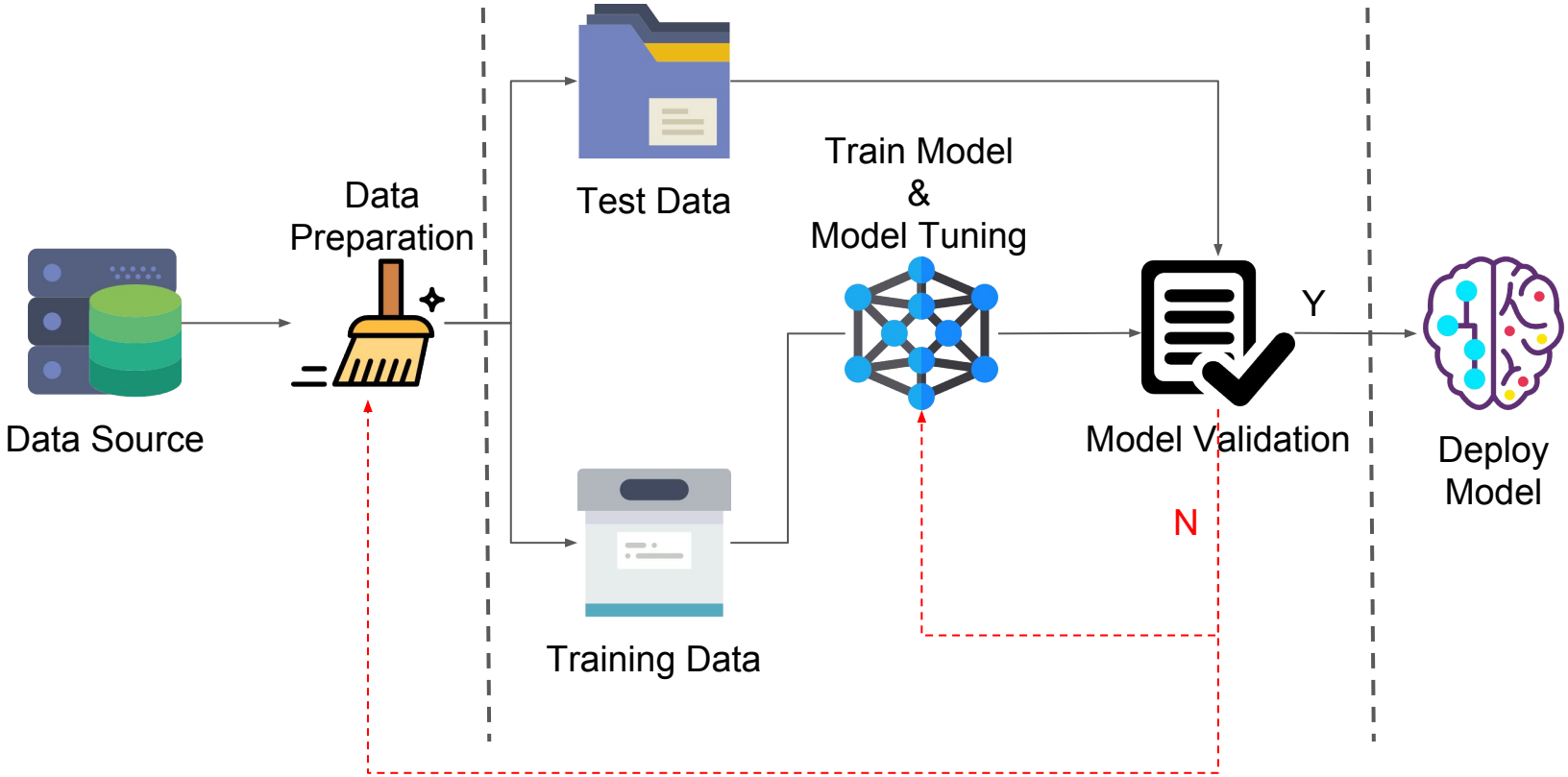
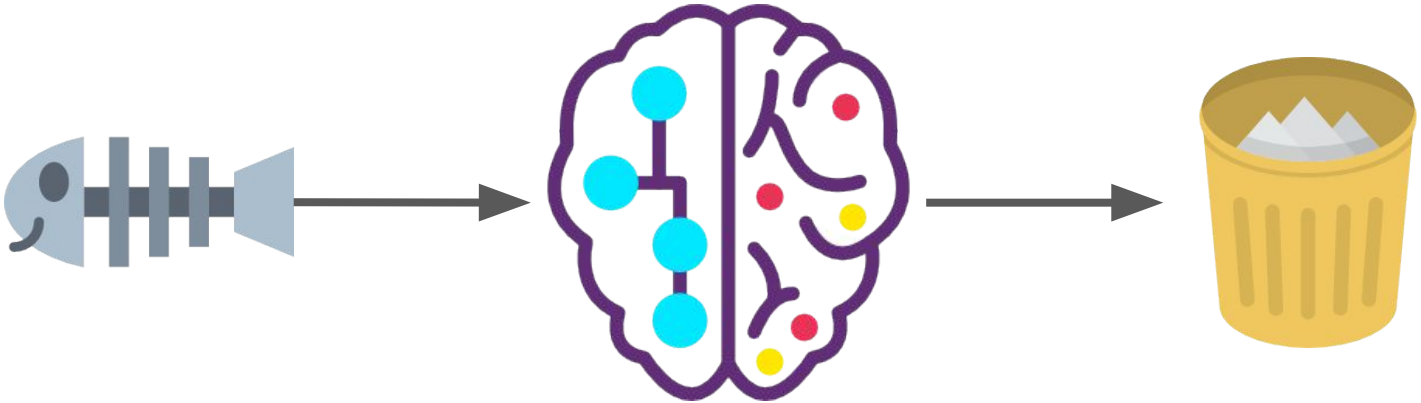
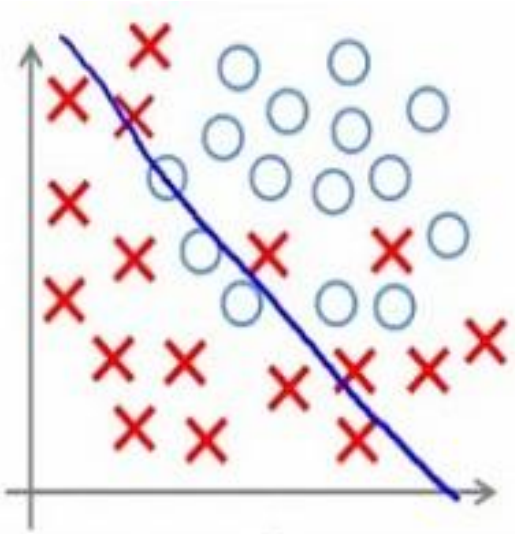
Machine Learning **Process**



