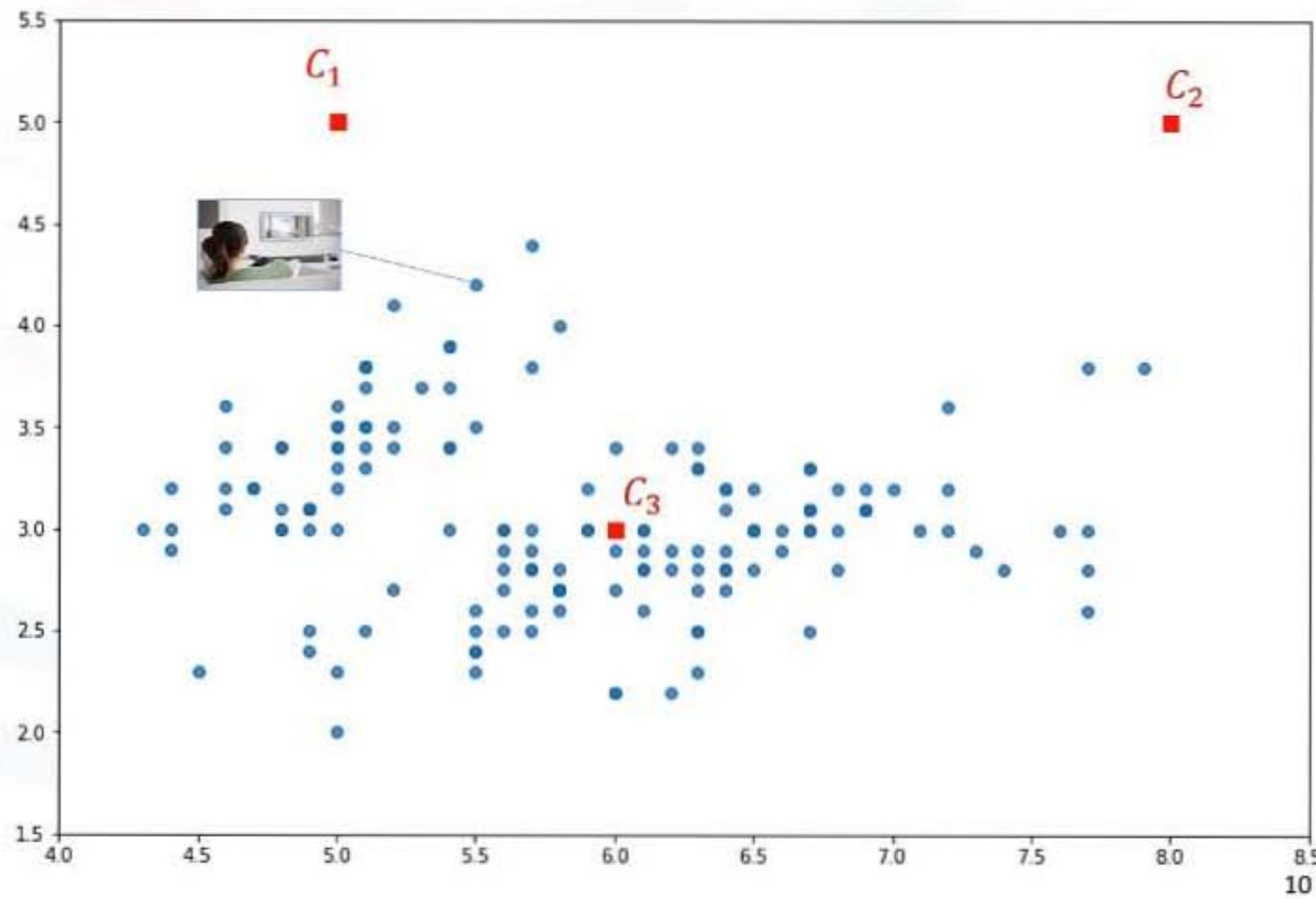
1) Initialize k=3
centroids randomly

$C_1 = [8., 5.]$
 $C_2 = [5., 5.]$
 $C_3 = [6., 3.]$



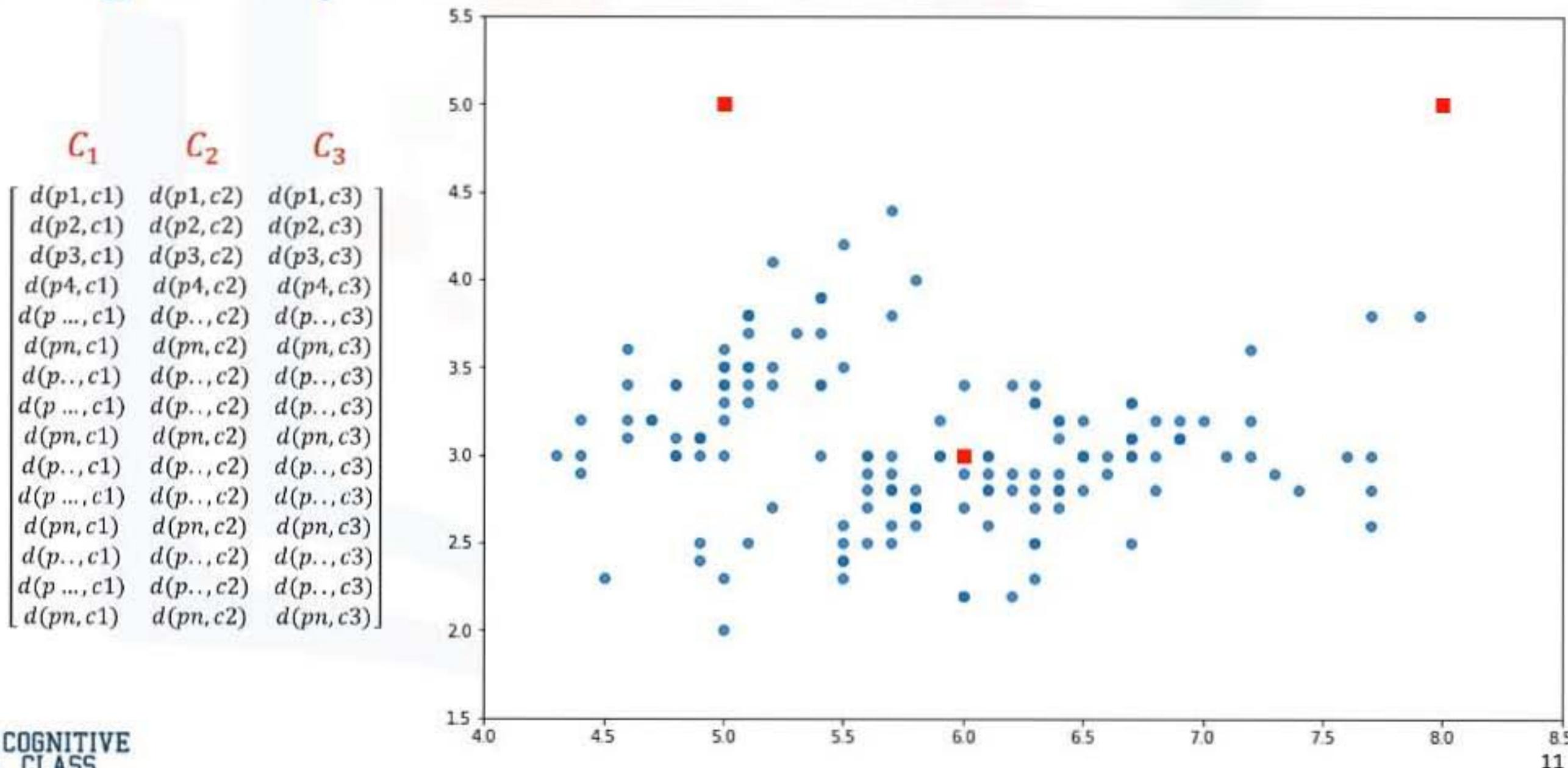
K-Means clustering – calculate the distance

2) Distance calculation



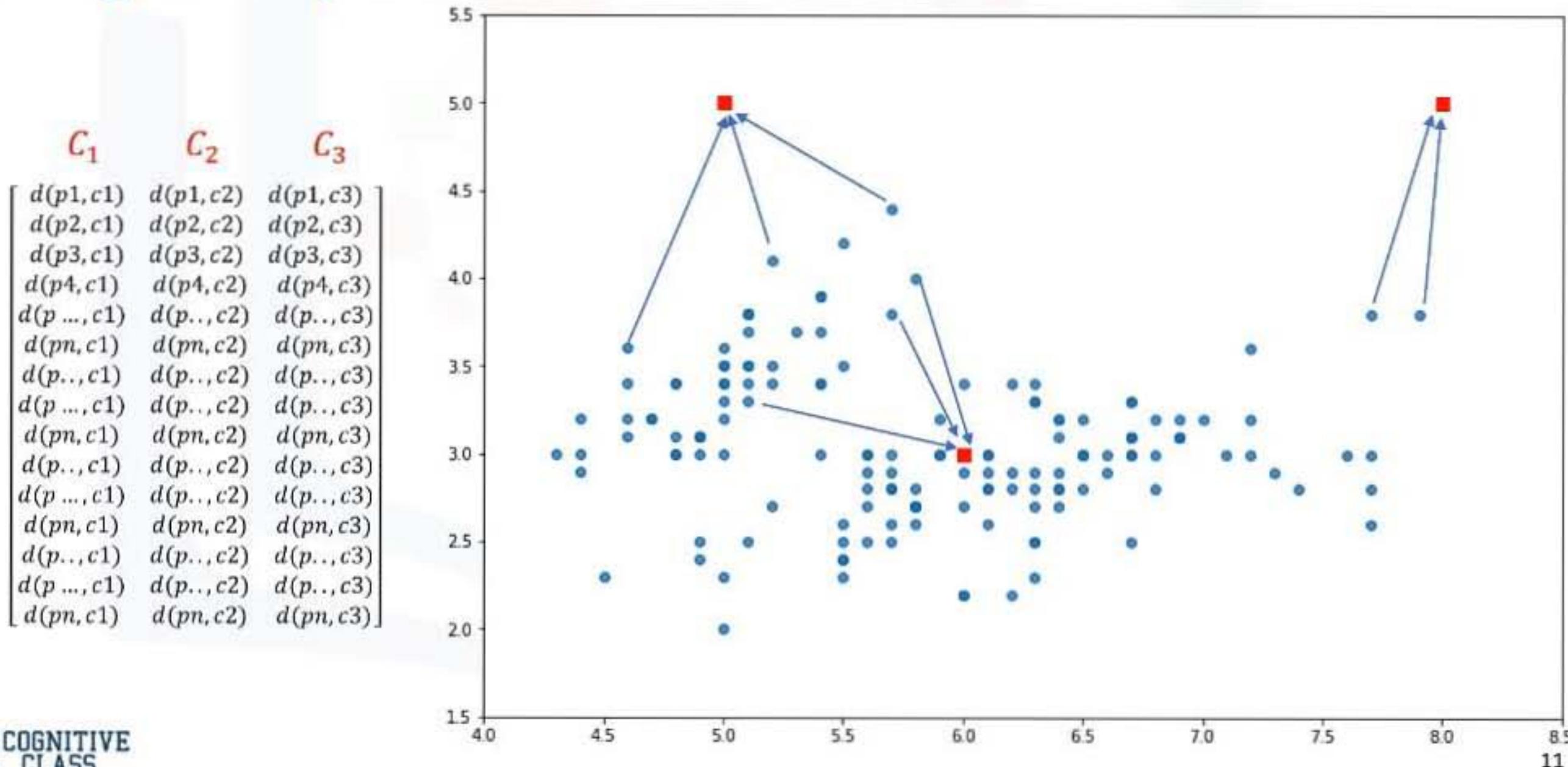
k-Means clustering – assign to centroid

3) Assign each point to the closest centroid



k-Means clustering – assign to centroid

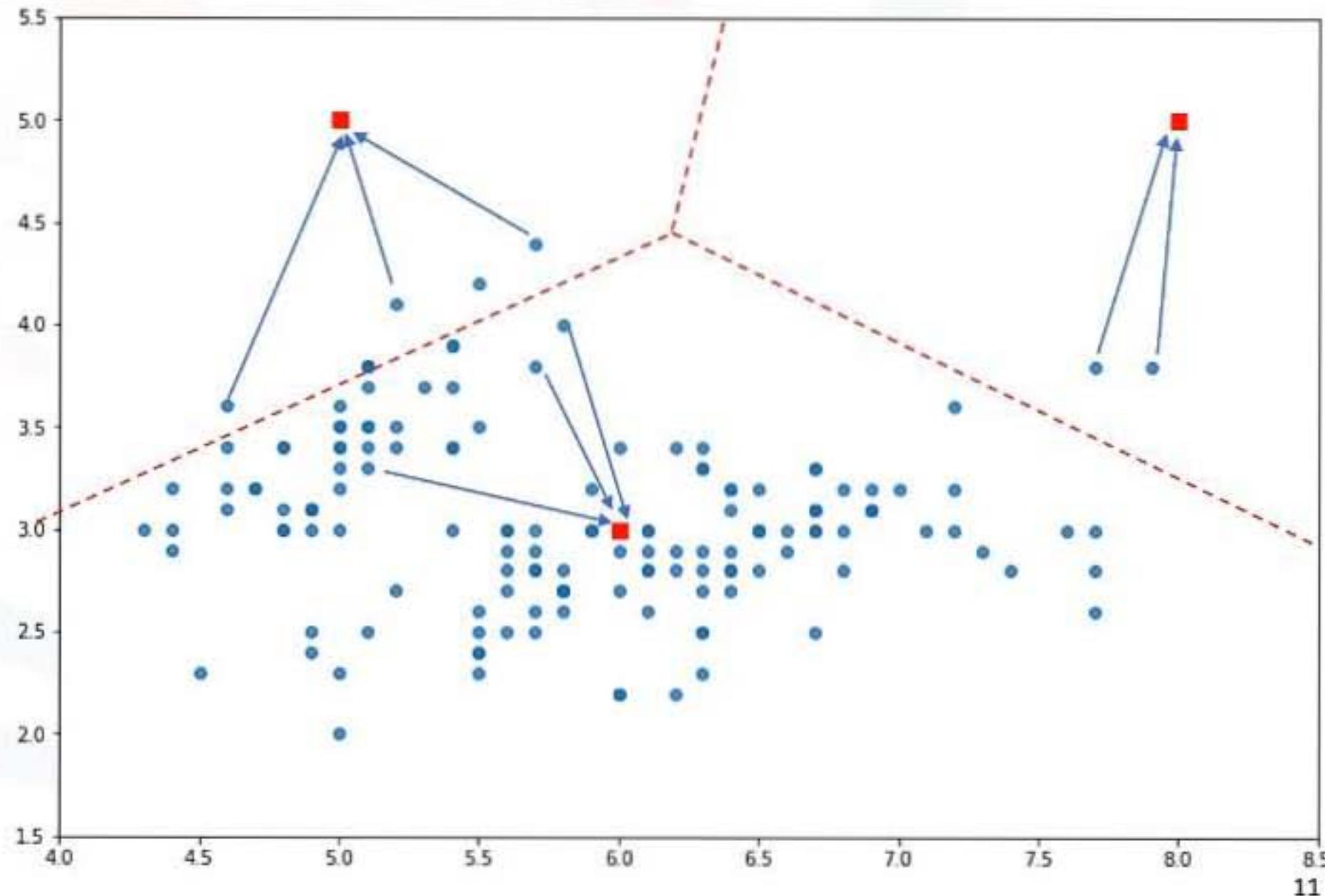
3) Assign each point to the closest centroid



k-Means clustering – assign to centroid

3) Assign each point to the closest centroid

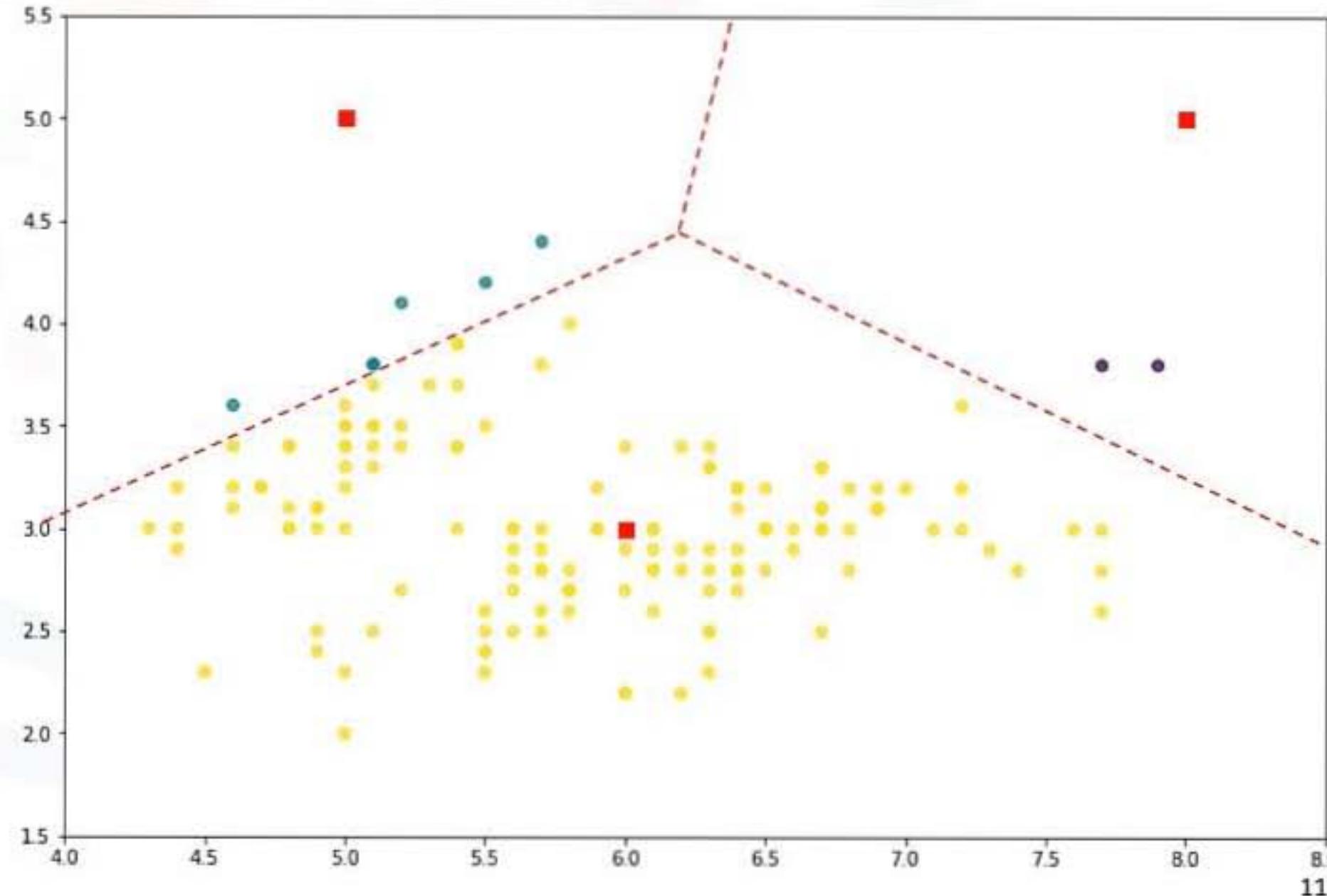
C_1	C_2	C_3
$d(p_1, c_1)$	$d(p_1, c_2)$	$d(p_1, c_3)$
$d(p_2, c_1)$	$d(p_2, c_2)$	$d(p_2, c_3)$
$d(p_3, c_1)$	$d(p_3, c_2)$	$d(p_3, c_3)$
$d(p_4, c_1)$	$d(p_4, c_2)$	$d(p_4, c_3)$
$d(p \dots, c_1)$	$d(p \dots, c_2)$	$d(p \dots, c_3)$
$d(p_n, c_1)$	$d(p_n, c_2)$	$d(p_n, c_3)$
$d(p \dots, c_1)$	$d(p \dots, c_2)$	$d(p \dots, c_3)$
$d(p \dots, c_1)$	$d(p \dots, c_2)$	$d(p \dots, c_3)$
$d(p_n, c_1)$	$d(p_n, c_2)$	$d(p_n, c_3)$
$d(p \dots, c_1)$	$d(p \dots, c_2)$	$d(p \dots, c_3)$
$d(p \dots, c_1)$	$d(p \dots, c_2)$	$d(p \dots, c_3)$
$d(p_n, c_1)$	$d(p_n, c_2)$	$d(p_n, c_3)$
$d(p \dots, c_1)$	$d(p \dots, c_2)$	$d(p \dots, c_3)$
$d(p \dots, c_1)$	$d(p \dots, c_2)$	$d(p \dots, c_3)$
$d(p_n, c_1)$	$d(p_n, c_2)$	$d(p_n, c_3)$
$d(p \dots, c_1)$	$d(p \dots, c_2)$	$d(p \dots, c_3)$
$d(p \dots, c_1)$	$d(p \dots, c_2)$	$d(p \dots, c_3)$
$d(p_n, c_1)$	$d(p_n, c_2)$	$d(p_n, c_3)$



k-Means clustering – assign to centroid

3) Assign each point to the closest centroid

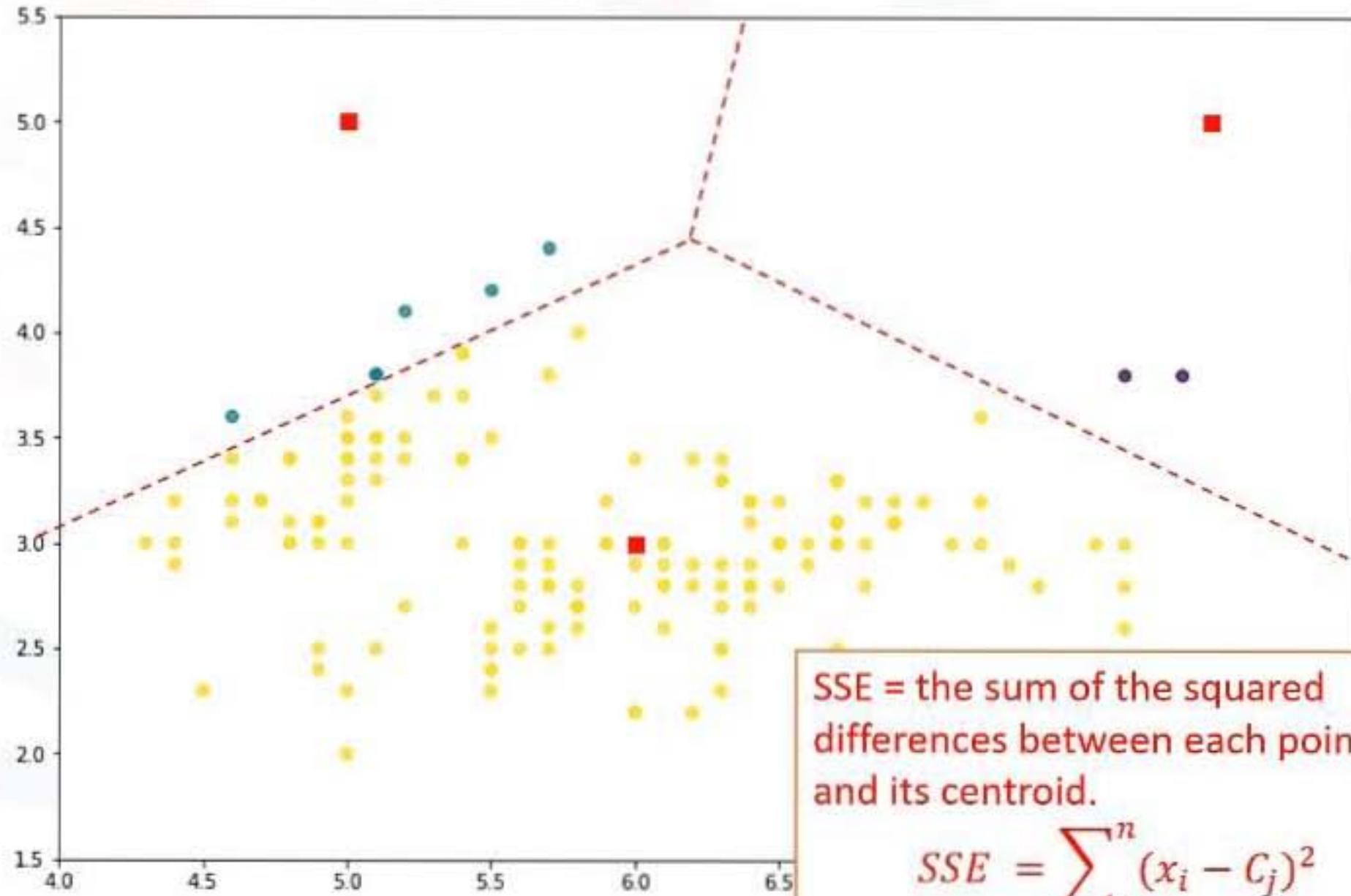
C_1	C_2	C_3
$d(p_1, c_1)$	$d(p_1, c_2)$	$d(p_1, c_3)$
$d(p_2, c_1)$	$d(p_2, c_2)$	$d(p_2, c_3)$
$d(p_3, c_1)$	$d(p_3, c_2)$	$d(p_3, c_3)$
$d(p_4, c_1)$	$d(p_4, c_2)$	$d(p_4, c_3)$
$d(p \dots, c_1)$	$d(p \dots, c_2)$	$d(p \dots, c_3)$
$d(p_n, c_1)$	$d(p_n, c_2)$	$d(p_n, c_3)$
$d(p \dots, c_1)$	$d(p \dots, c_2)$	$d(p \dots, c_3)$
$d(p \dots, c_1)$	$d(p \dots, c_2)$	$d(p \dots, c_3)$
$d(p_n, c_1)$	$d(p_n, c_2)$	$d(p_n, c_3)$
$d(p \dots, c_1)$	$d(p \dots, c_2)$	$d(p \dots, c_3)$
$d(p \dots, c_1)$	$d(p \dots, c_2)$	$d(p \dots, c_3)$
$d(p_n, c_1)$	$d(p_n, c_2)$	$d(p_n, c_3)$
$d(p \dots, c_1)$	$d(p \dots, c_2)$	$d(p \dots, c_3)$
$d(p \dots, c_1)$	$d(p \dots, c_2)$	$d(p \dots, c_3)$
$d(p_n, c_1)$	$d(p_n, c_2)$	$d(p_n, c_3)$
$d(p \dots, c_1)$	$d(p \dots, c_2)$	$d(p \dots, c_3)$
$d(p \dots, c_1)$	$d(p \dots, c_2)$	$d(p \dots, c_3)$
$d(p_n, c_1)$	$d(p_n, c_2)$	$d(p_n, c_3)$



k-Means clustering – assign to centroid

3) Assign each point to the closest centroid

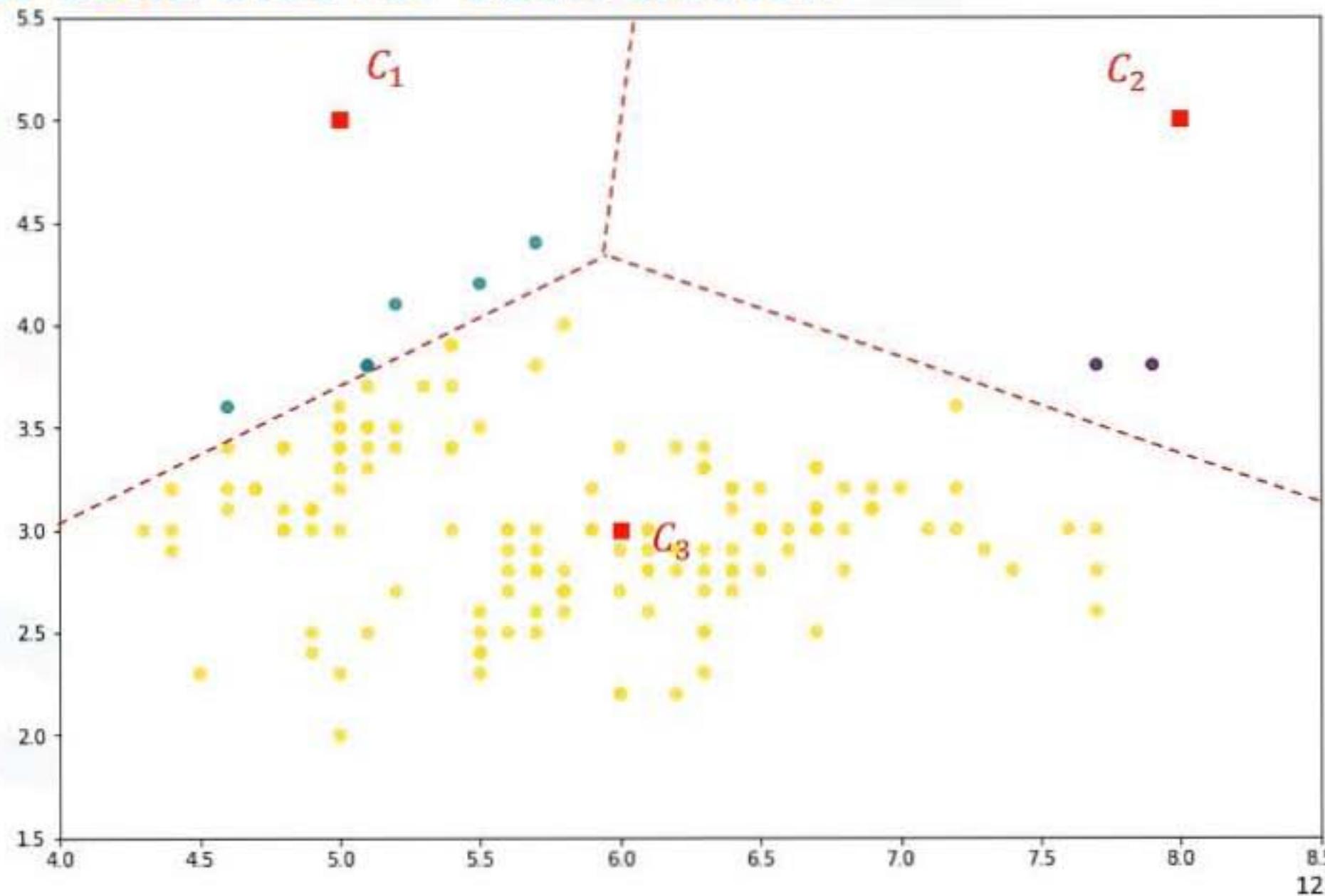
C_1	C_2	C_3
$d(p_1, c_1)$	$d(p_1, c_2)$	$d(p_1, c_3)$
$d(p_2, c_1)$	$d(p_2, c_2)$	$d(p_2, c_3)$
$d(p_3, c_1)$	$d(p_3, c_2)$	$d(p_3, c_3)$
$d(p_4, c_1)$	$d(p_4, c_2)$	$d(p_4, c_3)$
$d(p \dots, c_1)$	$d(p \dots, c_2)$	$d(p \dots, c_3)$
$d(p_n, c_1)$	$d(p_n, c_2)$	$d(p_n, c_3)$
$d(p \dots, c_1)$	$d(p \dots, c_2)$	$d(p \dots, c_3)$
$d(p \dots, c_1)$	$d(p \dots, c_2)$	$d(p \dots, c_3)$
$d(p_n, c_1)$	$d(p_n, c_2)$	$d(p_n, c_3)$
$d(p \dots, c_1)$	$d(p \dots, c_2)$	$d(p \dots, c_3)$
$d(p \dots, c_1)$	$d(p \dots, c_2)$	$d(p \dots, c_3)$
$d(p_n, c_1)$	$d(p_n, c_2)$	$d(p_n, c_3)$
$d(p \dots, c_1)$	$d(p \dots, c_2)$	$d(p \dots, c_3)$
$d(p \dots, c_1)$	$d(p \dots, c_2)$	$d(p \dots, c_3)$
$d(p_n, c_1)$	$d(p_n, c_2)$	$d(p_n, c_3)$
$d(p \dots, c_1)$	$d(p \dots, c_2)$	$d(p \dots, c_3)$
$d(p \dots, c_1)$	$d(p \dots, c_2)$	$d(p \dots, c_3)$
$d(p_n, c_1)$	$d(p_n, c_2)$	$d(p_n, c_3)$



k-Means clustering – compute new centroids

4) Compute the new centroids for each cluster.

