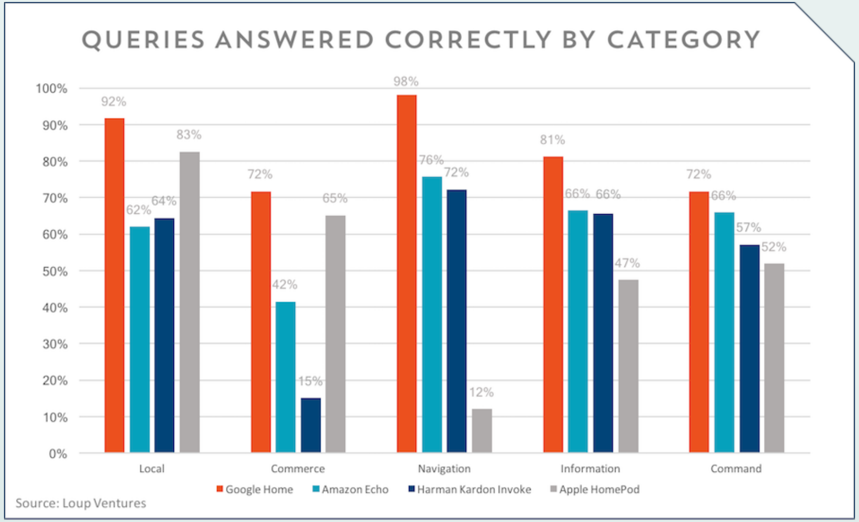
An Intro to Machine Learning

Rouhollah Abolhasani

What is Machine Learning?

Speech Recognition

- Alexa and Siri
  - Extract vowels and consonants from the **signals**
  - Recognize Words, Sentences, and Sentiment
  - Act accordingly.

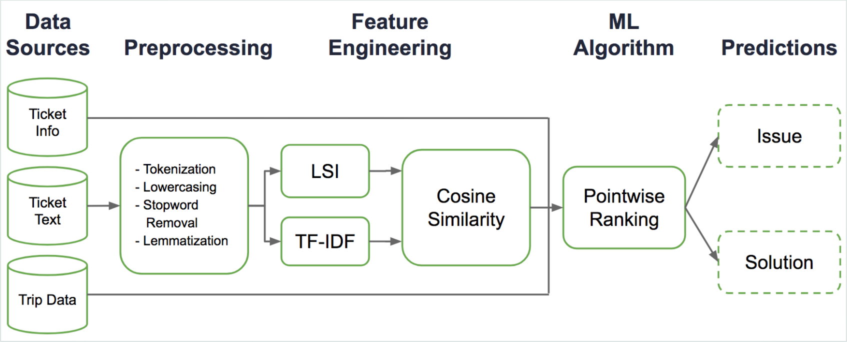

QUERIES ANSWERED CORRECTLY BY CATEGORY

Source: Loup Ventures

Time-series Forecasting

- Time-series
  - Stock Market
  - Texts, Tweets, …

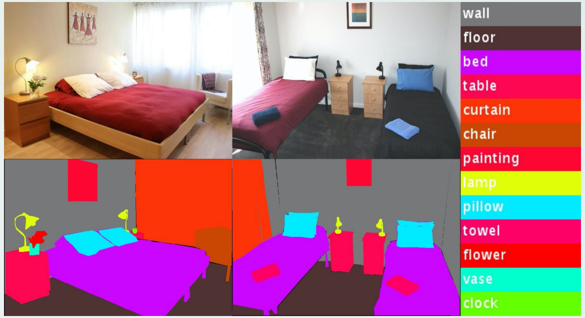
- COTA: Customer Obsession Ticket Assistant
  - Handling Uber support tickets
  - Processing tickets and proposing solutions

