

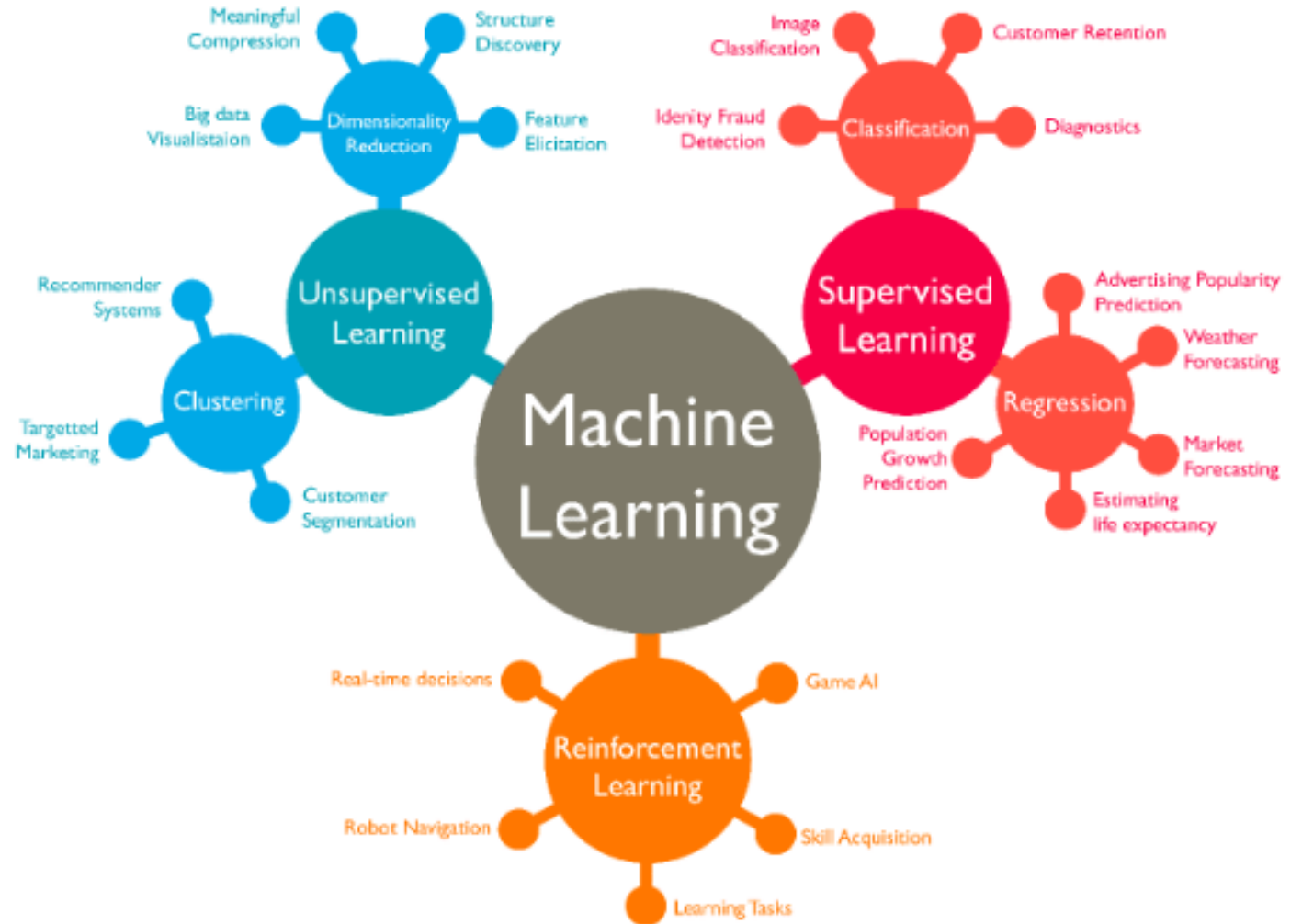


# Introduction

# Types of Machine learning

---

- Supervised learning
- Unsupervised learning
- Reinforcement learning



Source: <https://wordstream-files-prod.s3.amazonaws.com/s3fs-public/machine-learning.png>

# Importance of ML

- Machine learning is important because it gives enterprises a view of trends in customer behavior and business operational patterns, as well as supports the development of new products, with the ability to detect anomalies.

# Simple Linear Regression

- The simplest deterministic mathematical relationship between two variables  $x$  and  $y$  is a linear relationship:

Diagram illustrating the Simple Linear Regression equation:  $y = \beta_0 + \beta_1 x$ .