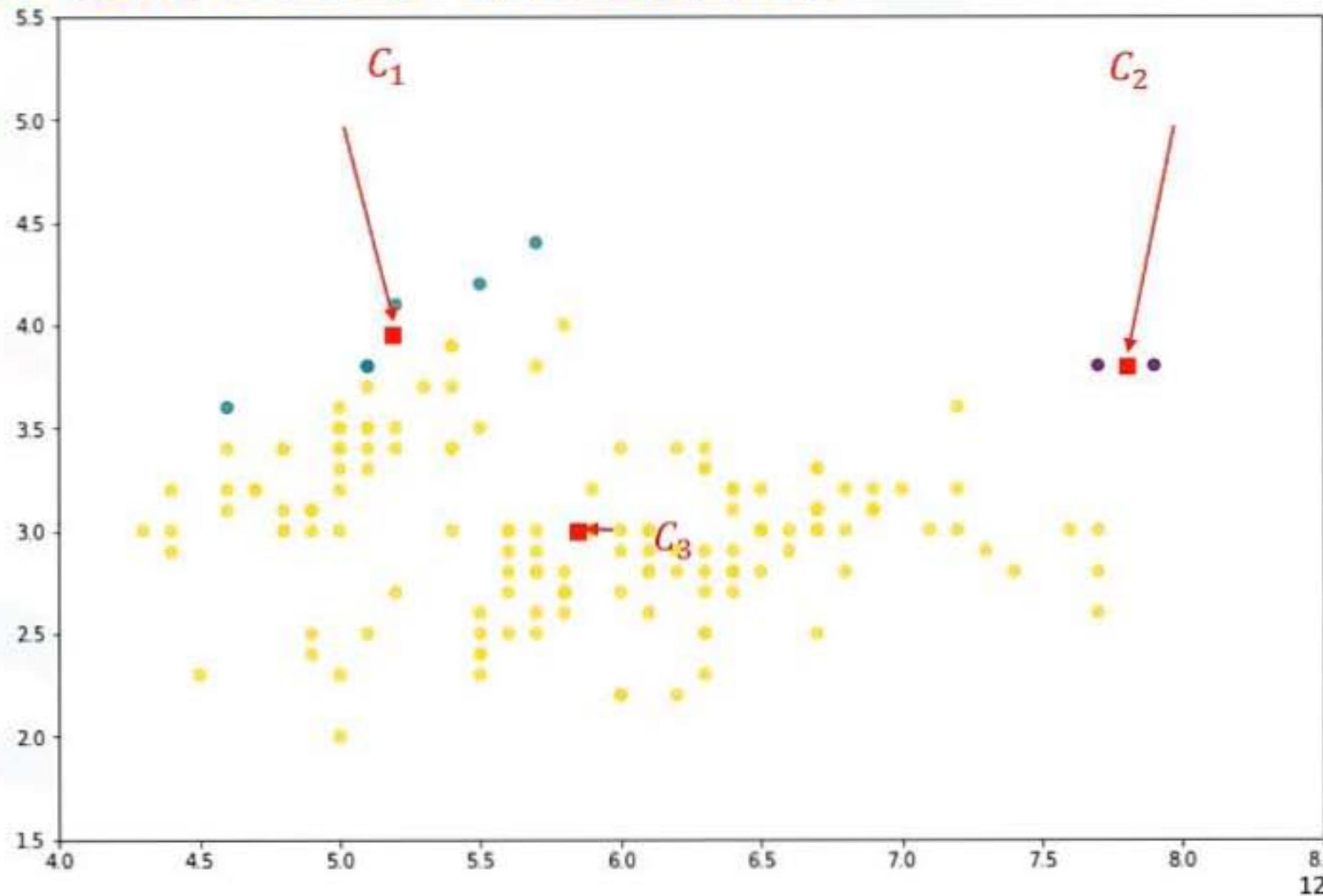
C_1	C_2	C_3
$d(p_1, c_1)$	$d(p_1, c_2)$	$d(p_1, c_3)$
$d(p_2, c_1)$	$d(p_2, c_2)$	$d(p_2, c_3)$
$d(p_3, c_1)$	$d(p_3, c_2)$	$d(p_3, c_3)$
$d(p_4, c_1)$	$d(p_4, c_2)$	$d(p_4, c_3)$
$d(p \dots, c_1)$	$d(p \dots, c_2)$	$d(p \dots, c_3)$
$d(p_n, c_1)$	$d(p_n, c_2)$	$d(p_n, c_3)$
$d(p \dots, c_1)$	$d(p \dots, c_2)$	$d(p \dots, c_3)$
$d(p \dots, c_1)$	$d(p \dots, c_2)$	$d(p \dots, c_3)$
$d(p_n, c_1)$	$d(p_n, c_2)$	$d(p_n, c_3)$
$d(p \dots, c_1)$	$d(p \dots, c_2)$	$d(p \dots, c_3)$
$d(p \dots, c_1)$	$d(p \dots, c_2)$	$d(p \dots, c_3)$
$d(p_n, c_1)$	$d(p_n, c_2)$	$d(p_n, c_3)$
$d(p \dots, c_1)$	$d(p \dots, c_2)$	$d(p \dots, c_3)$
$d(p \dots, c_1)$	$d(p \dots, c_2)$	$d(p \dots, c_3)$
$d(p_n, c_1)$	$d(p_n, c_2)$	$d(p_n, c_3)$
$d(p \dots, c_1)$	$d(p \dots, c_2)$	$d(p \dots, c_3)$
$d(p \dots, c_1)$	$d(p \dots, c_2)$	$d(p \dots, c_3)$
$d(p_n, c_1)$	$d(p_n, c_2)$	$d(p_n, c_3)$



k-Means clustering – compute new centroids

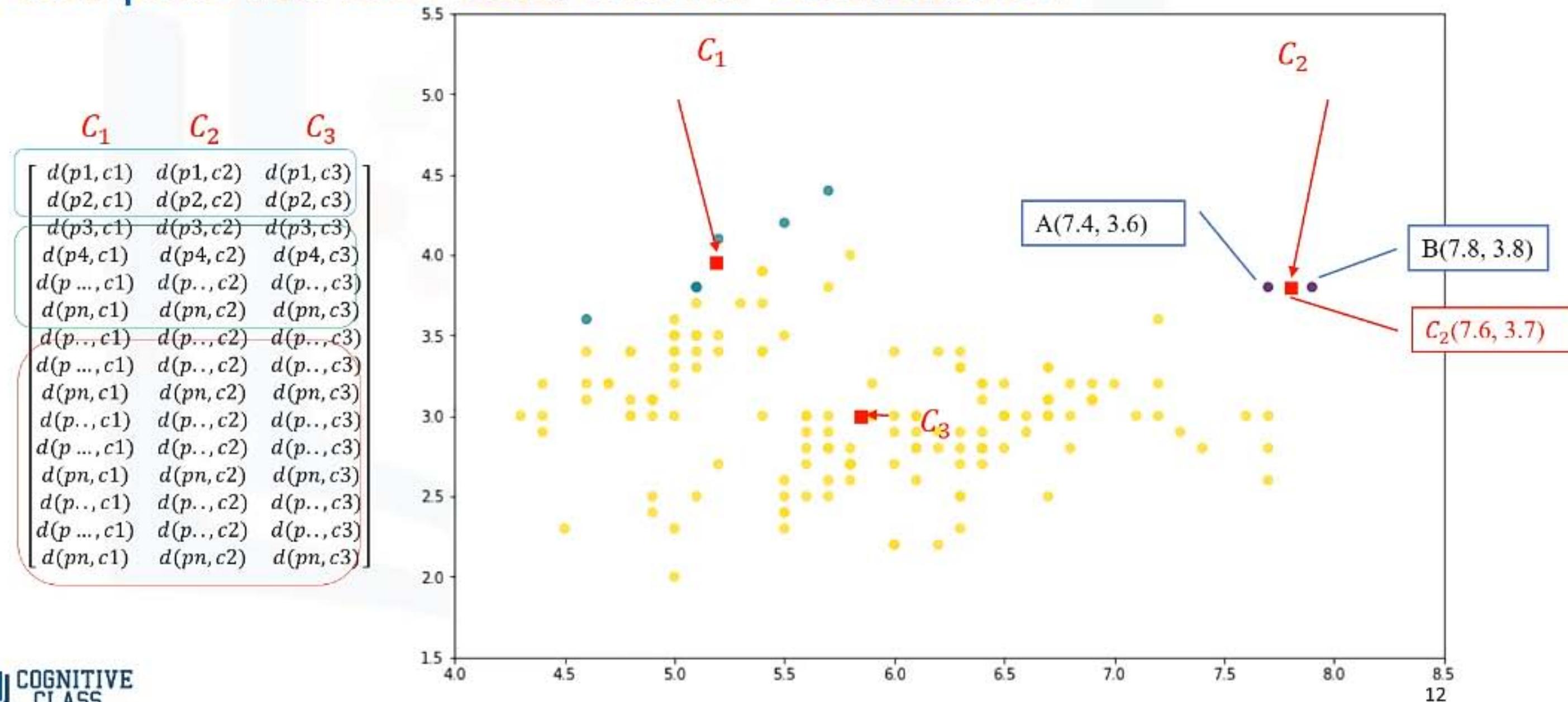
4) Compute the new centroids for each cluster.

C_1	C_2	C_3
$d(p_1, c_1)$	$d(p_1, c_2)$	$d(p_1, c_3)$
$d(p_2, c_1)$	$d(p_2, c_2)$	$d(p_2, c_3)$
$d(p_3, c_1)$	$d(p_3, c_2)$	$d(p_3, c_3)$
$d(p_4, c_1)$	$d(p_4, c_2)$	$d(p_4, c_3)$
$d(p \dots, c_1)$	$d(p \dots, c_2)$	$d(p \dots, c_3)$
$d(p_n, c_1)$	$d(p_n, c_2)$	$d(p_n, c_3)$
$d(p \dots, c_1)$	$d(p \dots, c_2)$	$d(p \dots, c_3)$
$d(p \dots, c_1)$	$d(p \dots, c_2)$	$d(p \dots, c_3)$
$d(p_n, c_1)$	$d(p_n, c_2)$	$d(p_n, c_3)$
$d(p \dots, c_1)$	$d(p \dots, c_2)$	$d(p \dots, c_3)$
$d(p \dots, c_1)$	$d(p \dots, c_2)$	$d(p \dots, c_3)$
$d(p_n, c_1)$	$d(p_n, c_2)$	$d(p_n, c_3)$
$d(p \dots, c_1)$	$d(p \dots, c_2)$	$d(p \dots, c_3)$
$d(p \dots, c_1)$	$d(p \dots, c_2)$	$d(p \dots, c_3)$
$d(p_n, c_1)$	$d(p_n, c_2)$	$d(p_n, c_3)$
$d(p \dots, c_1)$	$d(p \dots, c_2)$	$d(p \dots, c_3)$
$d(p \dots, c_1)$	$d(p \dots, c_2)$	$d(p \dots, c_3)$
$d(p_n, c_1)$	$d(p_n, c_2)$	$d(p_n, c_3)$



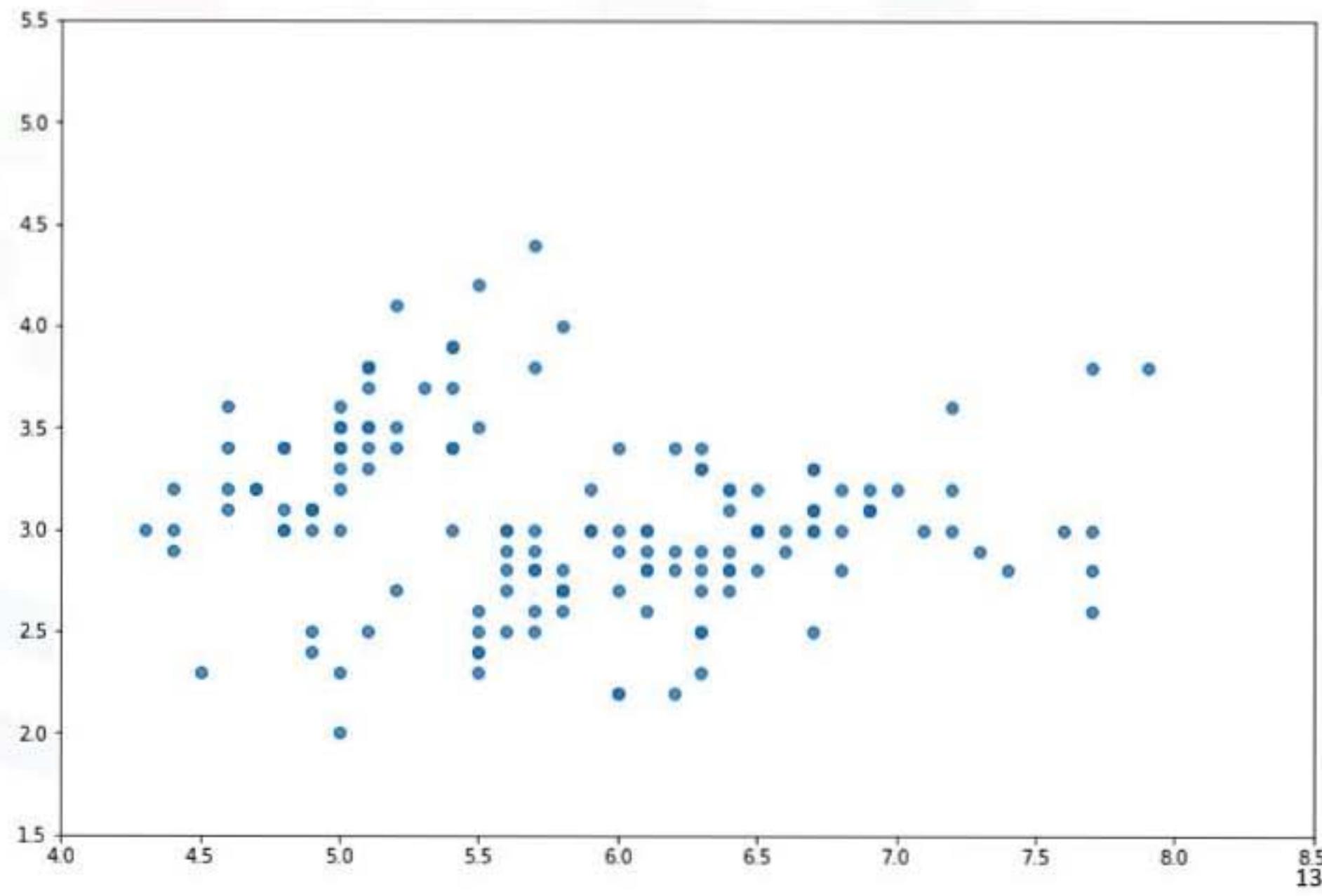
k-Means clustering – compute new centroids

4) Compute the new centroids for each cluster.



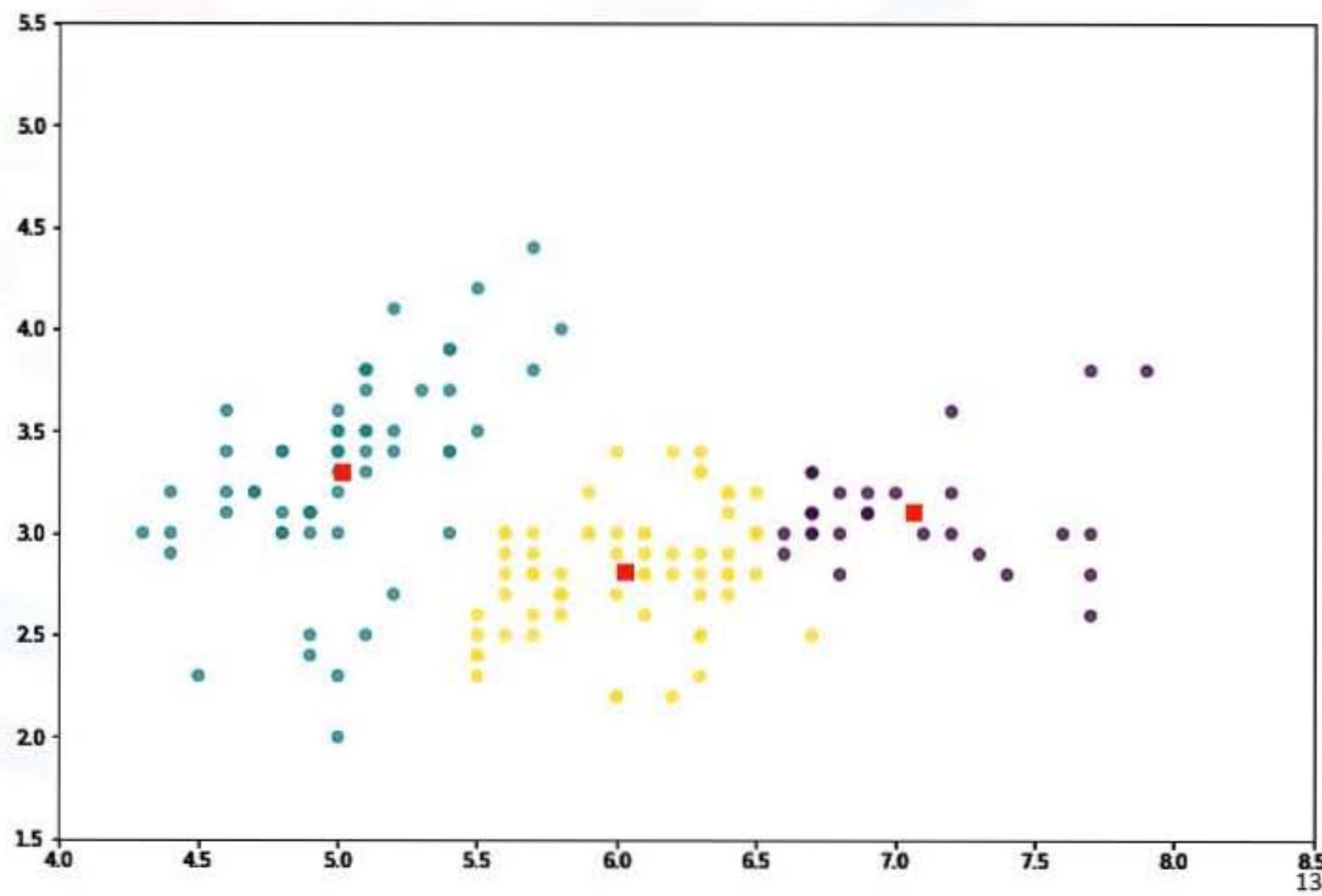
k-Means clustering – repeat

5) Repeat until there
are no more changes.



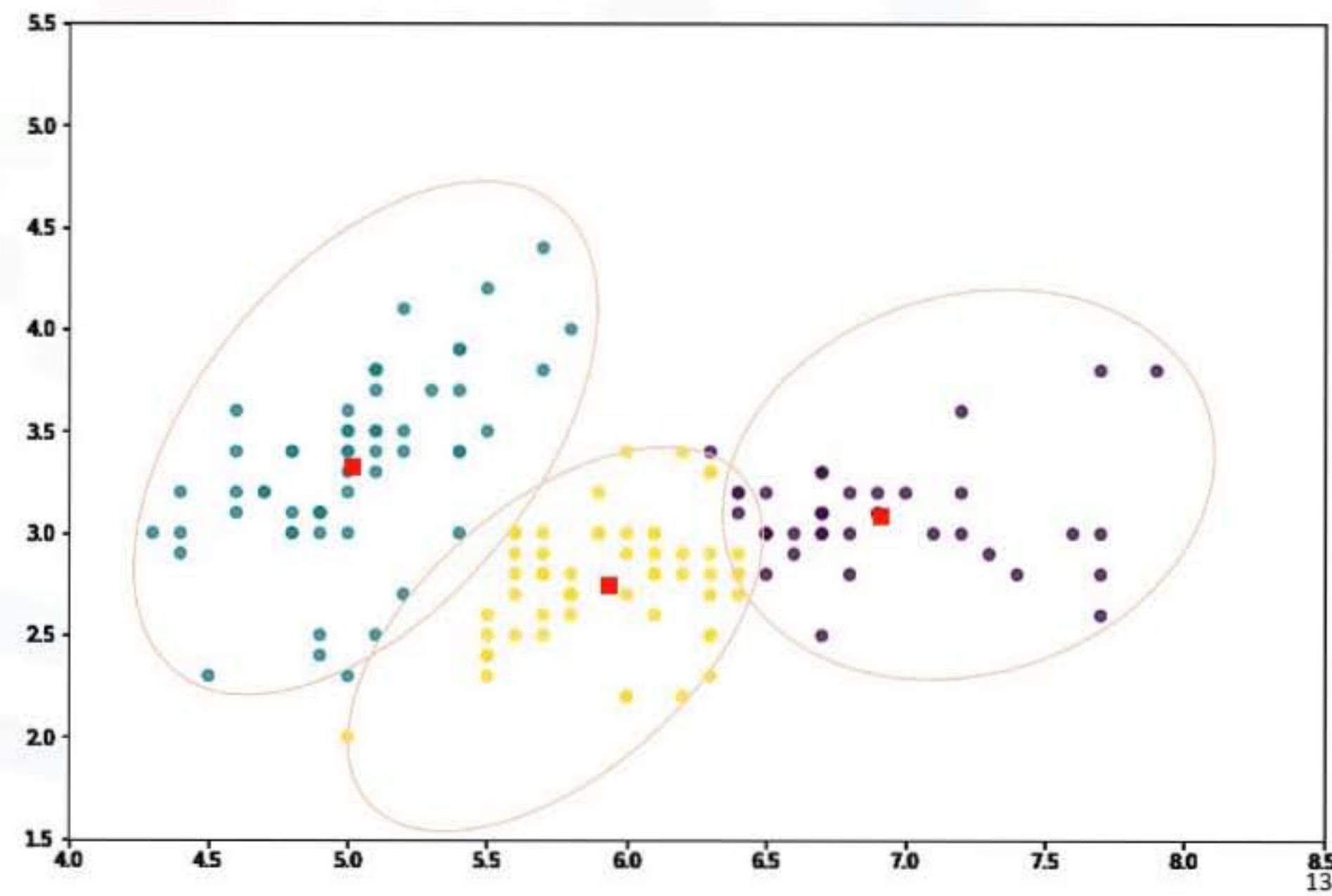
k-Means clustering – repeat

5) Repeat until there
are no more changes.



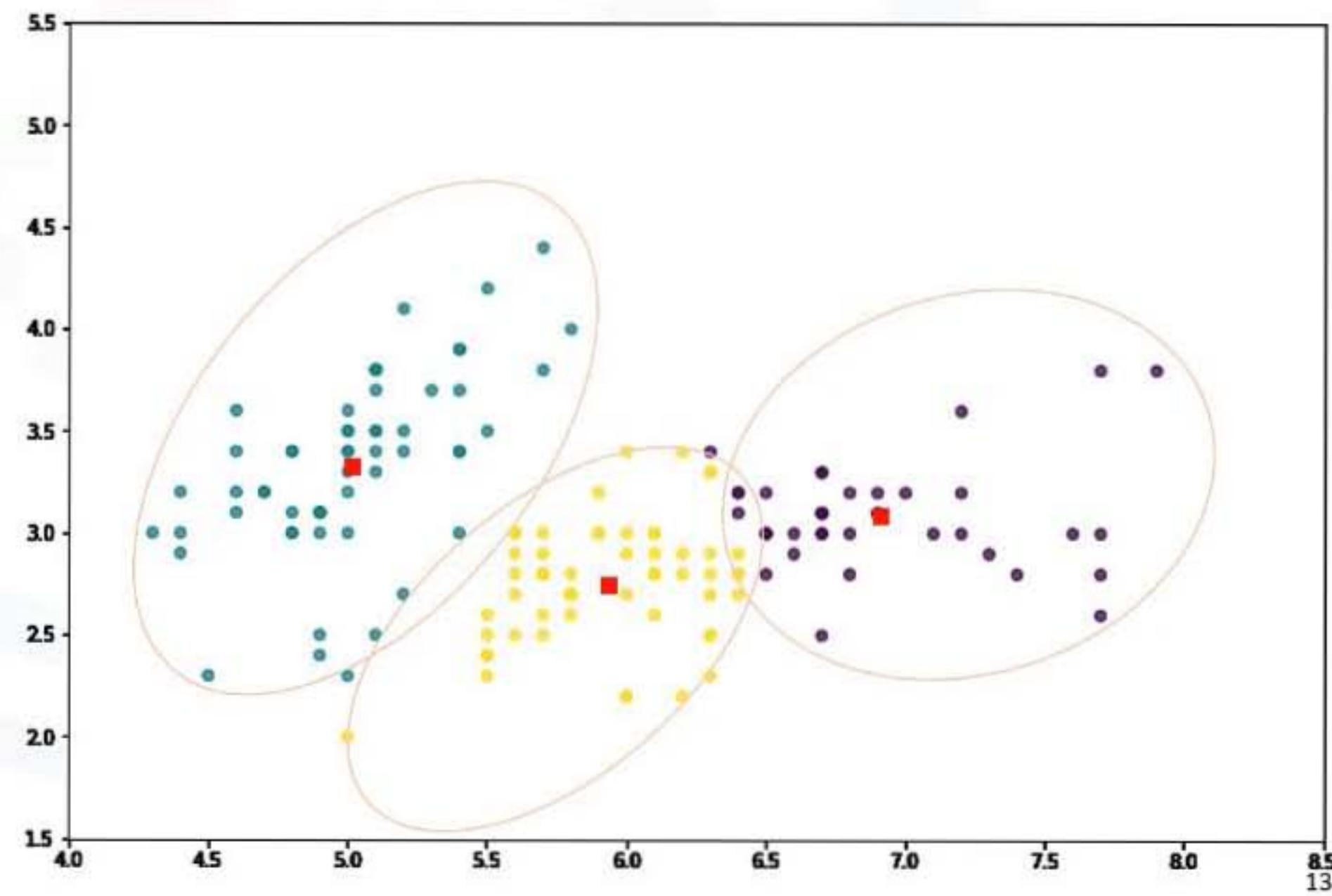
k-Means clustering – repeat

5) Repeat until there
are no more changes.



k-Means clustering – repeat

5) Repeat until there
are no more changes.



