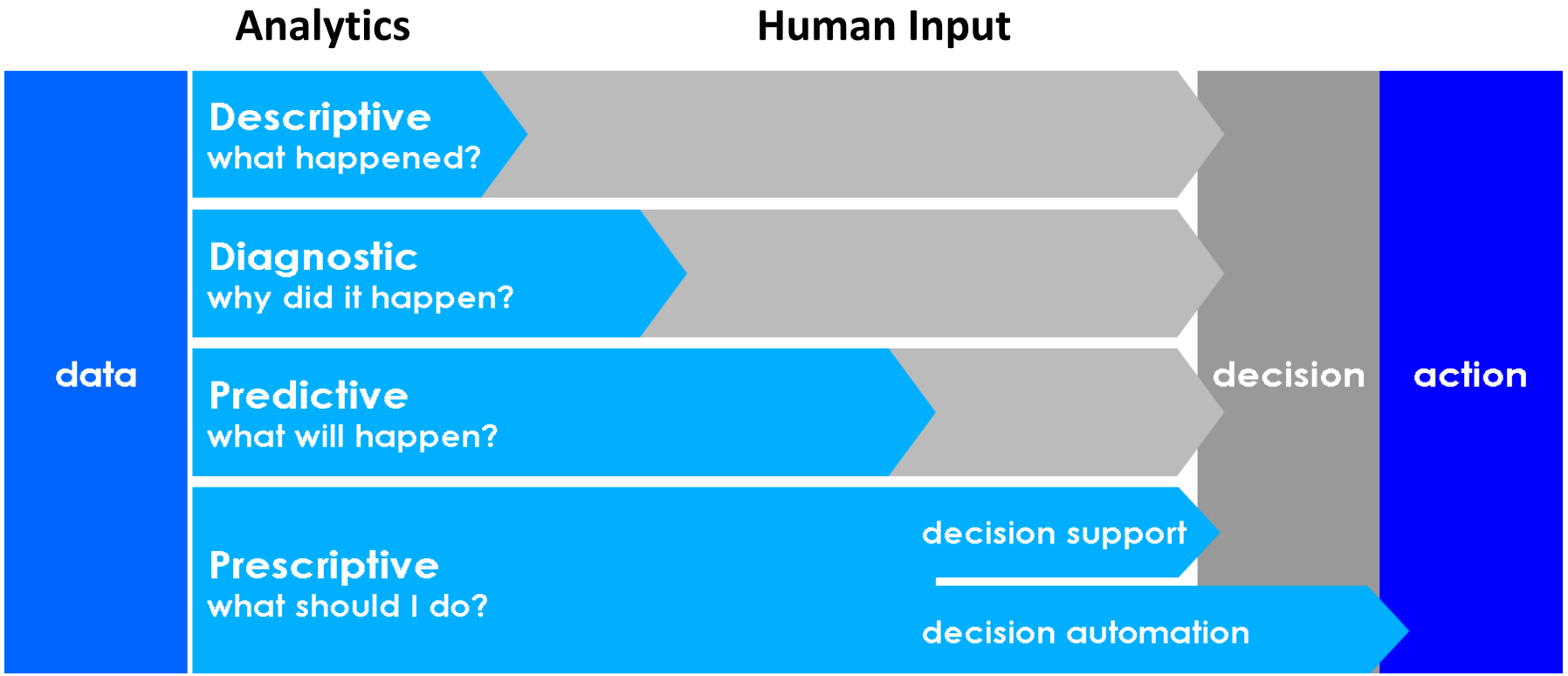
Practice Supervised Modelling

Red & White Consulting Partners LLP

Table of Content

Recap of Introduction to Predictive Modelling

Common Misunderstanding



Predictive analytics isn't the 'holy grail' of analytical culmination – it's about **optimization** and **not problem-solving**.

Problem-framing and **diagnostic** is **key to problem-solving**.

Figure out **root causes**. Map out which decisions you want to improve.

Common Misunderstanding

- Predictive models **do not solve problems**
- Predictive models are **solution optimization tools**
- Prediction is not about the future; **can be about the present**
- Not everything is **worth predicting**

Most Predictive Models are all about Correlation

WHEN CORRELATION IS ALL YOU NEED

Gaining an advantage from the outcome

- You don't need to care how you got there, but being 'there' gives a competitive advantage

Classic examples:

✓ Marketing propensity / response models

✓ Credit underwriting models

WHEN CAUSATION MATTERS

Need to intervening or prevent the outcome

- You care how you get there because you want to avoid it

Classic examples:

✓ Anti-attrition models

✓ Performance decline models

Classification vs Prediction

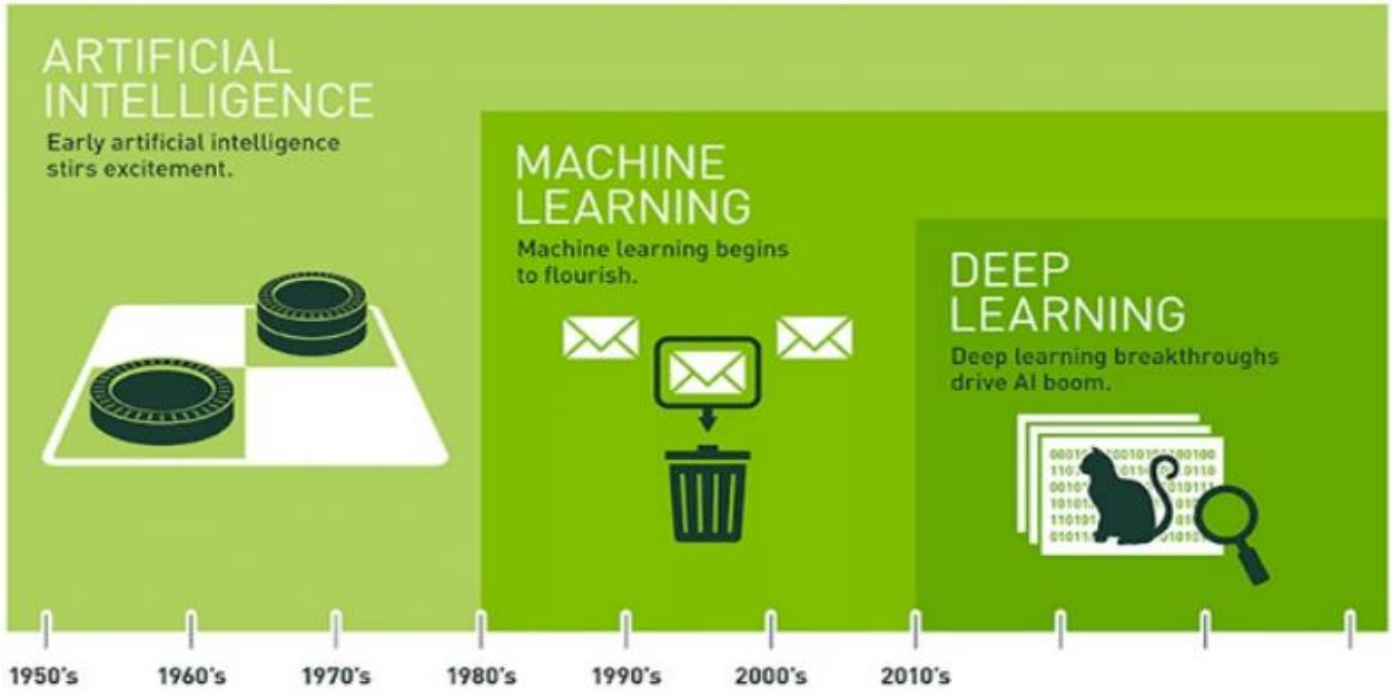
CLASSIFICATION

- A statement about today
- “Who are my top and bottom performers?”
- Classification models aim to remove uncertainty about a current state so that appropriate reaction can be taken.
- Model output can be discrete or continuous

PREDICTION

- A statement about the future
- “Who will default on a loan?”
- Provides the opportunity to take intervention for the future – exploit the predicted outcome or change the trajectory of the predicted outcome
- By analysing past and current data, predictive modelling helps to predict future outcomes
- Model output can be discrete or continuous

AI Vs Machine Learning Vs Deep Learning



AI

AI starts when scientists want to create a complex algorithm to perform task equally or better than a human's capacity.

