**Torc Robotics Machine Learning Questions:**

* *Regularization, Over-fitting, Unbalanced data, ResNet, CNNs, C++ Inheritance, Djikstra Algorithms*
* Multiple choice test on bayesian estimation.

**Resources:**

1/ Towards Data Science: <https://towardsdatascience.com/ten-machine-learning-concepts-you-should-know-for-data-science-interviews-70107ca84754>

2/ Cheat Sheet: <https://www.kdnuggets.com/2018/09/machine-learning-cheat-sheets.html>

3/ ML Mastery: <https://machinelearningmastery.com/basic-concepts-in-machine-learning/>

4/ ML Fundamentals: <https://towardsdatascience.com/machine-learning-basics-part-1-a36d38c7916>

5/ Youtube 10 Hours: <https://www.youtube.com/watch?v=GwIo3gDZCVQ&ab_channel=edureka%21>

**Basics Concepts:**

a/ SL vs UL

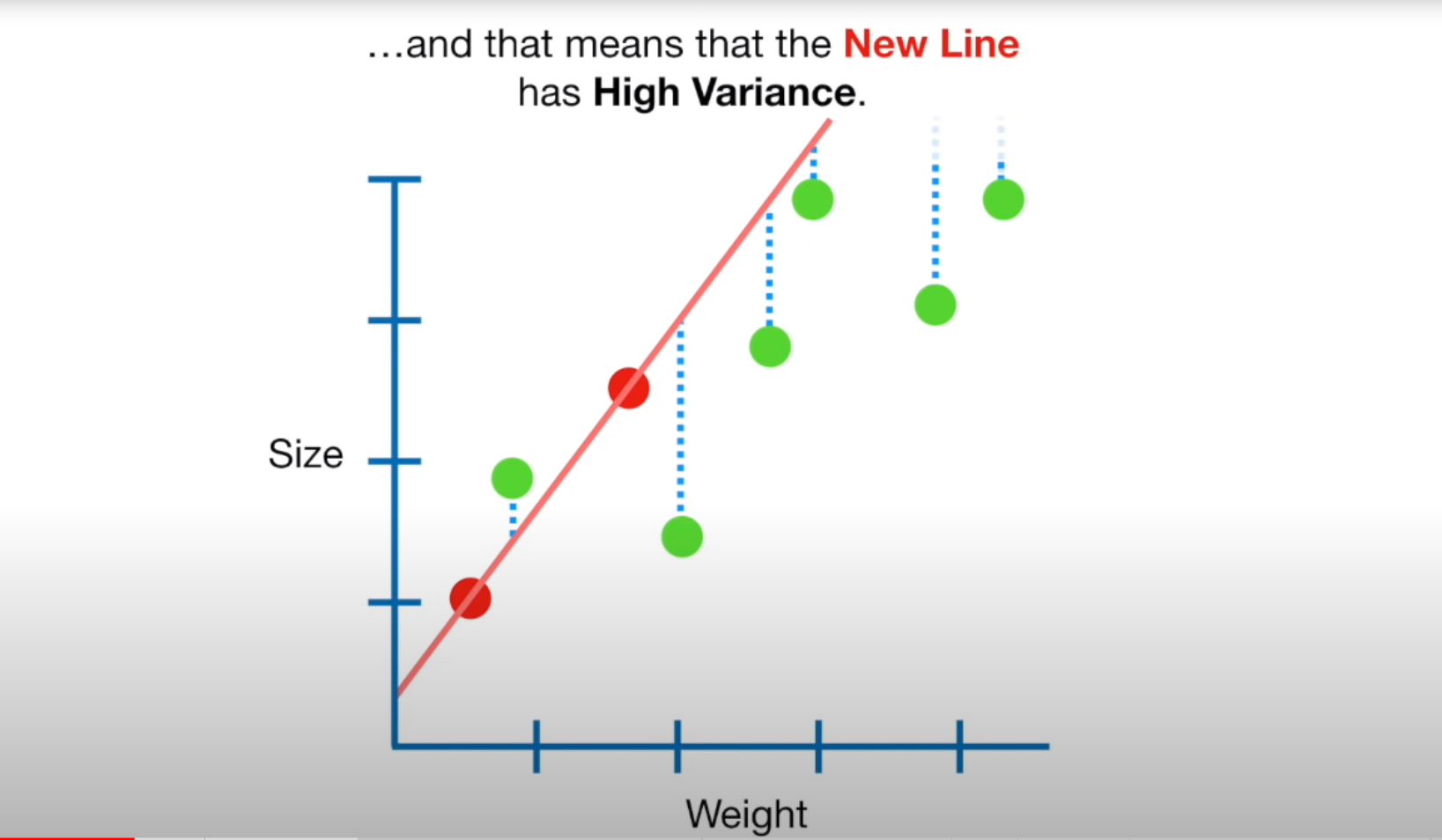
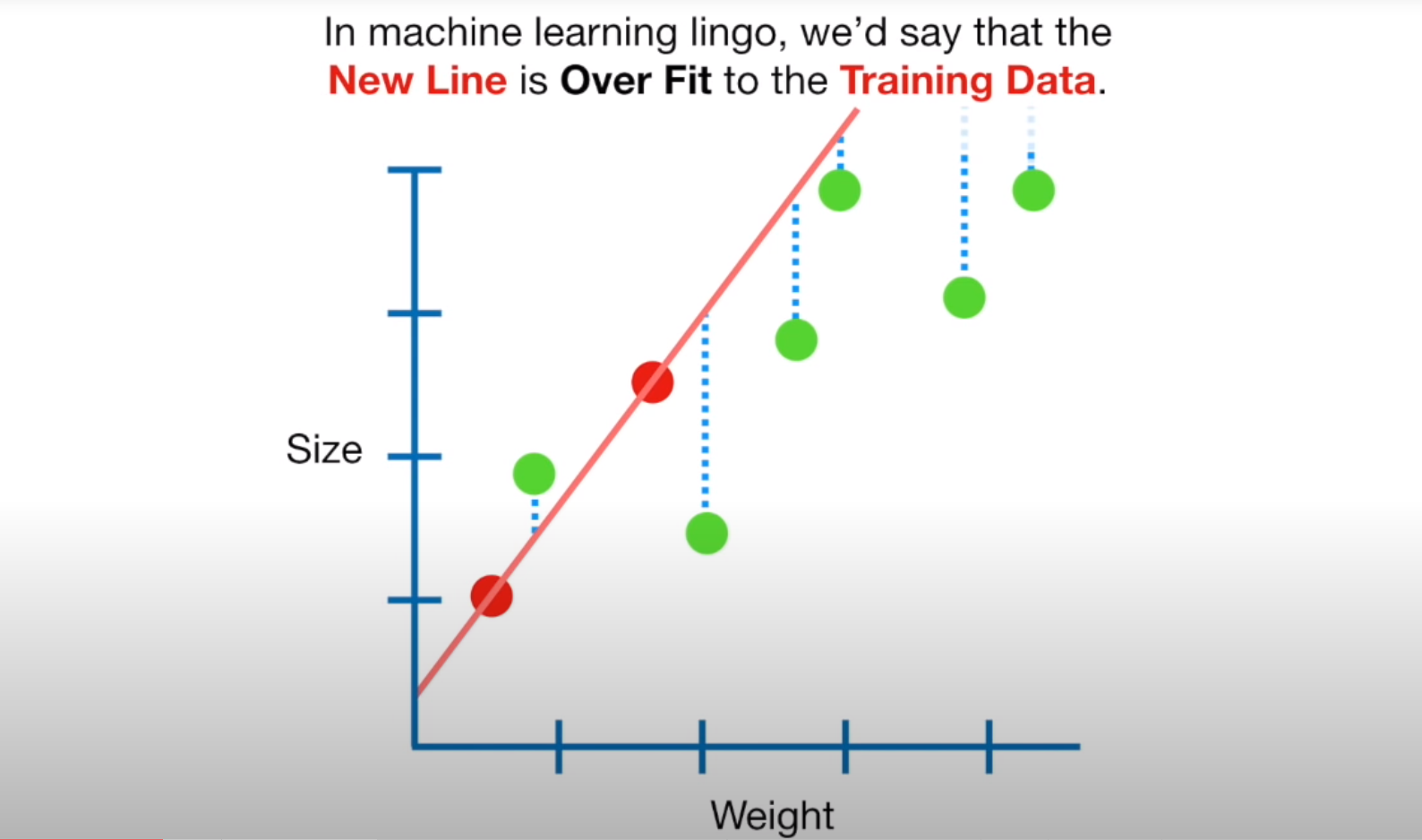
* Supervised = learning on a labeled dataset where the target variable is known
* Unsupervised = draw inferences and find pattern from input data without reference to labeled outcomes. No target variable.