More on k-Means

Saeed Aghabozorgi

k-Means clustering algorithm

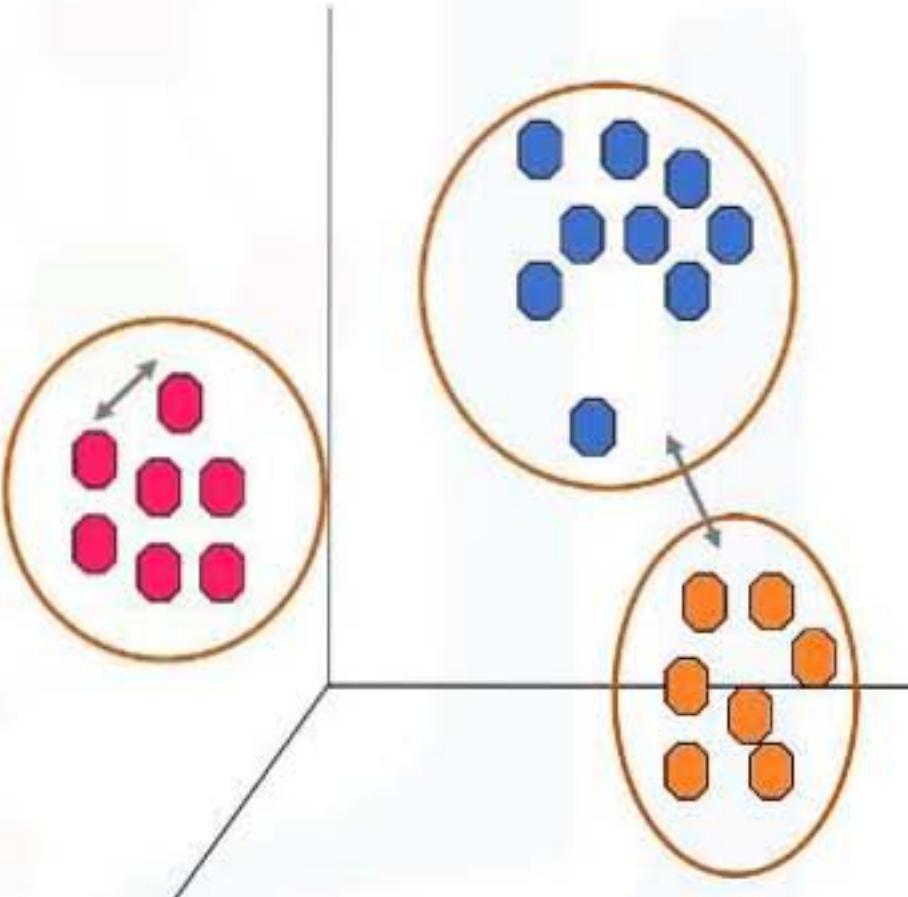
1. Randomly placing k centroids, one for each cluster.
2. Calculate the distance of each point from each centroid.
3. Assign each data point (object) to its closest centroid, creating a cluster.
4. Recalculate the position of the k centroids.

k-Means clustering algorithm

1. Randomly placing k centroids, one for each cluster.
2. Calculate the distance of each point from each centroid.
3. Assign each data point (object) to its closest centroid, creating a cluster.
4. Recalculate the position of the k centroids.
5. Repeat the steps 2-4, until the centroids no longer move.

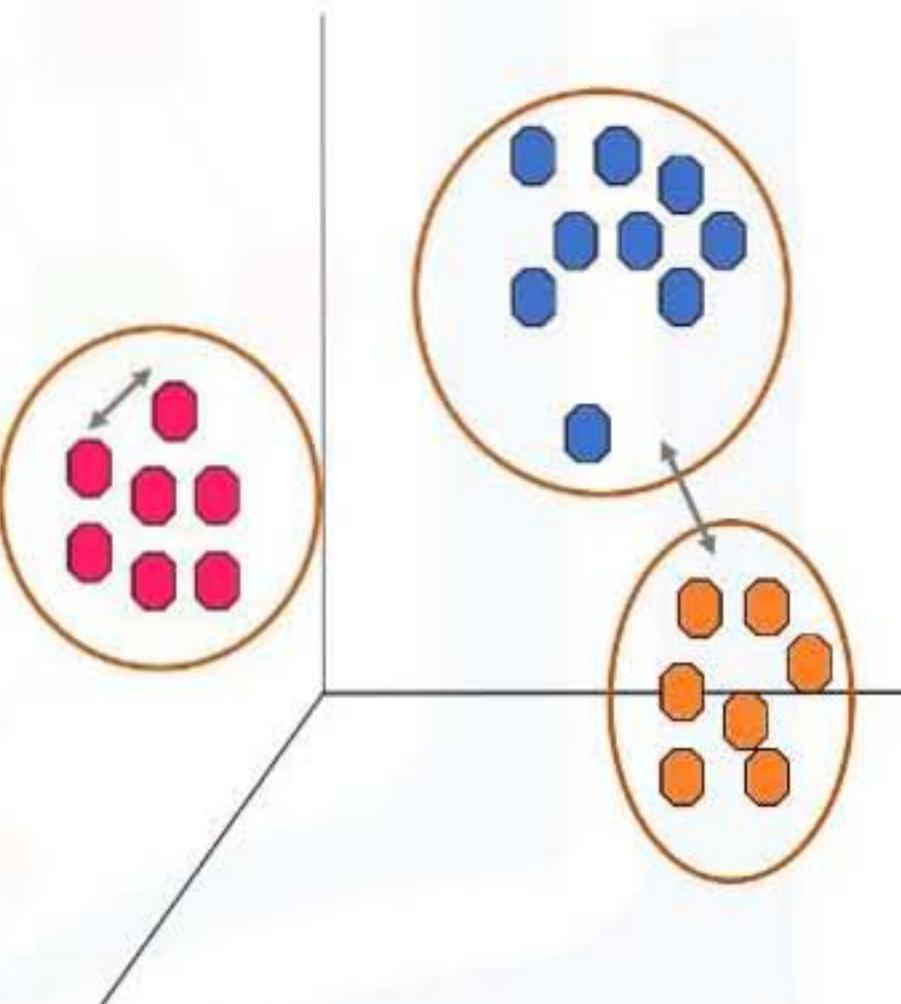
k-Means accuracy

- External approach
- Internal approach

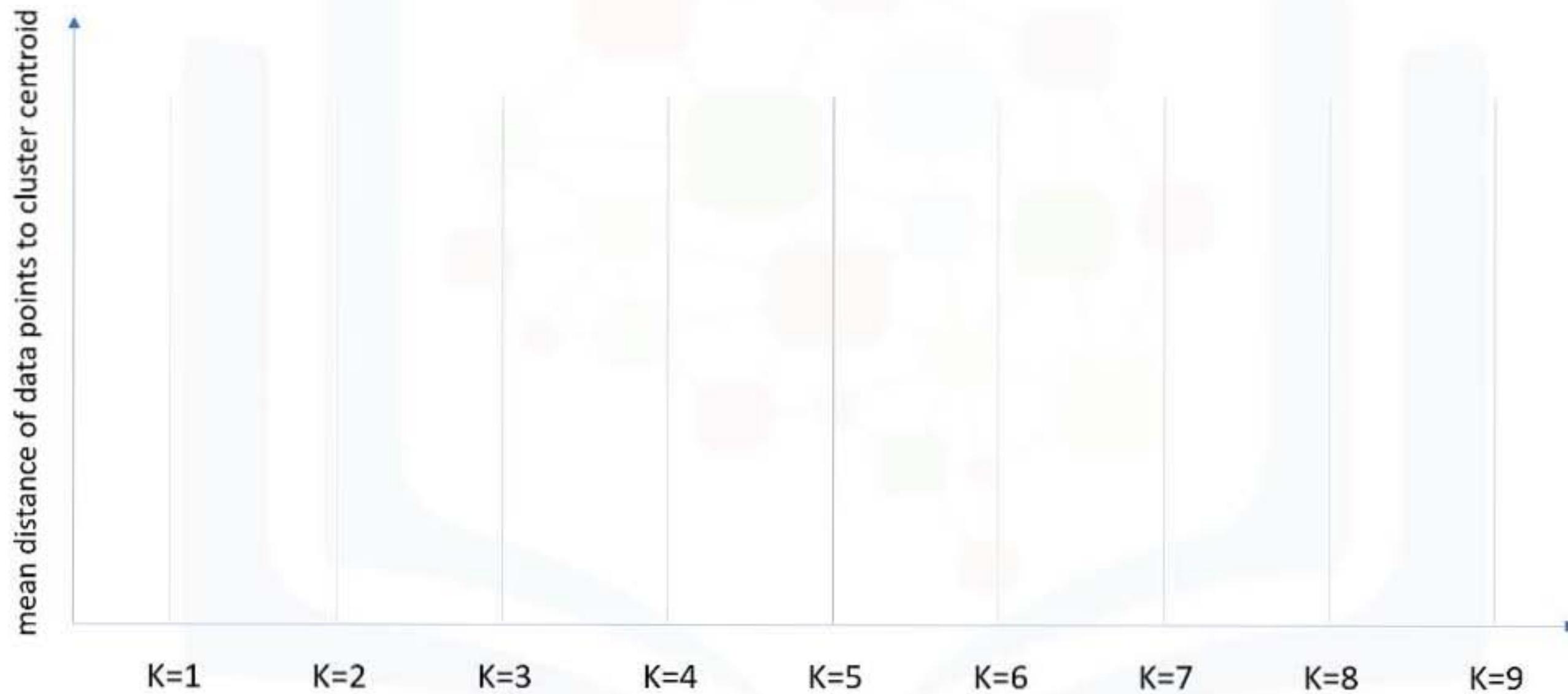


k-Means accuracy

- External approach
 - Compare the clusters with the ground truth, if it is available.
- Internal approach
 - Average the distance between data points within a cluster.



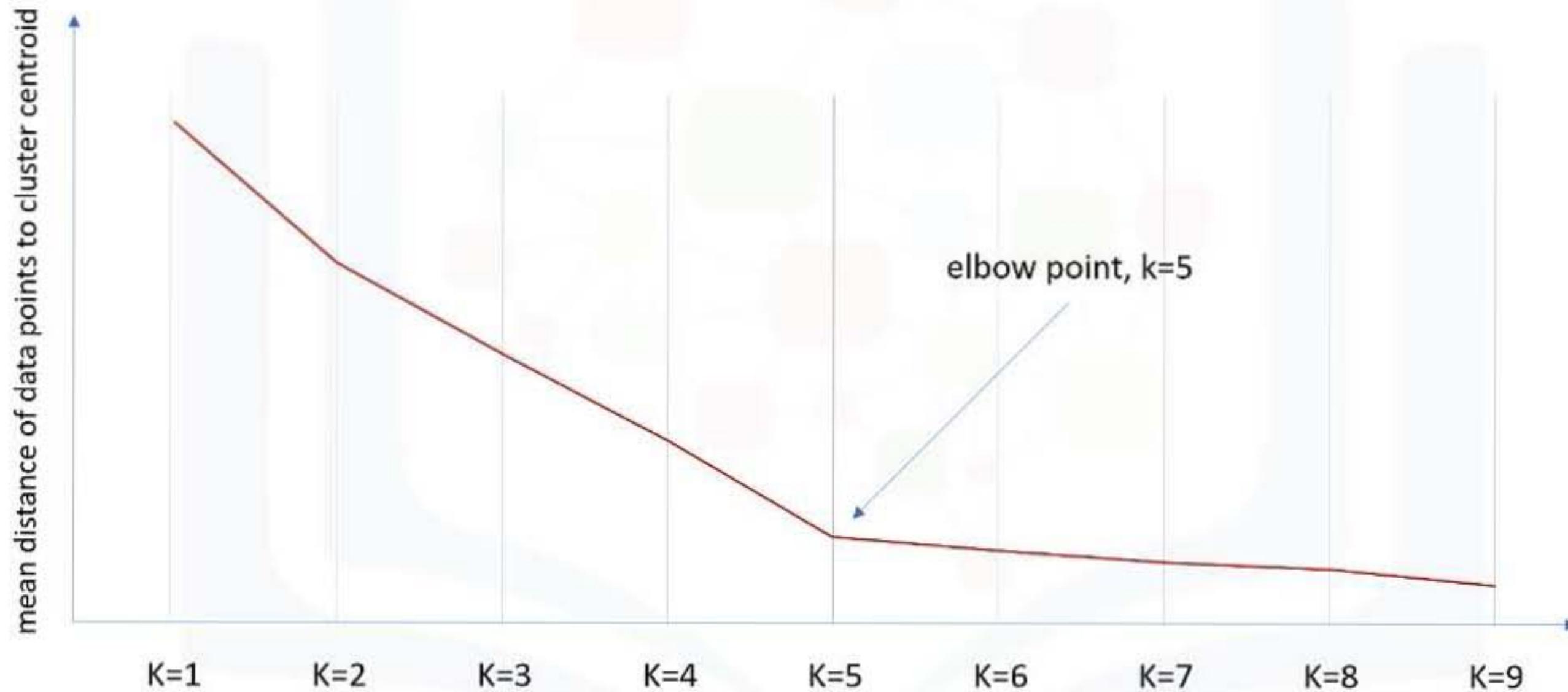
Choosing k



Choosing k



Choosing k



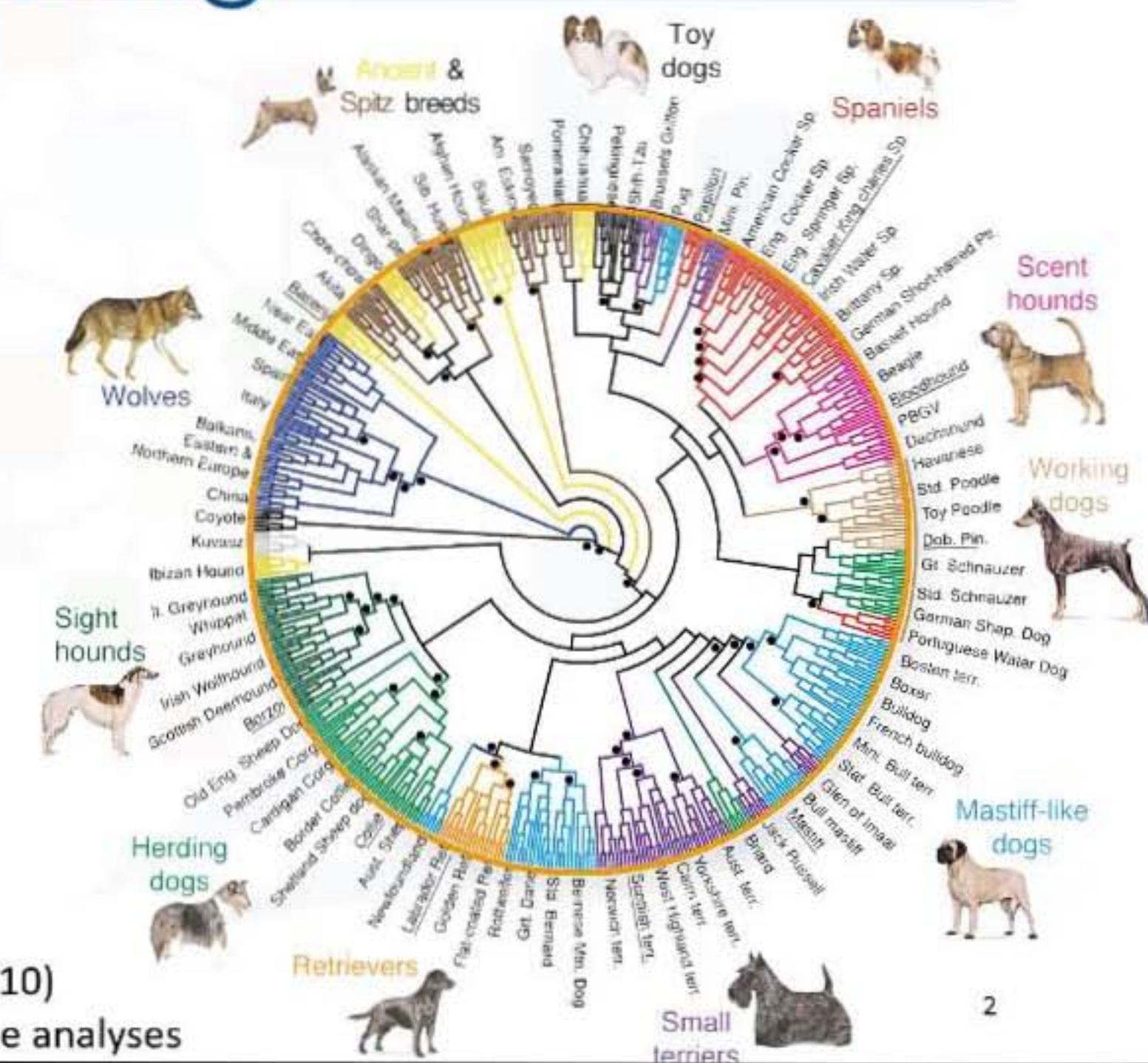
k-Means recap

- Med and Large sized databases (*Relatively efficient*)
- Produces sphere-like clusters
- Needs number of clusters (k)

Hierarchical Clustering

Saeed Aghabozorgi

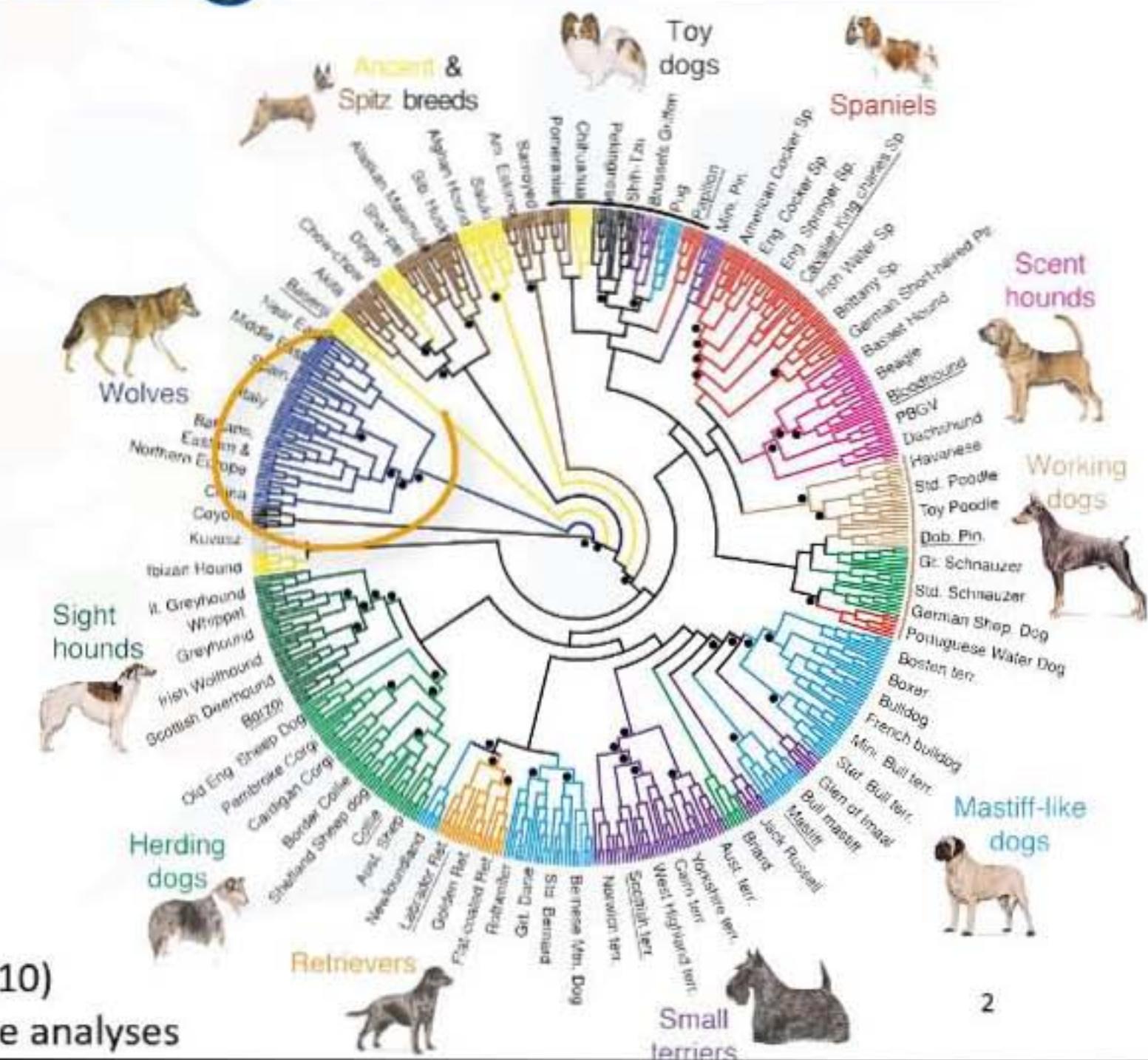
Hierarchical clustering



Source: von Holdt B.M. et al. (2010)
Genome-wide SNP and haplotype analyses

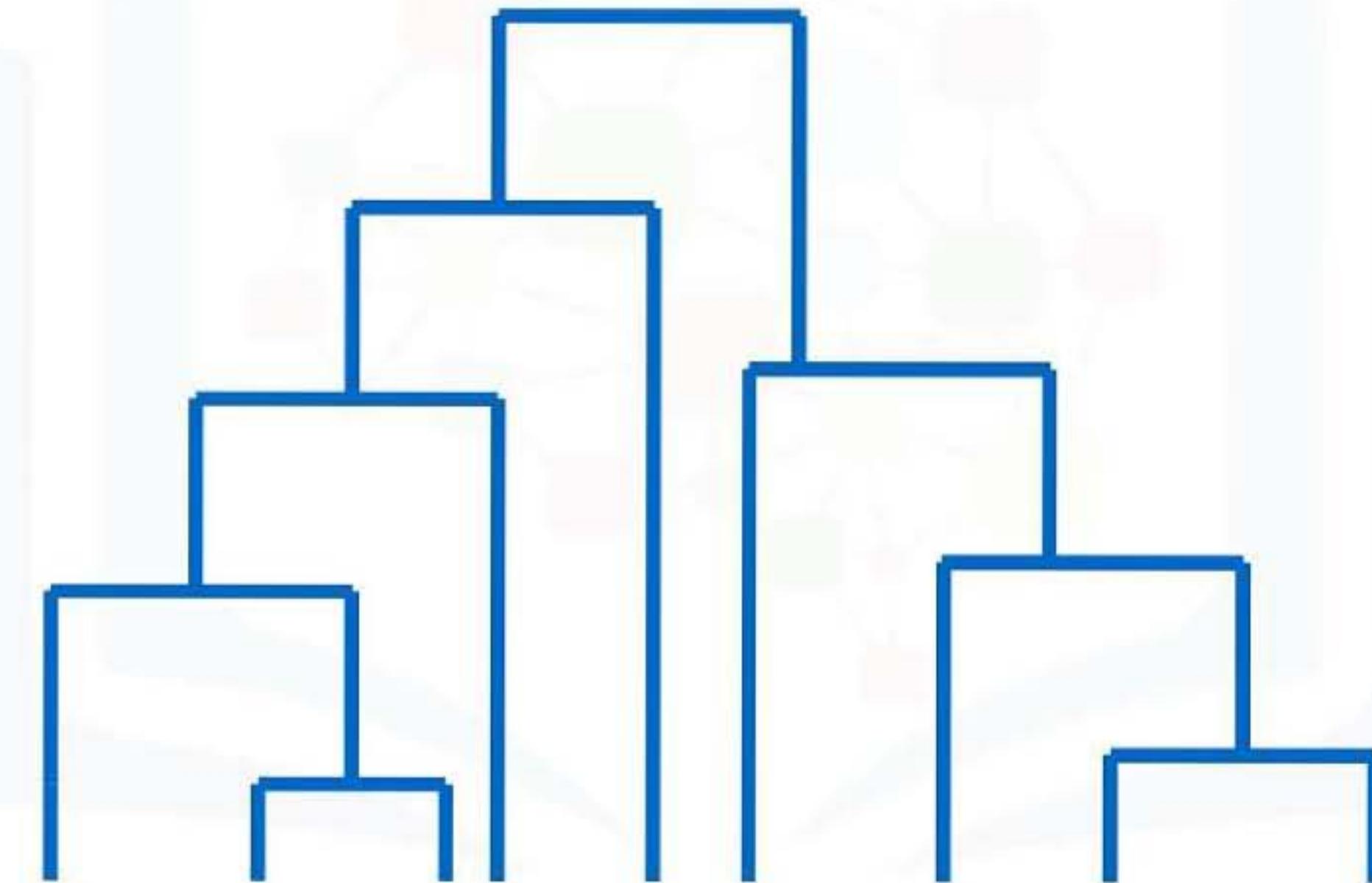
Hierarchical clustering

Hierarchical clustering algorithms build a hierarchy of clusters where each node is a cluster consists of the clusters of its daughter nodes.

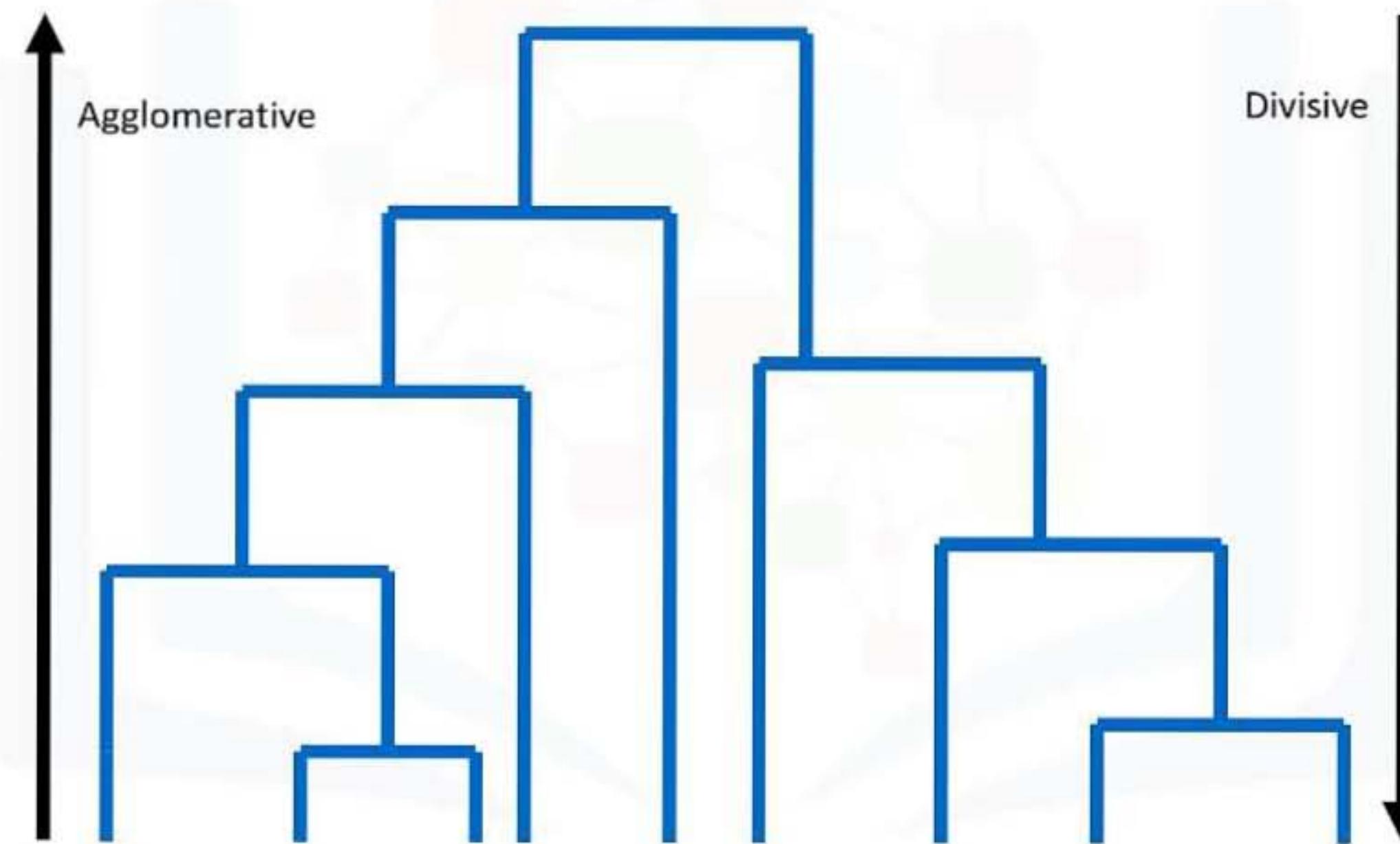


Source: von Holdt B.M. et al. (2010)
Genome-wide SNP and haplotype analyses

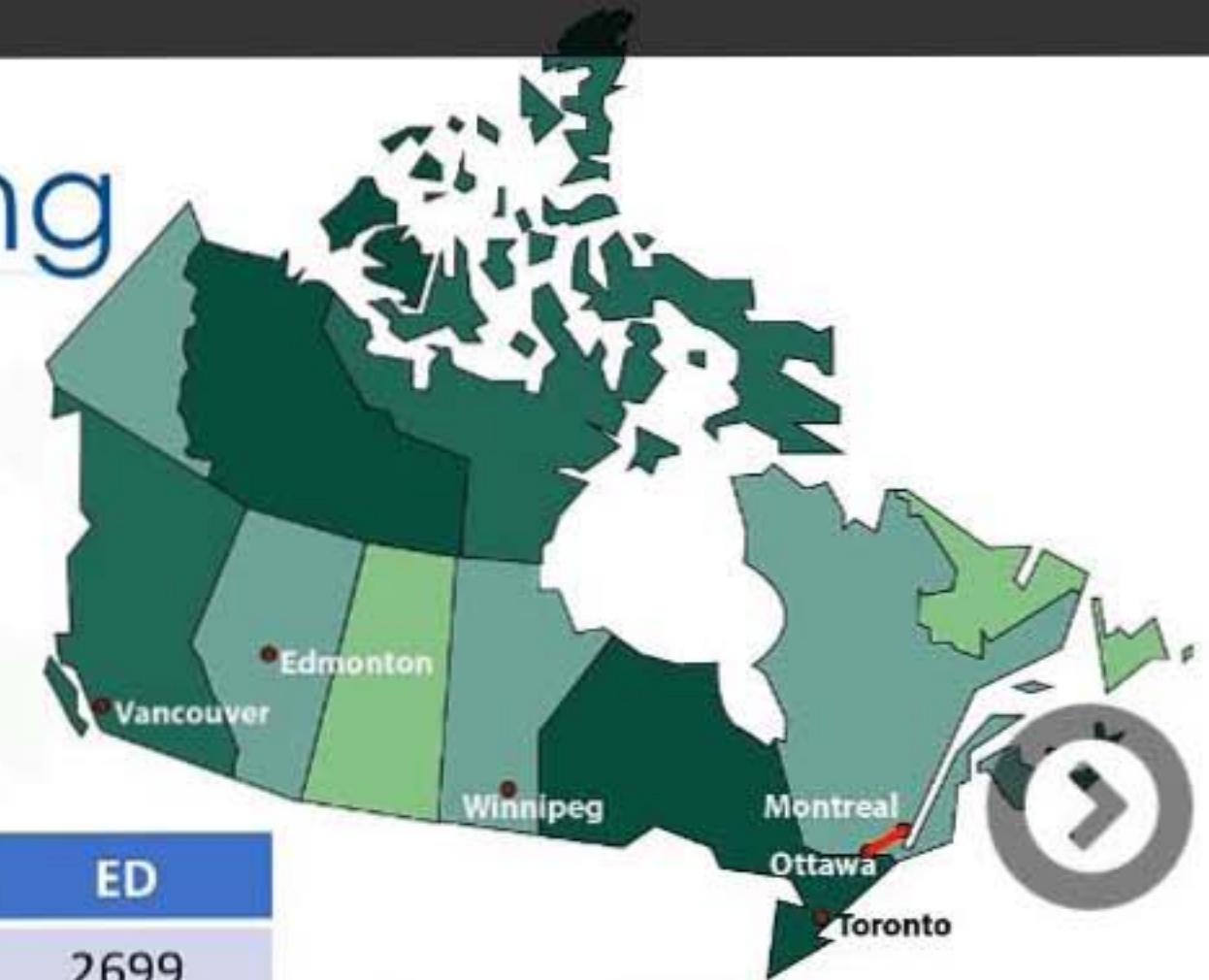
Hierarchical clustering



Hierarchical clustering



Agglomerative clustering



TO OT MO VA ED WI

	TO	OT	VA	MO	WI	ED
TO		351	3363	505	1510	2699
OT			3543	167	1676	2840
VA				3690	1867	819
MO					1824	2976
WI						1195
ED						