Machine Learning

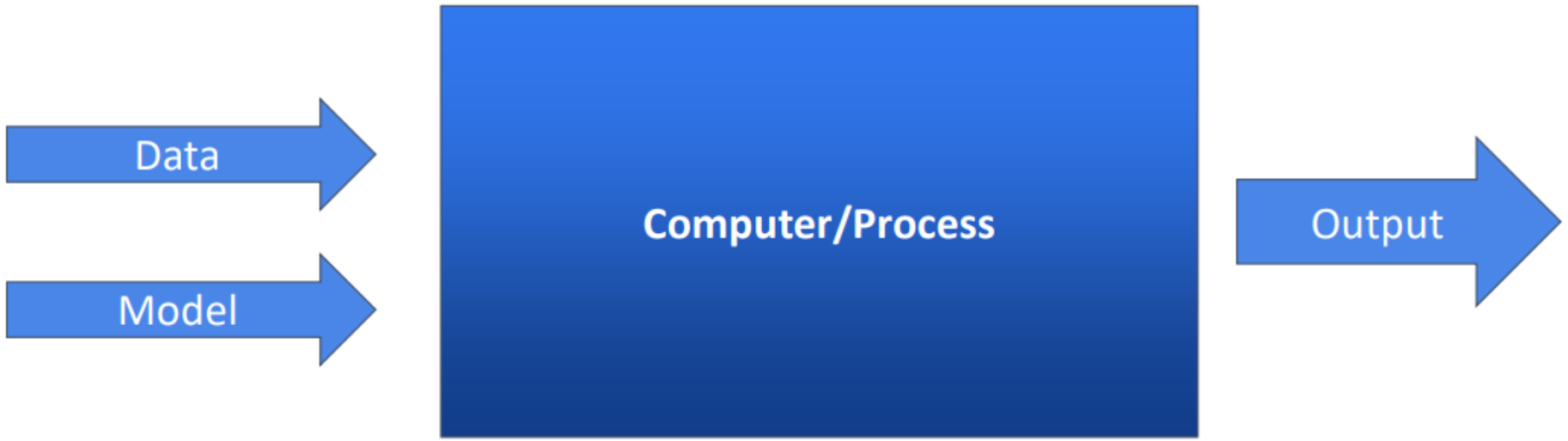
The practice of applying algorithms in such a way that it enables the model to perform task such as classification, prediction, etc.

Deep Learning

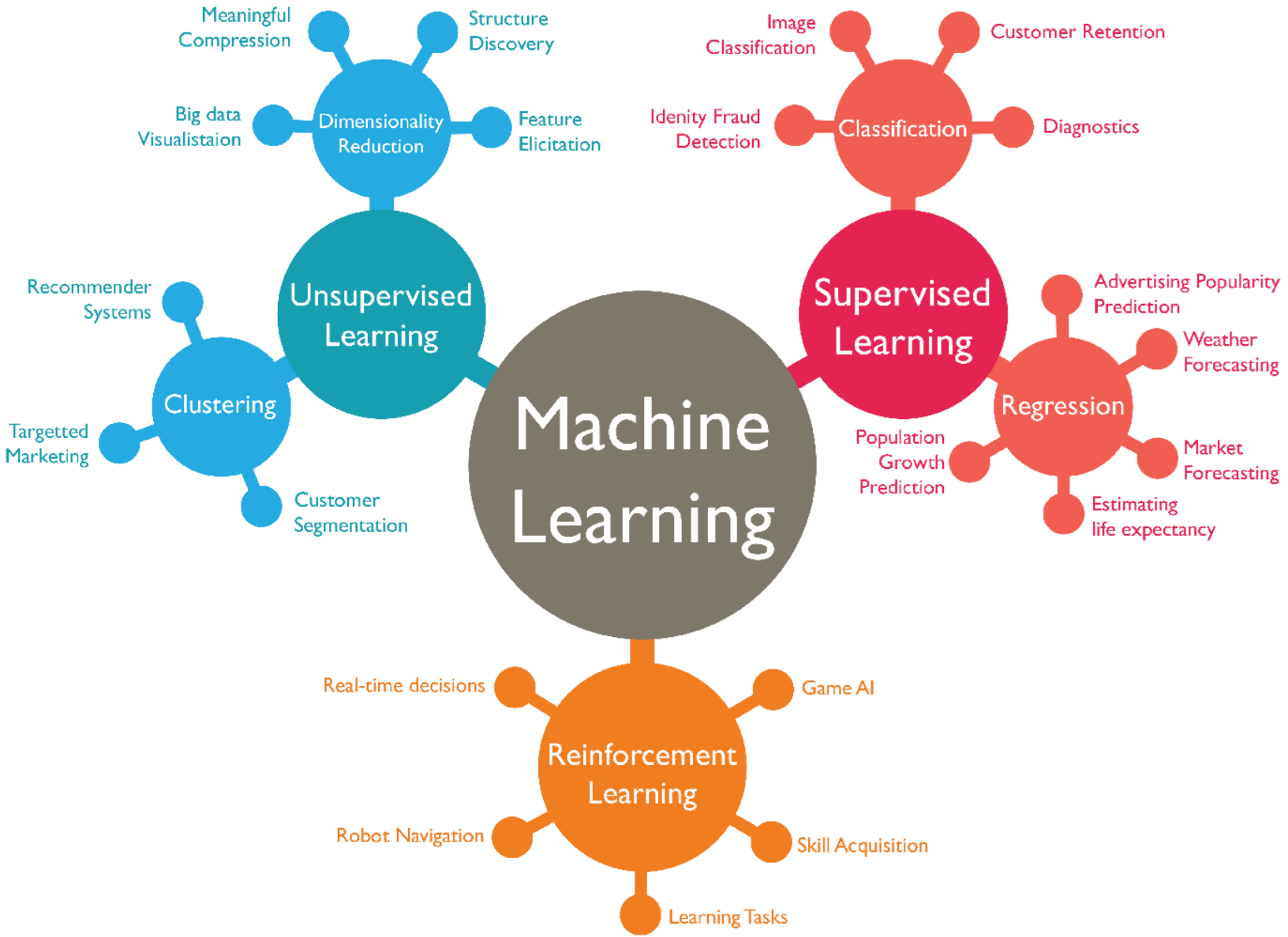
Application of model to perform task by making a very complex and putting very big data (millions or billions data) to make the model learning "very deeply".

Machine Learning 101

- By putting the model and data we provide, this machine learning will "learn" the data and performing task which we want to be the output.



Types of Machine Learning



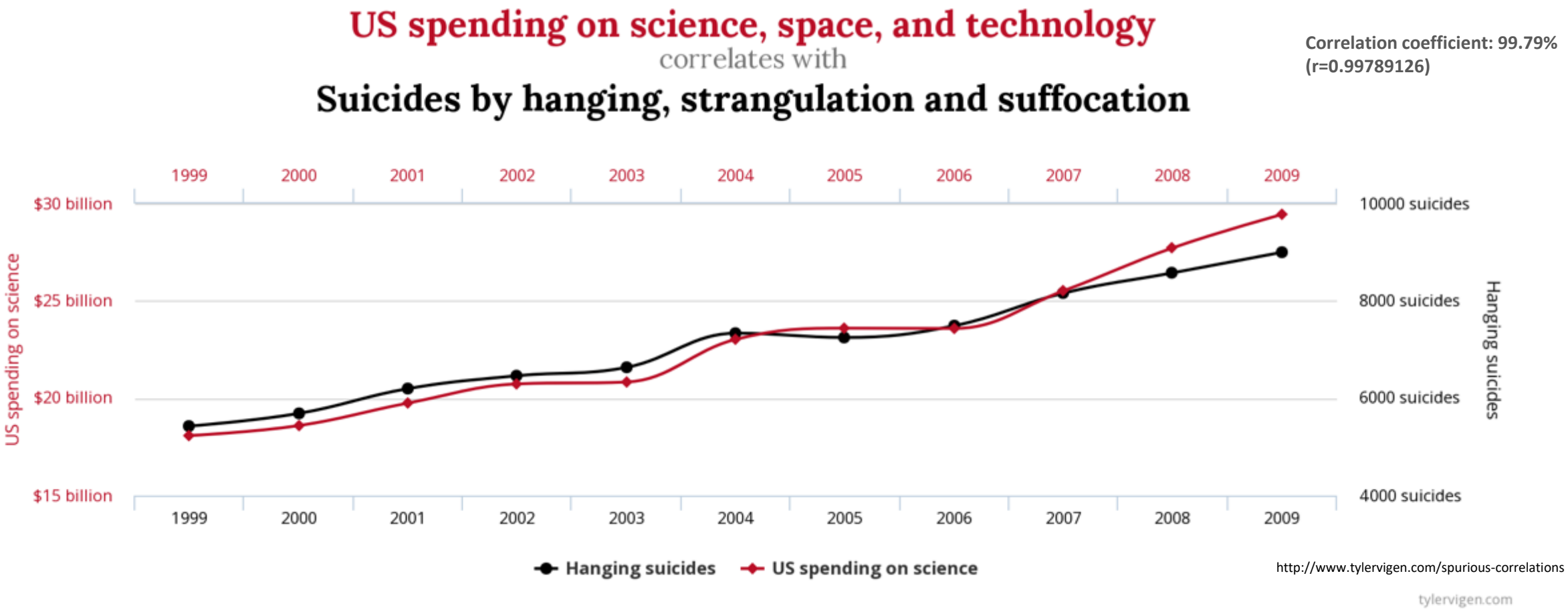
Understanding Bias in Data

ALL EXISTING ORGANIZATION DATA IS BIASED:
They are constrained by how the business operates (*e.g. only certain segments are represented)*