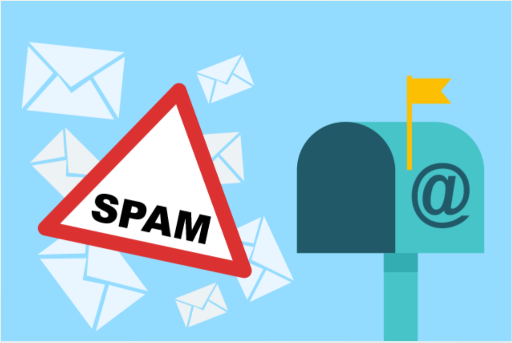


https://eng.uber.com/cota/

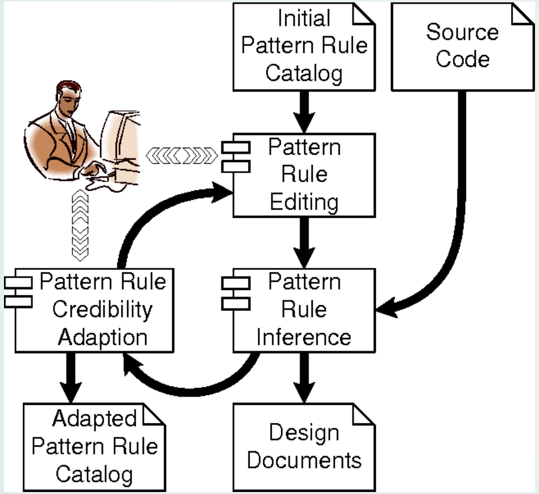
Classification

- Image Classification
  - Cat vs. dog
  - Image segmentation
- Old-fashioned Spam Detector
  - What would we do?
  - Writing rules
  - More rules
  - And more rules …
- Better Spam Detector
  - Automatic feature extraction
  - Model training
  - Classifying new emails





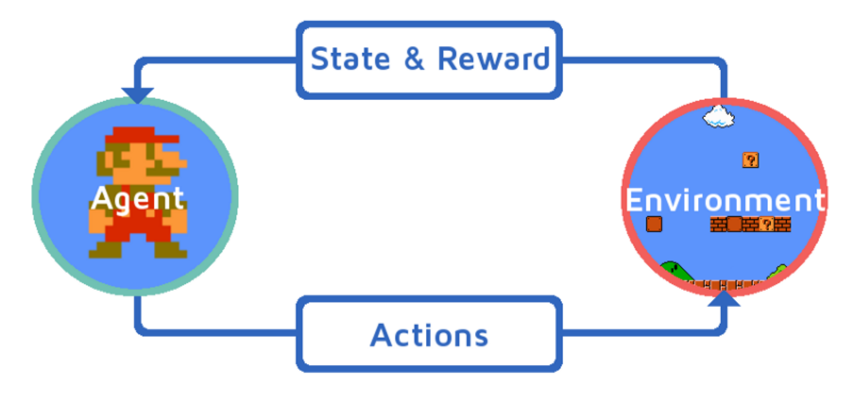
Rule-based vs. Learning



Initial Pattern Rule Catalog
Source Code
Pattern Rule Editing
Pattern Rule Credibility Adaption
Pattern Rule Inference
Adapted Pattern Rule Catalog
Design Documents

Rule –based Programming

Write some rules. Evaluate them. Add some again, and again, …



Learning

Machine programming machine with experience/data.

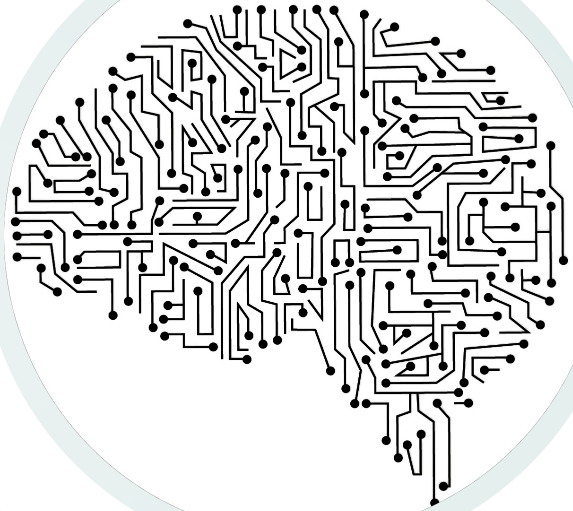
Two Definitions of ML

Arthur Samuel(1959):

- "Machine Learning is a field of study that gives computers, the ability to learn without explicitly being programmed."