Garbage In - **Garbage** Out

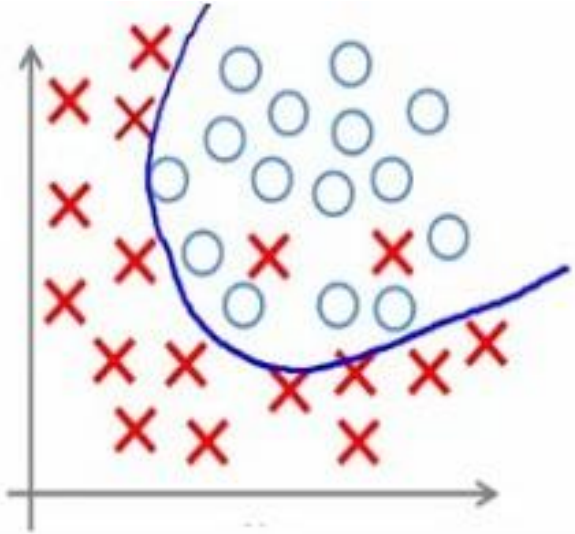


Why Feature Correlation Matter?

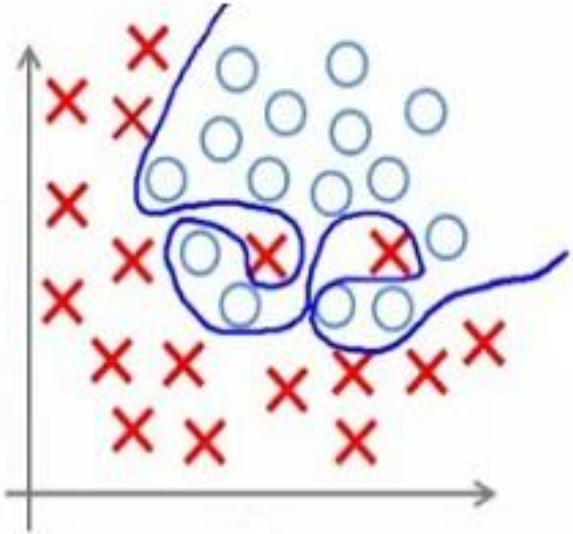


Under-fitting

(too simple to explain the variance)



Appropriate-fitting

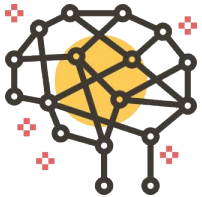


Over-fitting

(forcefitting -- too good to be true)