EXISTING DATA	MAKING NEW DATA (IF DATA IS NOT AVAILABLE)
• Is the data sufficient? • What is the bias in the data? • Should the data be sampled? • Must the data be transformed?	• Can I find proxy data? • Can I collect data through observations - e.g. time motion study? • Can I collect data through surveys & questionnaires? • Can I conduct experiments to create the necessary contextualised data?

More Data -> More Spurious Correlation

Figure: US sector spending on science correlated against suicides



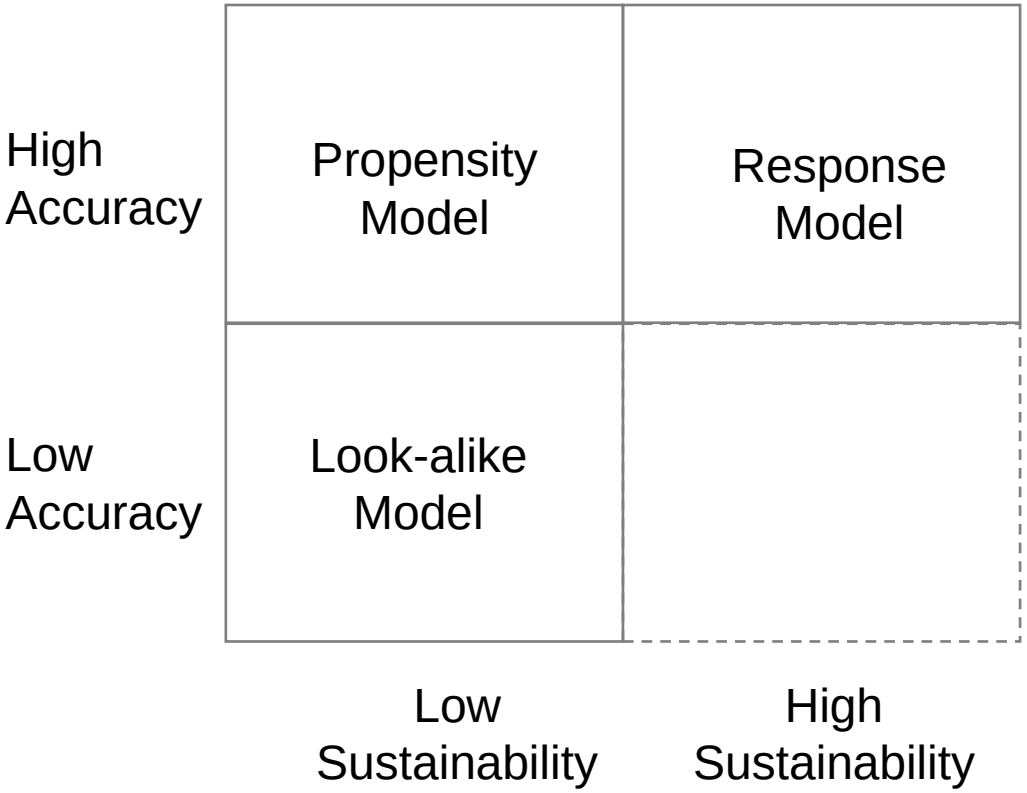
Additional article on spurious correlation

- https://hbr.org/2015/06/beware-spurious-correlations

Website capturing various spurious correlations

- https://www.tylervigen.com/spurious-correlations

Model Accuracy and Sustainability



MODEL ACCURACY

- Ability to separate between target variable and non-target variable

Model Sustainability

- How long will the model be accurate?
- Models decay due to changes in environment and target population, causing a break in correlations
- If models decay faster, they need to be refreshed / rebuilt

Validating Predictive Models

How do we know the predictive model works well? We conduct 2 types of validation – in-time validation and out-of-time validation.

In-time validation	Out-of-time validation
• Only use a sample of the data during modelling process (hold back 30-50% of the data which contains 'good' and 'bad' accounts)	• Select data from another time period (preferably a more recent one)
• Predictive model is built on 50-70% of the data	
• Derived formula from the model is tested to see if it predicts similarly well on the hold-back data	• Derived formula from the model is tested to see if it predicts well

Coffee Break

10:00 - 10:15

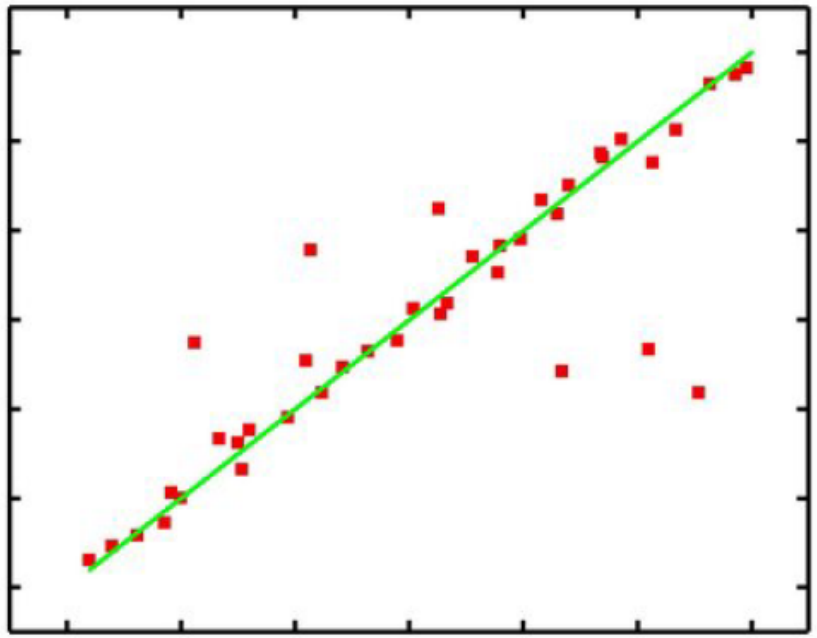
Supervised Learning: Linear Regression

What is Regression

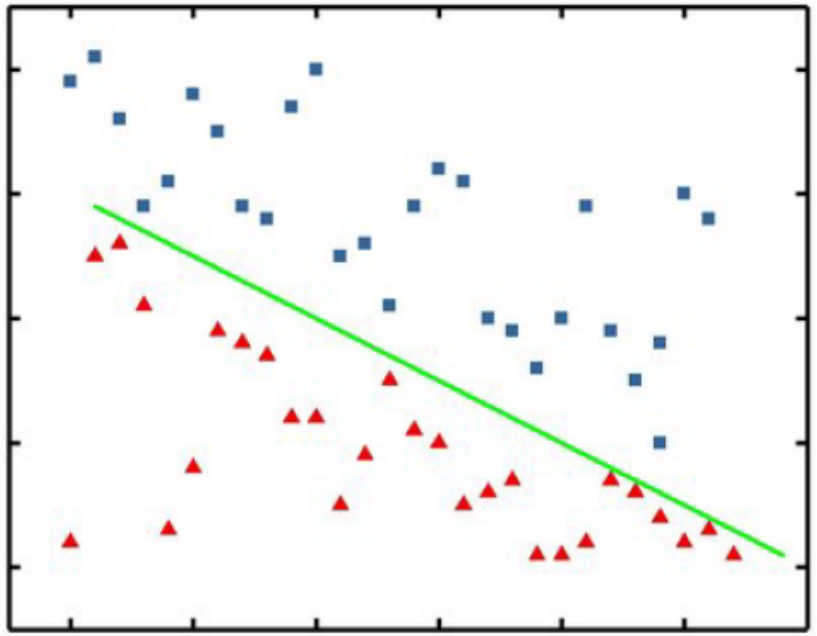
- Regression is an analysis where it analyzes relationship between 2 or more variables by creating a line to figure out the relationship