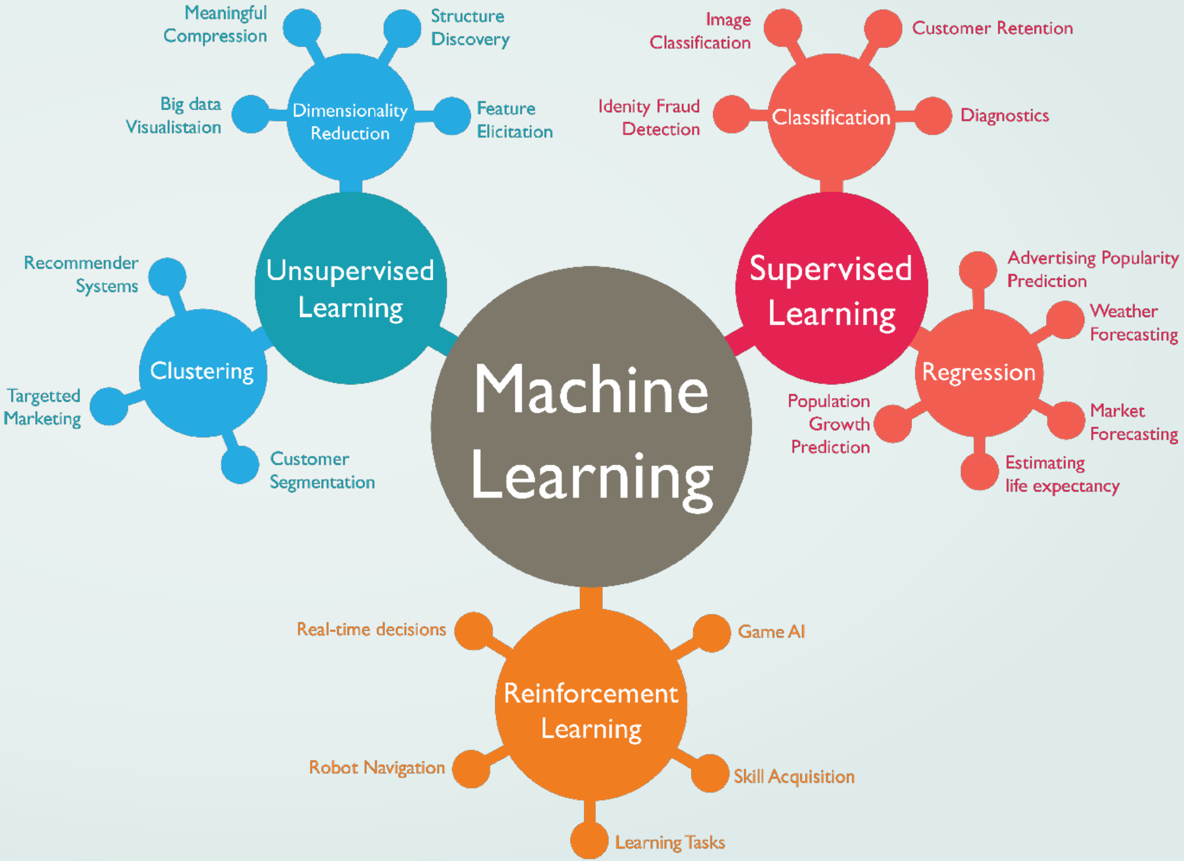
Tom Michel(1999)

- "A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E."

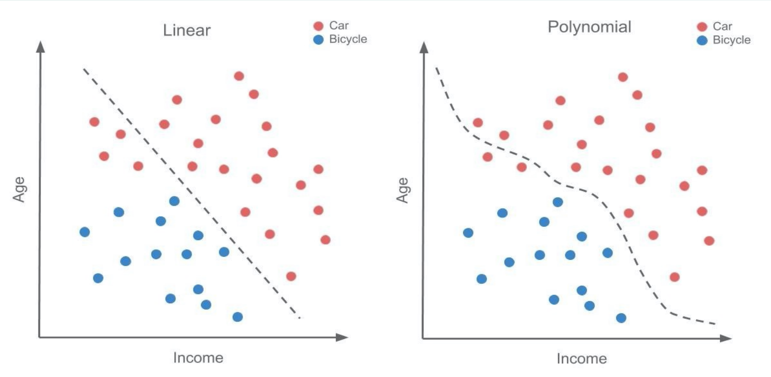


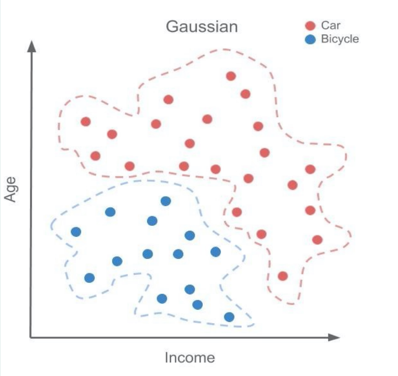
The Big Picture

- Supervised Learning
  - Labeled data
  - Classification
  - Regression
- Unsupervised Learning
  - Unlabeled data
  - Clustering
  - Dimensionality Reduction
- Reinforcement Learning
  - Learning by living
  - Entirely different realm
  - Not covered in this course