Model **Improvement**

Brain Power



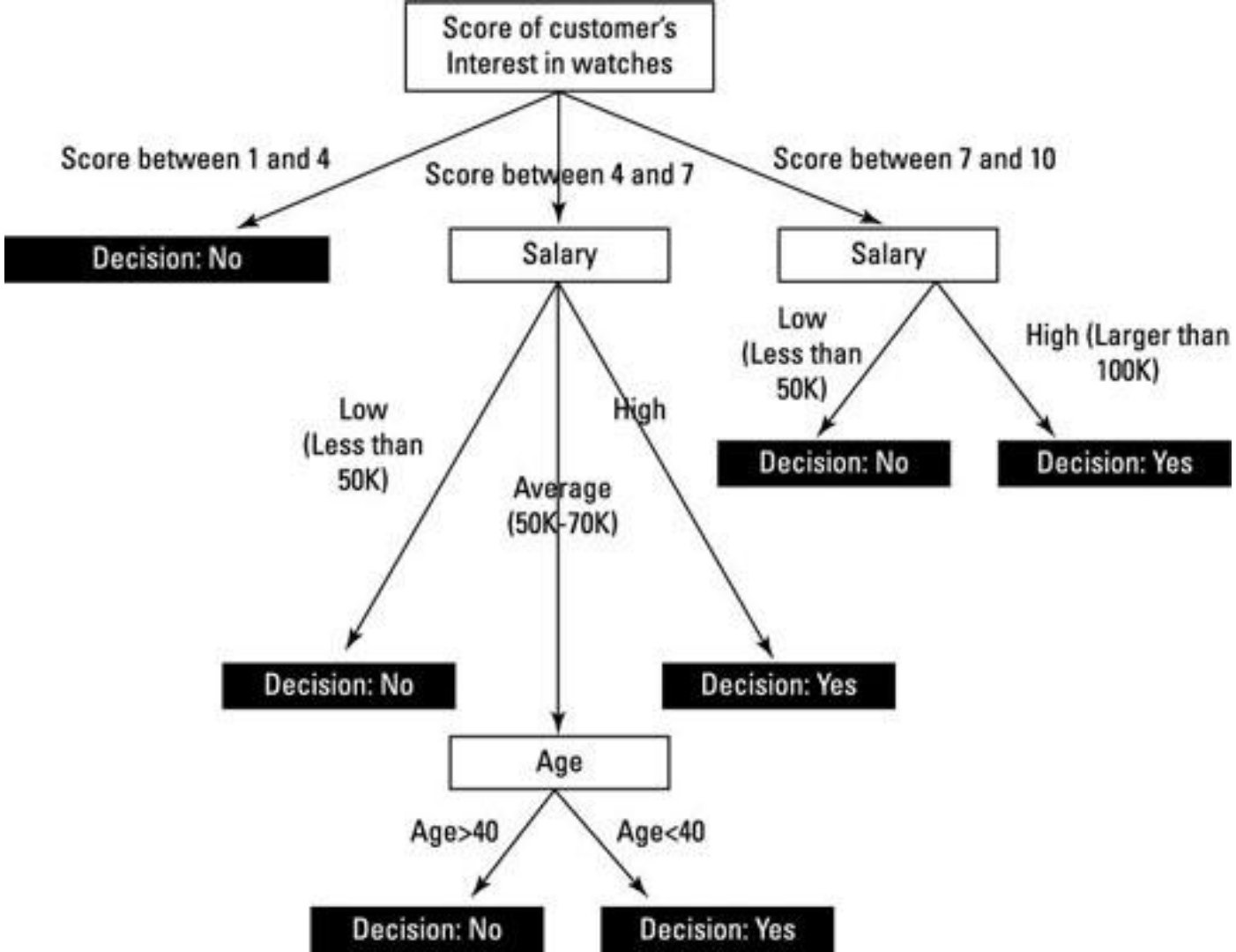
Feature Engineering

Brute Force

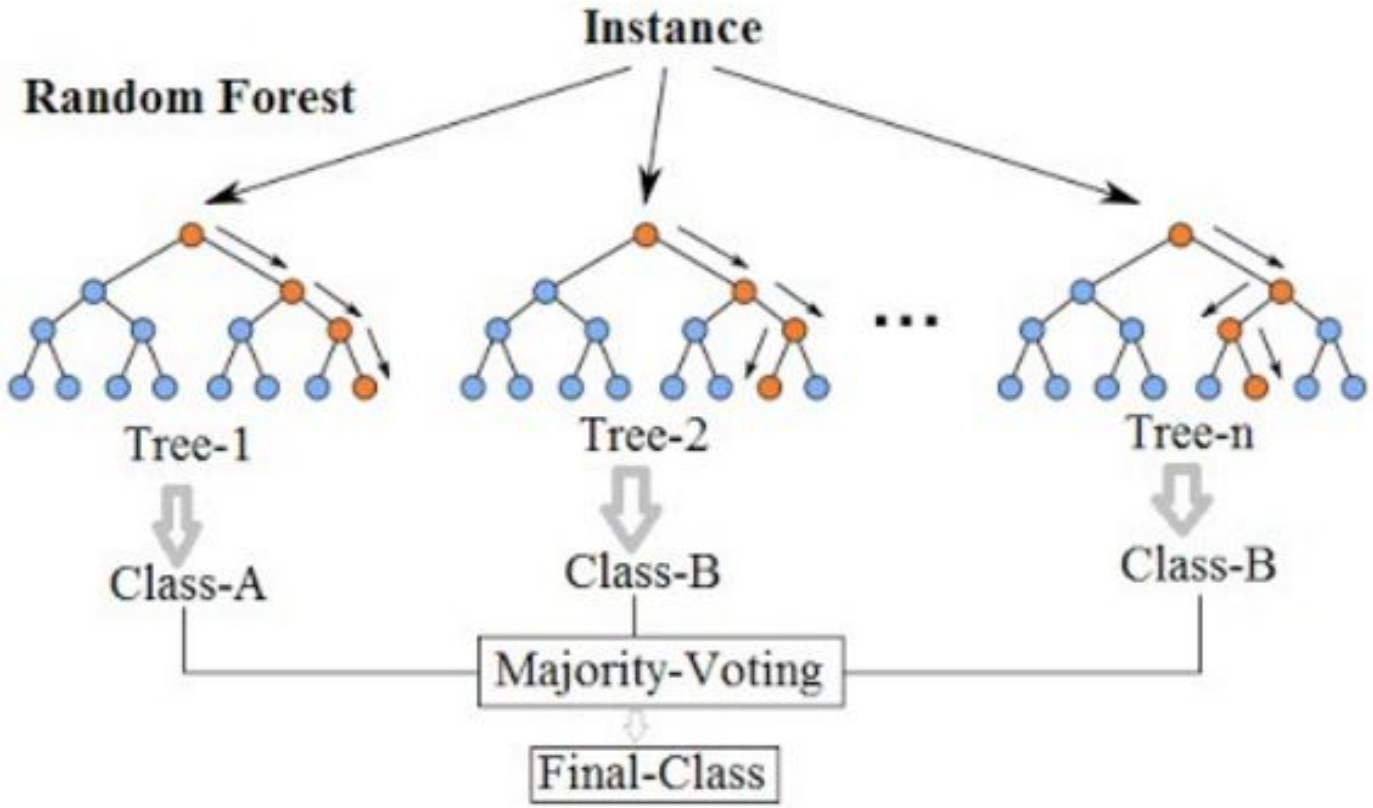
Other ML Algo

Hyper Parameter Tuning