Agglomerative clustering



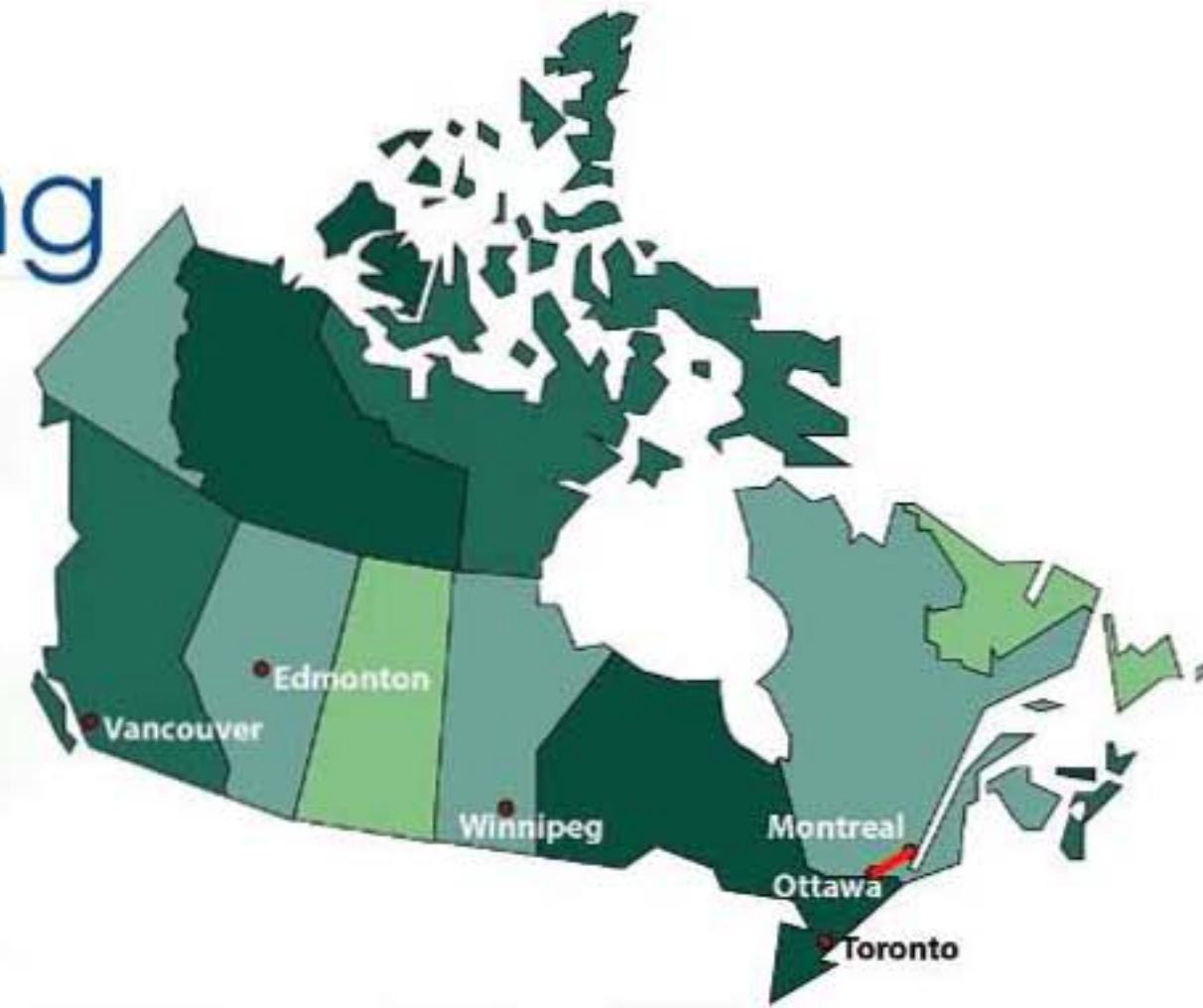
	TO	OT	VA	MO	WI	ED
TO		351	3363	505	1510	2699
OT			3543	167	1676	2840
VA				3690	1867	819
MO					1824	2976
WI						1195
ED						



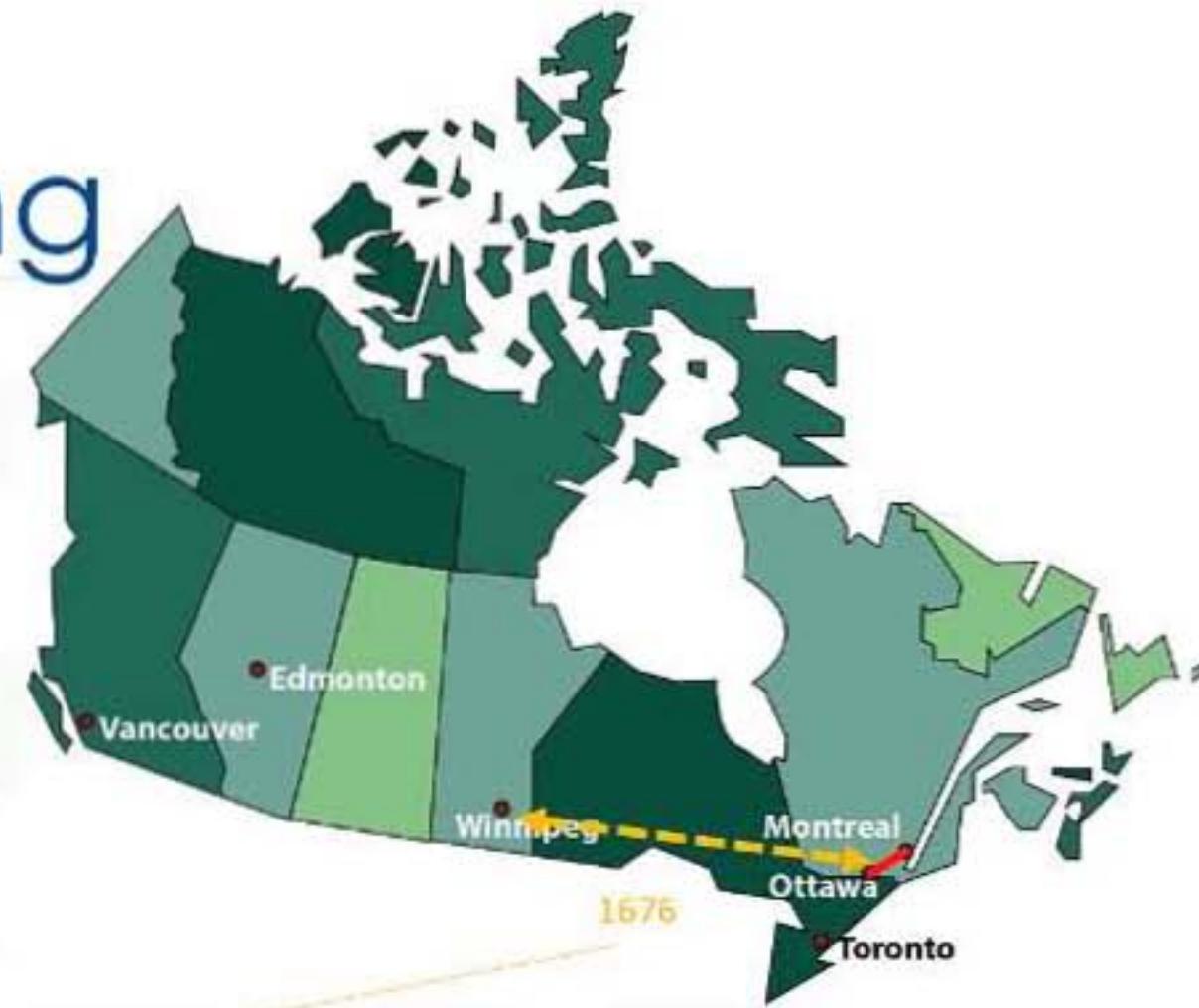
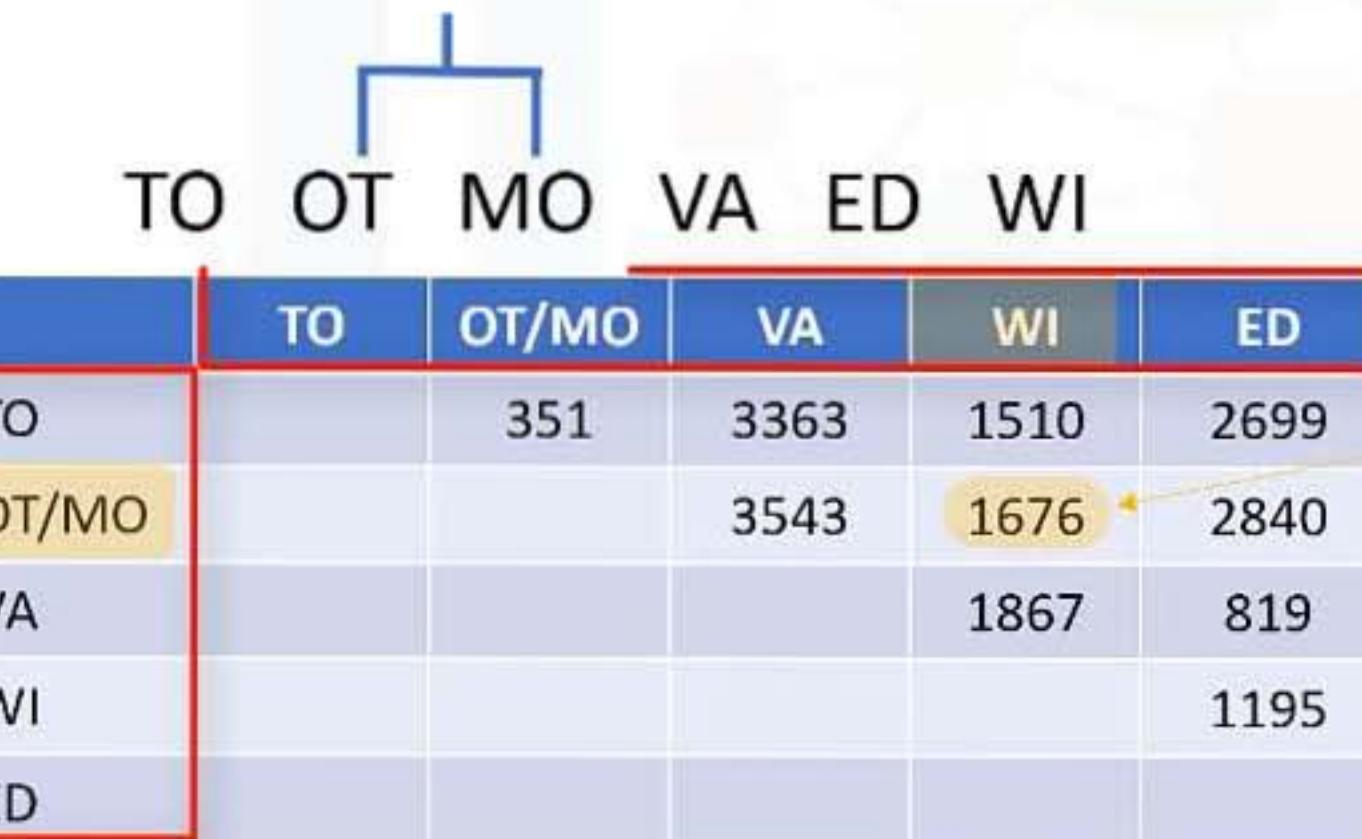
Agglomerative clustering