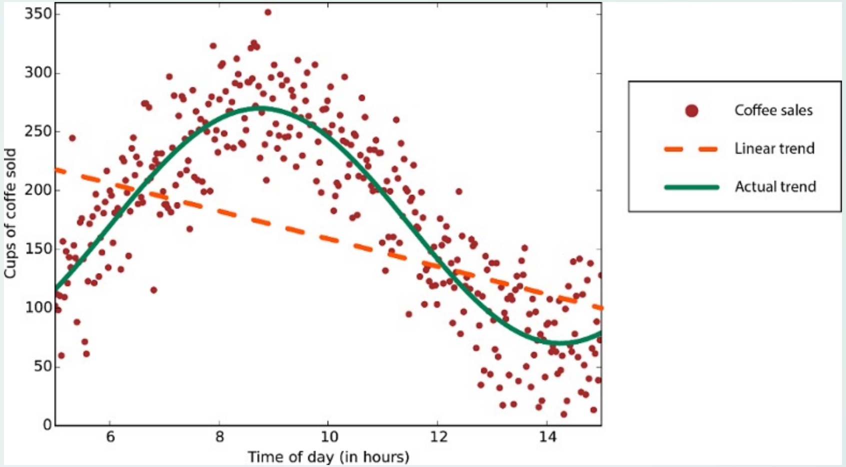
Supervised Learning





Classification

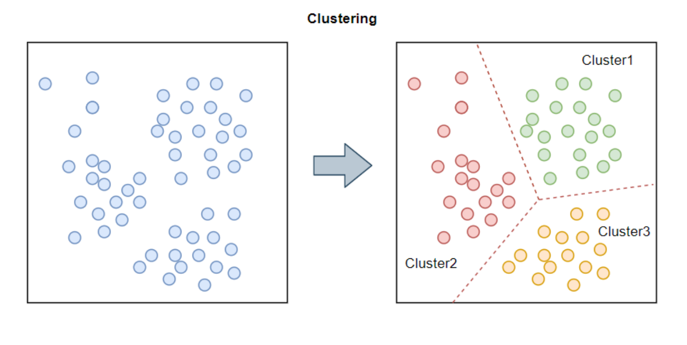
For categorical variables

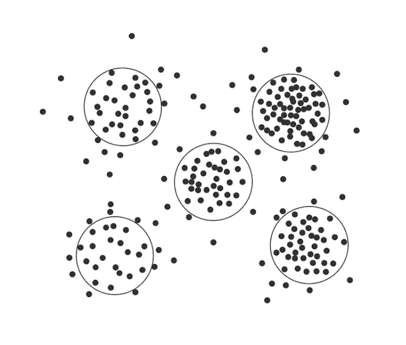


Regression

For numeric variables

Unsupervised Learning

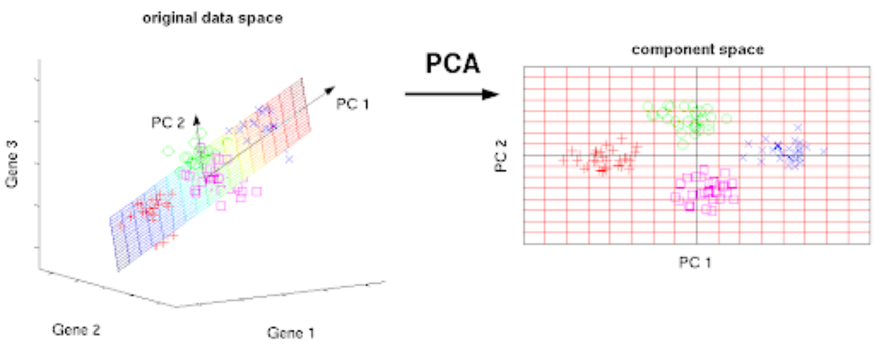




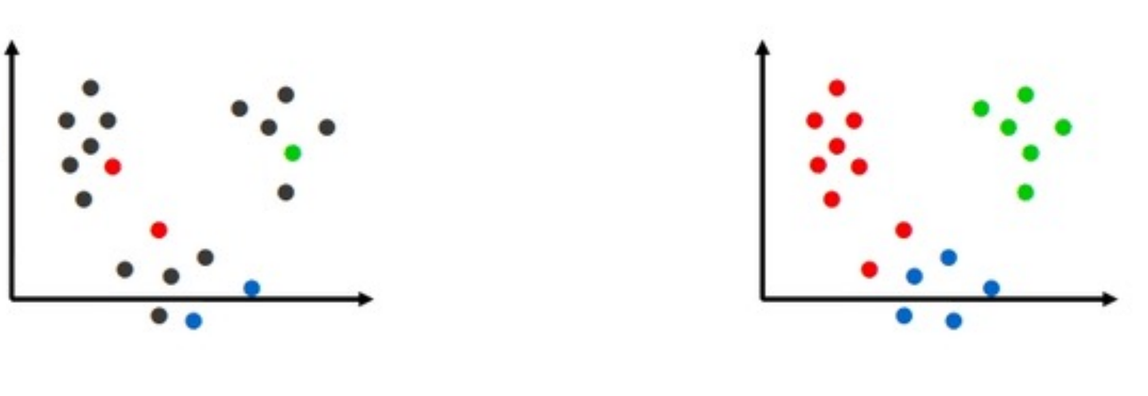
Clustering

Dimensionality Reduction

To avoid the curse of dimensionality



Semi-supervised Learning



Clustering and Classifying