Purpose of Regression

- Regression can be used for two different purposes:

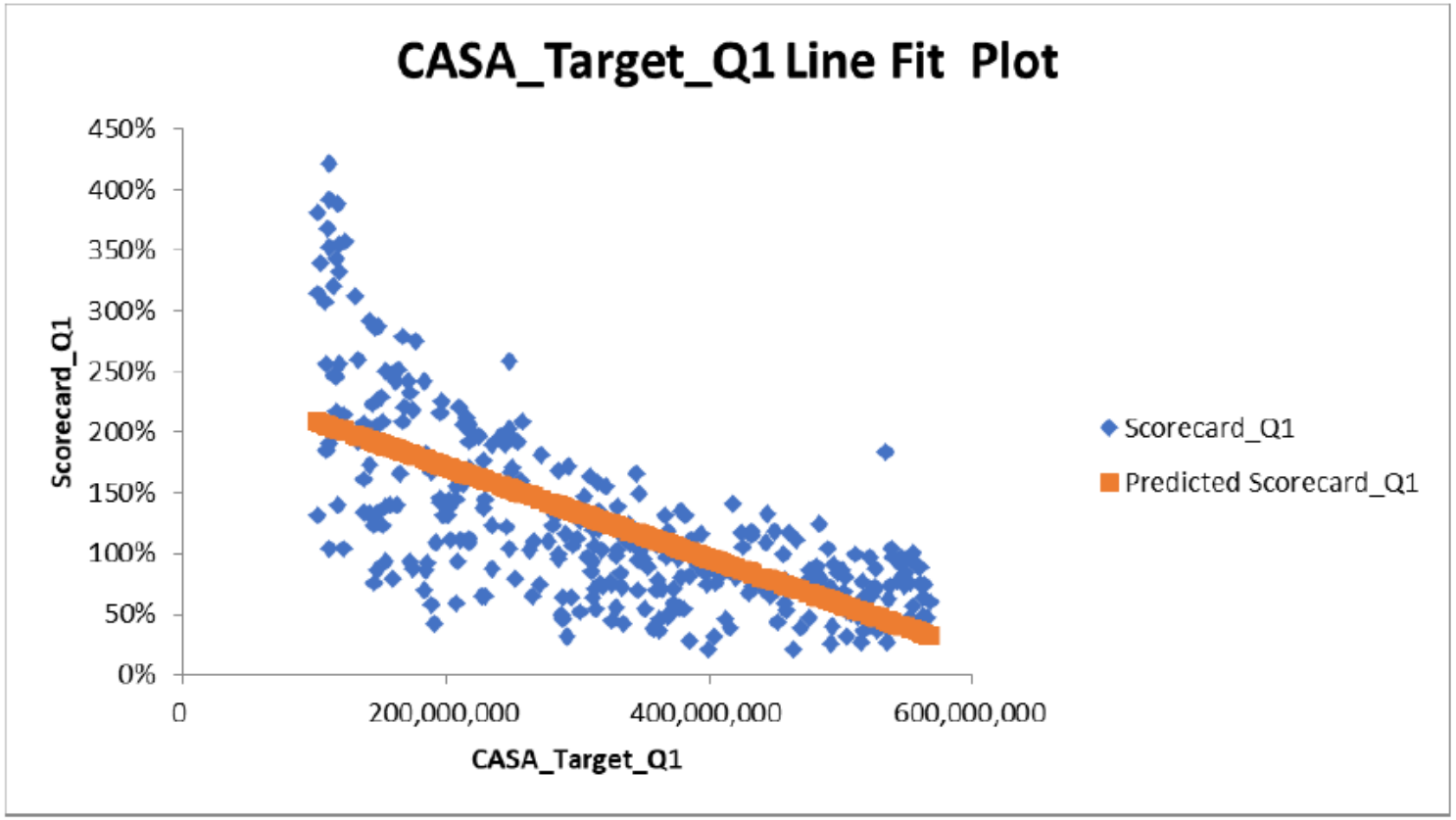


Prediction



Classification

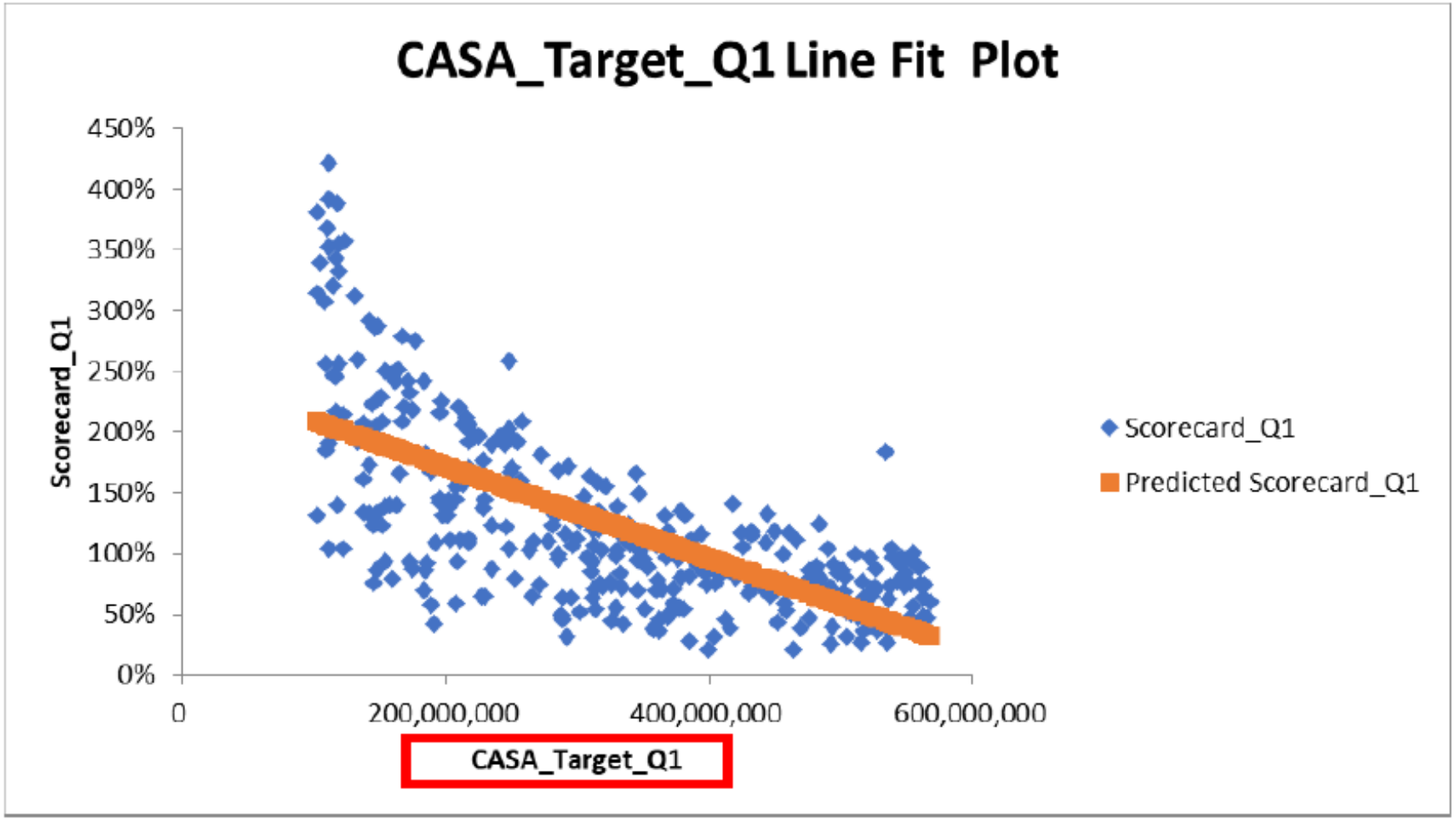
Purpose of Regression: Prediction



- Regression can be used for predicting two variables when we are given values of the other variables
- Examples:
  - Linear Regression
  - Non-Linear Regression