	TO	OT/MO	VA	WI	ED
TO		351	3363	1510	2699
OT/MO			3543	1676	2840
VA				1867	819
WI					1195
ED					



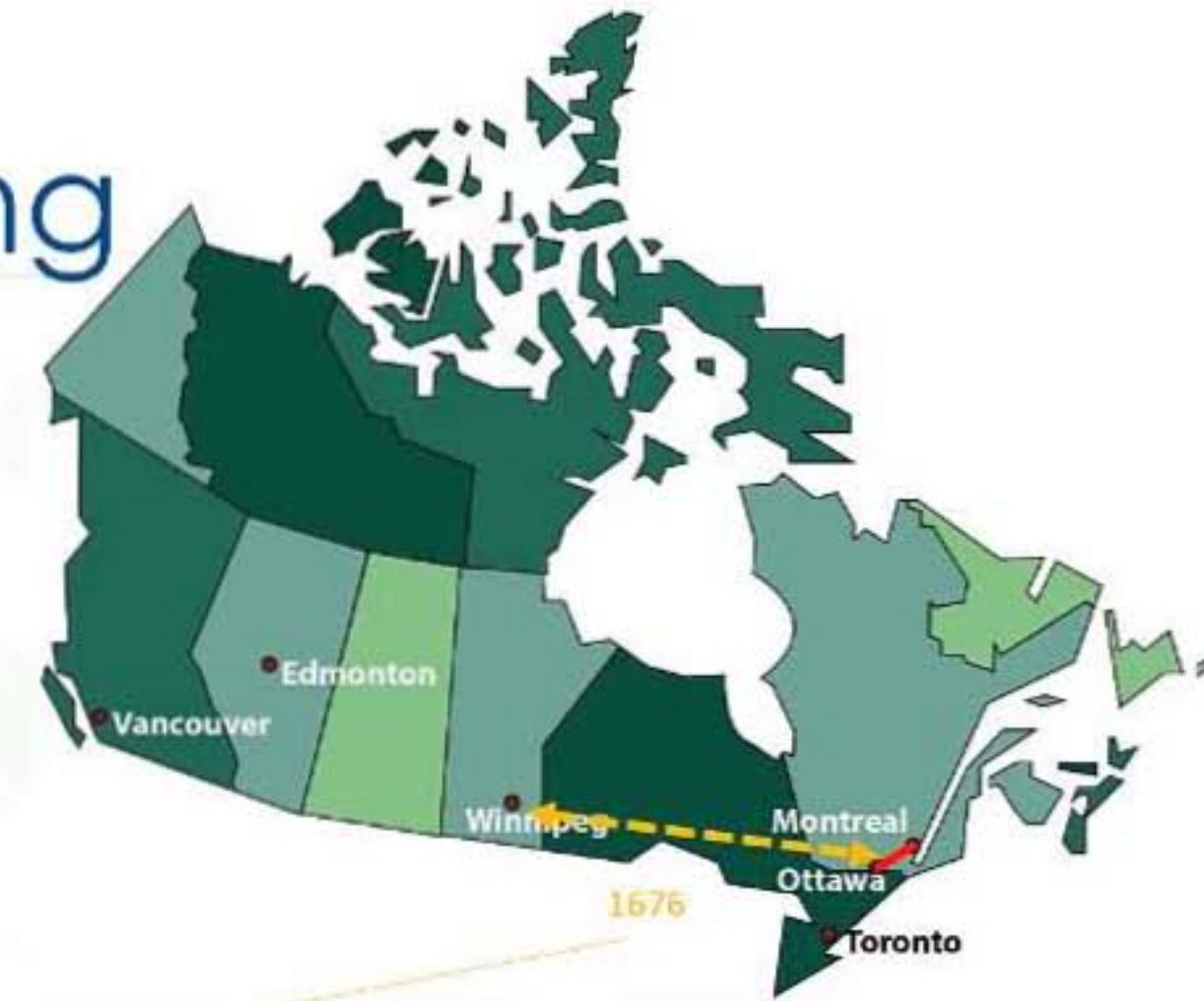
Agglomerative clustering



Agglomerative clustering



	TO	OT/MO	VA	WI	ED
TO		351	3363	1510	2699
OT/MO			3543	1676	2840
VA				1867	819
WI					1195
ED					



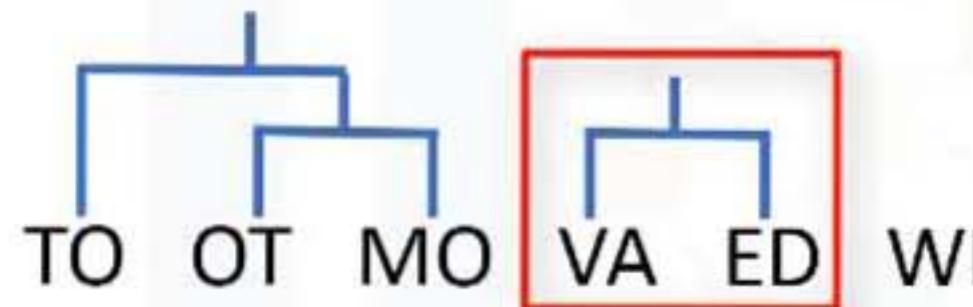
Agglomerative clustering



	TO/OT/MO	VA	WI	ED
TO/OT/MO		3543	1676	2840
VA			1867	819
WI				1195
ED				



Agglomerative clustering



	TO/OT/MO	VA	WI	ED
TO/OT/MO		3543	1676	2840
VA			1867	819
WI				1195
ED				



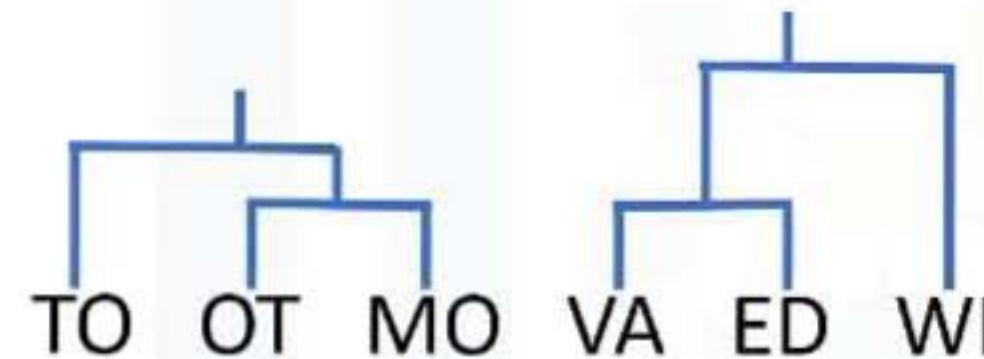
Agglomerative clustering



	TO/OT/MO	VA	WI	ED
TO/OT/MO		3543	1676	2840
VA			1867	819
WI				1195
ED				



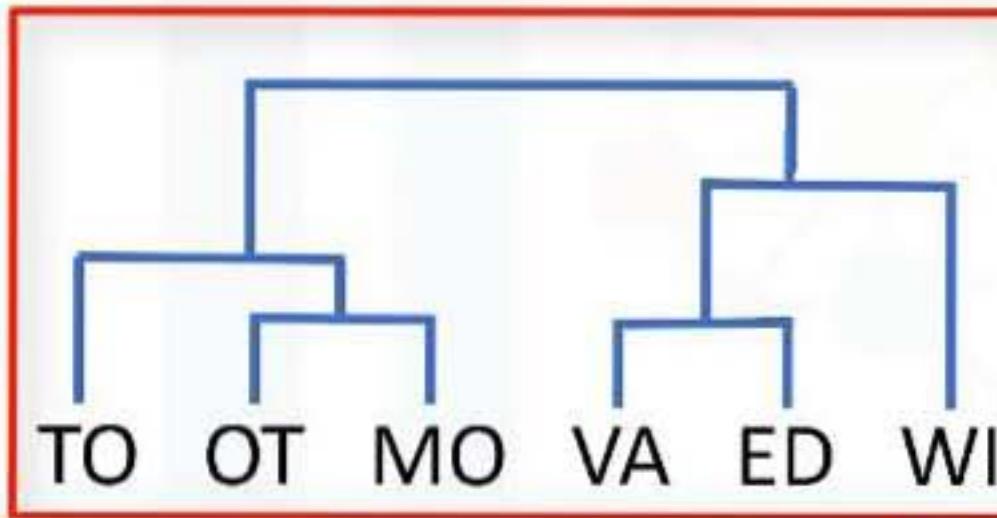
Agglomerative clustering



	TO/OT/MO	VA/ED	WI
TO/OT/MO		2840	1676
VA/ED			1667
WI			

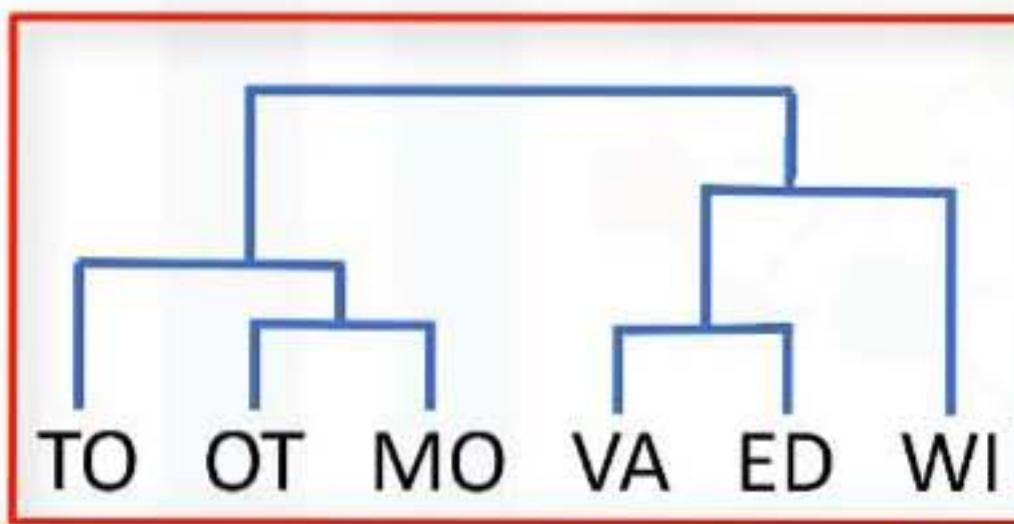


Hierarchical clustering



	TO/OT/MO	VA/ED/WI
TO/OT/MO		1676
VA/ED/WI		

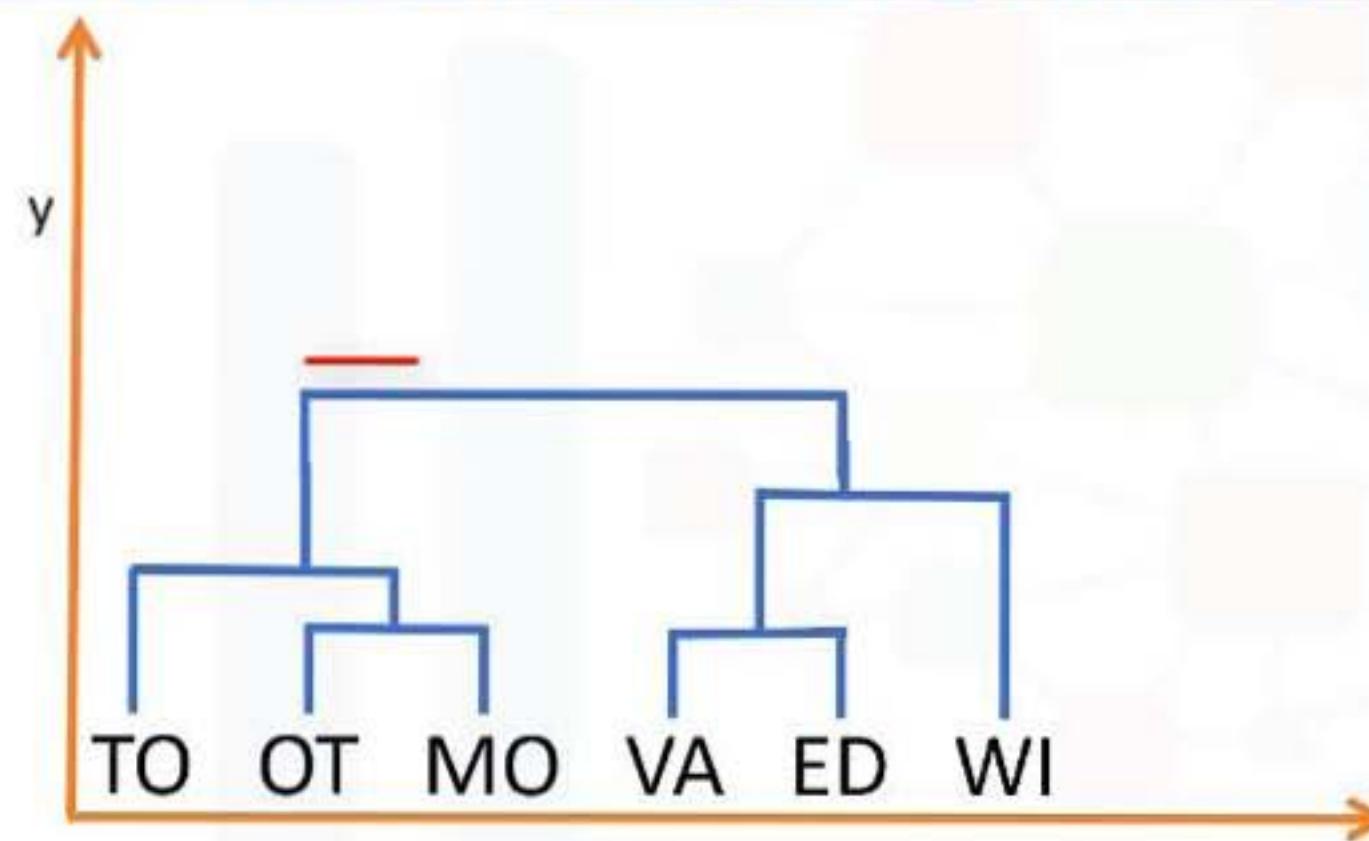
Hierarchical clustering



	TO/OT/MO	VA/ED/WI
TO/OT/MO		1676
VA/ED/WI		



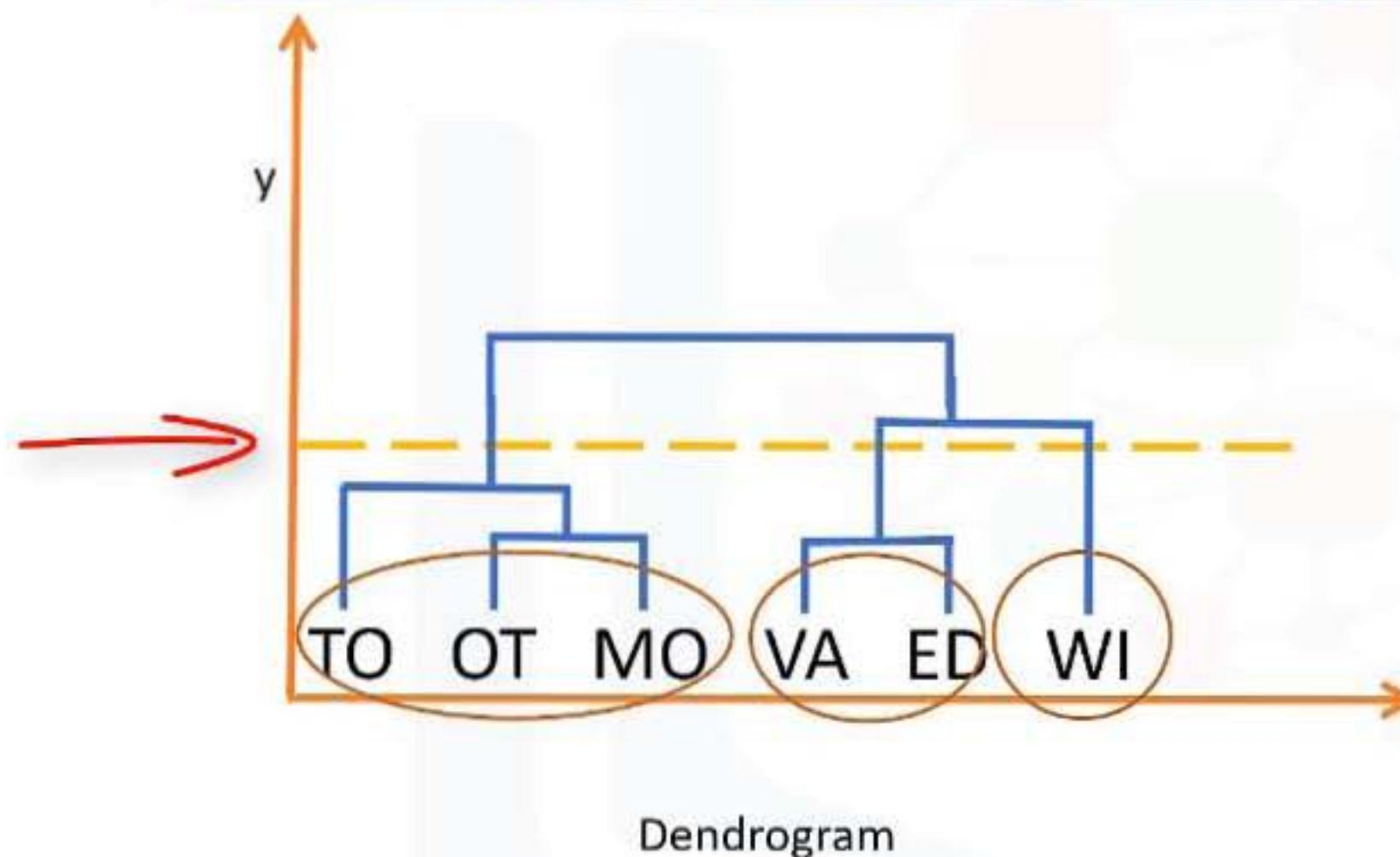
Hierarchical clustering



Dendrogram



Hierarchical clustering

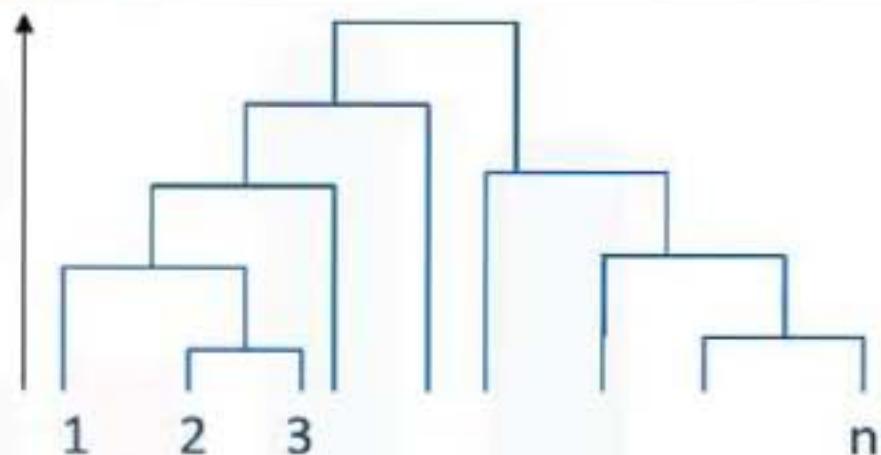


More on Hierarchical Clustering

Saeed Aghabozorgi

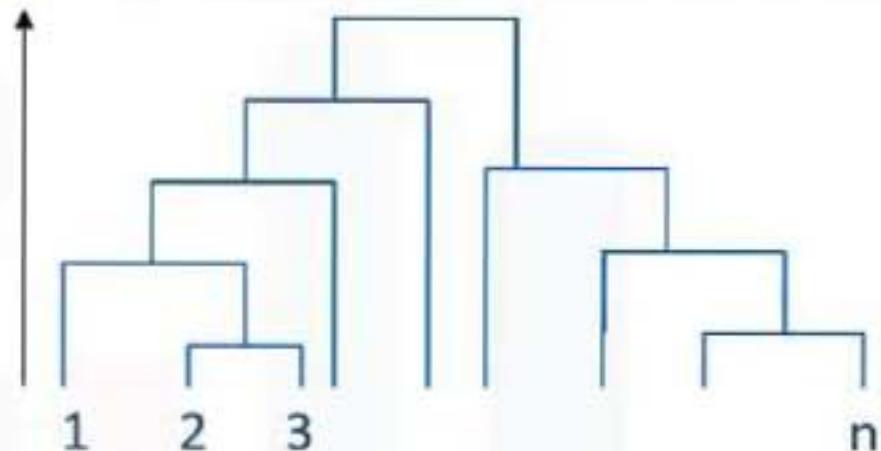


Agglomerative algorithm



Agglomerative algorithm

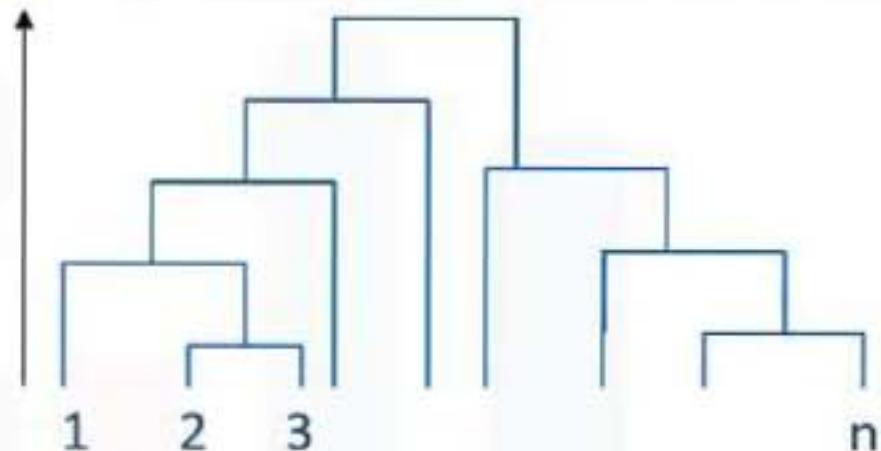
1. Create n clusters, one for each data point
2. Compute the Proximity Matrix
- 3. Repeat**
 - i. Merge the two closest clusters
 - ii. Update the proximity matrix



$$\begin{bmatrix} 0 \\ d(2,1) & 0 \\ d(3,1) & d(3,2) & 0 \\ \vdots & \vdots & \vdots \\ d(n,1) & d(n,2) & \dots & \dots & 0 \end{bmatrix}$$

Agglomerative algorithm

1. Create n clusters, one for each data point
2. Compute the Proximity Matrix
3. Repeat
 - i. Merge the two closest clusters
 - ii. Update the proximity matrix
4. Until only a single cluster remains



$$\begin{bmatrix} 0 \\ d(2,1) & 0 \\ d(3,1) & d(3,2) & 0 \\ \vdots & \vdots & \vdots \\ d(n,1) & d(n,2) & \dots & \dots & 0 \end{bmatrix}$$

Similarity/Distance



Patient 1		
Age	BMI	BP
54	190	120

Patient 2		
Age	BMI	BP
50	200	125

Dis (p1,p2)

$$\begin{aligned} &= \sqrt{\sum_{i=0}^n (x_i - y_i)^2} \\ &= \sqrt{(54 - 50)^2 + (190 - 200)^2 + (120 - 125)^2} \\ &= 11.87 \end{aligned}$$

Similarity/Distance

Dataset



x_{11}	...	x_{1f}	...	x_{1p}
...
x_{i1}	...	x_{if}	...	x_{ip}
...
x_{n1}	...	x_{nf}	...	x_{np}



Dissimilarity matrix

0				
$d(2,1)$	0			
$d(3,1)$	$d(3,2)$	0		
:	:	:		
$d(n,1)$	$d(n,2)$	0

Similarity/Distance

Dataset

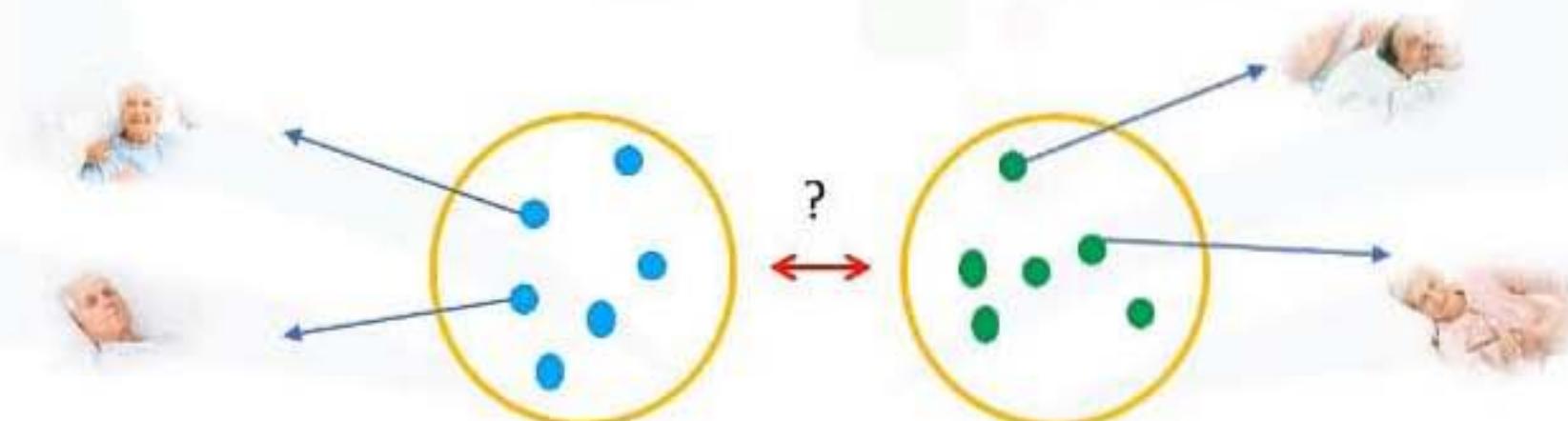


$$\begin{bmatrix} x_{11} & \dots & x_{1f} & \dots & x_{1p} \\ \dots & \dots & \dots & \dots & \dots \\ x_{i1} & \dots & x_{if} & \dots & x_{ip} \\ \dots & \dots & \dots & \dots & \dots \\ x_{n1} & \dots & x_{nf} & \dots & x_{np} \end{bmatrix}$$



Dissimilarity matrix

$$\begin{bmatrix} 0 & & & & \\ d(2,1) & 0 & & & \\ d(3,1) & d(3,2) & 0 & & \\ \vdots & \vdots & \vdots & \ddots & \\ d(n,1) & d(n,2) & \dots & \dots & 0 \end{bmatrix}$$

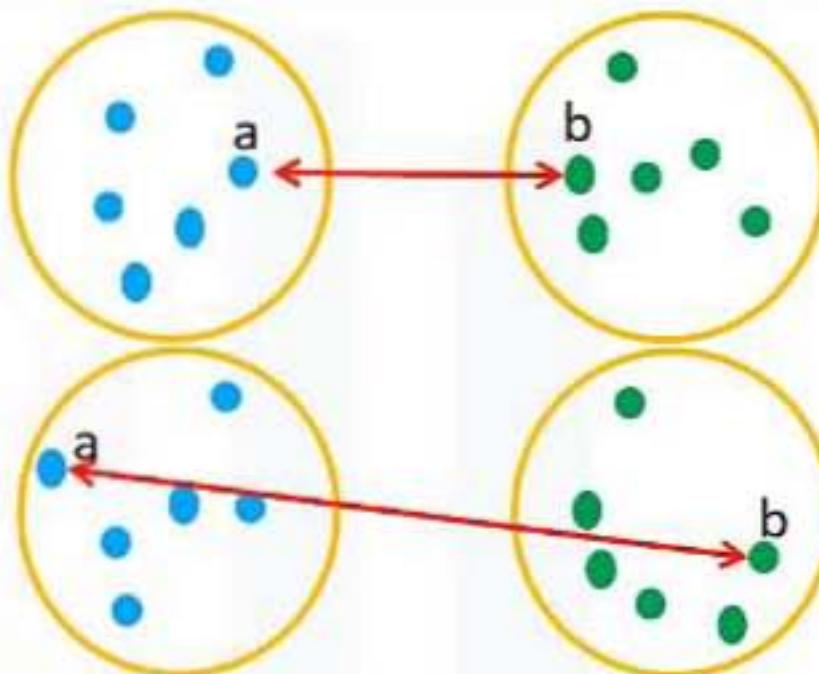


Distance between clusters

- Single-Linkage Clustering
 - Minimum distance between clusters
- Complete-Linkage Clustering
 - Maximum distance between clusters
- Average Linkage Clustering
 - Average distance between clusters
- Centroid Linkage Clustering
 - Distance between cluster centroids

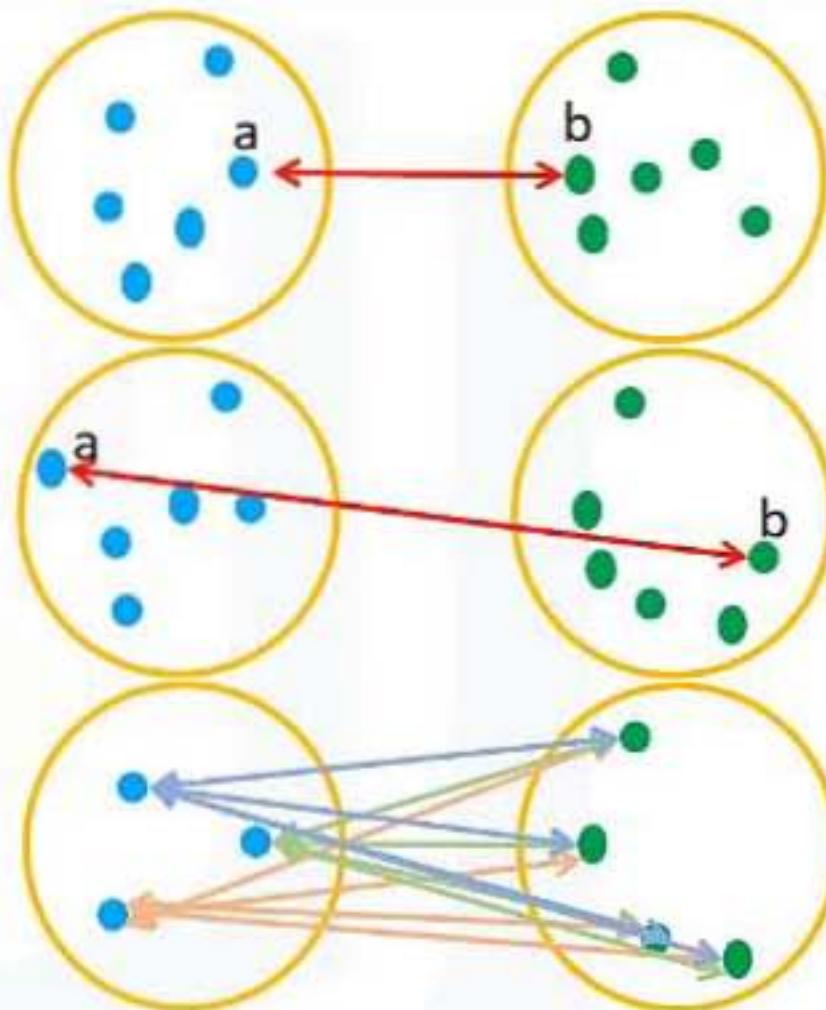
Distance between clusters

- Single-Linkage Clustering
 - Minimum distance between clusters
- ★ • Complete-Linkage Clustering
 - Maximum distance between clusters
- Average Linkage Clustering
 - Average distance between clusters
- Centroid Linkage Clustering
 - Distance between cluster centroids



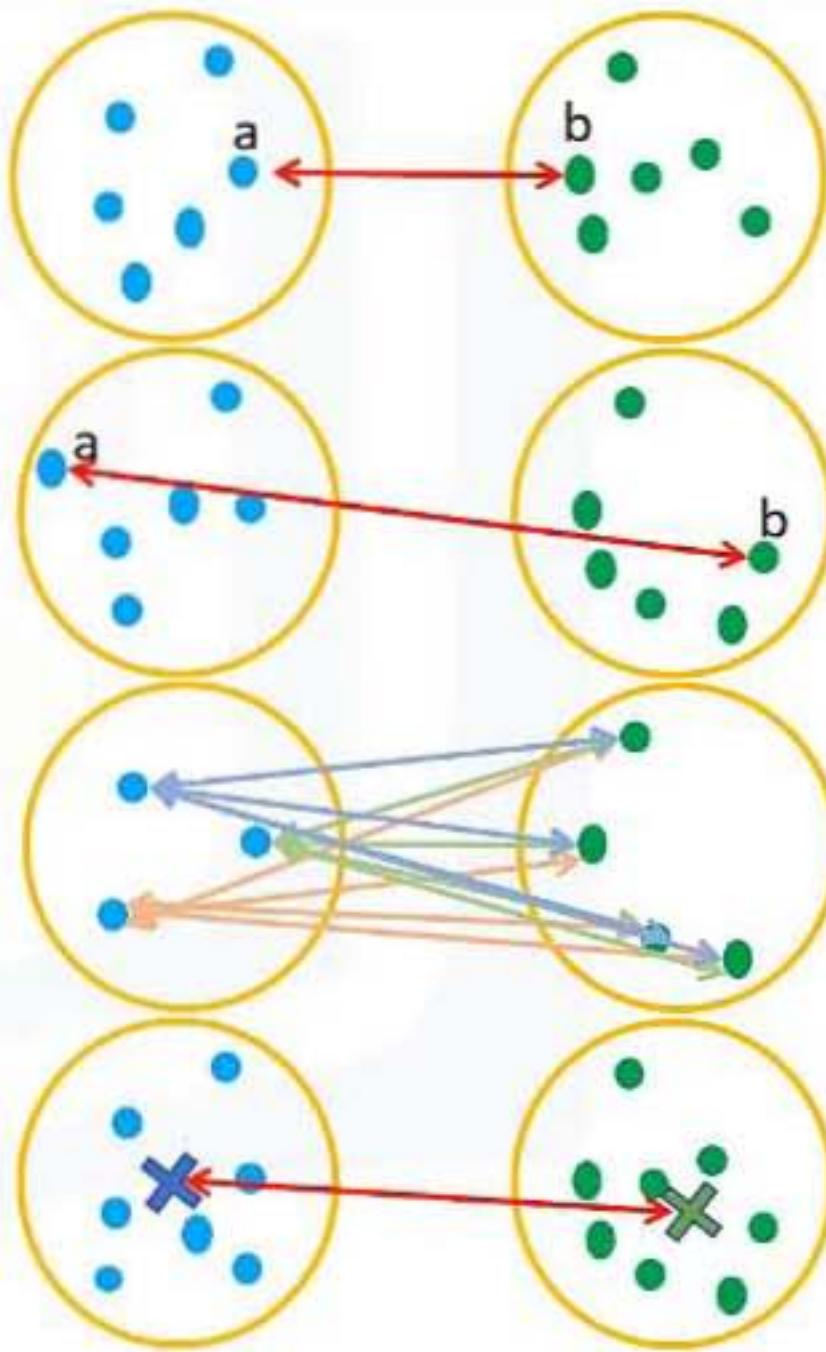
Distance between clusters

- Single-Linkage Clustering
 - Minimum distance between clusters
- Complete-Linkage Clustering
 - Maximum distance between clusters
- ★ • Average Linkage Clustering
 - Average distance between clusters
- Centroid Linkage Clustering
 - Distance between cluster centroids



Distance between clusters

- Single-Linkage Clustering
 - Minimum distance between clusters
- Complete-Linkage Clustering
 - Maximum distance between clusters
- Average Linkage Clustering
 - Average distance between clusters
- Centroid Linkage Clustering
 - Distance between cluster centroids



Advantages vs. disadvantages

Advantages	Disadvantages
Doesn't required number of clusters to be specified.	Can never undo any previous steps throughout the algorithm.
Easy to implement. Produces a dendrogram, which helps with understanding the data.	Generally has long runtimes. Sometimes difficult to identify the number of clusters by the dendrogram.

Hierarchical clustering Vs. K-means

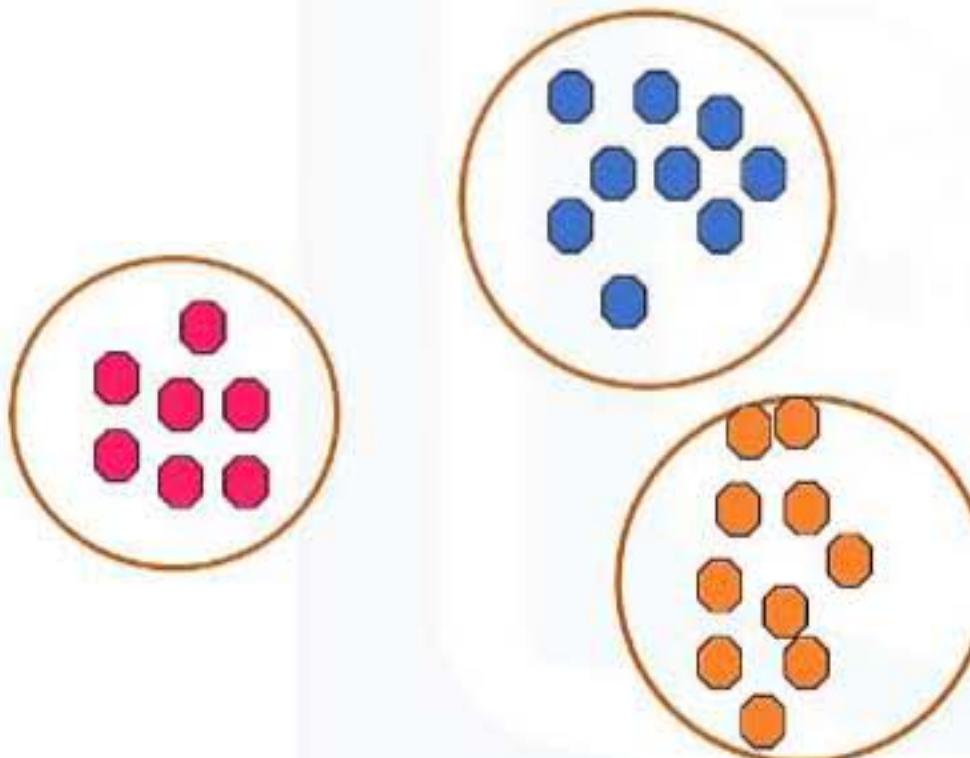
K-means	Hierarchical Clustering
1. Much more efficient	1. Can be slow for large datasets
2. Requires the number of clusters to be specified	2. Does not require the number of clusters to run
3. Gives only one partitioning of the data based on the predefined number of clusters	3. Gives more than one partitioning depending on the resolution
4. Potentially returns different clusters each time it is run due to random initialization of centroids	4. Always generates the same clusters

DBSCAN Clustering

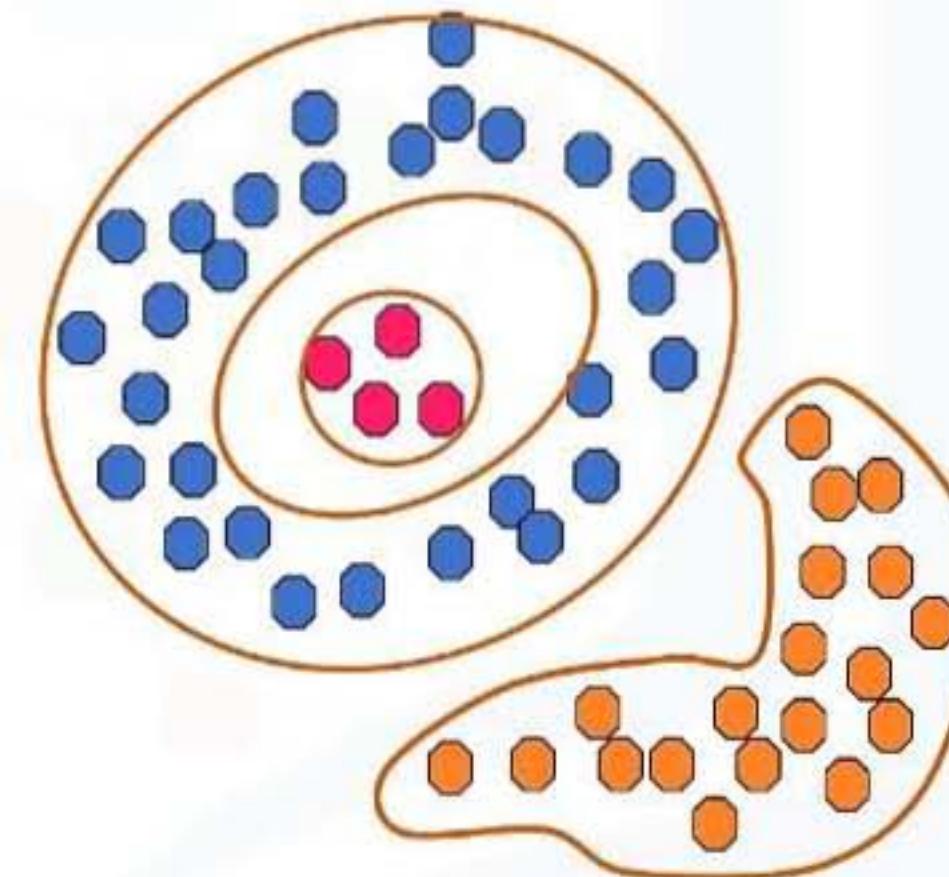
Saeed Aghabozorgi

Density-based clustering

- Spherical-shape clusters

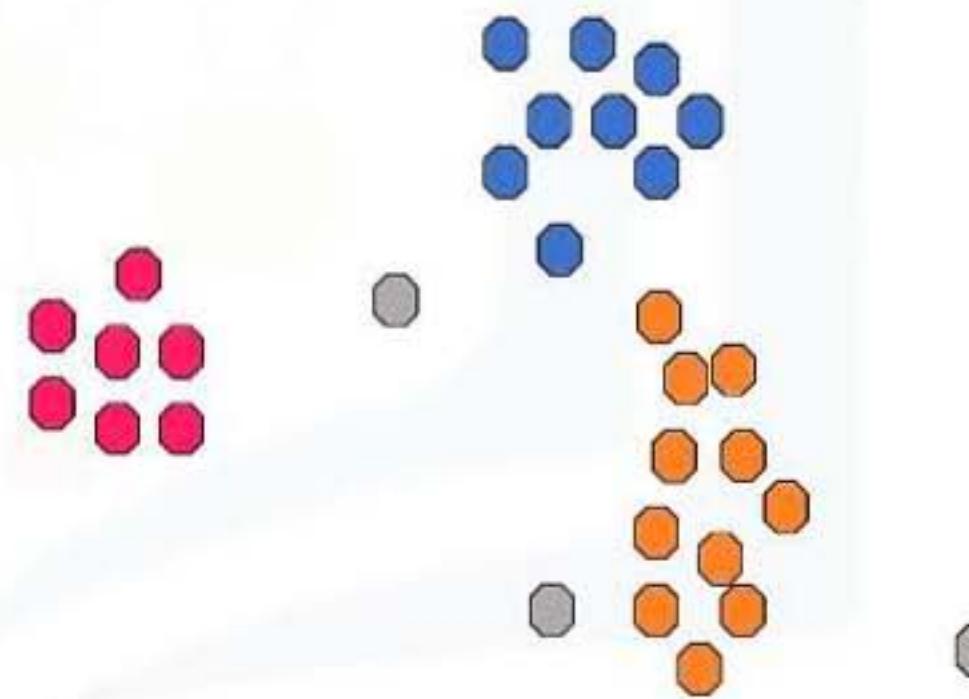
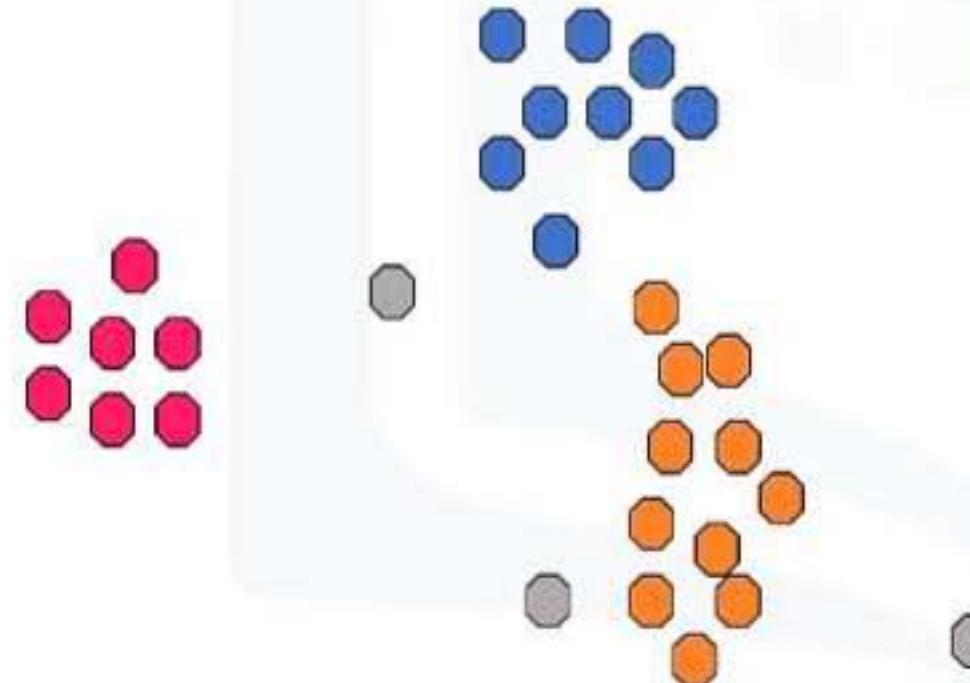