Deep Belief Networks for image classification, video recognition, ...

SciKit-Learn

2

What is SciKit-Learn

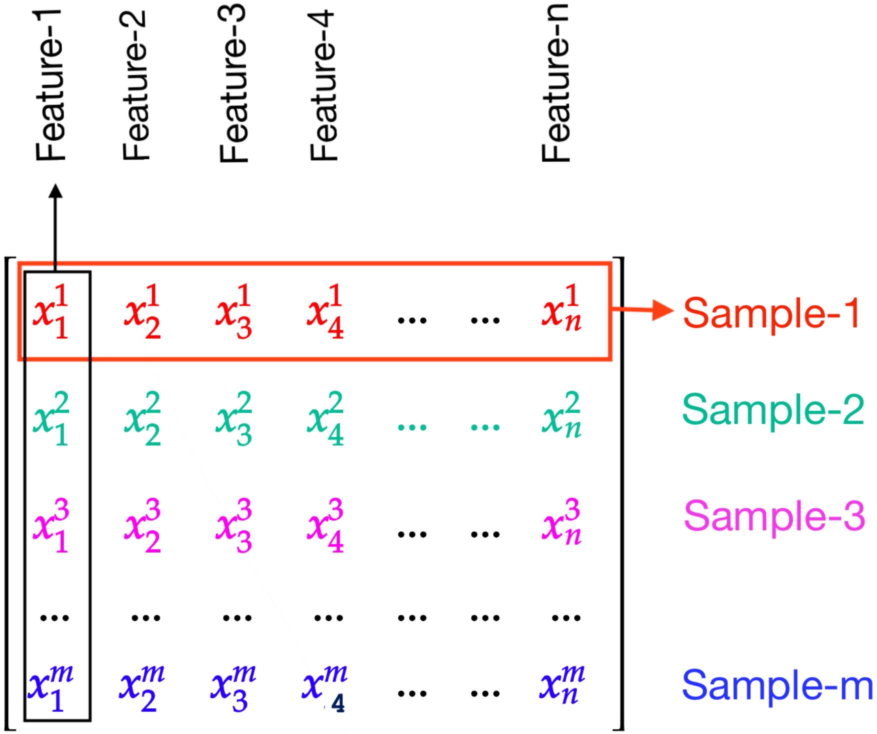
- An open-source machine-learning library
- Built on Numpy, Scipy, and Matplotlib
- Contains many ML algorithms and models
- Good documentation and support
- Good for beginners

Data in Scikit-Learn

- Data In ML typically consists of:
    - Features
    - Labels

- Features are stored in a 2D matrix
    - Shape=(n features, n data)