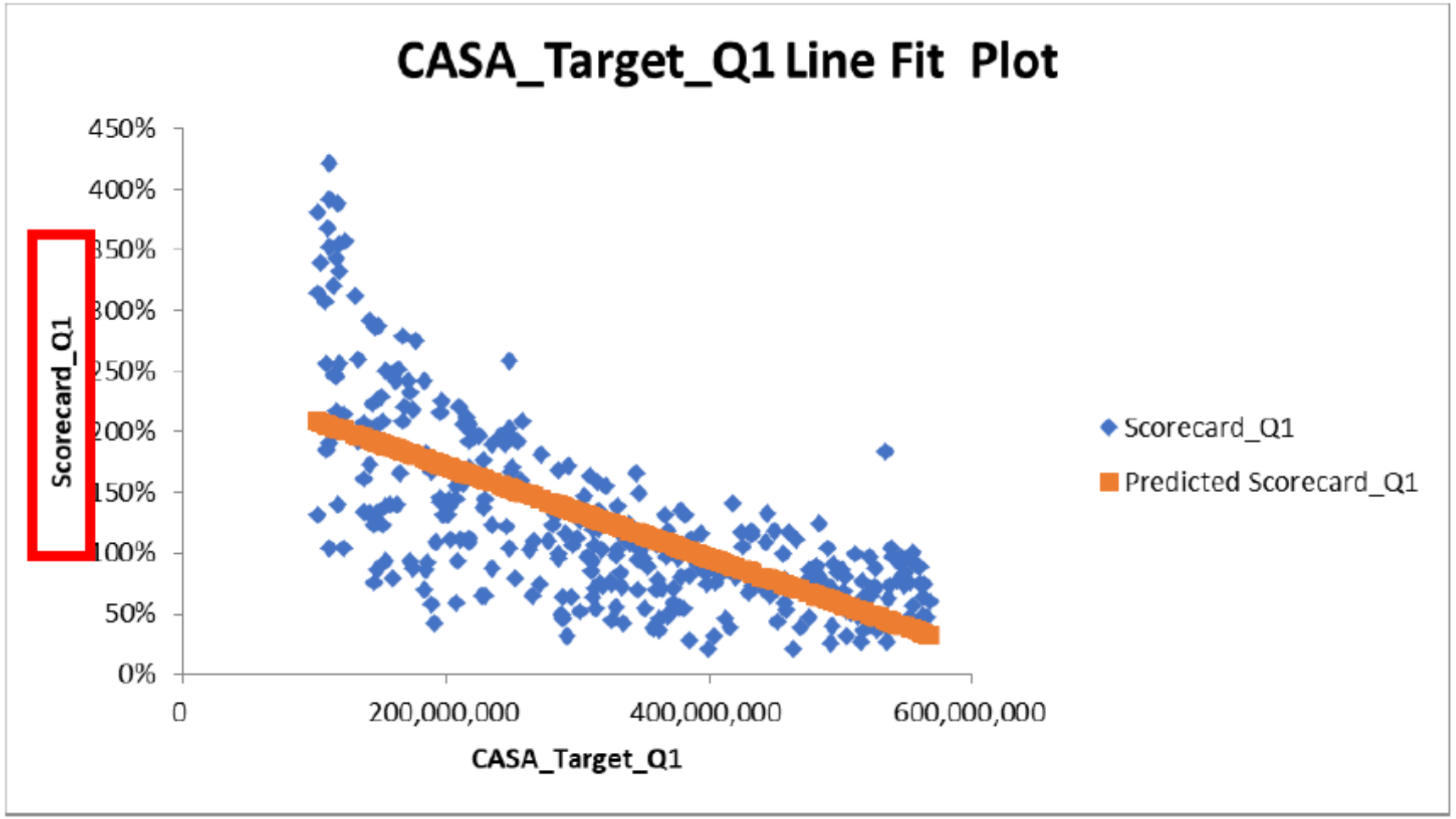
Purpose of Regression: Prediction

- The values which we have the values are called **independent variable.**

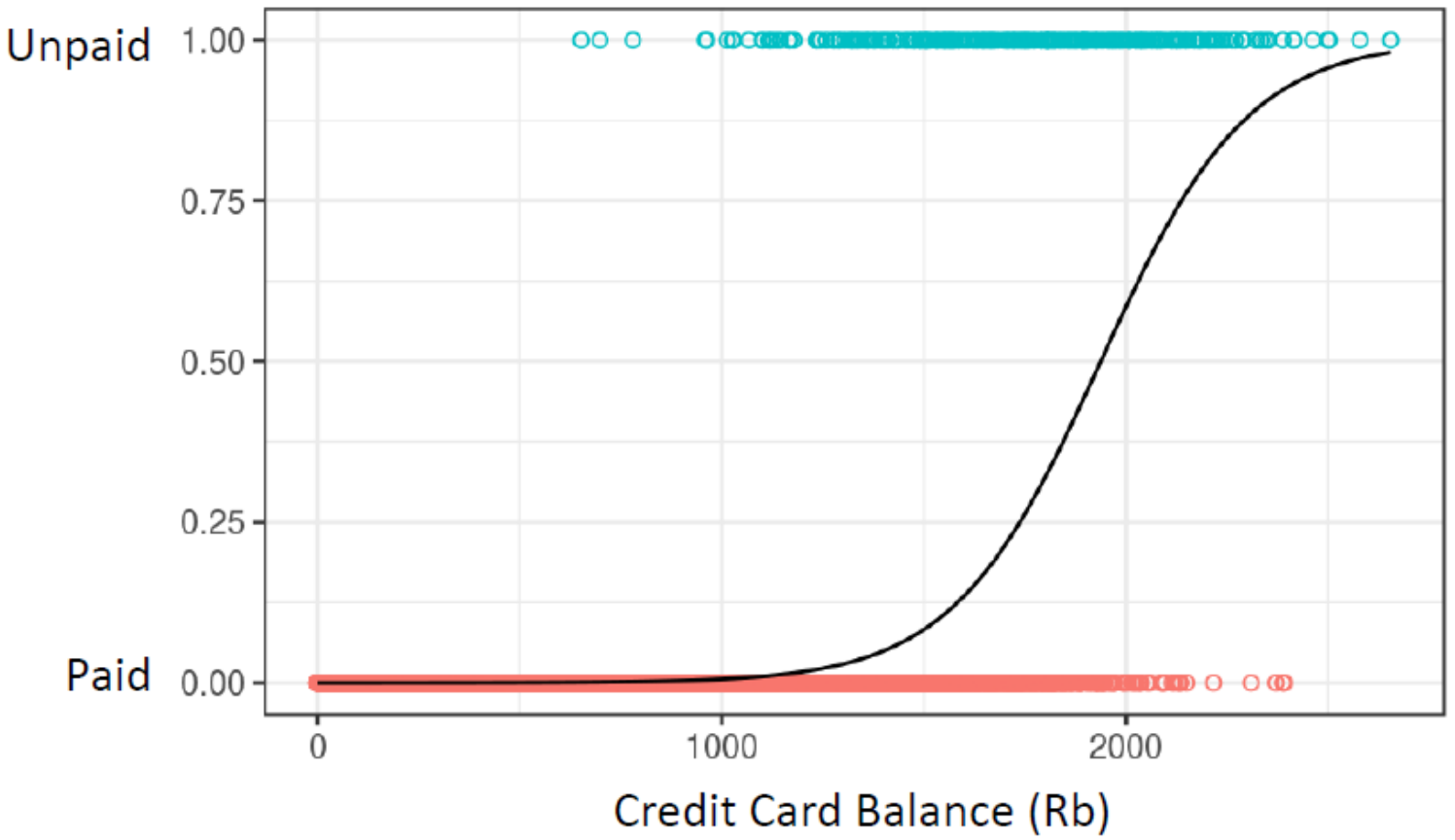


Purpose of Regression: Prediction

- The values which we want to predict are called **dependent variable.**



Purpose of Regression: Classification

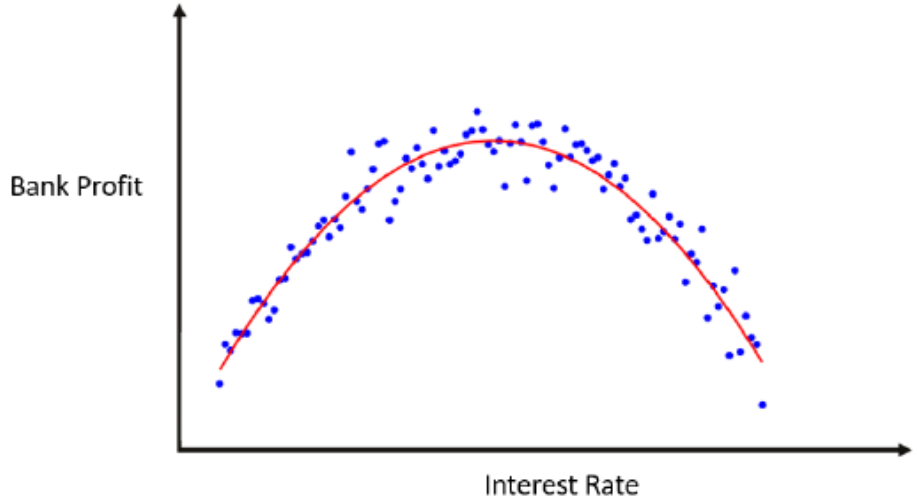


- The other purpose of Regression is to classify between two separate events and counting the likelihood of each data belongs to.
- Examples:
  - Logistics Regression

What is Linear Regression?

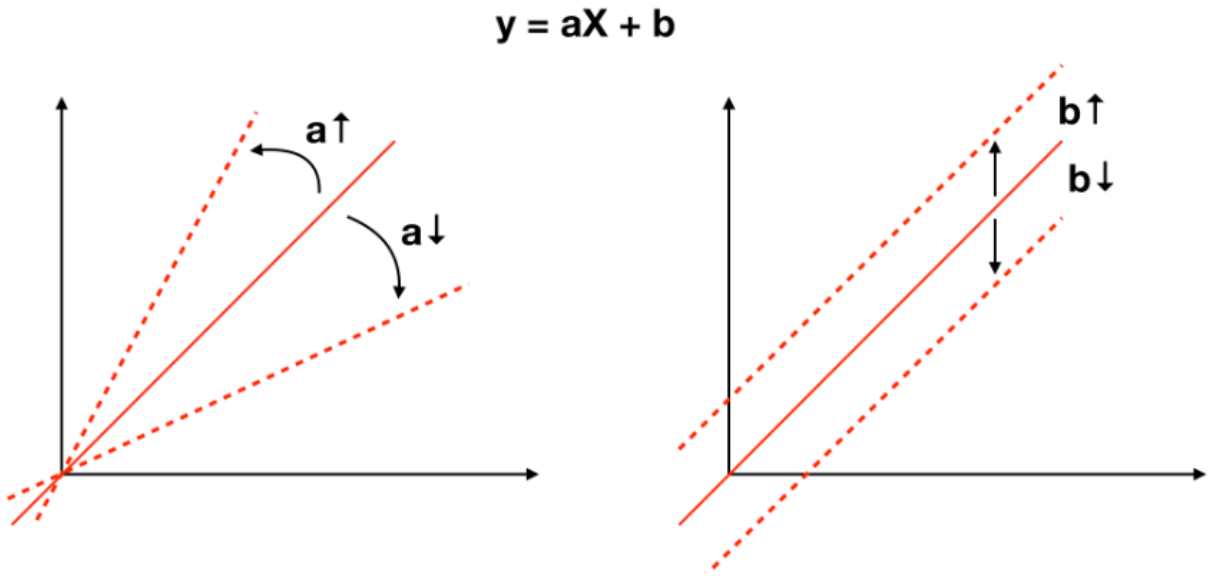
- Linear Regression is part of regression that focuses on finding linear relationship between two or more variables.
- Linear Regression only works with linear relationship variables.