Feature Engineering Concept

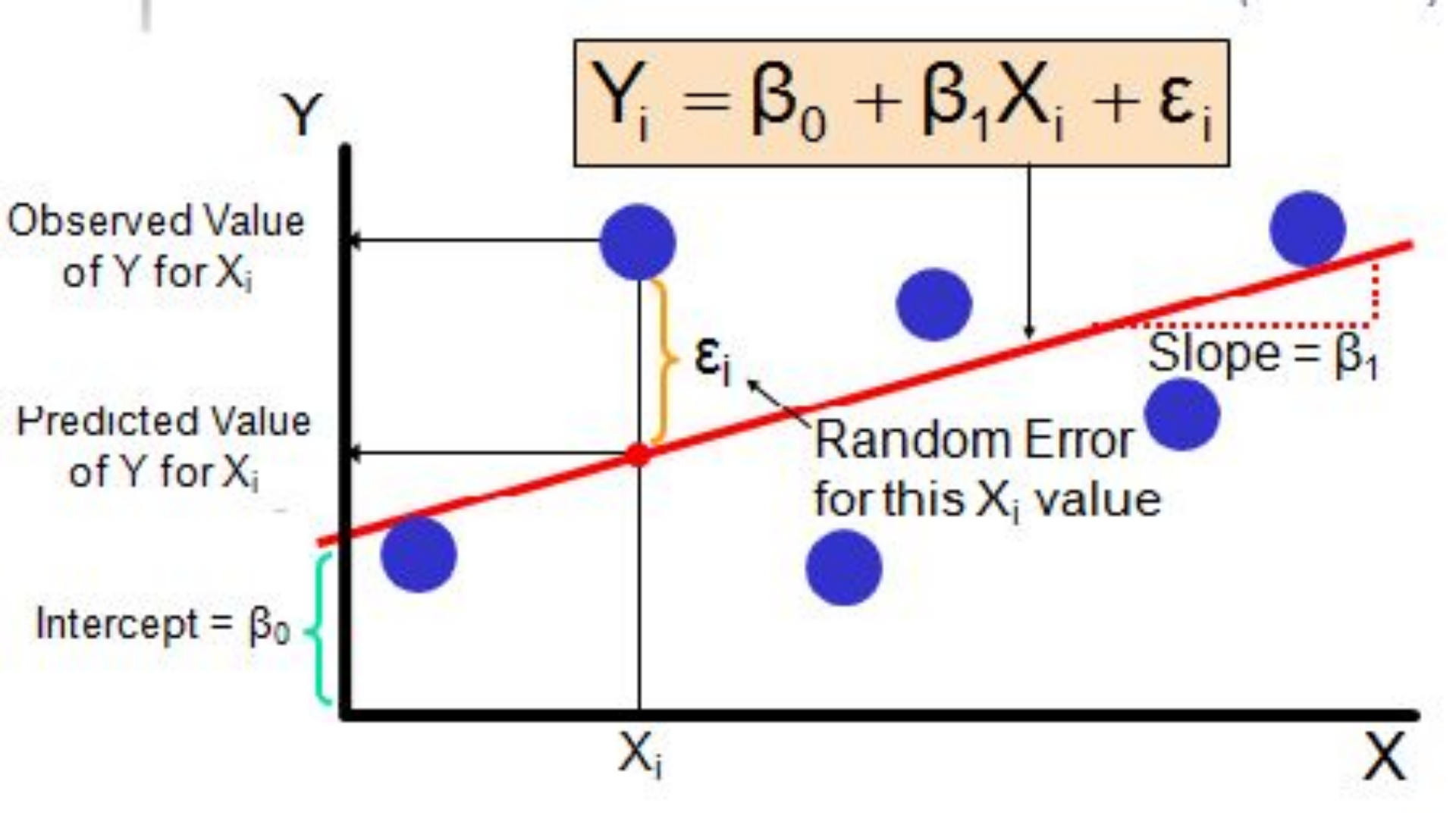

Score of customer's Interest in watches
Score between 1 and 4
Score between 4 and 7
Score between 7 and 10
Decision: No
Salary
Salary
Low (Less than 50K)
High (Larger than 100K)
Low (Less than 50K)
High
Average (50K-70K)
Decision: No
Decision: No
Decision: Yes
Decision: Yes
Age
Age>40
Age<40
Decision: No
Decision: Yes