The equation is annotated with labels and arrows:

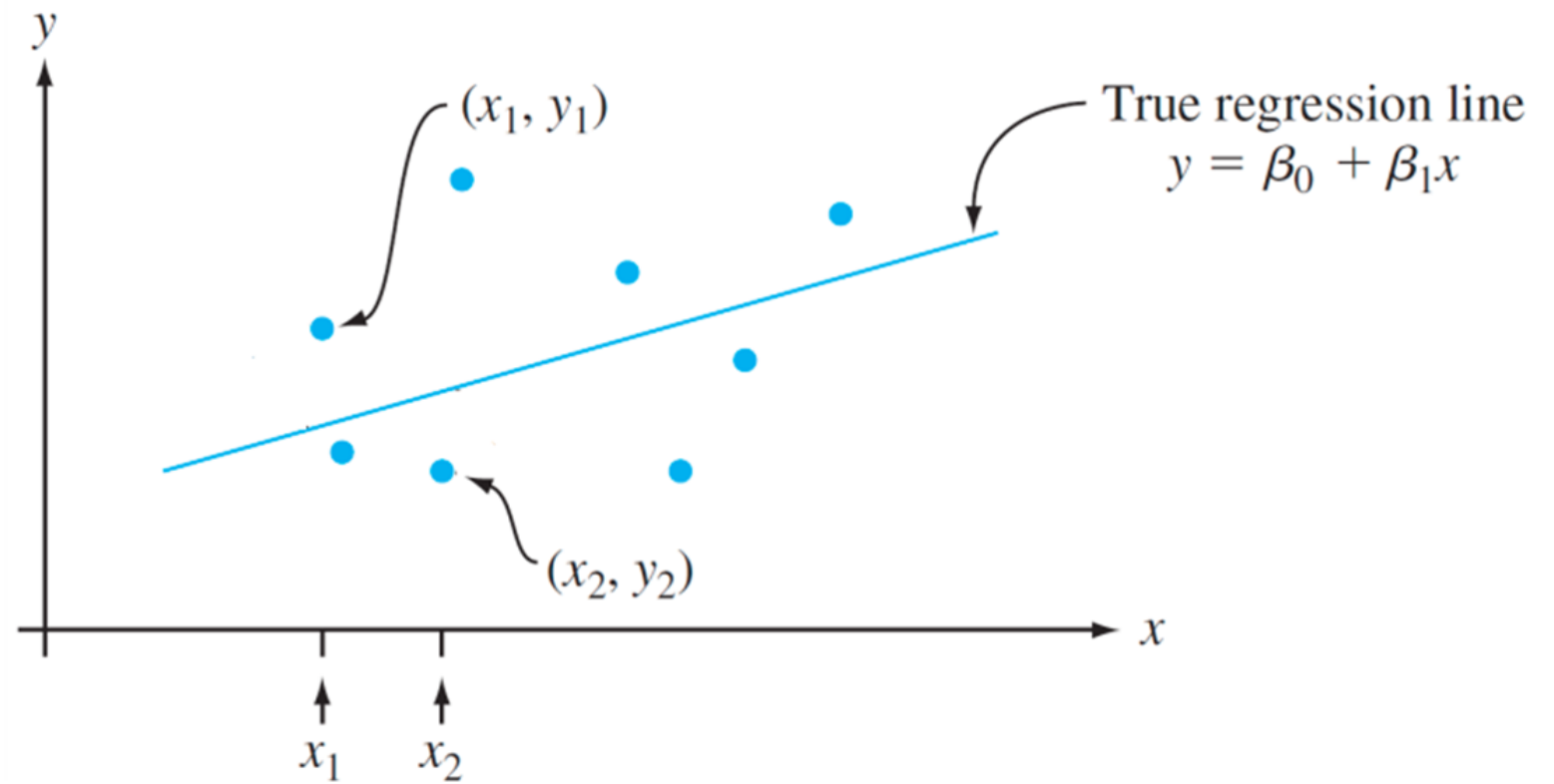
- $y$  is labeled as the **Dependent** variable.
- $\beta_0$  is labeled as the **Intercept**.
- $\beta_1$  is labeled as the **Slope**.
- $x$  is labeled as the **Predictor** variable.
- $\beta_0$  and  $\beta_1$  are collectively labeled as **coefficients**.

Observation Number	Temperature ( $x_i$ )	Yield ( $y_i$ )
1	50	122
2	53	118
3	54	128
4	55	121
5	56	125
6	59	136
7	62	144
8	65	142
9	67	149
10	71	161
11	72	167
12	74	168
13	75	162
14	76	171
15	79	175
16	80	182
17	82	180
18	85	183
19	87	188
20	90	200
21	93	194
22	94	206
23	95	207
24	97	210