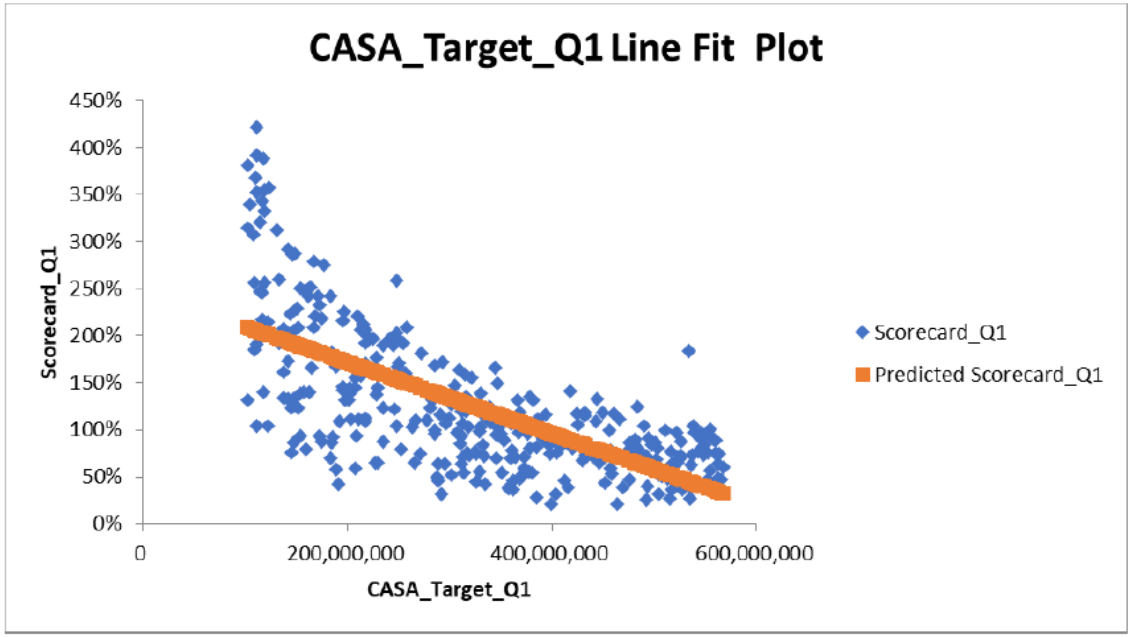
Linear Relationship



Non-Linear Relationship

Regression Formula





Regression Evaluation Metrics: R-Squared

R-Squared Formula:

$$R^2 = 1 - \frac{SS_{RES}}{SS_{TOT}} = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \overline{y})^2}$$

R-Squared Range:

- inf < R-Squared <= 1

Definition:

y_i : Actual Data

$\hat{y}_i$: Prediction from Regression Result

$\overline{y}_i$: Average of Data

Regression Evaluation Metrics: R-Squared

RMSE Formula:

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

R-Squared Criteria:

The lower the RMSE the better

Definition:

y_i : **Actual Data**

$\hat{y}_i$: **Prediction from Regression Result**

Linear Regression Requirement

- Linear Regression must be given with continuous data:

Continuous Data

- Balance
- Age
- Scorecard
- Number of Product

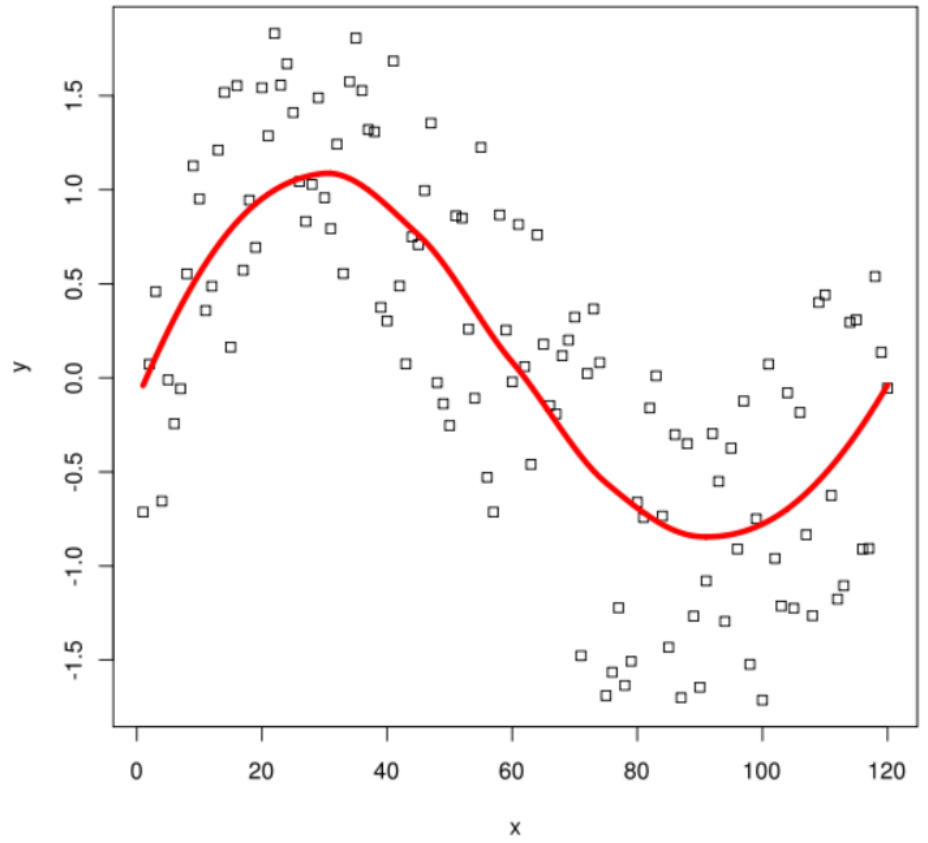
Non-Continuous Data

- Unpaid Tagging
- Credit Card Tag
- Customer ID
- Branch Code

Supervised Learning: Non-Linear Regression

LOWESS Regression

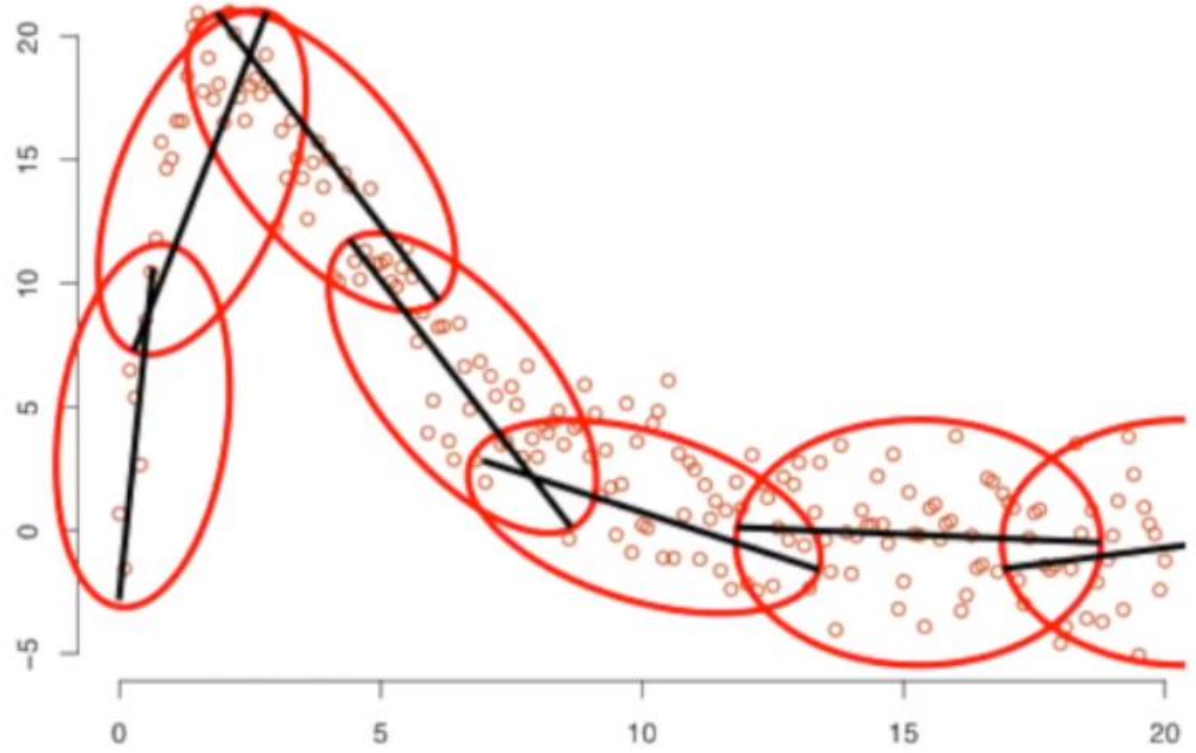
- **LOWESS Regression** is one of the non-parametric methods to calculate non-linear regressions
- It is initially developed for **scatterplot smoothing** (Locally Weighted Scatterplot Smoothing)



Logic of LOWESS Regression

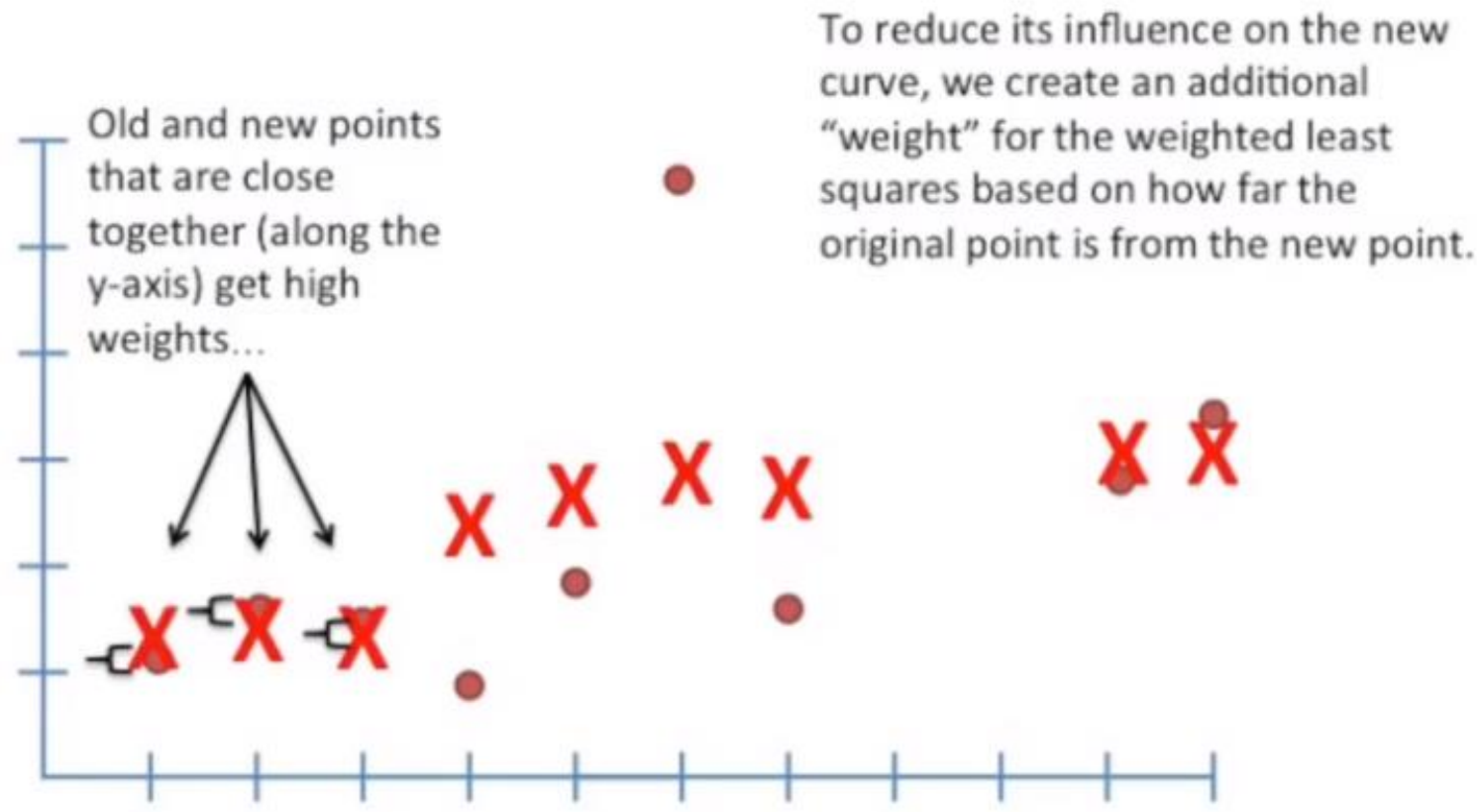
- The main idea to **fit a curve to data point** is to use a type of sliding window to **divide the data into smaller blobs**.

- The second main idea is that **each data point use the least squares to fit a line**.