Random Forest Simplified



Random Forest API

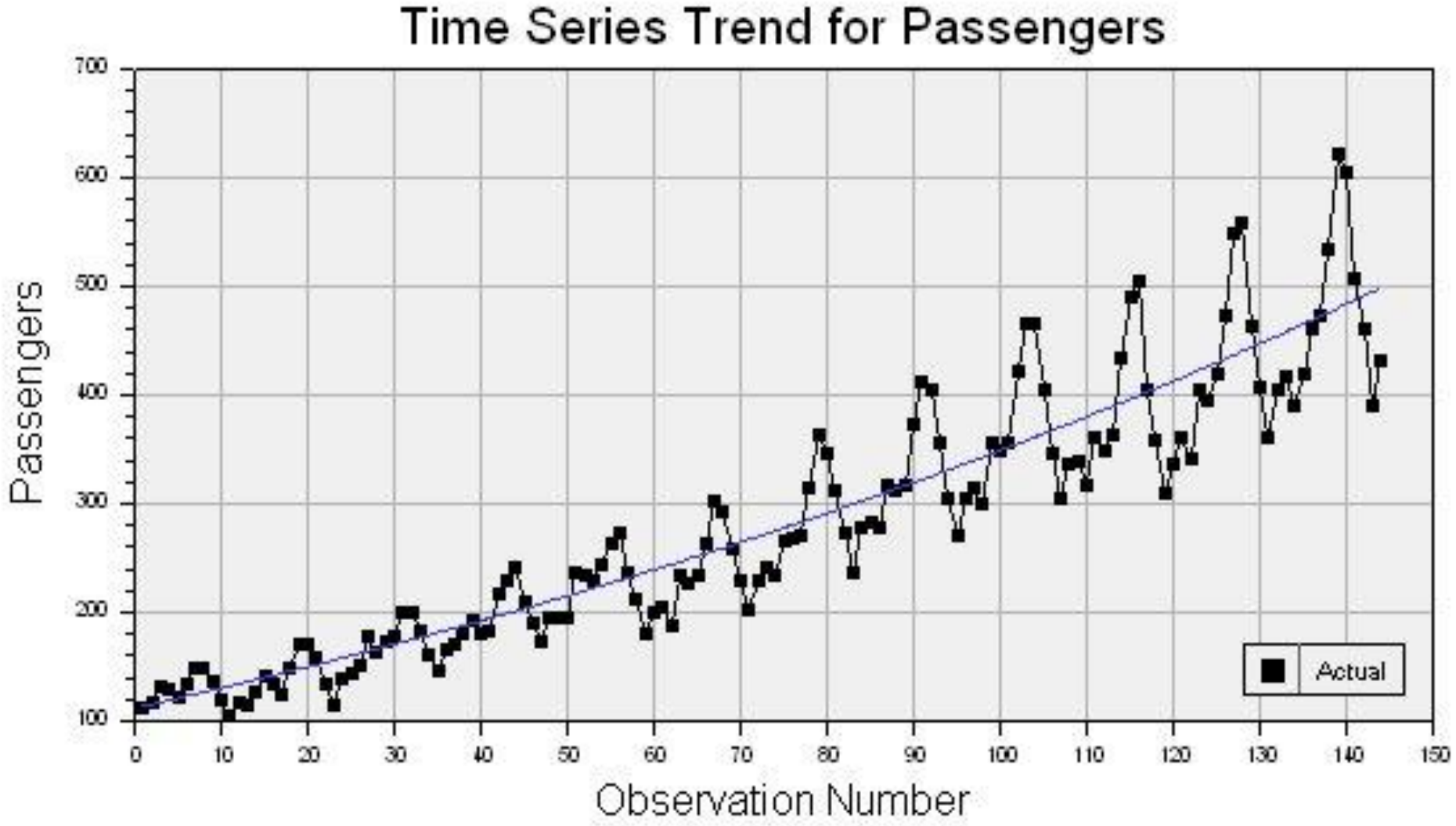
RandomForestClassifier(*n_estimators=10*, *criterion='gini'*, *max_depth=None*, *min_samples_split=2*, *min_samples_leaf=1*, *min_weight_fraction_leaf=0.0*, *max_features='auto'*, *max_leaf_nodes=None*, *min_impurity_decrease=0.0*, *min_impurity_split=None*, *bootstrap=True*, *oob_score=False*, *n_jobs=1*, *random_state=None*, *verbose=0*, *warm_start=False*, *class_weight=None*)


$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$
Y
Observed Value of Y for X_i
Predicted Value of Y for X_i
Intercept = β_0
ε_i
Random Error for this X_i value
Slope = β_1
X_i
X