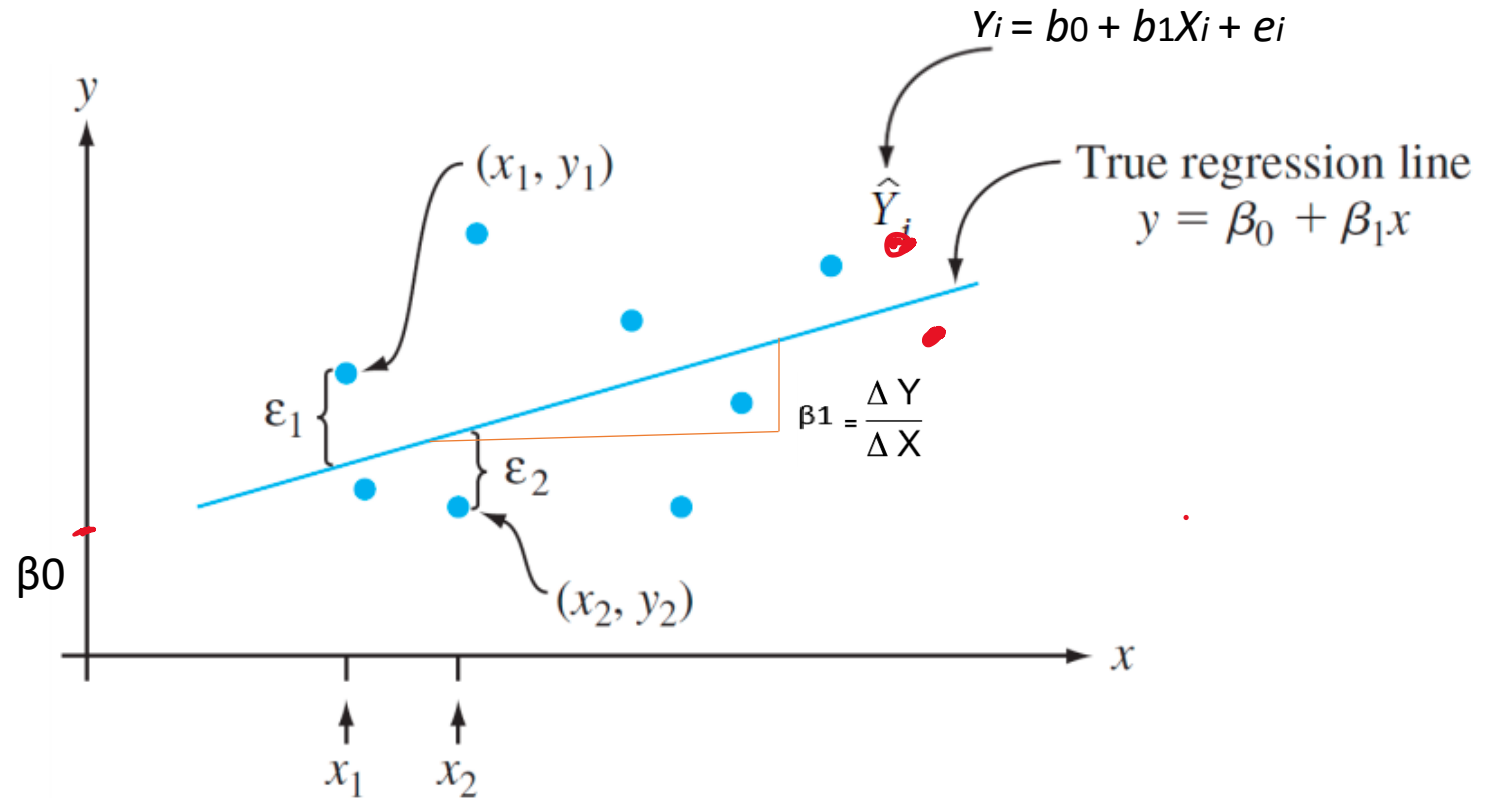
# Terms of simple linear regression

Term	Synonyms
<b>response:</b> The variable we are trying to predict.	dependent variable, Y variable, target, outcome
<b>Independent variable:</b> The variable used to predict the response.	X variable, feature, attribute, predictor
<b>Record:</b> The vector of predictor and outcome values for a specific individual or case.	row, case, instance, example
<b>Intercept:</b> The intercept of the regression line—that is, the predicted value when $X = 0$ .	$b_0$ , $\beta_0$
<b>Regression coefficient:</b> The slope of the regression line.	slope, $b_1$ , $\beta_1$ , parameter estimates, weights
<b>Fitted values:</b> The estimates $\hat{Y}_i$ obtained from the regression line.	predicted values
<b>Residuals:</b> The difference between the observed values and the fitted values.	errors
<b>Least squares:</b> The method of fitting a regression by minimizing the sum of squared residuals.	ordinary least squares, OLS

# Example



# Real situations



- In general, the data doesn't fall exactly on a line, so the regression equation should include an explicit error term  $e_i$ :

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \text{ where } \epsilon_i: \epsilon_1, \epsilon_2, \dots, \epsilon_n,$$

$Y_1, Y_2, \dots, Y_n,$   
 $x_1, x_2, \dots, x_n$

# The purpose of regression

The main goal is to understand a relationship and explain it using the data that the regression was fit to. With the focus is not on predicting individual cases but rather on understanding the overall relationship among variables.

