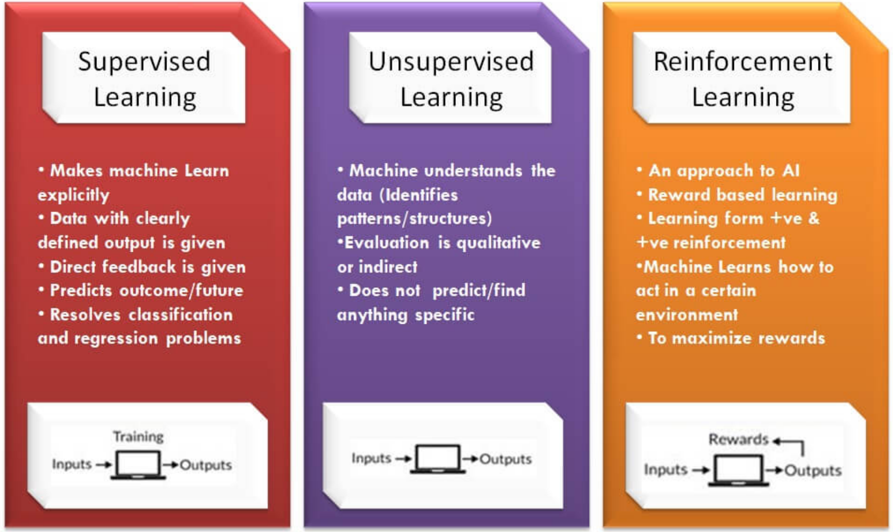
```r
library(dplyr)

rladies_global %>%
  filter(city == ’Johannesburg')
```

House prices :
basic EDA & prediction

Type of Machine Learning



Supervised Learning

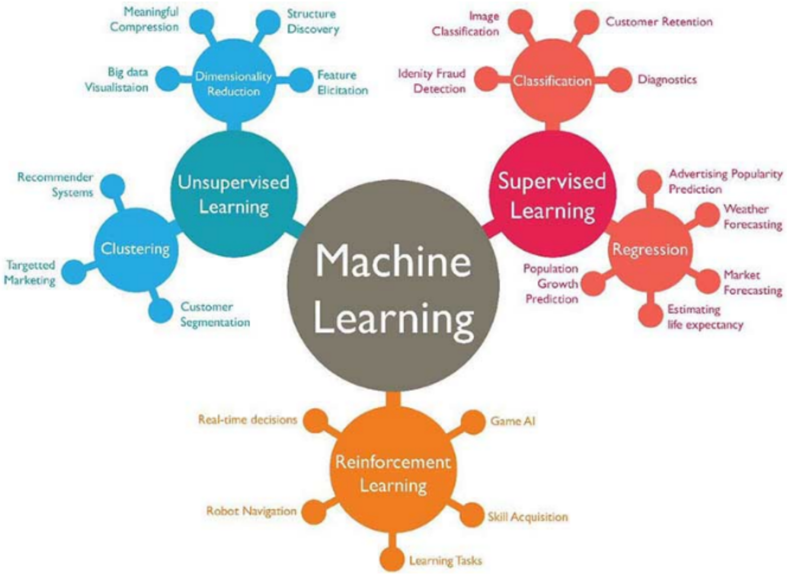
- Makes machine Learn explicitly
- Data with clearly defined output is given
- Direct feedback is given
- Predicts outcome/future
- Resolves classification and regression problems

Unsupervised Learning

- Machine understands the data (Identifies patterns/structures)
- Evaluation is qualitative or indirect
- Does not predict/find anything specific

Reinforcement Learning

- An approach to AI
- Reward based learning
- Learning form +ve & +ve reinforcement
- Machine Learns how to act in a certain environment
- To maximize rewards



CheatSheets: https://www.analyticsvidhya.com/blog/2015/09/full-cheatsheet-machine-learning-algorithms/

1.
Overview

House prices

source: https://www.kaggle.com/c/iowa-house-price-prediction

- "Ask a home buyer to describe their dream house, and they probably won't begin with the height of the basement ceiling or the proximity to an east-west railroad. But this playground competition's dataset proves that much more influences price negotiations than the number of bedrooms or a white-picket fence.

- With 79 explanatory variables describing (almost) every aspect of residential homes in Ames, Iowa, this competition challenges you to predict the final price of each home.

- The potential for creative feature engineering provides a rich opportunity for fun and learning. This dataset lends itself to advanced regression techniques like random forests and gradient boosting with the popular XGBoost library."

- Goal: Predict sale price for each house. For every id in the test set, predict the SalesPrice variable

2.
About the Data

Data Loading & Preparation

- Load Libraries
  - library(readr)
  - library(ggplot2)
  - library(gridExtra)
  - library(tabplot)
  - library(lsr)
  - library(corrplot)
  - library(dplyr)
  - library(magrittr)
  - library(caret)

- Load data from csv
- Understand data
  - Get factor levels
  - Check factor levels
  - Fix level names
  - Convert column data types

3.
Visualisation

Lets see what we can see in the data

- Histogram
- Plot all features sorted by SalesPrice
- Correlation of variables
- Ordinal vs continuous vs nominal against predictor variable SalesPrice

4. Pre-processing

What data can be fixed

- Understand missingness
- Imputation of missing data
- Transformation of data
- Near zero variance checks

5.
Model training & parameter tuning

Lets get to the good part

- Splitting data
- Train set
- Test set

7.
Summary

Story-telling

- How did you begin, what problem are you solving for,
- describe your data, approach that you use (supervised vs unsupervised)
- What patterns emerged before you got to modelling
- Analyse the results/ouputs
- Add context to the visualisations produced
- Include your recommendation on future work that can be done/opinion on results