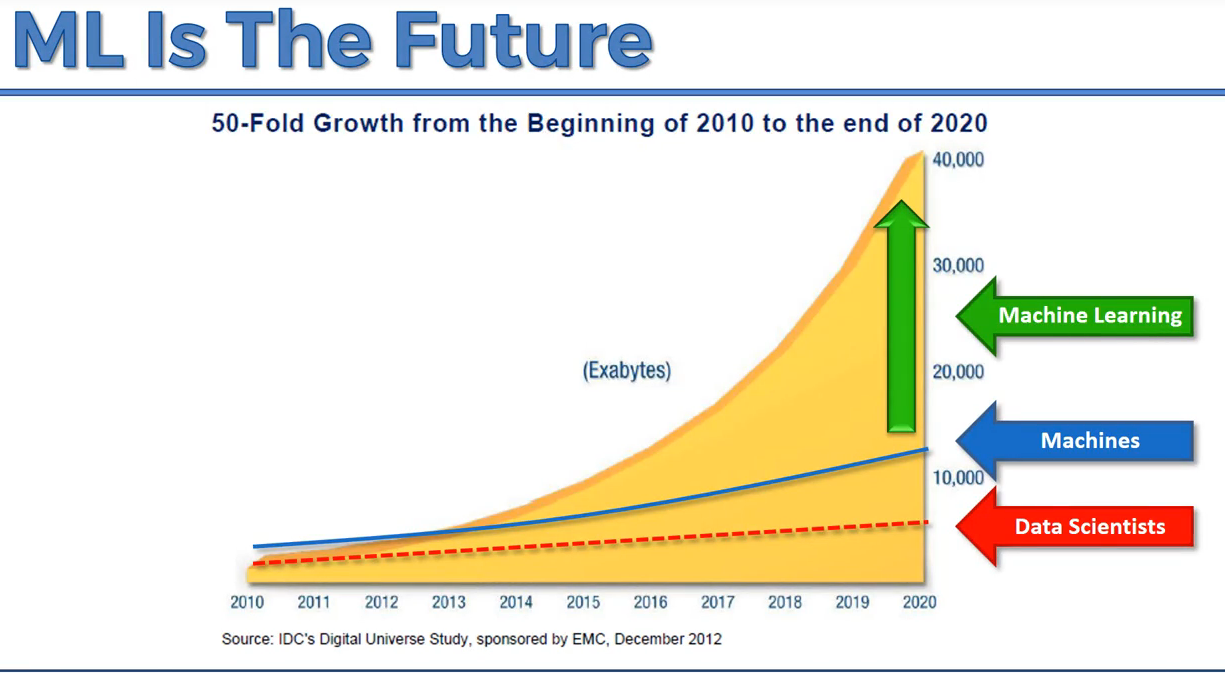
**Applications of ML:**

1. Facebook face tagging
2. Motion sensor
3. Virtual Machine
4. Voice recognition
5. Robot dogs(Reinforcement Learning-Robots walk on their own)
6. Facebook Ads
7. Amazon, Netflix
8. Medical Purpose to save lives
9. Maps
10. Satellites

**ML is the future:**

* We are generating huge amount of data (exponential Growth)
* 2005🡪 130 Exabytes of Data. 2010 🡪1,200 Exabytes. 2015🡪7900 Exabytes. 2020🡪 4000 EB.



**Data Preprocessing:**

Without Preprocessing, the model won’t work well.

* Get the dataset
* Importing the libraries
* Import the Dataset
* Missing Data
* Categorical Data
* Splitting the Dataset into the training set and Test Set
* Feature Scaling
* Data Preprocessing Template

**Missing Data:**

1. To **remove the row**, but it is quite dangerous situation to loose information
2. To take **mean** to fill the missing value.

**Categorical Variable:**

Country and Purchased are categorical Variable

Since Machine Learning model are mathematical, we cannot fit the text .

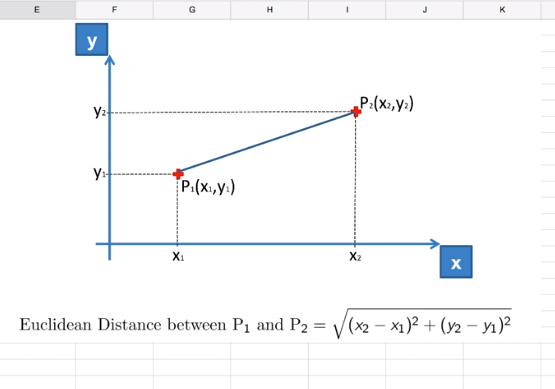
So we will convert the text into numeric. <http://queirozf.com/entries/one-hot-encoding-a-feature-on-a-pandas-dataframe-an-example>

**Splitting the Dataset:**

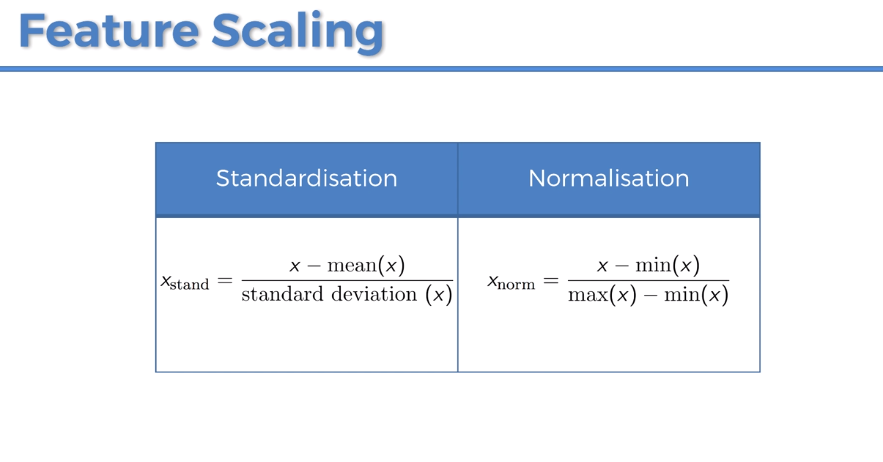
1. For training and testing the system.
2. Correlation based on the train and test set. The better system learns it predicts the test set.
   1. If it byhearts the data not understanding, it cannot predict well(**Over fitting**🡪 use **regularization** to prevent it)

**Feature Scaling:**

1. Euclidean Distance(Distance Formula) will be calculated by the system.



There are two types:



To improve the run time and performance.

Questions:

1. Do we need fit and transform dummy variable? **Yes**

2. Do we need to fit and scale Y\_test and Y\_train? **No**