- Arbitrary-shape clusters



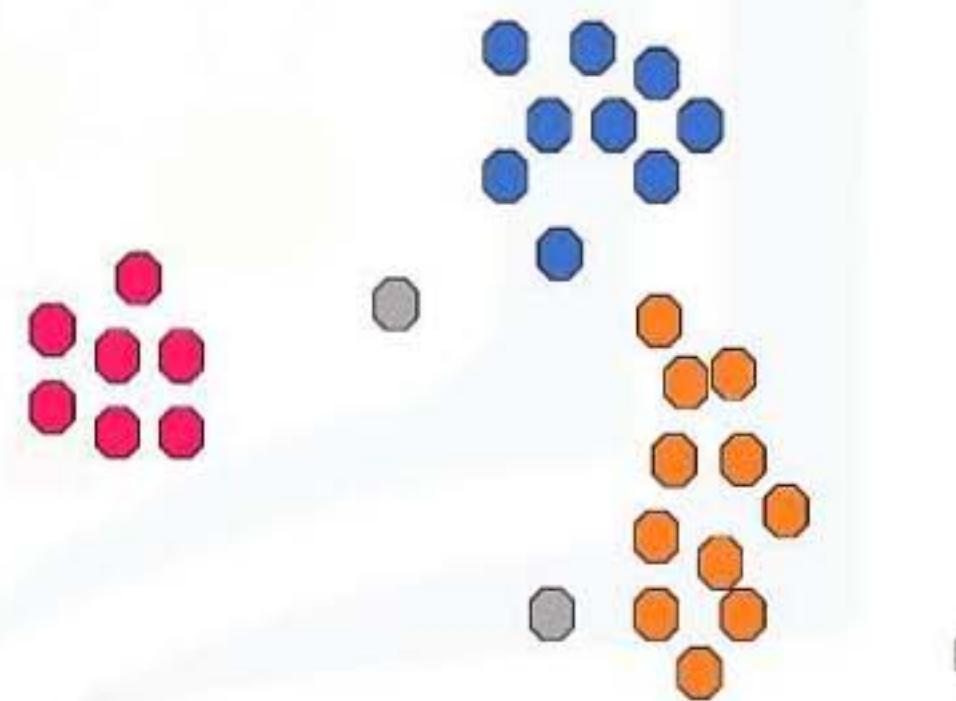
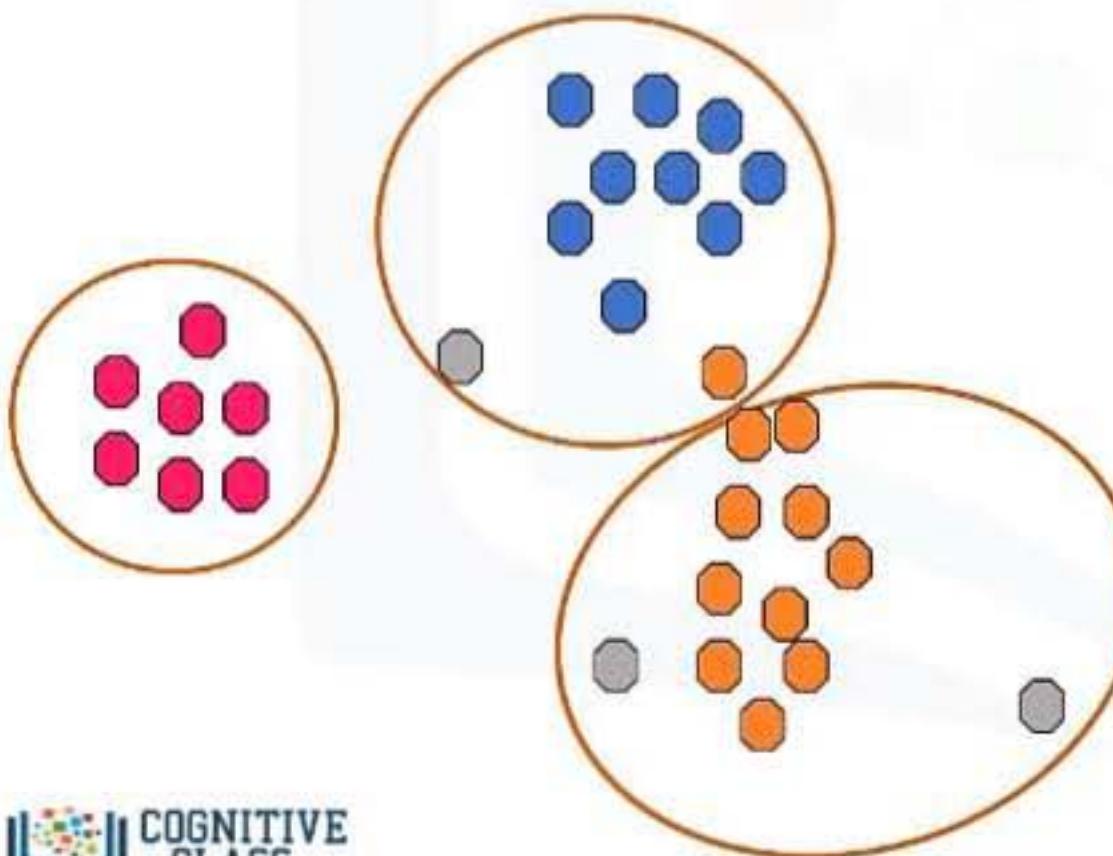
k-Means Vs. density-based clustering

- k-Means assigns all points to a cluster even if they do not belong in any
- Density-based Clustering locates regions of **high density**, and separates outliers

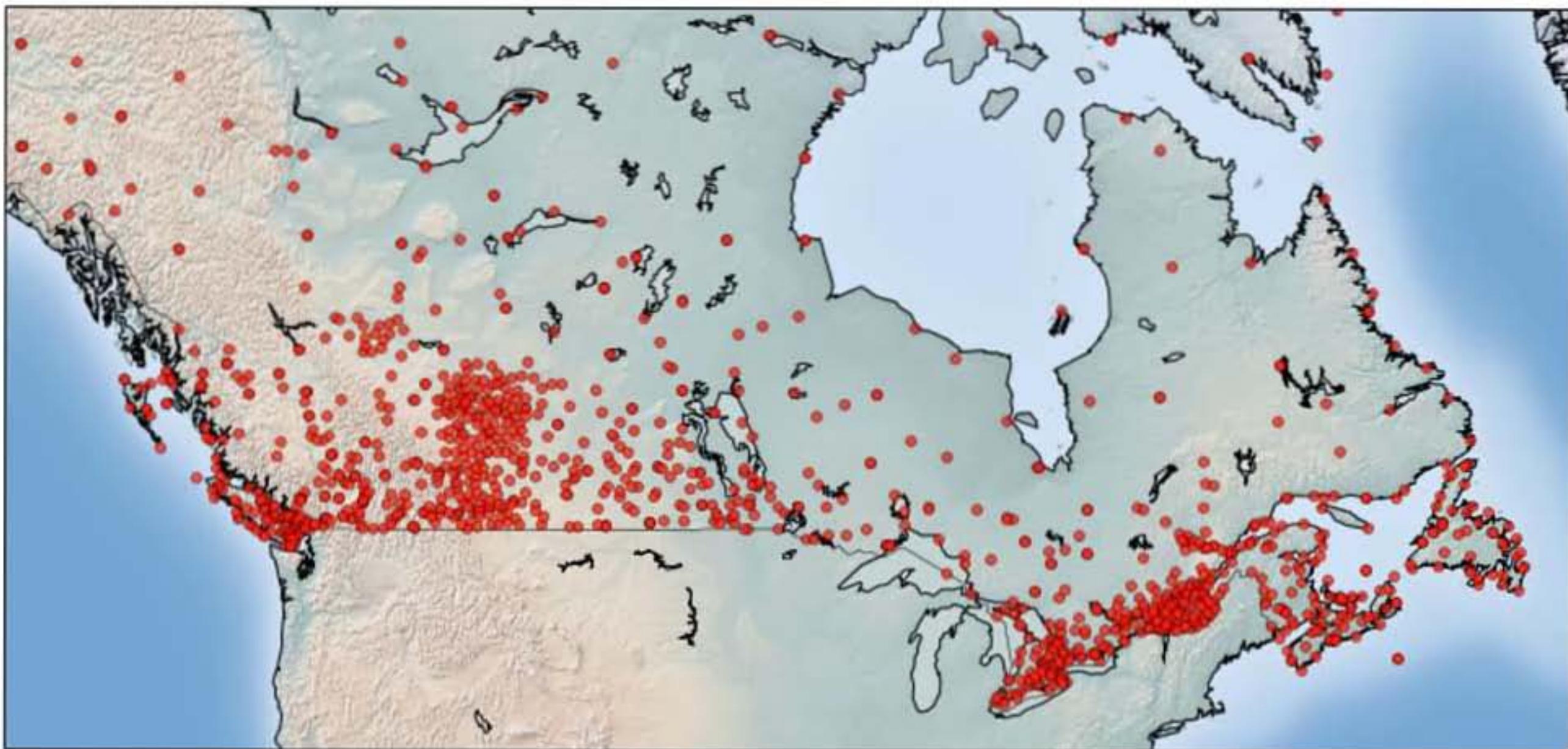


k-Means Vs. density-based clustering

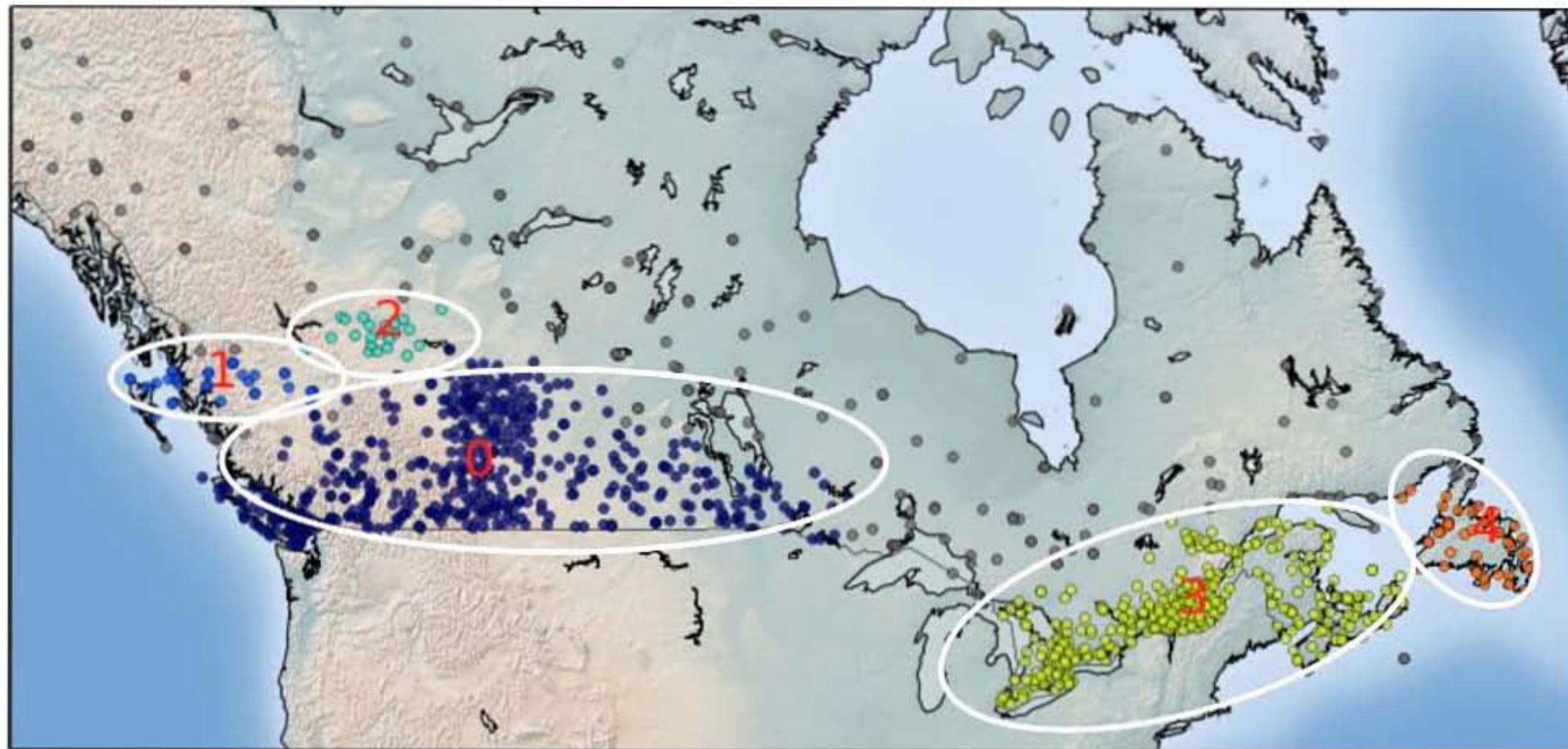
- k-Means assigns all points to a cluster even if they do not belong in any
- Density-based Clustering locates regions of **high density**, and separates outliers



DBSCAN for class identification

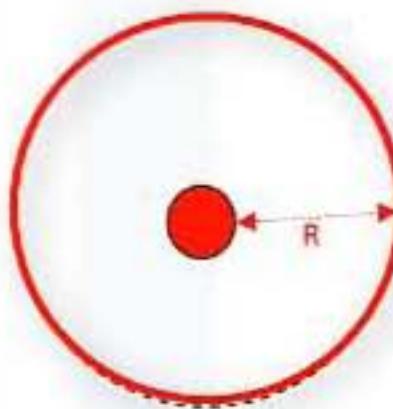


DBSCAN for class identification



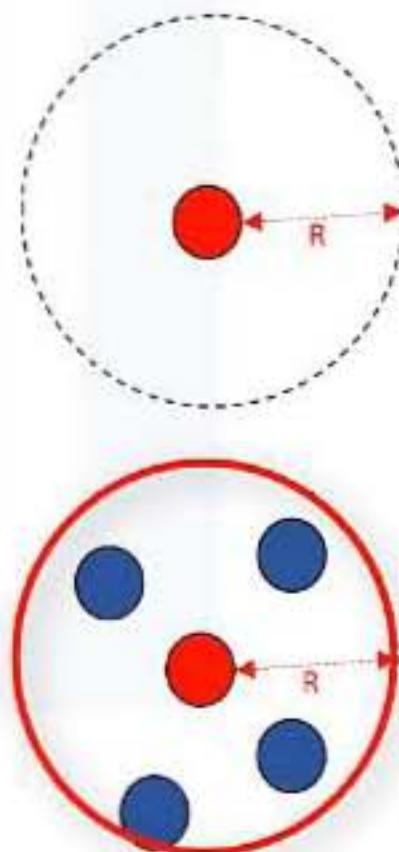
What is DBSCAN?

- DBSCAN (Density-Based Spatial Clustering of Applications with Noise)
 - Is one of the most common clustering algorithms
 - Works based on density of objects
- R (Radius of neighborhood)
 - Radius (R) that if includes enough number of points within, we call it a dense area
- M (Min number of neighbors)

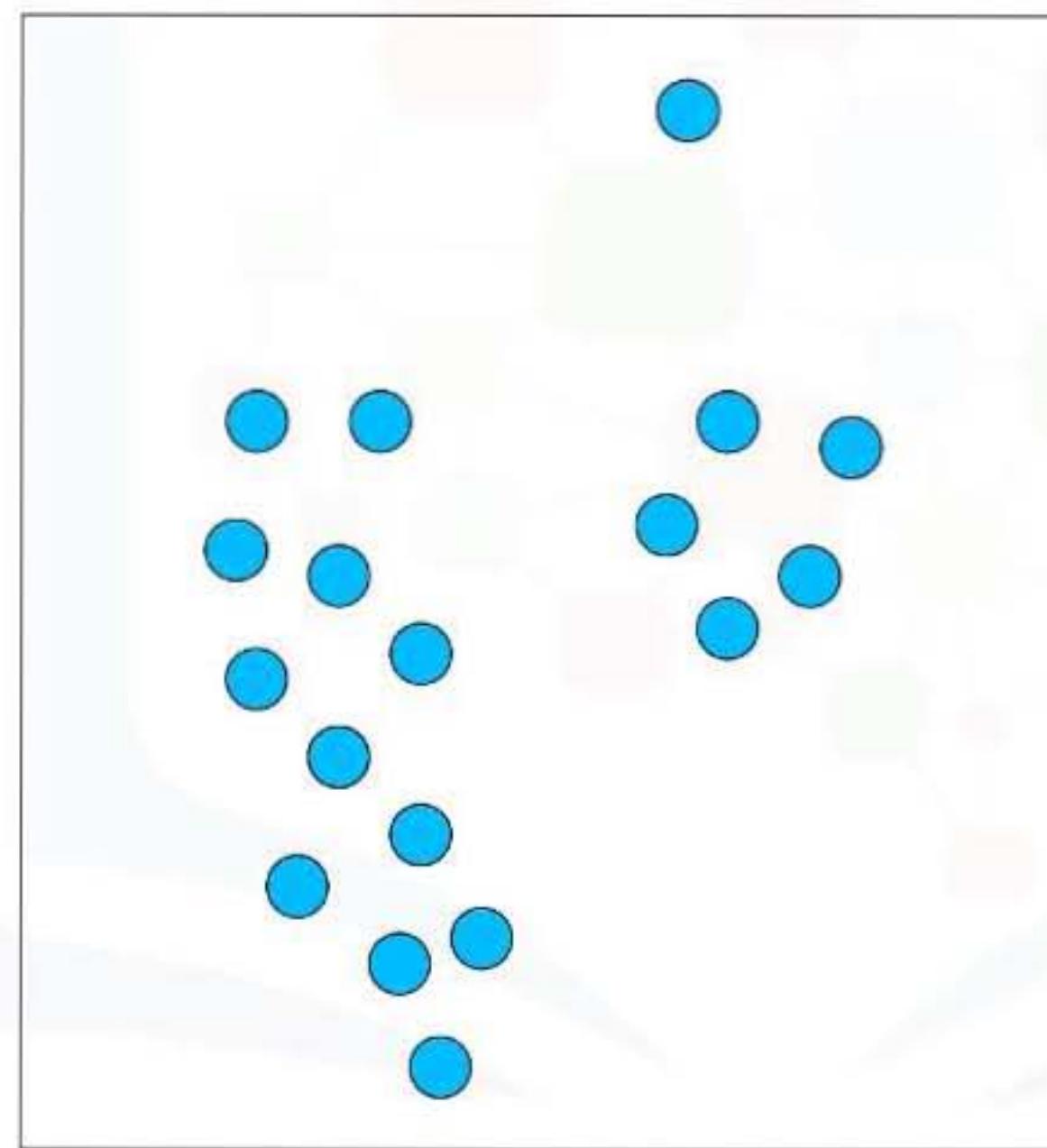


What is DBSCAN?

- DBSCAN (Density-Based Spatial Clustering of Applications with Noise)
 - Is one of the most common clustering algorithms
 - Works based on density of objects
- R (Radius of neighborhood)
 - Radius (R) that if includes enough number of points within, we call it a dense area
- M (Min number of neighbors)
 - The minimum number of data points we want in a neighborhood to define a cluster

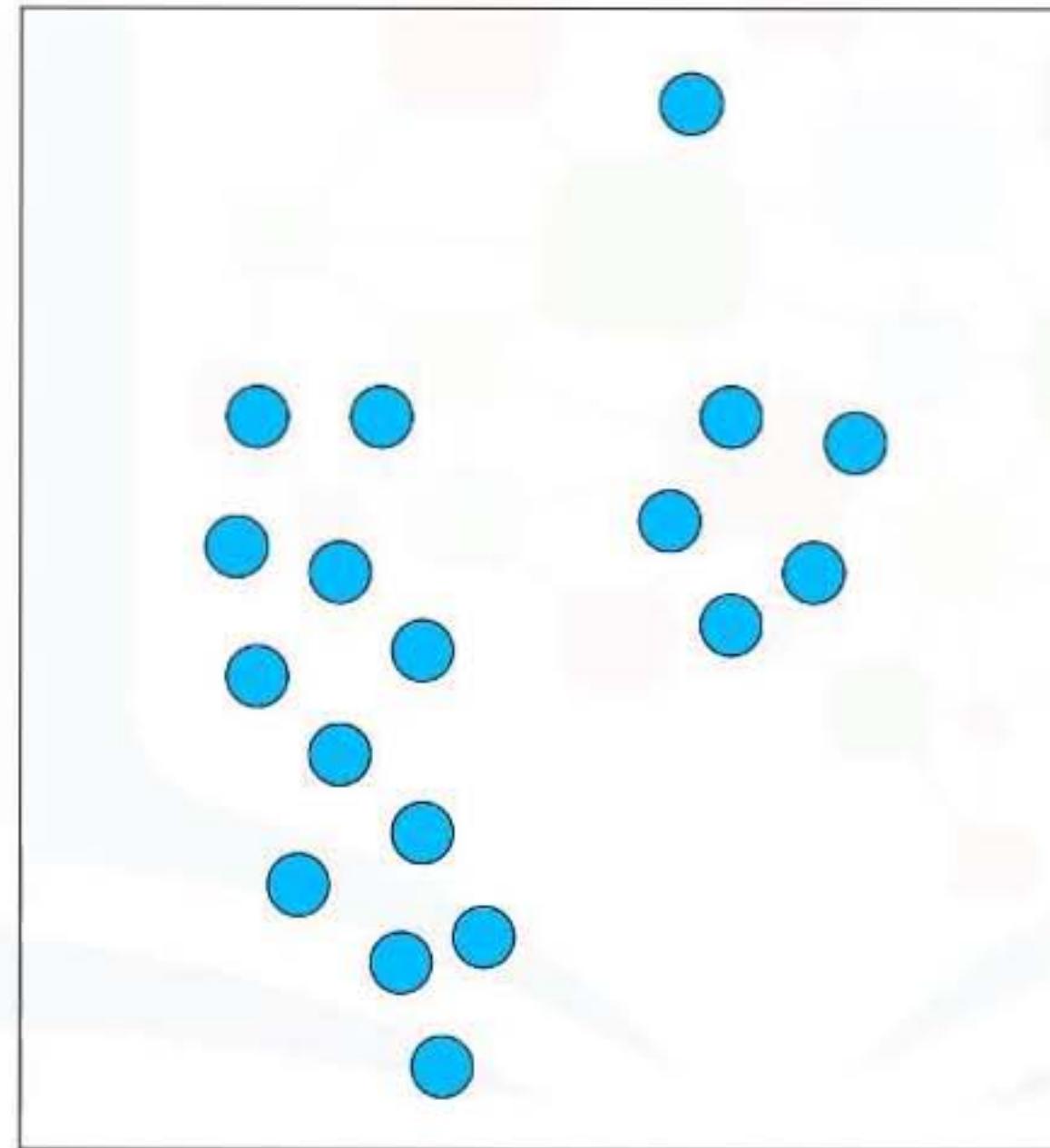


How DBSCAN works



$R = 2\text{unit}$, $M = 6$

How DBSCAN works

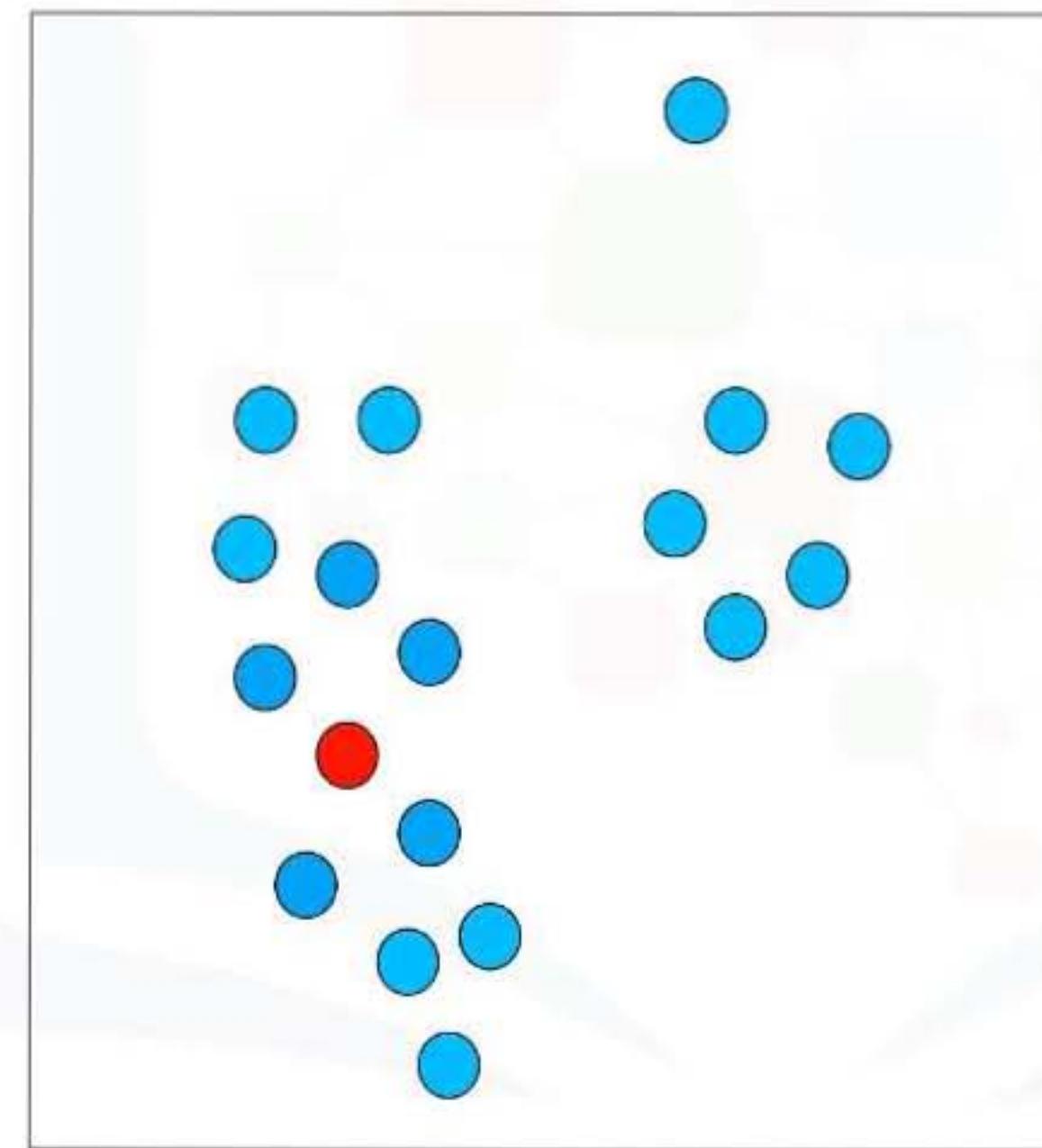


Each point is either:

- *core point*
- *border point*
- *outlier point*

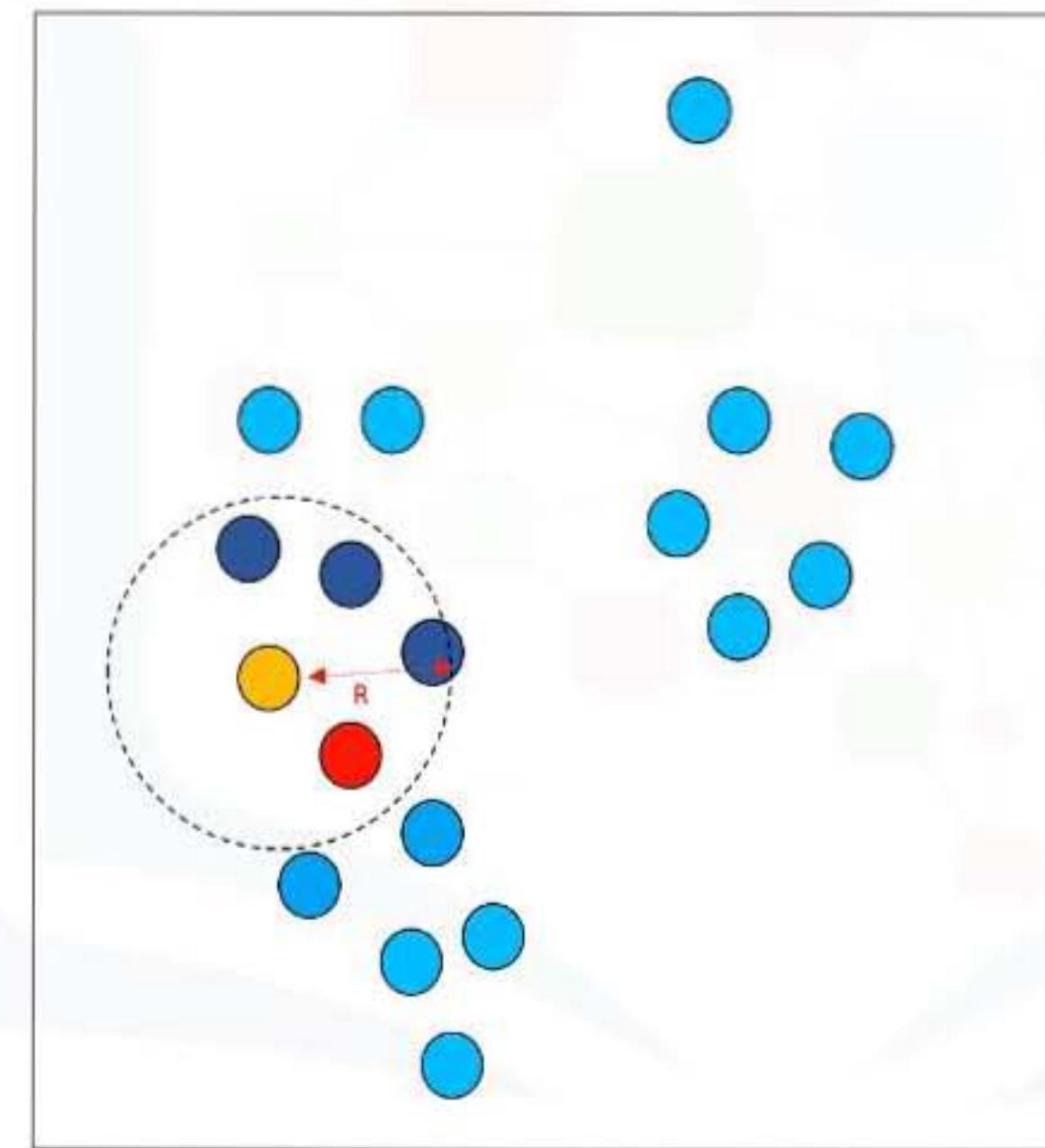
$R = 2\text{unit}$, $M = 6$

DBSCAN algorithm – core point?