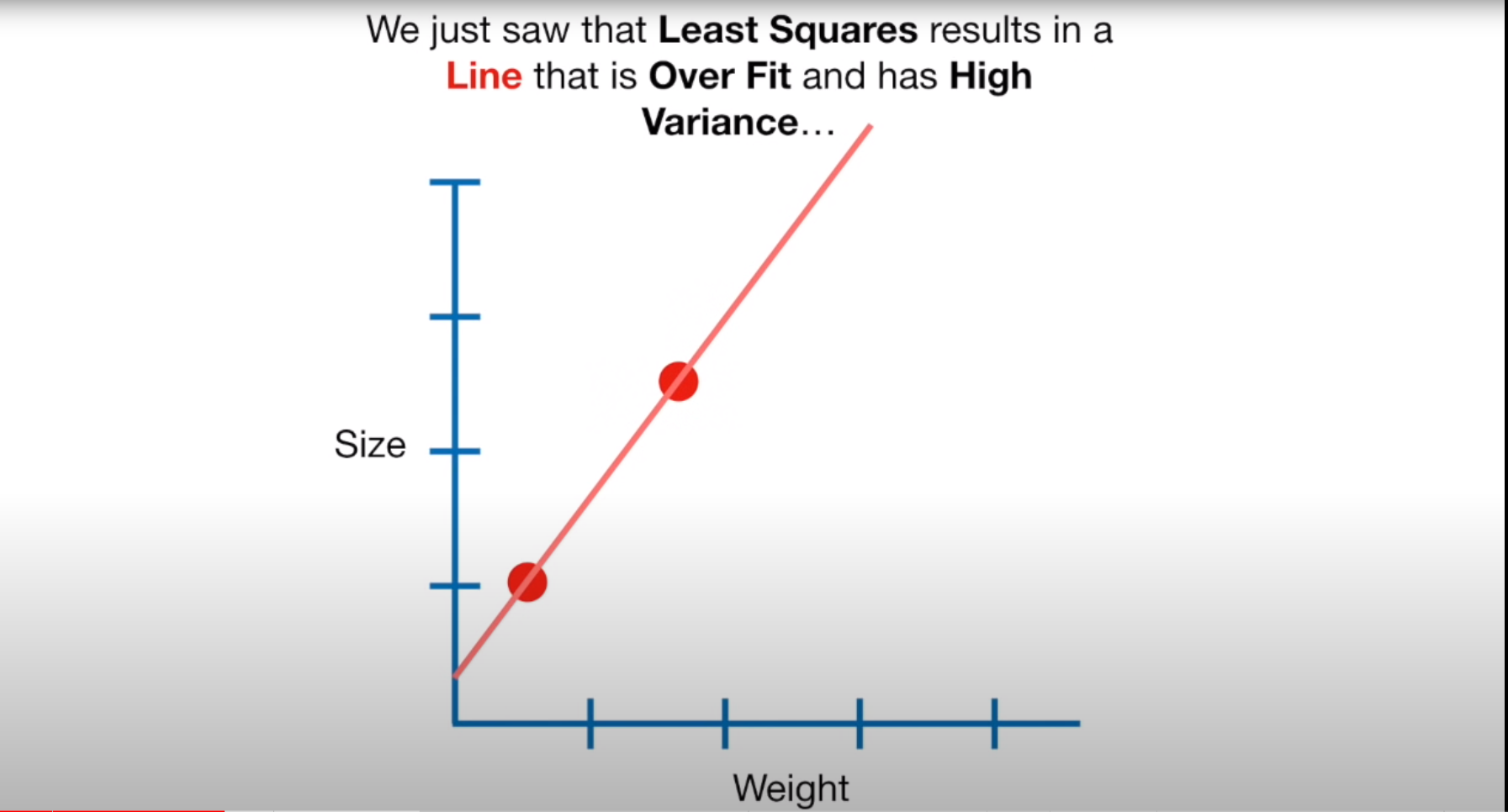