In Business the main items of interest are the fitted values  $\hat{Y}$ .




# Simple linear regression evaluation metric

sensitive to outliers


## 1. Least Squares:

$$\begin{aligned}RSS &= \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 \\ &= \sum_{i=1}^n (Y_i - \hat{b}_0 - \hat{b}_1 X_i)^2\end{aligned}$$

A good model when the regression line is the estimate that minimizes the sum of squared residual values.



## Key Ideas

- The regression equation models the relationship between a response variable  $Y$  and a predictor variable  $X$  as a line.
  - A regression model yields fitted values and residuals—predictions of the response and the errors of the predictions.
  - Regression models are typically fit by the method of least squares.
  - Regression is used both for prediction and explanation.
- 
- 



# Multiple linear regression

---

$$Y = b_0 + b_1X_1 + b_2X_2 + \dots + b_pX_p + e$$

Observation Number	Factor 1 ( $x_{i1}$ )	Factor 2 ( $x_{i2}$ )	Yield ( $y_i$ )
1	41.9	29.1	251.3
2	43.4	29.3	251.3
3	43.9	29.5	248.3
4	44.5	29.7	267.5
5	47.3	29.9	273.0
6	47.5	30.3	276.5
7	47.9	30.5	270.3
8	50.2	30.7	274.9
9	52.8	30.8	285.0
10	53.2	30.9	290.0
11	56.7	31.5	297.0
12	57.0	31.7	302.5
13	63.5	31.9	304.5
14	65.3	32.0	309.3
15	71.1	32.1	321.7
16	77.0	32.5	330.7
17	77.8	32.9	349.0

# Key Terms for Multiple Linear Regression

Term	Synonym	
Root mean squared error: The square root of the average squared error of the regression (this is the most widely used metric to compare regression models).	RMSE	
Residual standard error: The same as the root mean squared error, but adjusted for degrees of freedom.	RSE	
R-squared: The proportion of variance explained by the model, from 0 to 1.	coefficient of determination, $R^2$	
t-statistic: The coefficient for a predictor, divided by the standard error of the coefficient, giving a metric to compare the importance of variables in the model. See “t-Tests” on page 110.		
Weighted regression: Regression with the records having different weights.		

# Multiple linear regression evaluation metrics

1. **RMSE**: root mean squared error, or RMSE.

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

The difference between RMSE and RSE is very small, particularly for big data applications.

2. **RSE**: residual standard error.

$$RSE = \sqrt{\frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{(n - p - 1)}} \text{ ,where } p \text{ number of predictors.}$$

### 3. $R^2$ : The coefficient of determination

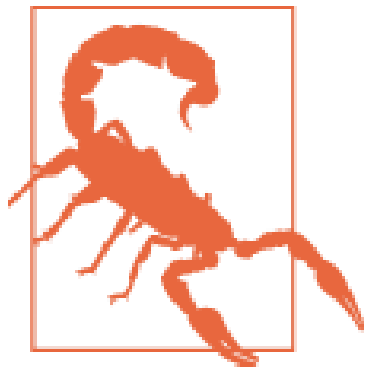
$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2}$$