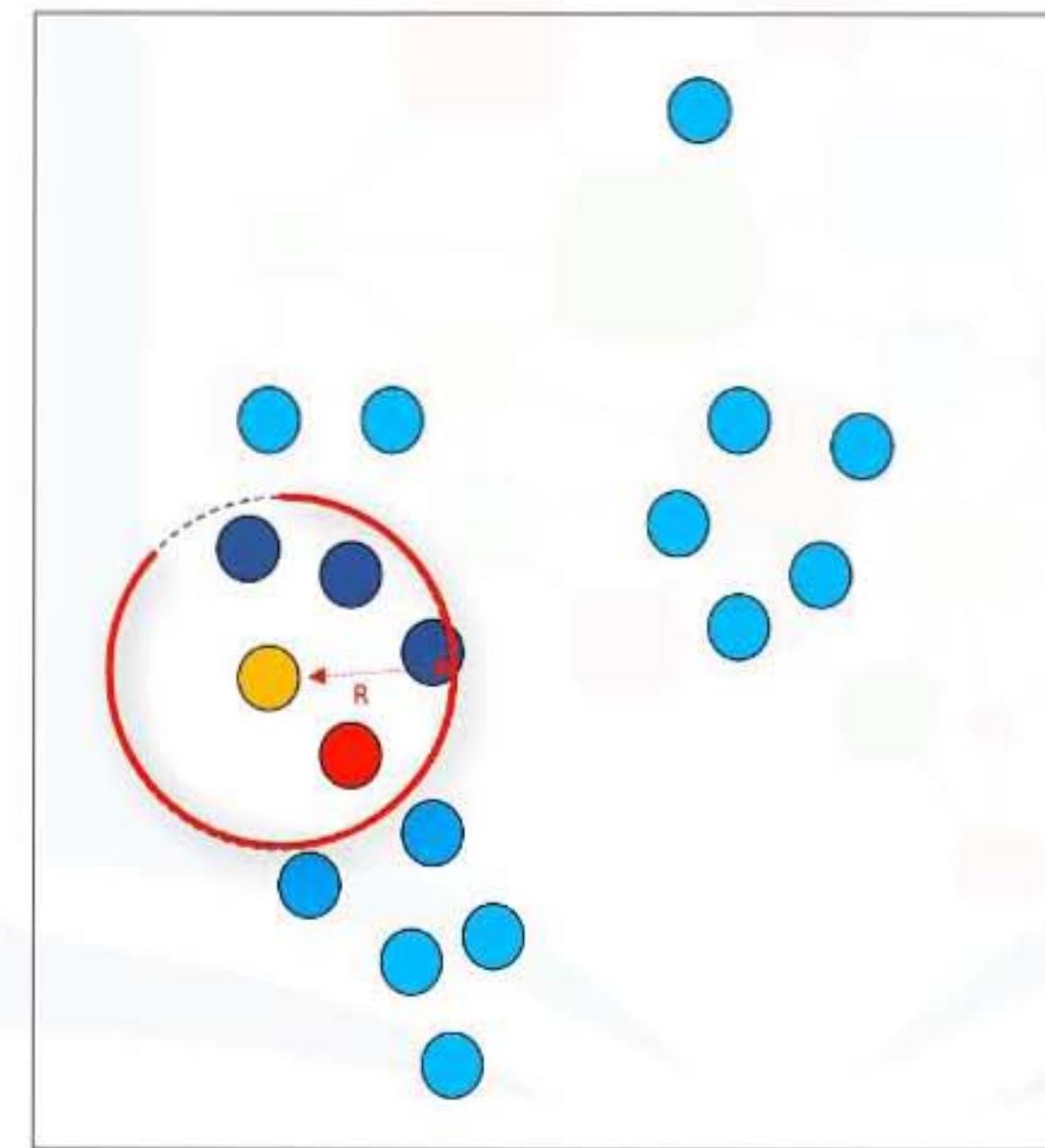
$R = 2\text{unit}$, $M = 6$

DBSCAN algorithm – border points?



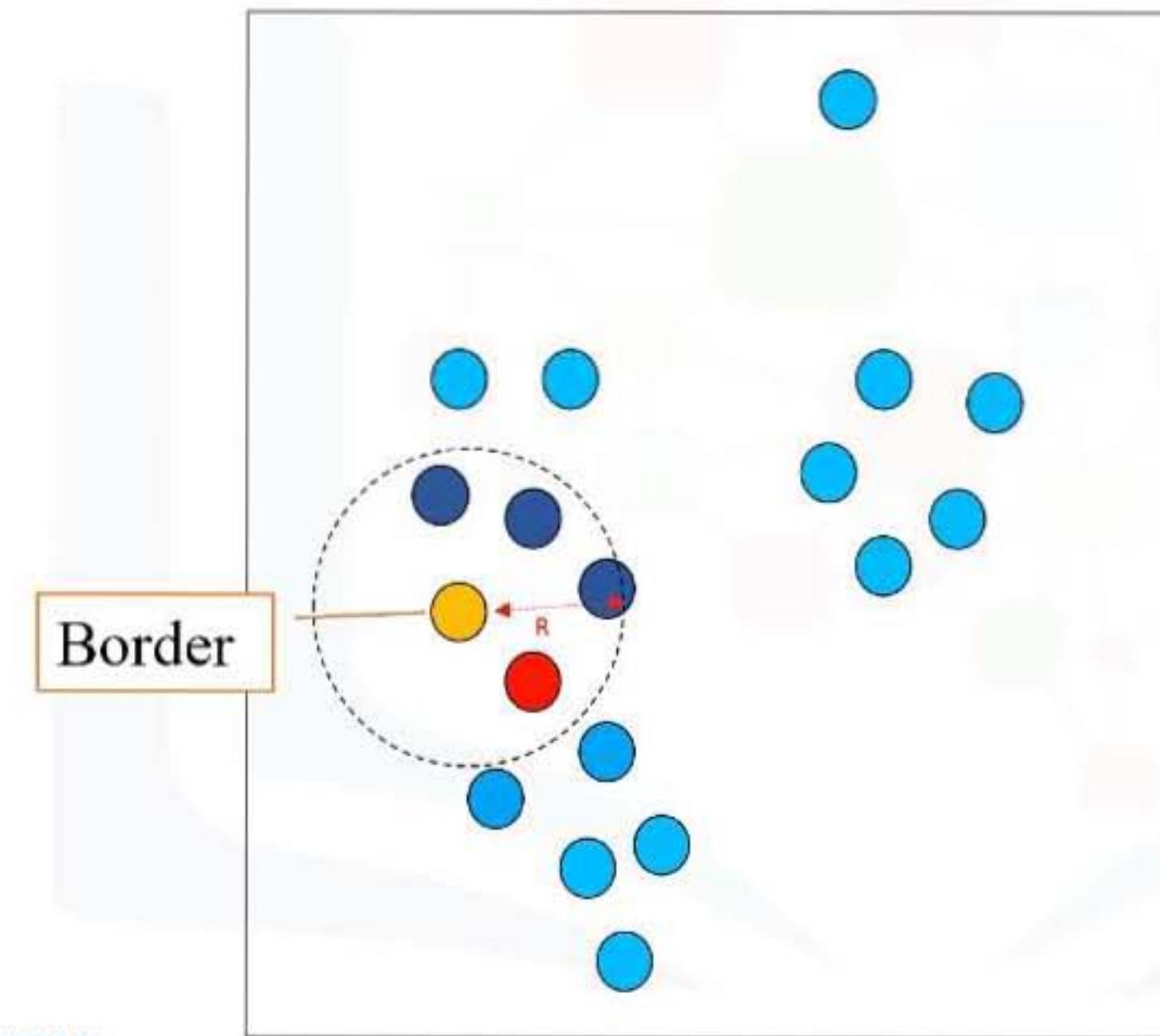
$R = 2\text{unit}$, $M = 6$

DBSCAN algorithm – border points?



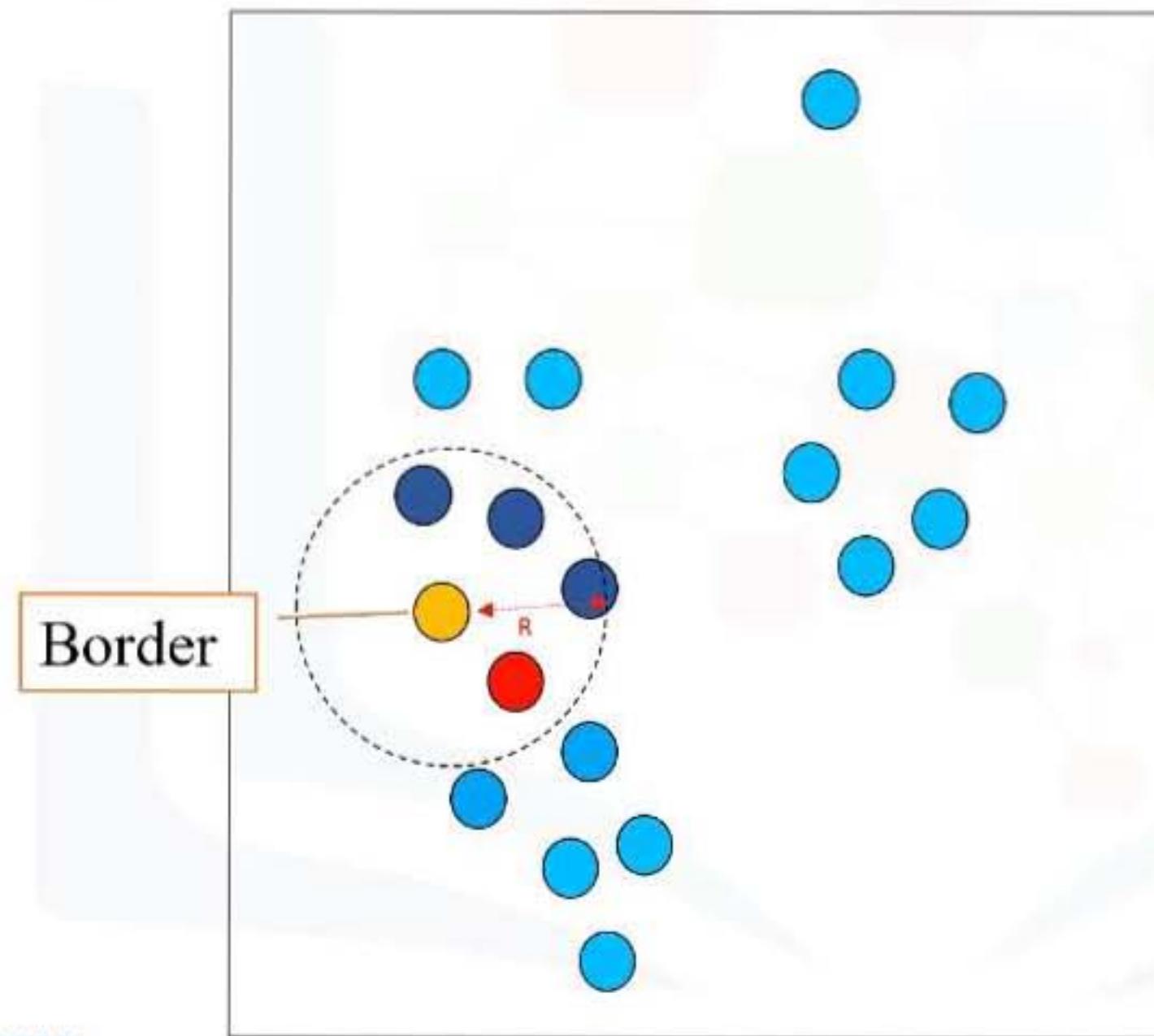
$R = 2\text{unit}$, $M = 6$

DBSCAN algorithm – border points?



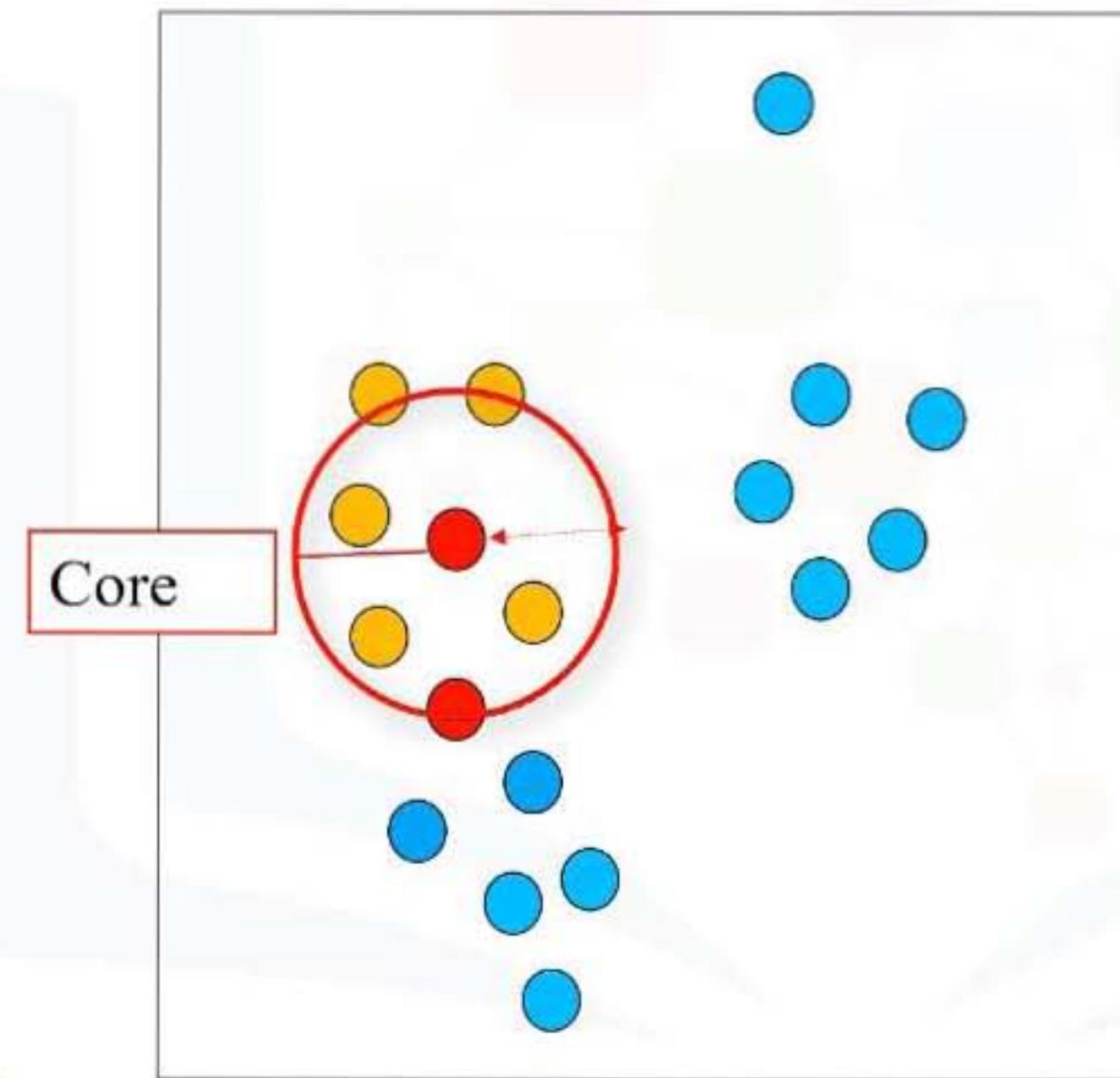
$R = 2\text{unit}$, $M = 6$

DBSCAN algorithm – border points?



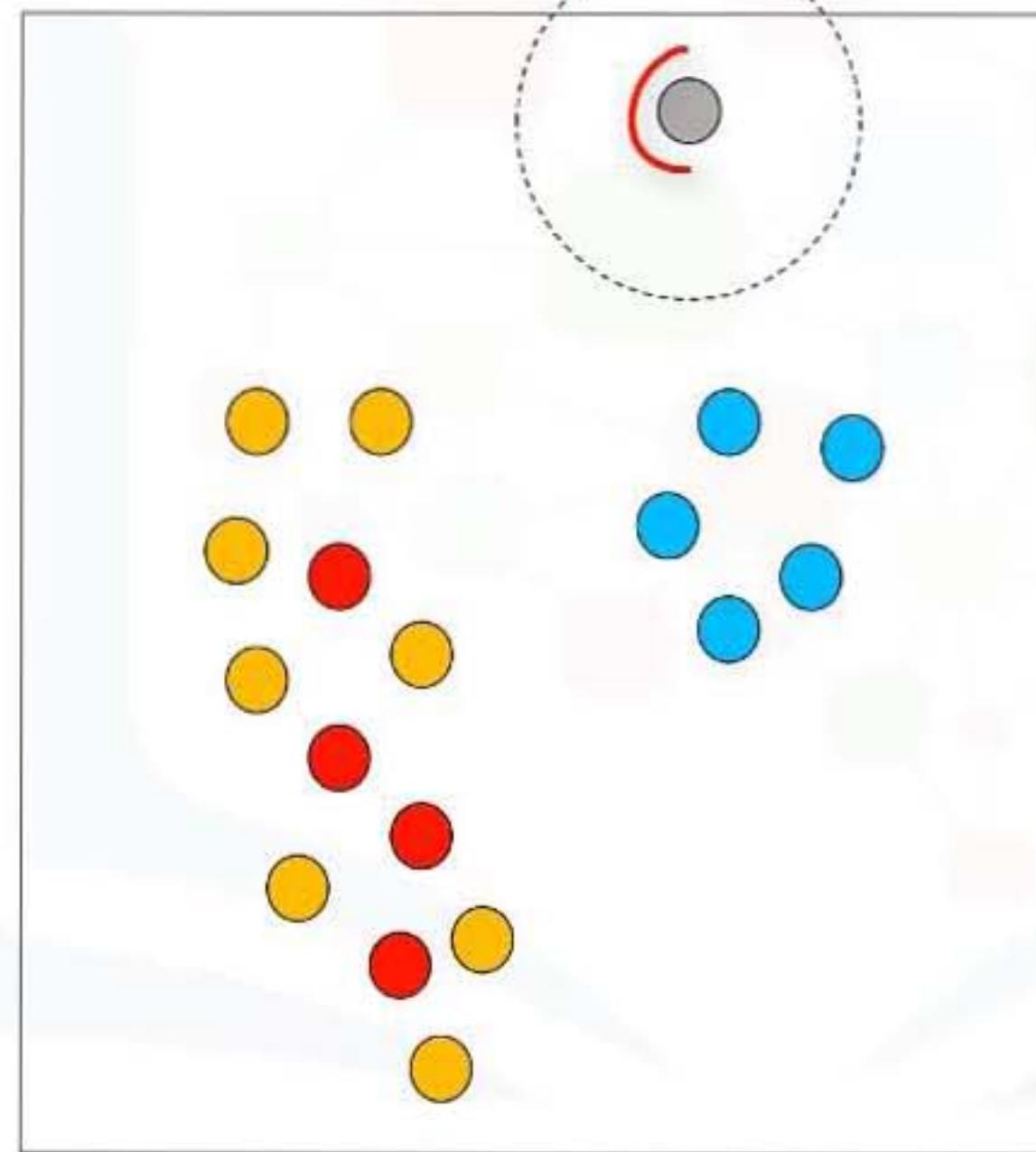
$R = 2\text{unit}$, $M = 6$

DBSCAN algorithm – border points



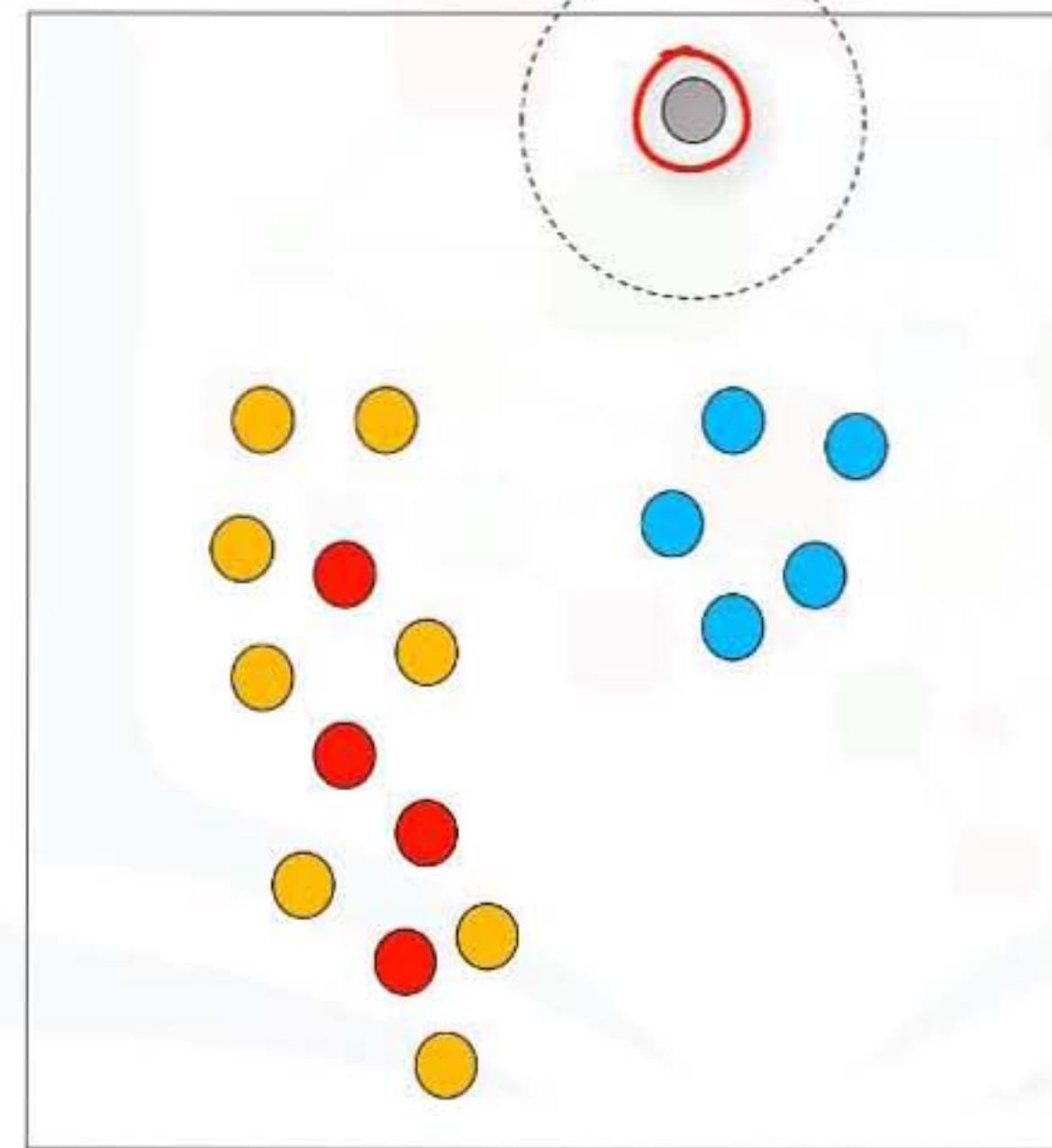
$R = 2\text{unit}$, $M = 6$

DBSCAN algorithm – outliers



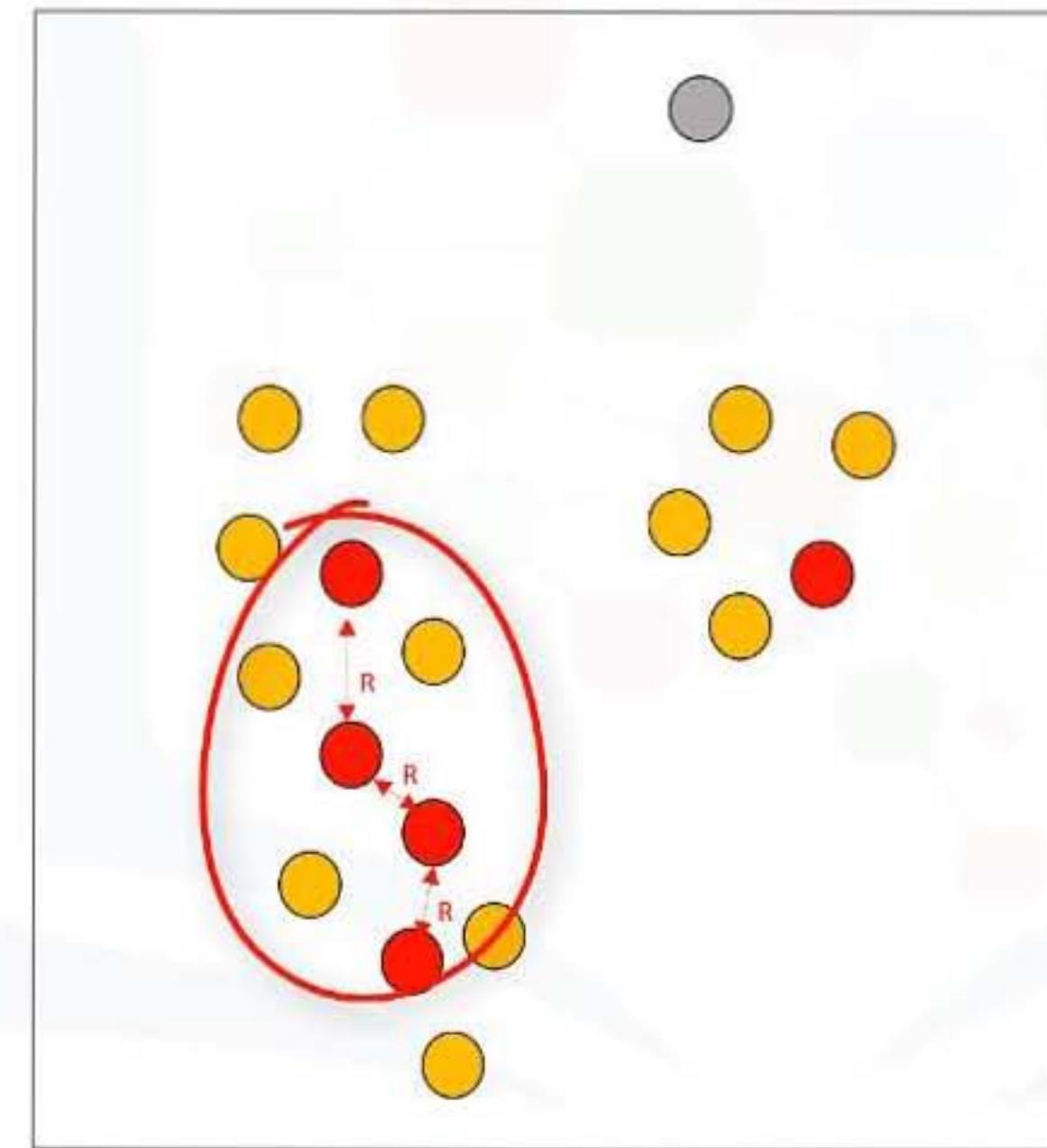
$R = 2\text{unit}$, $M = 6$

DBSCAN algorithm – outliers



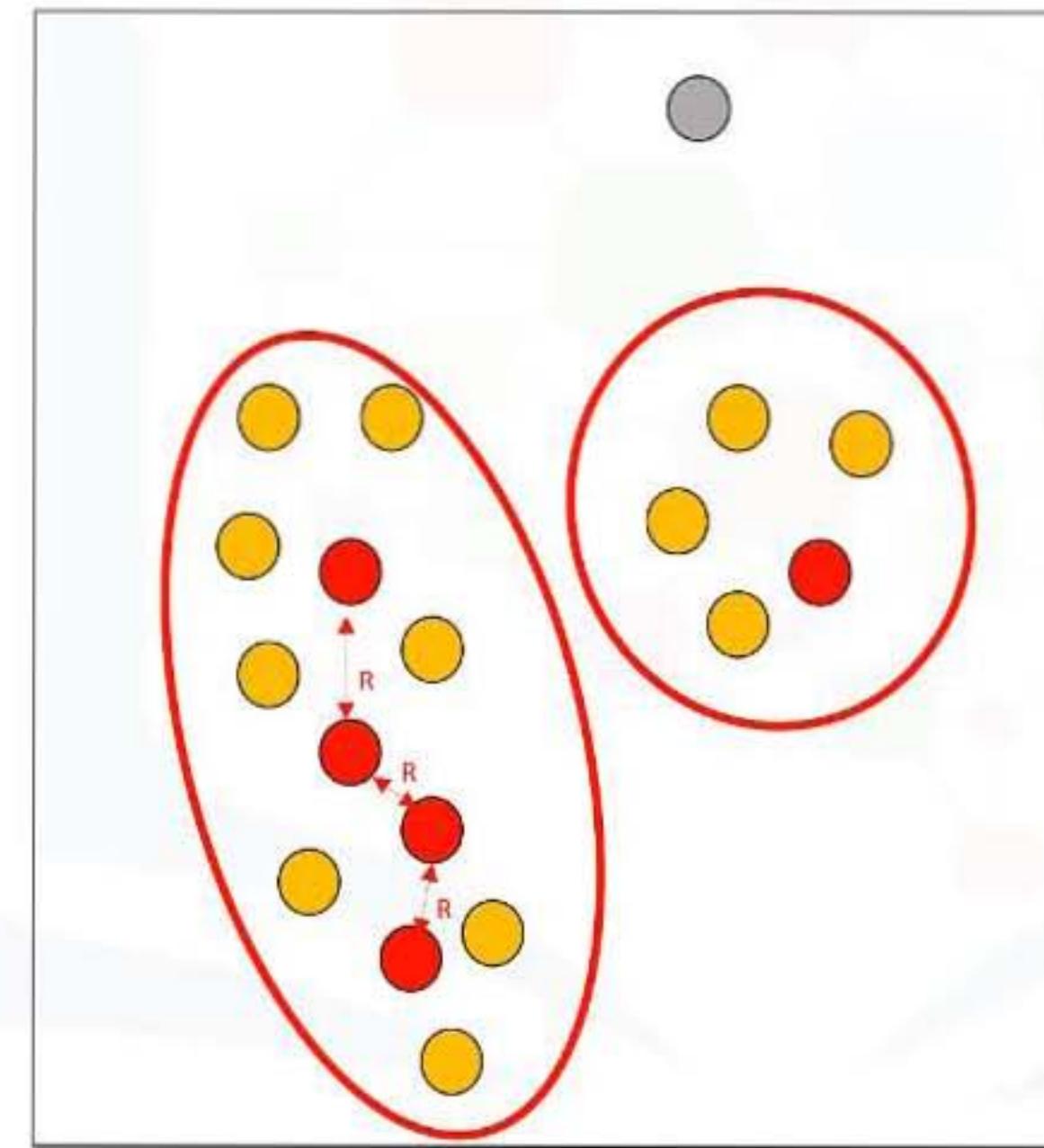
$R = 2\text{unit}$, $M = 6$

DBSCAN algorithm – clusters?