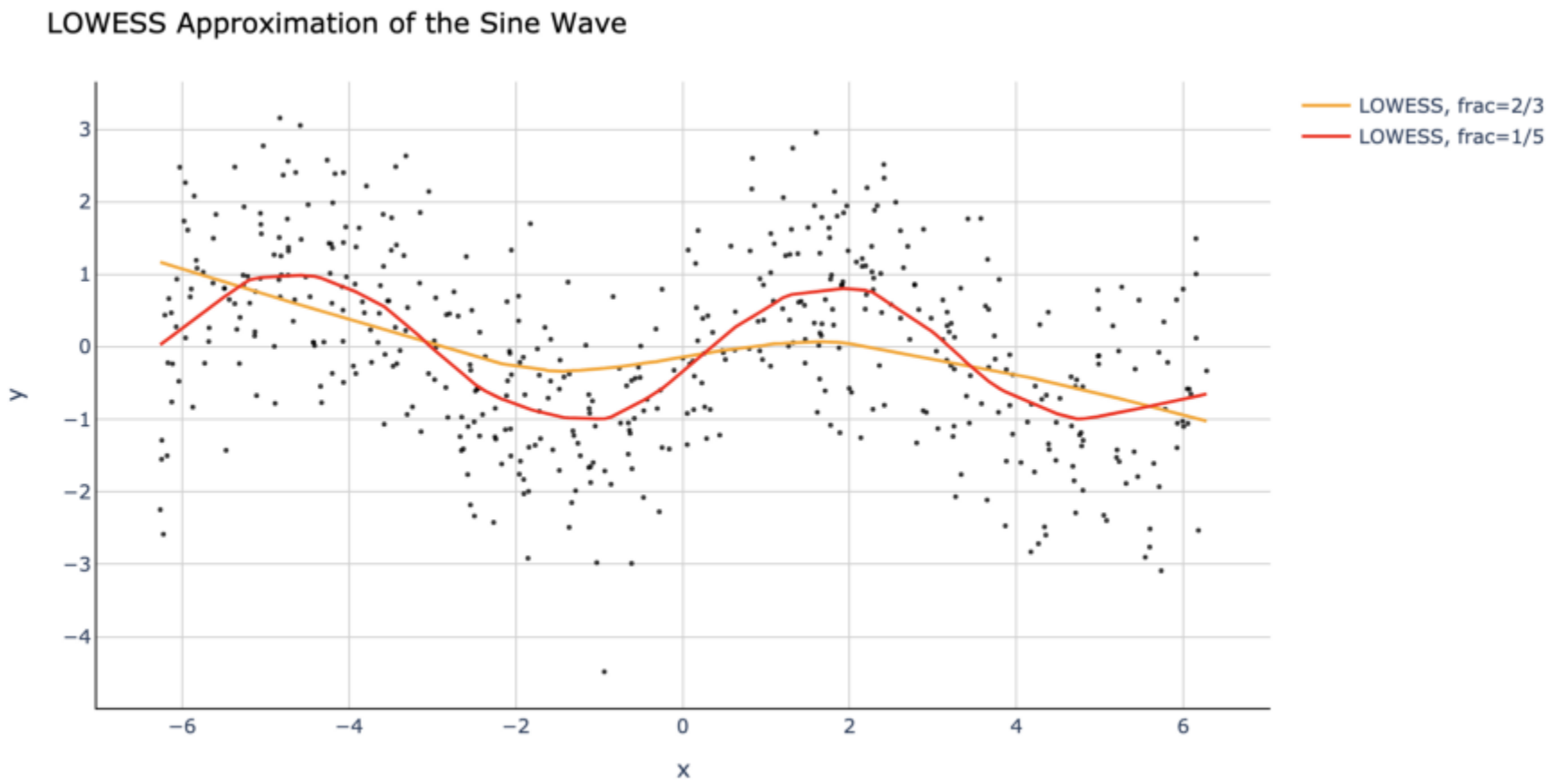
Purpose of Weight in LOWESS Regression

- Weight is used for representing the significant level of outlier data

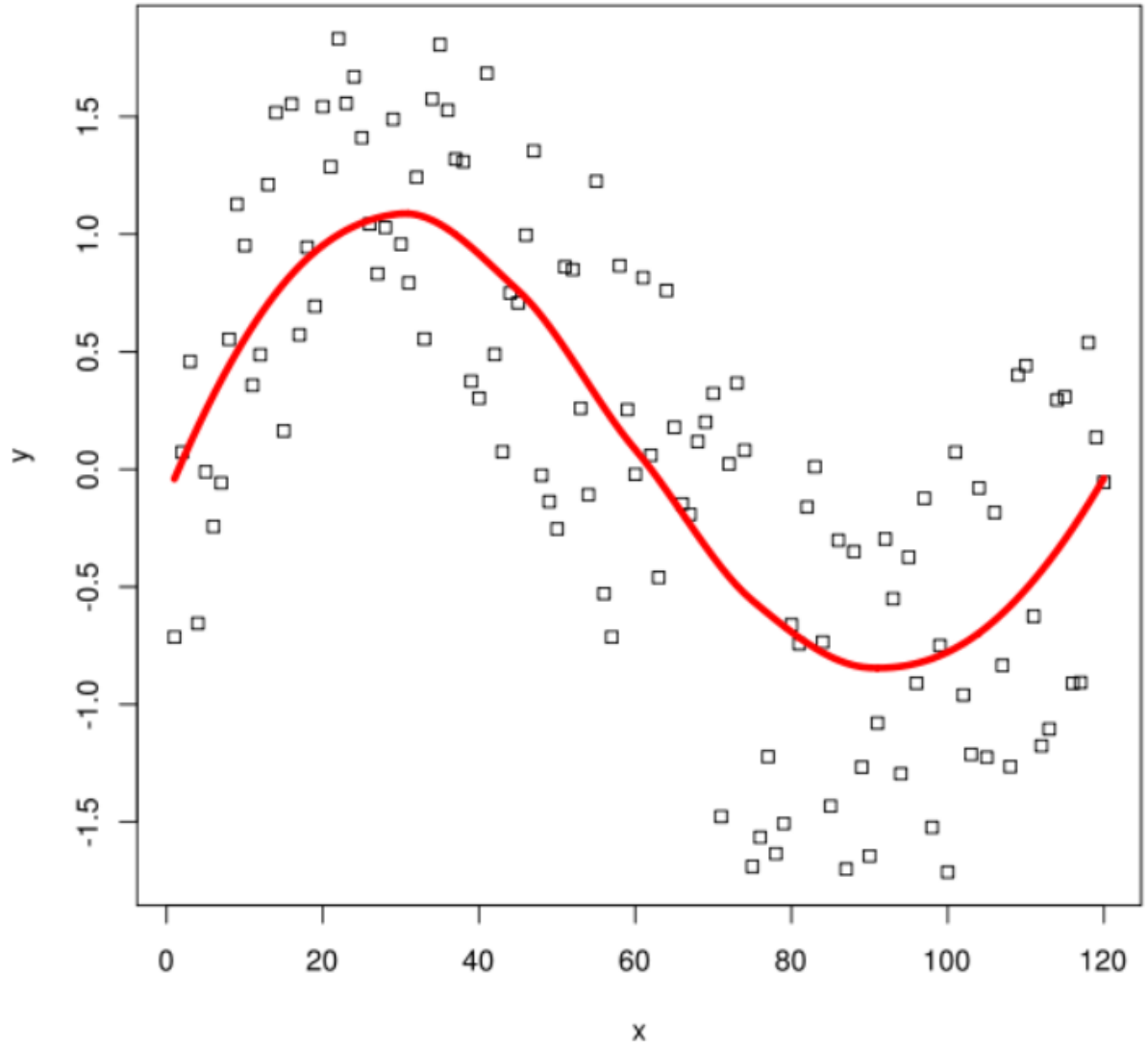


Purpose of Weight in LOWESS Regression

- The higher the weightage, more sensitive the curve will become



LOWESS Regression: Pros & Cons



Benefits:

- Provides a flexible approach to representing data
- Easy to Use
- Computations are relatively easy

Disadvantage:

- Can’t be used to obtain a simple equations
- Less well understood than linear regression
- Requires the analyst to use a little guesswork to obtain a result

Ishoma

12:00 - 13:00

Case Study: Regression Analysis Exercise

Let’s Share!


Q & A


Thank
you