Linear Regression API

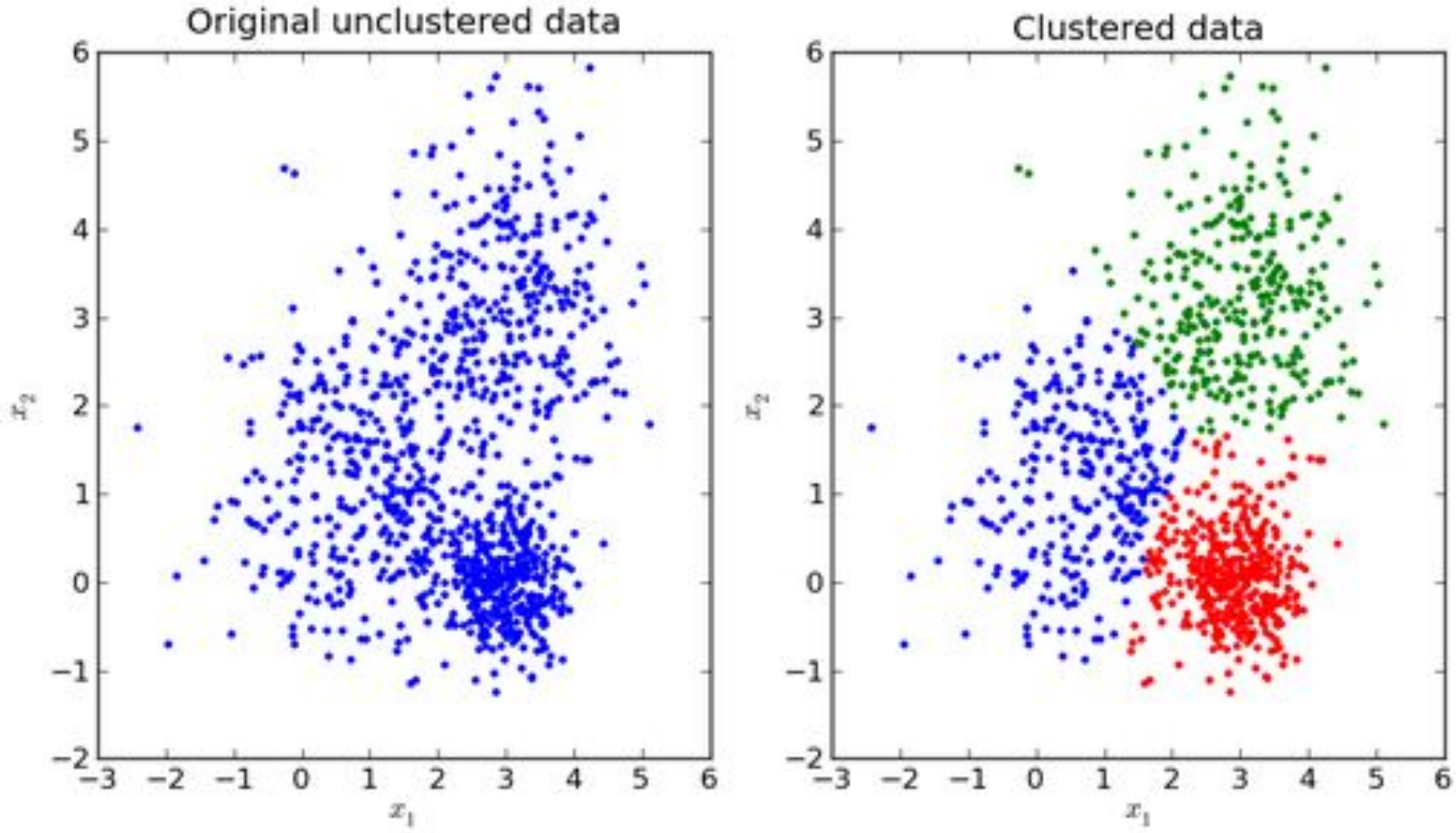
OLS(*endog*, *exog=None*, *missing='none'*, *hasconst=None*, ***kwargs*)

OLS(formula, data, subset=None, *args,**kwargs)


Time Series Trend for Passengers
Passengers
Observation Number
Actual
100
200
300
400
500
600
700
0
10
20
30
40
50
60
70
80
90
100
110
120
130
140
150

Time Series (SARIMAX) API

SARIMAX(*endog*, *exog=None*, *order=(1, 0, 0)*, *seasonal_order=(0, 0, 0, 0)*, *trend=None*, *measurement_error=False*, *time_varying_regression=False*, *mle_regression=True*, *simple_differencing=False*, *enforce_stationarity=True*, *enforce_invertibility=True*, *hamilton_representation=False*, ***kwargs*)


Original unclustered data
Clustered data
x_1
x_2

K-Means API

KMeans(*n_clusters=8*, *init='k-means++'*, *n_init=10*, *max_iter=300*, *tol=0.0001*, *precompute_distances='auto'*, *verbose=0*, *random_state=None*, *copy_x=True*, *n_jobs=1*, *algorithm='auto'*)

Putting It Together

Case Study: Retail Pricing Strategy

Backup Slides