# Understanding The Concept of Bias : 01

## Problem Statement (As Defined By The Client)

### Client Requirement

- The Client, A Healthcare Analytics Firm, Wants To Develop A Predictive Model That Accurately Forecasts Patient Length of Stay (LOS) in Hospitals Based on Multiple Factors, Such As Patient Demographics And Health Conditions
- The Current Predictive Model (Predicted LOS) Has Been Observed To Consistently Overestimate The Actual Length of Stay For Patients, Particularly For Those With Mild Health Conditions OR Shorter Stays
- The Client Has Provided Historical Patient Data, Including Actual Length of Stay And Various Attributes Such As Age, Gender, And Diagnosis. The Client Wishes To Understand Why The Current Model Deviates From The Actual Observations And How A New Model Can Be Developed To Provide A More Accurate Forecast of Patient LOS
- The Client Seeks To Identify The Shortcomings of The Existing Model And Determine An Improved Model That Minimizes Prediction Errors, Thereby Helping in Better Resource Allocation And Hospital Management

### Specific Goals

1. Identify And Explain The Trend Differences Between The Actual Length of Stay And Predicted Length of Stay Using Visual Representations
2. Evaluate The Accuracy of The Existing Linear Model For Predicting Patient LOS And Identify The Reasons For High Variance And Overfitting
3. Propose A Revised Model That Fits The Patient Data More Accurately

## Problem Statement (As Defined By The Data Scientist)

### Title

Evaluating and Improving the Predictive Accuracy of a Hospital Length of Stay Model Using Machine Learning Techniques

### Objective

- To Analyze And Improve The Existing Predictive Model For Patient Length of Stay That Currently Exhibits High Variance And Overfitting
- The Project Aims To Identify Discrepancies Between Actual And Predicted Patient LOS And Propose An Enhanced Model That Minimizes Prediction Error, Thus Improving The Alignment of Predicted LOS With Actual Hospital Stay Durations

### Problem Description

- The Current Predictive Model Used By The Healthcare Firm is Based on A Linear Formula

  **Predicted LOS =** BaseLOS + LinearCoefficient * (Age + Severity Index)

- However, When Compared To The Actual Length of Stay, It Appears That The Predicted Values Show Considerable Overestimation For Patients With Lower Severity Scores And Shorter Actual Stays

  **Actual LOS =** BaseLOS + ExponentialScalingFactor * $e^{(Age/SeverityIndex)}$

  Where **"e"** is Eulers Number Which is Approximately Equal To **2.71828**, and is The Base of The Natural Logarithm

**Understanding The Concept of Bias : 01**

- The Exponential Nature of The Actual LOS Formula Indicates That Length of Stay Increases Exponentially With Age And Health Severity, While The Linear Model Only Captures A Direct Linear Relationship
- This Results in A Model With High Variance, Missing The Mark on Shorter Stays And Overestimating Them, Which is Detrimental For Hospital Planning And Resource Allocation

**Remarks on The Current Model**

- The Predicted Values Consistently Overestimate Length of Stay For Patients With Shorter Actual Stays And Mild Health Conditions
- The Discrepancy Suggests That The Existing Linear Model is Insufficient To Capture The True Relationship Between Patient Age, Health Severity, And LOS
- The High Variance Leads To Significant Overfitting, Indicating The Need For A More Complex, Non-Linear Model That Can Capture The Exponential Growth of Length of Stay For More Severe Cases

**Approach**

1. **Data Analysis**

   Perform Exploratory Data Analysis (EDA) To Understand The Distribution of Patient Demographics And Health Conditions, Along With The Corresponding Actual Length of Stay

2. **Model Evaluation**

   Plot And Analyze The Actual Versus Predicted Length of Stay To Visualize And Quantify The Variance

3. **Model Redesign**

   Implement And Test Alternative Models (e.g., Exponential Regression) To Capture The Non-Linear Relationship Between Patient Characteristics And LOS

4. **Model Validation**

   Evaluate The Revised Model's Performance Using Appropriate Metrics Such As

   - Mean Absolute Error (MAE)
   - Root Mean Squared Error (RMSE)
   - R-Squared Value.

**Expected Outcome**

- An Improved Predictive Model That Significantly Reduces Variance And Error, Accurately Predicting Patient LOS Across Different Demographics And Health Conditions
- This Will Provide The Client With A More Reliable Tool For Estimating Patient LOS, Leading To Better Decision-Making in Hospital Management And Resource Allocation

**Understanding The Concept of Bias : 01**

**Sample Data**

| Age (Years) | Severity Index | Actual LOS (Days) | Predicted LOS (Days) |
|---|---|---|---|
| **25** | 0.5 | 2 | 6 |
| **30** | 1 | 4 | 9 |
| **40** | 1.5 | 5 | 12 |
| **45** | 2 | 7 | 15 |
| **50** | 2.5 | 8 | 17 |
| **55** | 3 | 10 | 20 |
| **60** | 3.5 | 12 | 22 |
| **65** | 4 | 15 | 25 |
| **70** | 4.5 | 18 | 28 |
| **75** | 5 | 20 | 30 |

| **Variables Assumed** | |
|---|---|
| **Exponential Scaling Factor** | 1.2 |
| **Linear Coefficient** | 3.0 |
| **Base LOS (Actual LOS)** | 1 Day |