### 4. t- statics

$$t_b = \frac{\hat{b}}{\text{SE}(\hat{b})}$$

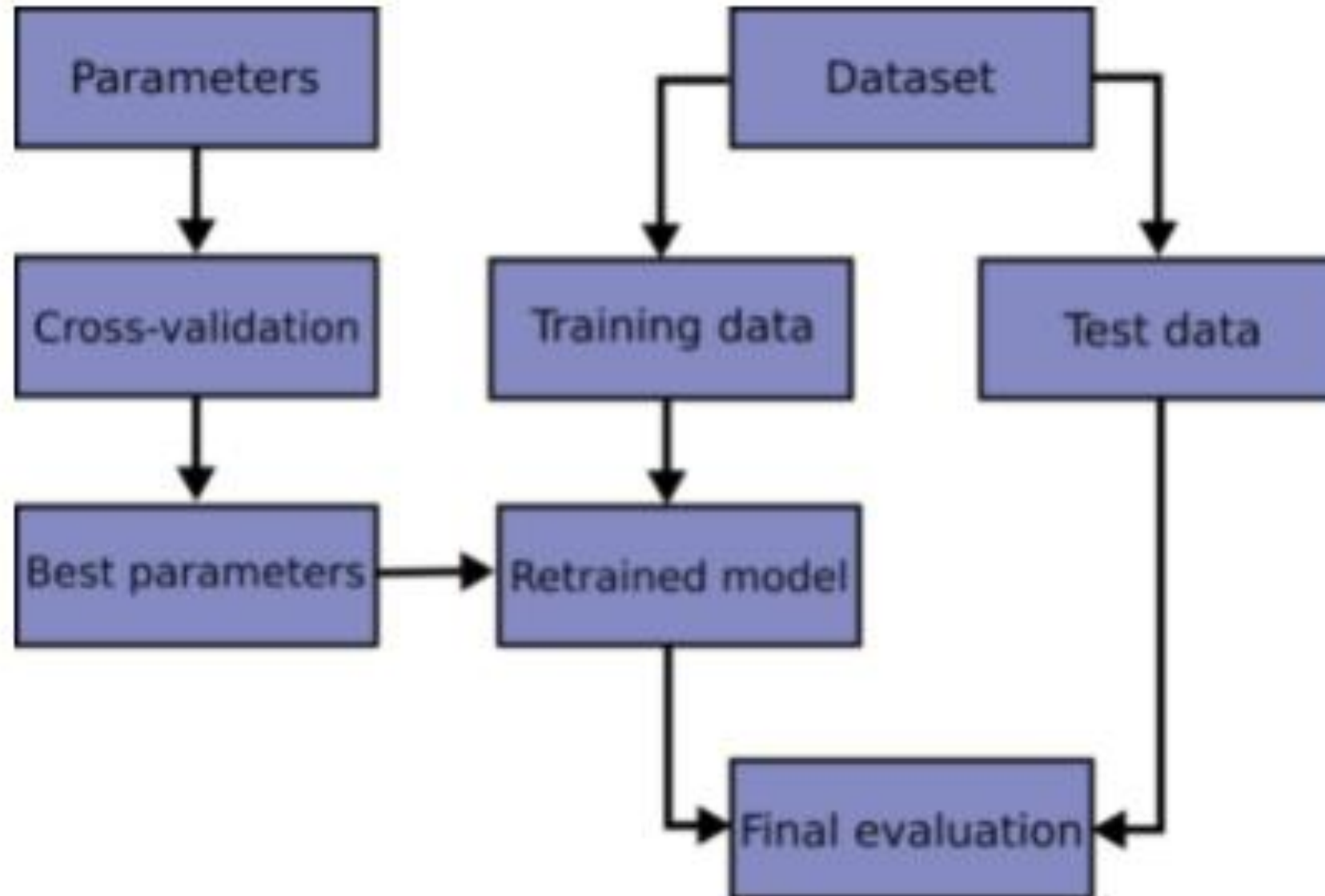
the p-value—measures the extent to which a coefficient is “statistically significant”

The higher t-statistic (and the lower the p-value), the more significant the predictor