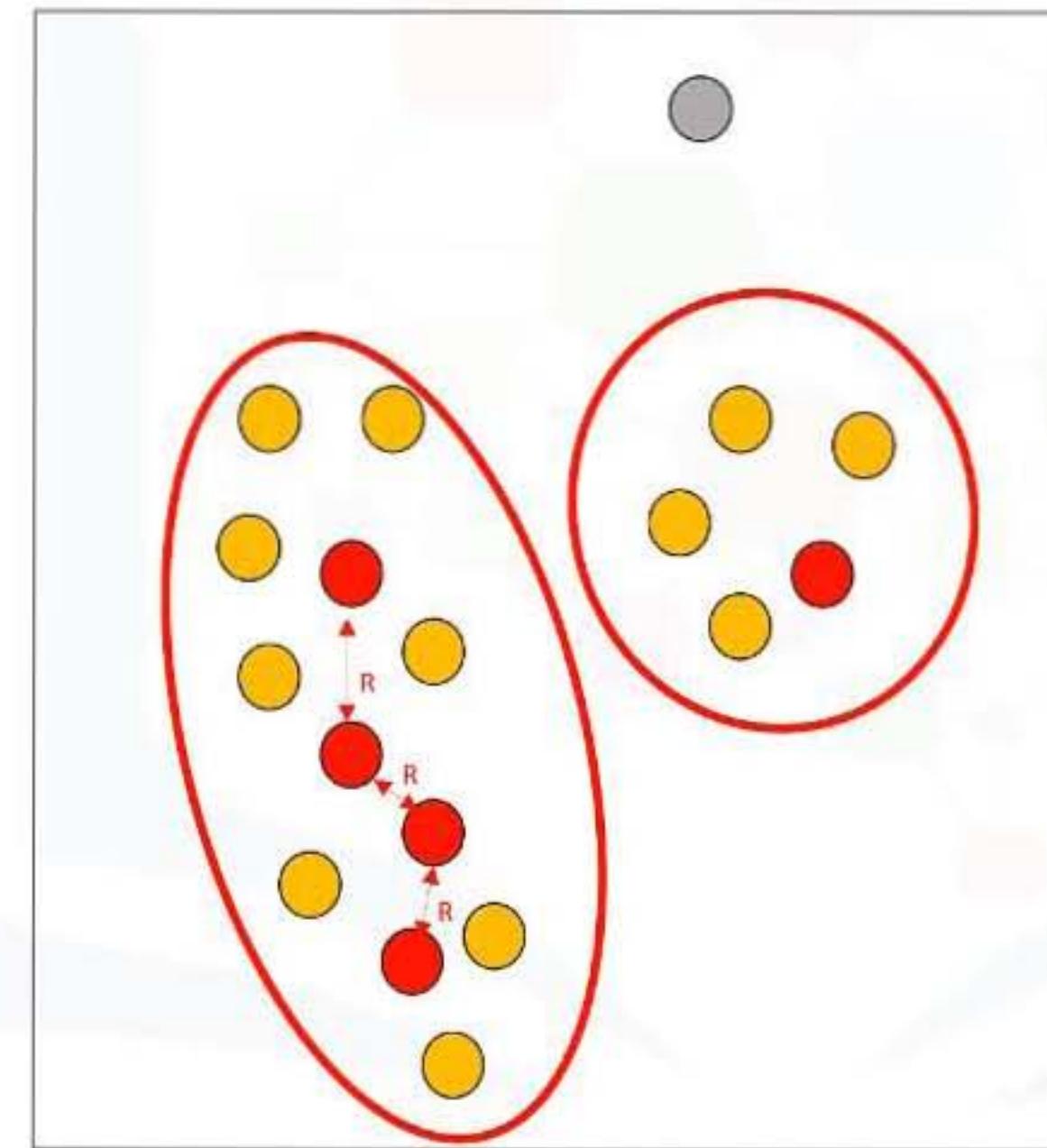
$R = 2\text{unit}$, $M = 6$

DBSCAN algorithm – clusters?



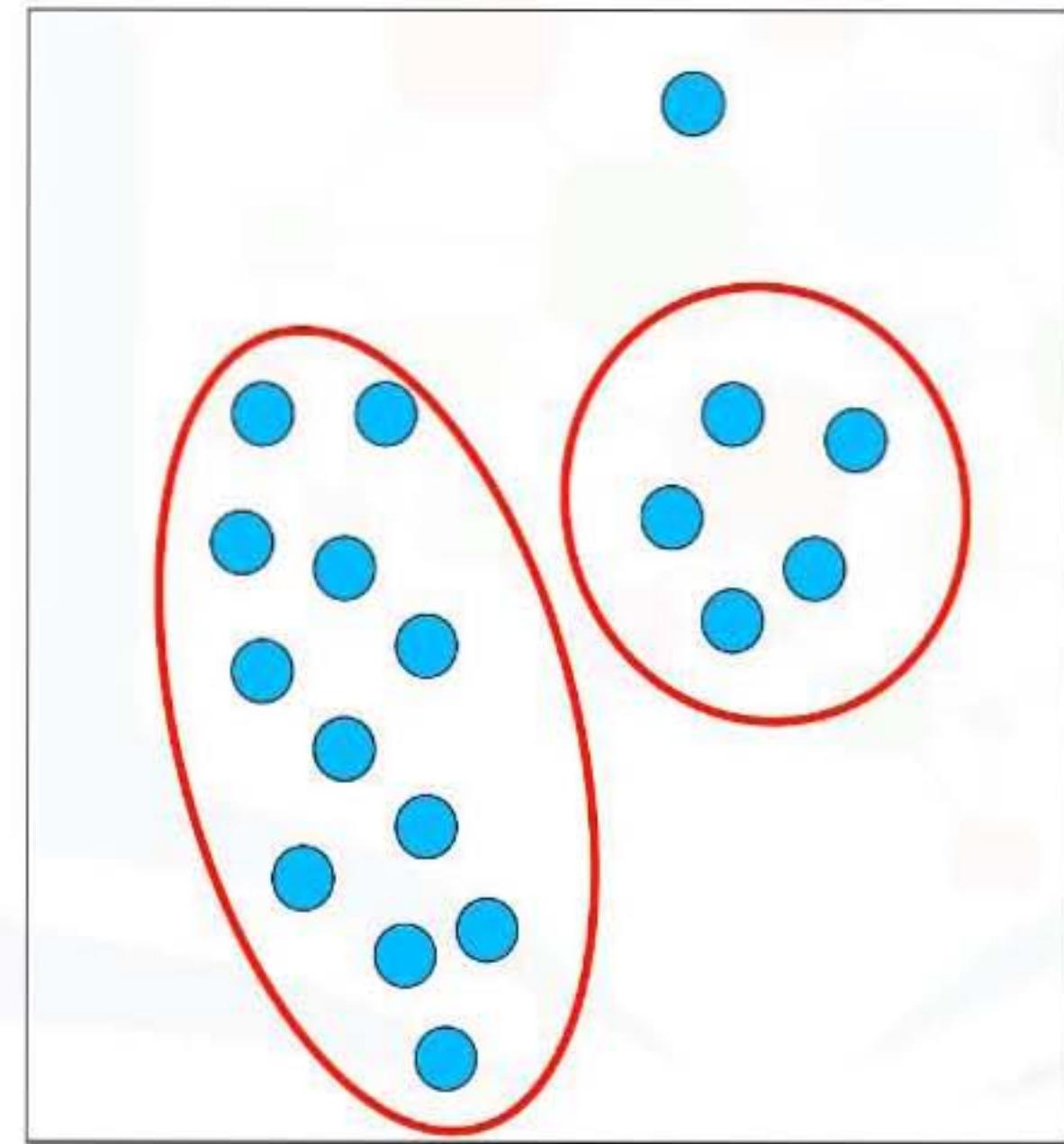
$R = 2\text{unit}$, $M = 6$

DBSCAN algorithm – clusters?

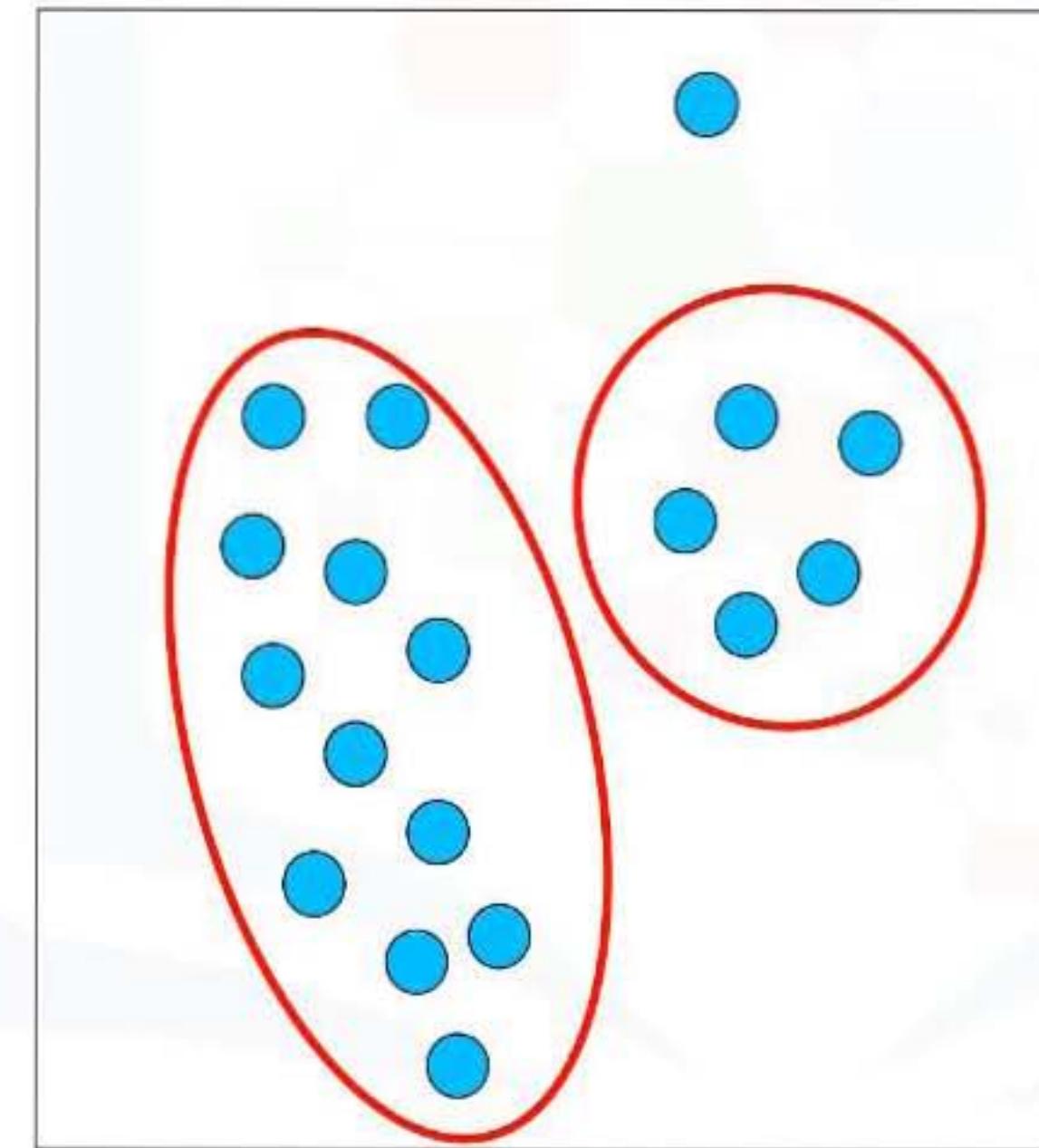


R = 2unit , M = 6

Advantages of DBSCAN



Advantages of DBSCAN



1. Arbitrarily shaped clusters
2. Robust to outliers
3. Does not require specification of the number of clusters

Recommender Systems

Saeed Aghabozorgi



© IBM Corporation. All rights reserved.

1



What are recommender systems?

Recommender systems capture the pattern of peoples' behavior and use it to predict what else they might want or like.



2



Applications

- What to buy?
 - E-commerce, books, movies, beer, shoes
- Where to eat?
- Which job to apply to?
- Who you should be friends with?
 - LinkedIn, Facebook, ...
- Personalize your experience on the web
 - News platforms, news personalization



3

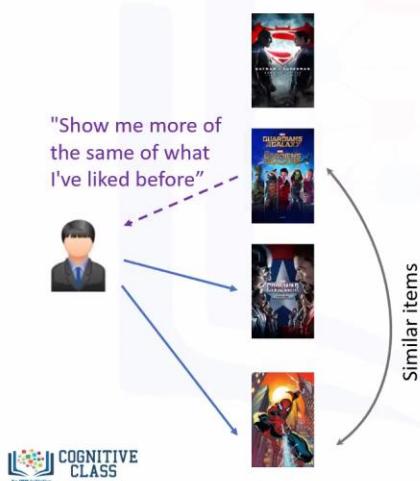
COGNITIVE CLASS

Advantages of recommender systems

- Broader exposure
- Possibility of continual usage or purchase of products
- Provides better experience

Two types of recommender systems

Content-Based



Collaborative Filtering

5

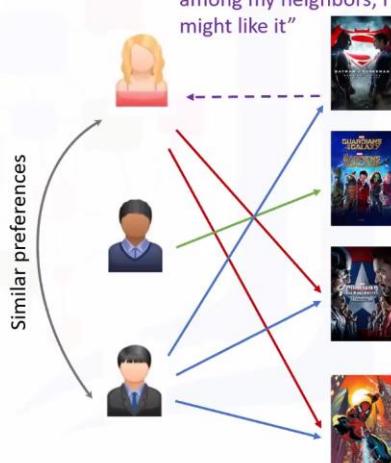


Content-Based



Collaborative Filtering

"Tell me what's popular among my neighbors, I also might like it"



5



Implementing recommender systems

- Memory-based

- Uses the entire user-item dataset to generate a recommendation
- Uses statistical techniques to approximate users or items
e.g., Pearson Correlation, Cosine Similarity, Euclidean Distance, etc.

- Model-based

- Develops a model of users in an attempt to learn their preferences
- Models can be created using Machine Learning techniques like regression, clustering, classification, etc.

Content-Based Recommender Systems

Saeed Aghabozorgi

Content-based recommender systems



Content-based recommender systems



Content-based recommender systems



 COGNITIVE
CLASS
An IBM Initiative

Batman v Superman		(Adventure, Super Hero)
Guardians of the Galaxy		(Comedy, Adventure, Super Hero, Sci-Fi)
Captain America: Civil War		(Comedy, Super Hero)
Hitchhiker's guide to the galaxy		(Comedy, Adventure, Sci-Fi)
Batman begins		(Super Hero)
Spiderman		(Comedy, Super Hero)

3



Content-based recommender systems



 COGNITIVE
CLASS
An IBM Initiative

3



Weighing the genres

		Weighted Genre Matrix			
		Comedy	Adventure	Super Hero	Sci-Fi
		0	2	2	0
	2	10	10	10	10
	10	8	0	8	0
	8				

Input User Ratings X Movies Matrix = User Profile



Weighing the genres

		Weighted Genre Matrix			
		Comedy	Adventure	Super Hero	Sci-Fi
		0	2	2	0
	2	10	10	10	10
	10	8	0	8	0
	8				

Input User Ratings X Movies Matrix = User Profile



Candidate movies for recommendation

	Comedy	Adventure	Super Hero	Sci-Fi
Hitchhiker's Guide to the Galaxy	1	1	0	1
Star Wars: Episode III - Revenge of the Sith	0	0	1	0
Spider-Man 2	1	0	1	0

Finding the recommendation

User Profile:

	Comedy	Adventure	Super Hero	Sci-Fi
	0.3	0.2	0.33	0.16

Movies Matrix:

	1	1	0	1
	0	0	1	0
	1	0	1	0

Weighted Movies Matrix:

	0.3	0.2	0	0.16
	0	0	0.33	0
	0.3	0	0.33	0

Recommendation Matrix:

Weighted Average	0.66
	0.33
	0.63

Σ

Content-based recommender systems



Content-based recommender systems



Collaborative Filtering

Saeed Aghabozorgi



© IBM Corporation. All rights reserved.

1



Collaborative filtering

- **User-based collaborative filtering**
 - Based on users' neighborhood
- **Item-based collaborative filtering**
 - Based on items' similarity

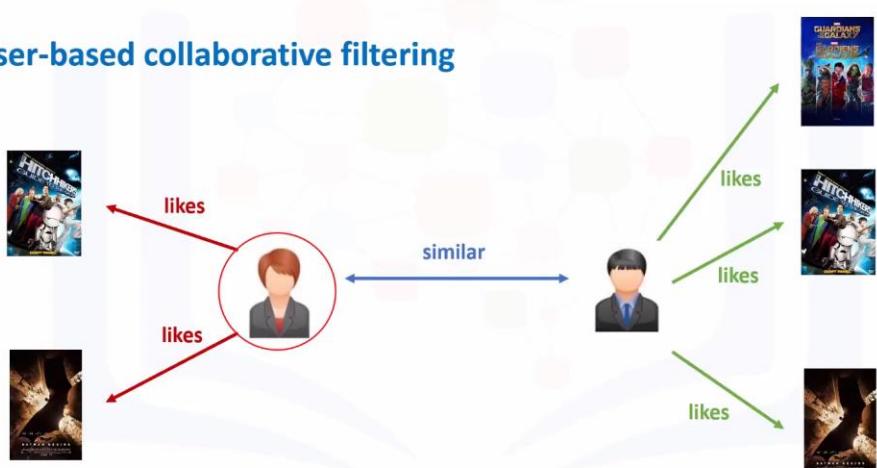


2



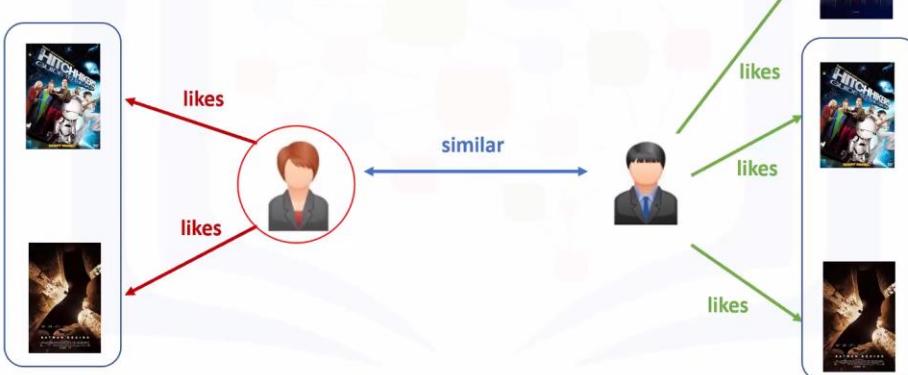
Collaborative filtering

- User-based collaborative filtering



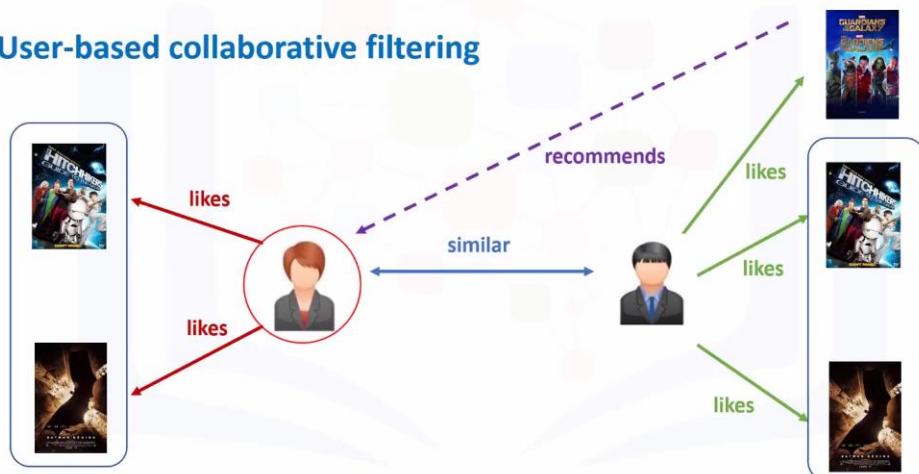
Collaborative filtering

- User-based collaborative filtering



Collaborative filtering

- User-based collaborative filtering



COGNITIVE CLASS
An IBM Initiative

3



User ratings matrix

	Superman Returns	Doctor Strange	Star Wars	Guardians of the Galaxy	The Martian
Man	9	6	8	4	
Woman	2	10	6		8
Active user	5	9		10	7
Active user	?	10	7	8	?

Ratings Matrix

COGNITIVE CLASS
An IBM Initiative

4



Learning the similarity weights

	9	6	8	4	
	2	10	6		8
	5	9		10	7
	?	10	7	8	?

Ratings Matrix

Learning the similarity weights

	9	6	8	4	
	2	10	6		8
	5	9		10	7
	?	10	7	8	?

Ratings Matrix

Creating the weighted ratings matrix

	9	
	2	8
	5	7

Ratings Matrix Subset

Creating the weighted ratings matrix

	9	
	2	8
	5	7

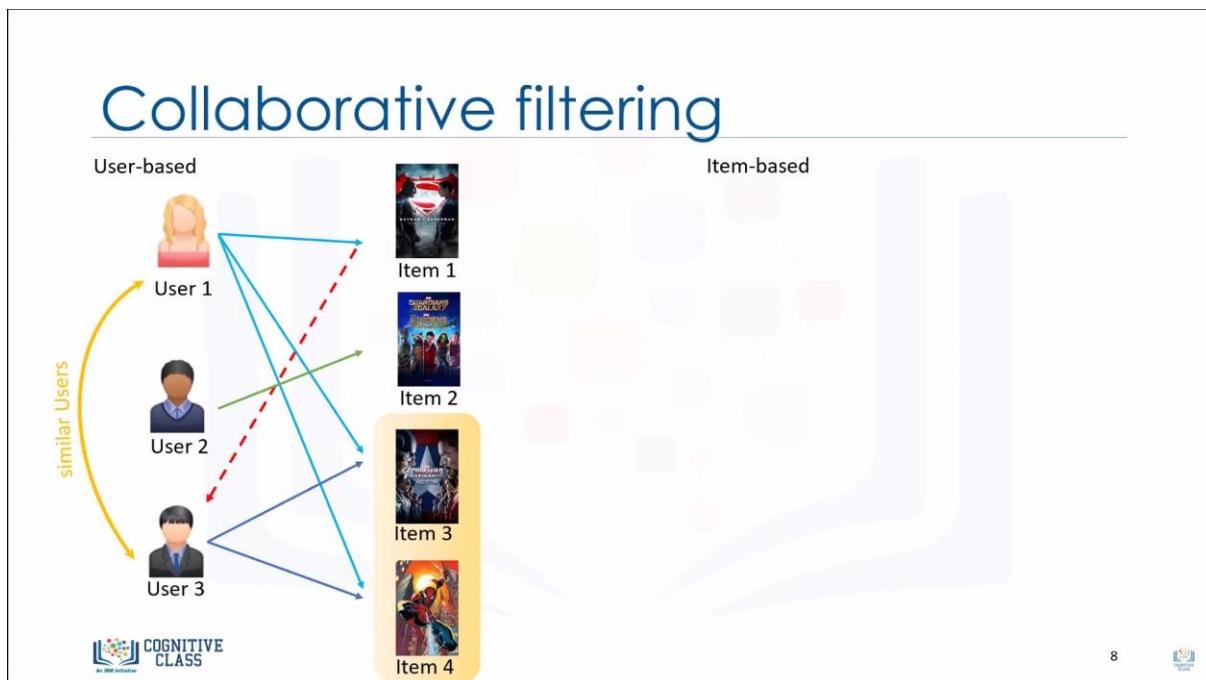
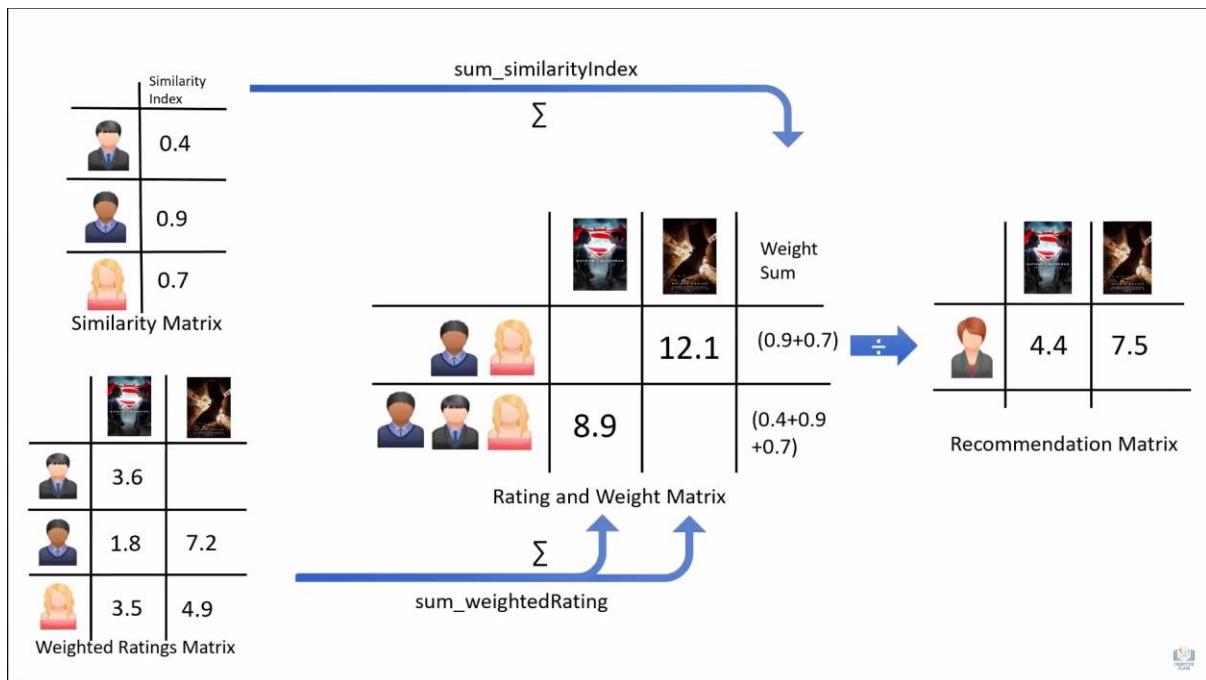
Ratings Matrix Subset

		Similarity Index
	x →	0.4
	x →	0.9
	x →	0.7

Similarity Matrix

3.6	
=	
1.8	7.2
3.5	4.9

Weighted Ratings Matrix



Collaborative filtering



Challenges of collaborative filtering

- **Data Sparsity**
 - Users in general rate only a limited number of items
- **Cold start**
 - Difficulty in recommendation to new users or new items
- **Scalability**
 - Increase in number of users or items