- Labels are stored in a 1D vector
    - Shape=(n data,)

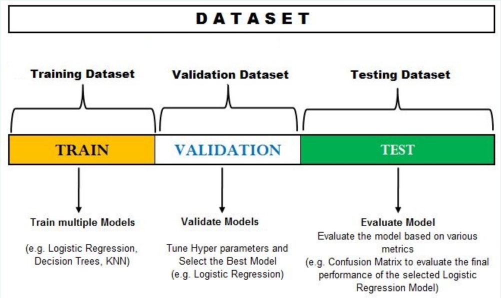


$$\begin{array}{cccccccc}
\text{Feature-1} & \text{Feature-2} & \text{Feature-3} & \text{Feature-4} & & & \text{Feature-n} & \\
x_1^1 & x_2^1 & x_3^1 & x_4^1 & \dots & \dots & x_n^1 & \rightarrow \text{Sample-1} \\
x_1^2 & x_2^2 & x_3^2 & x_4^2 & \dots & \dots & x_n^2 & \text{Sample-2} \\
x_1^3 & x_2^3 & x_3^3 & x_4^3 & \dots & \dots & x_n^3 & \text{Sample-3} \\
\dots & \dots & \dots & \dots & \dots & \dots & \dots & \\
x_1^m & x_2^m & x_3^m & x_4^m & \dots & \dots & x_n^m & \text{Sample-m}
\end{array}$$

Train-Test Data

- We always split data to:
  - Train data
  - Validation data
  - Test data: We only see in production
- Cross-validation
  - Holdout
  - K-fold
- In Scikit-Learn:

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



Train multiple Models

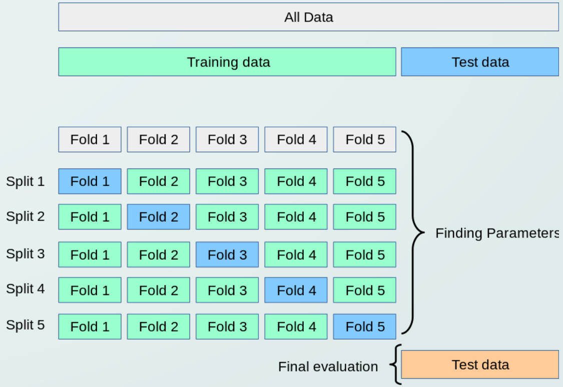
(e.g. Logistic Regression, Decision Trees, KNN)

Validate Models

Tune Hyper parameters and Select the Best Model (e.g. Logistic Regression)

Evaluate Model

Evaluate the model based on various metrics (e.g. Confusion Matrix to evaluate the final performance of the selected Logistic Regression Model)



All Data

Training data | Test data

	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
Split 1	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
Split 2	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
Split 3	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
Split 4	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
Split 5	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5

Finding Parameters

Final evaluation — Test data

Some Basic Regression Model

- Using LinearRegression model

```
from sklearn.linear_model import LinearRegression
```

- Constructing the model

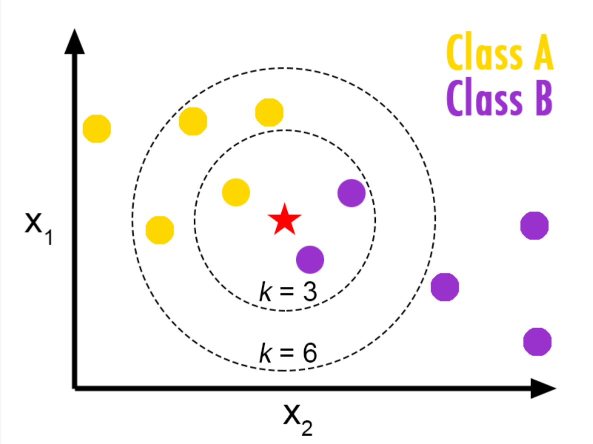
```
reg = LinearRegression()
```

Model API In Scikit-Learn

- Train the model using .fit(X, y) method
- Transform the data using .transform(X) method
- Do the combination using .fit_transform(X, [y])
- Test the model using .predict(X) method

K-Nearest Neighbor Classifier

- For each new data:
  - Find the k-closest data
  - Take majority vote
  - Assign the label


Class A
Class B
x_1
x_2
$k = 3$
$k = 6$