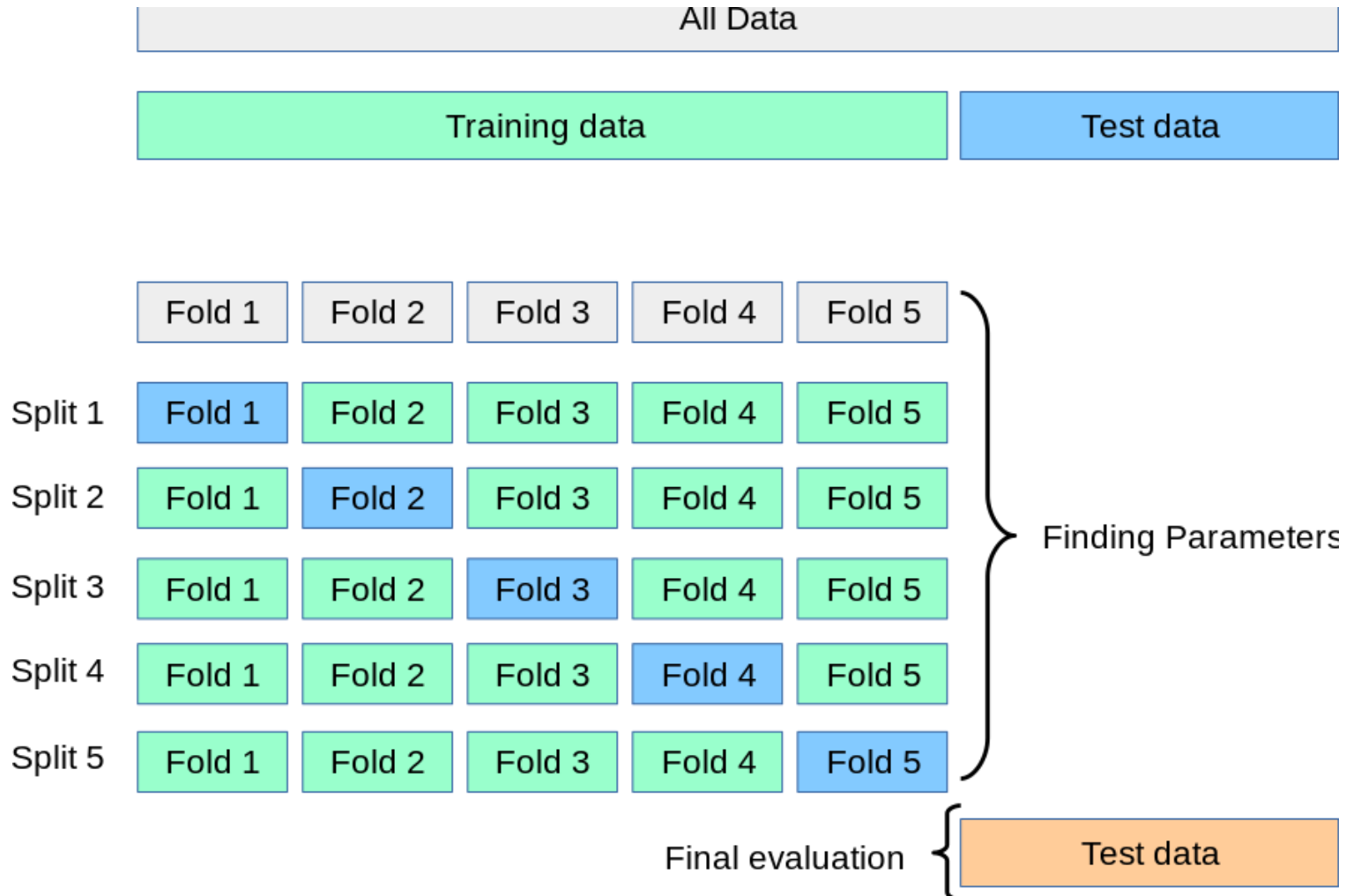
In addition to the t-statistic, *R* and other packages will often report a *p-value* ( $\Pr(>|t|)$  in the *R* output) and *F-statistic*. Data scientists do not generally get too involved with the interpretation of these statistics, nor with the issue of statistical significance. Data scientists primarily focus on the t-statistic as a useful guide for whether to include a predictor in a model or not. High t-statistics (which go with p-values near 0) indicate a predictor should be retained in a model, while very low t-statistics indicate a predictor could be dropped. See “p-Value” on page 106 for more discussion.

# Model evaluation approach



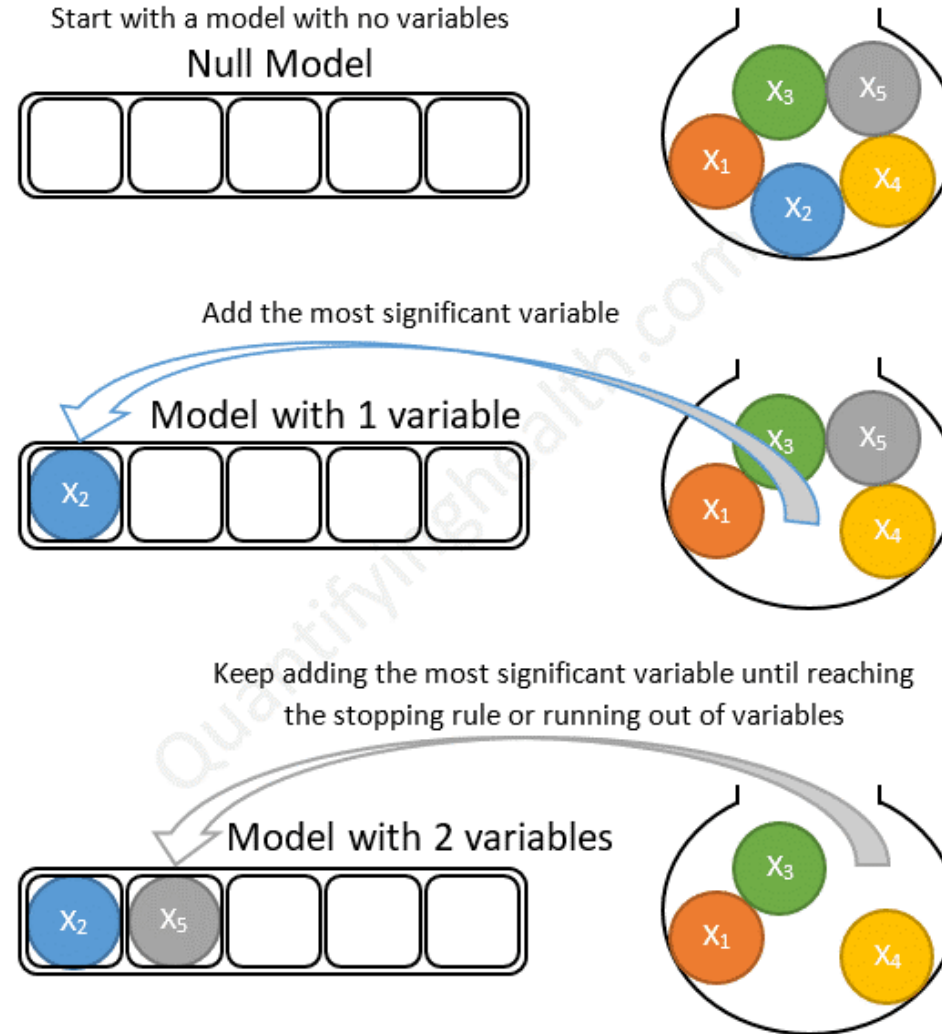


# Cross Validation



# Model selection and Forward stepwise Linear regression

Forward stepwise selection example with 5 variables:

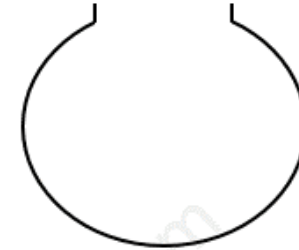
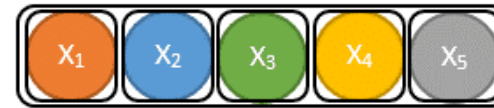


## Model selection and backward stepwise Linear regression

### Backward stepwise selection example with 5 variables

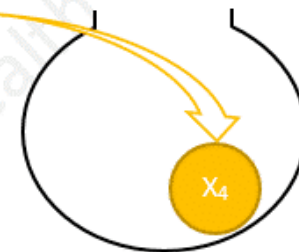
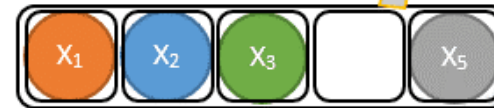
Start with a model that contains all the variables

Full Model



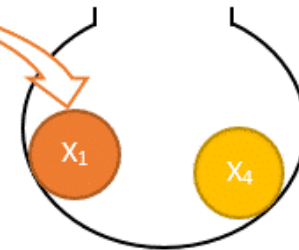
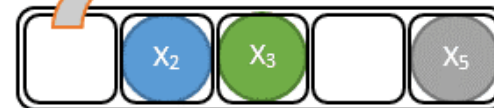
Remove the least significant variable

Model with 4 variables



Keep removing the least significant variable until reaching the stopping rule or running out of variables

Model with 3 variables



## How to determine the most significant variable to add at each step?

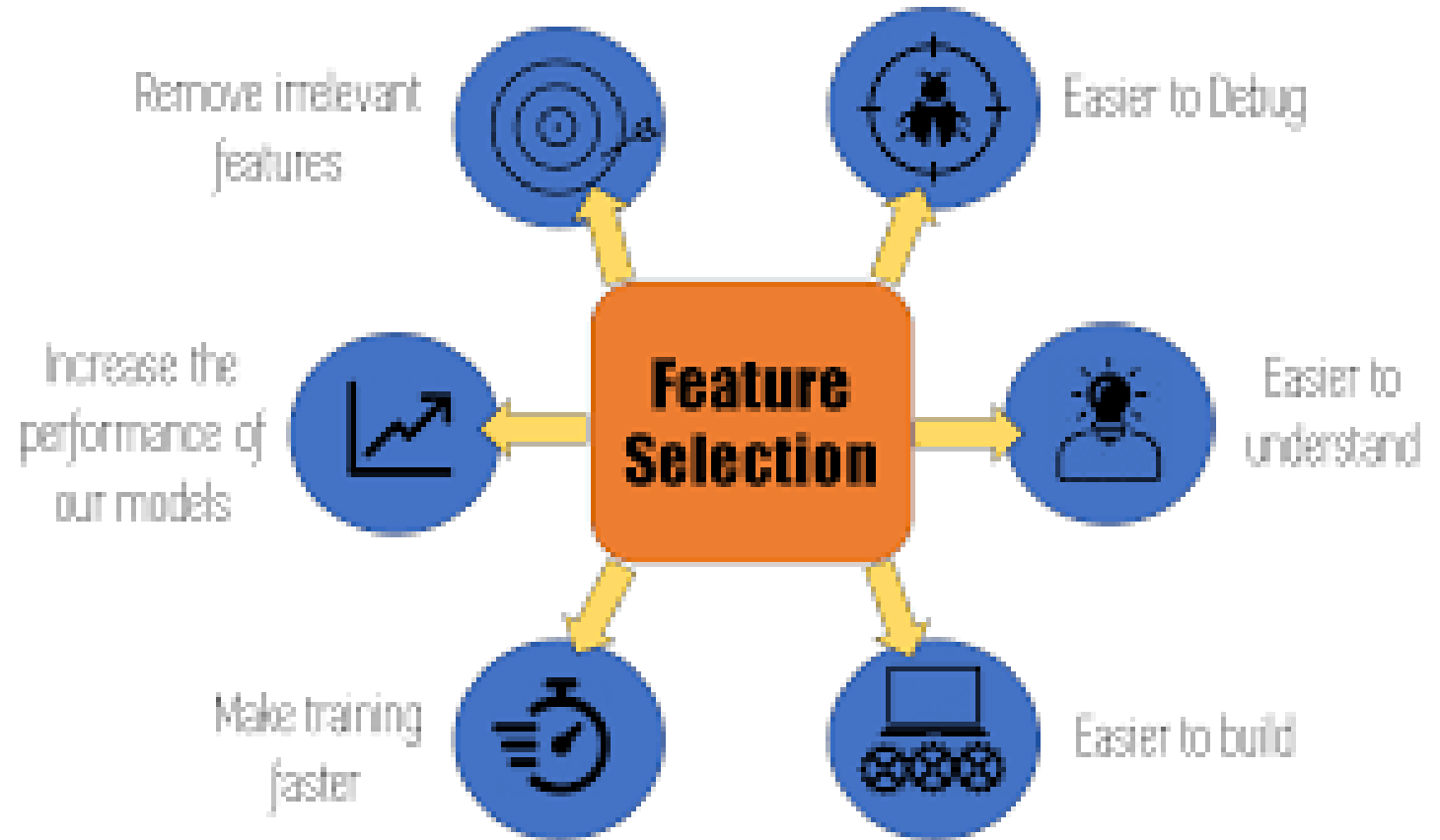
The most significant variable can be chosen so that, when added to the model:

It has the smallest p-value, or

It provides the highest increase in  $R^2$ , or

It provides the highest drop in model RSS (Residuals Sum of Squares) compared to other predictors under consideration.

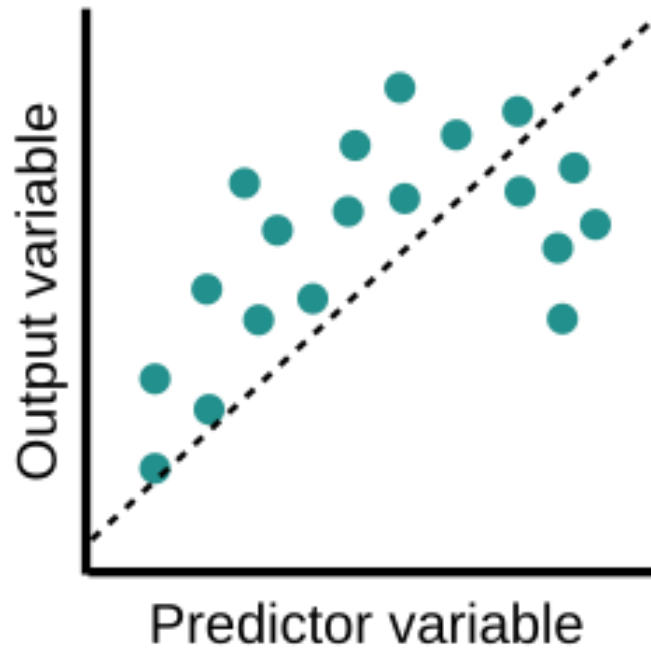
Why we do  
feature  
selection?



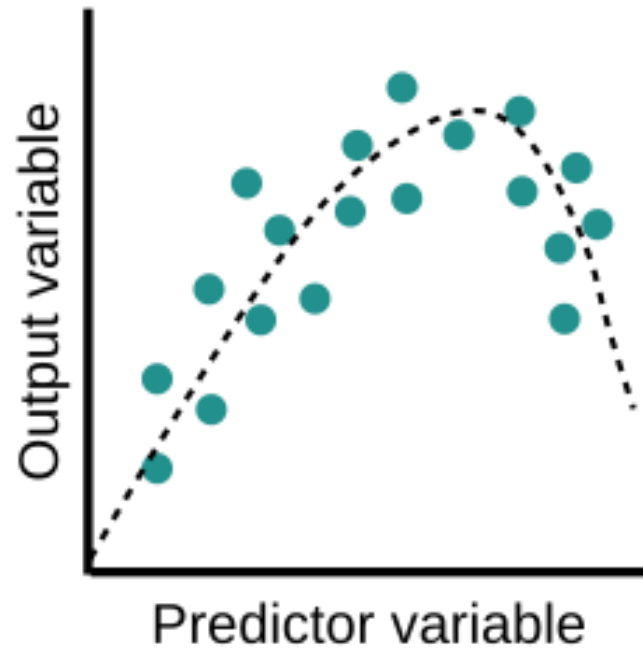
# Underfitting vs overfitting

---

Underfit



Optimal



Overfit