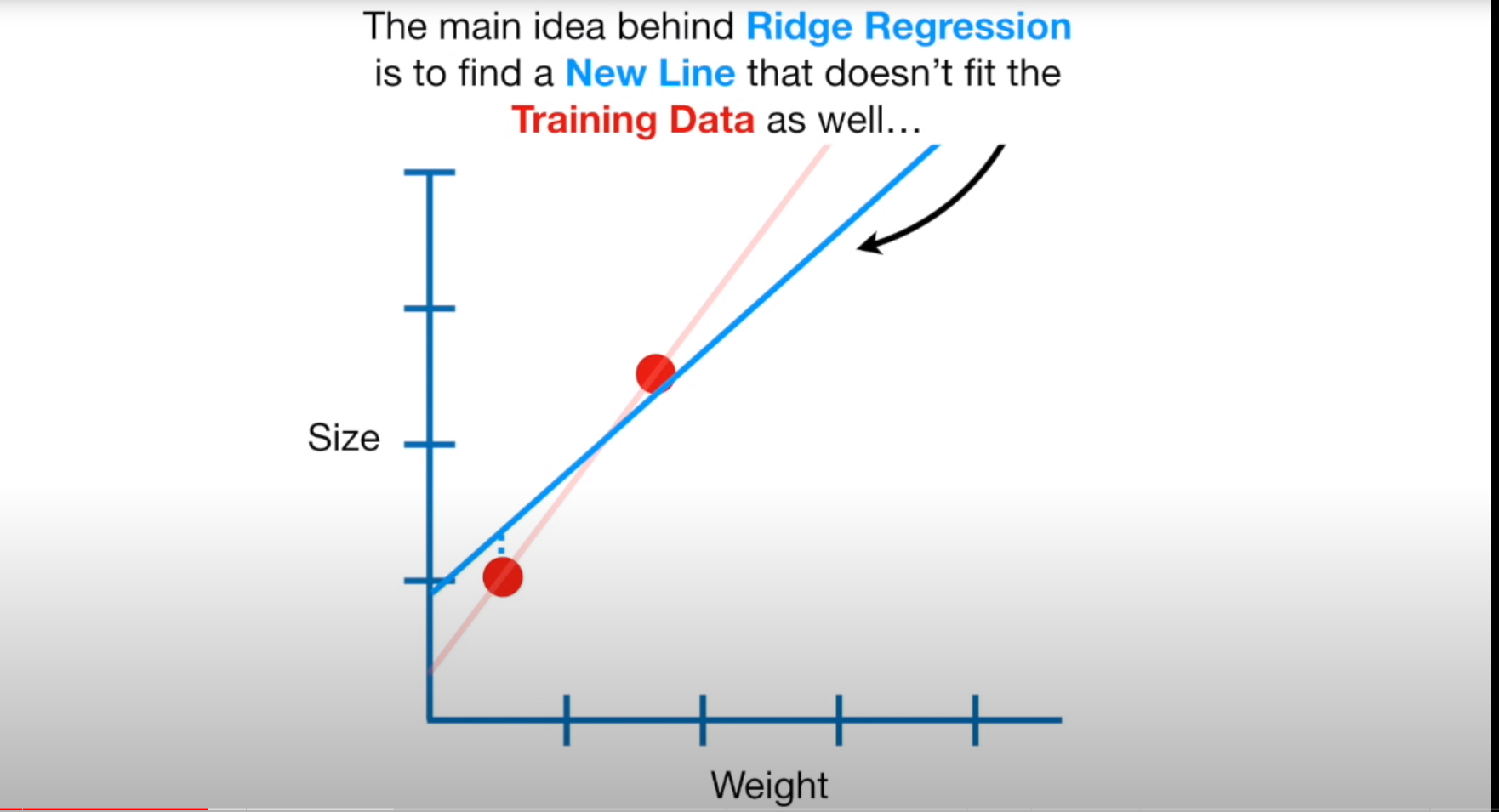
b/ Bias-Variance Tradeoff:

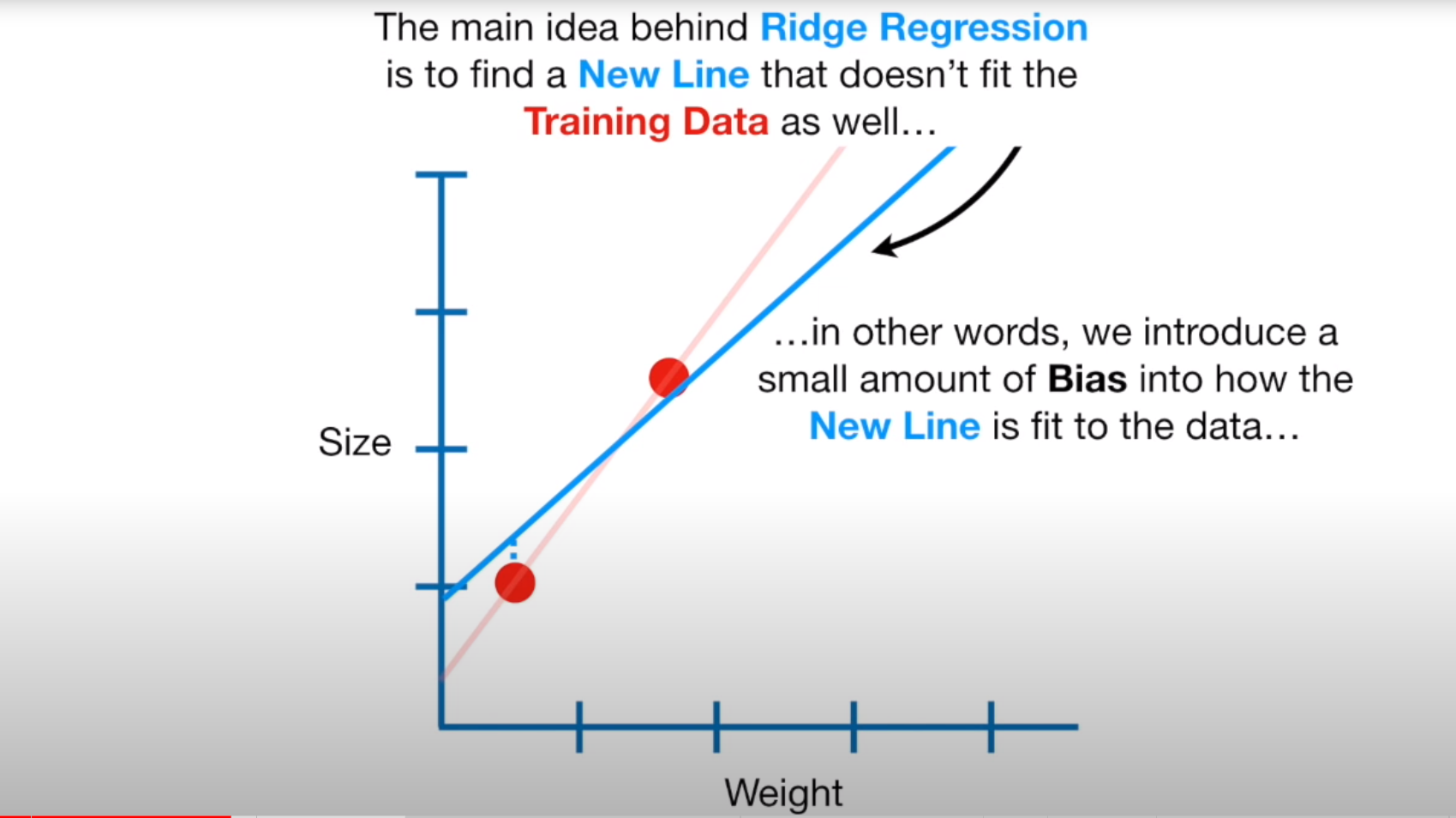
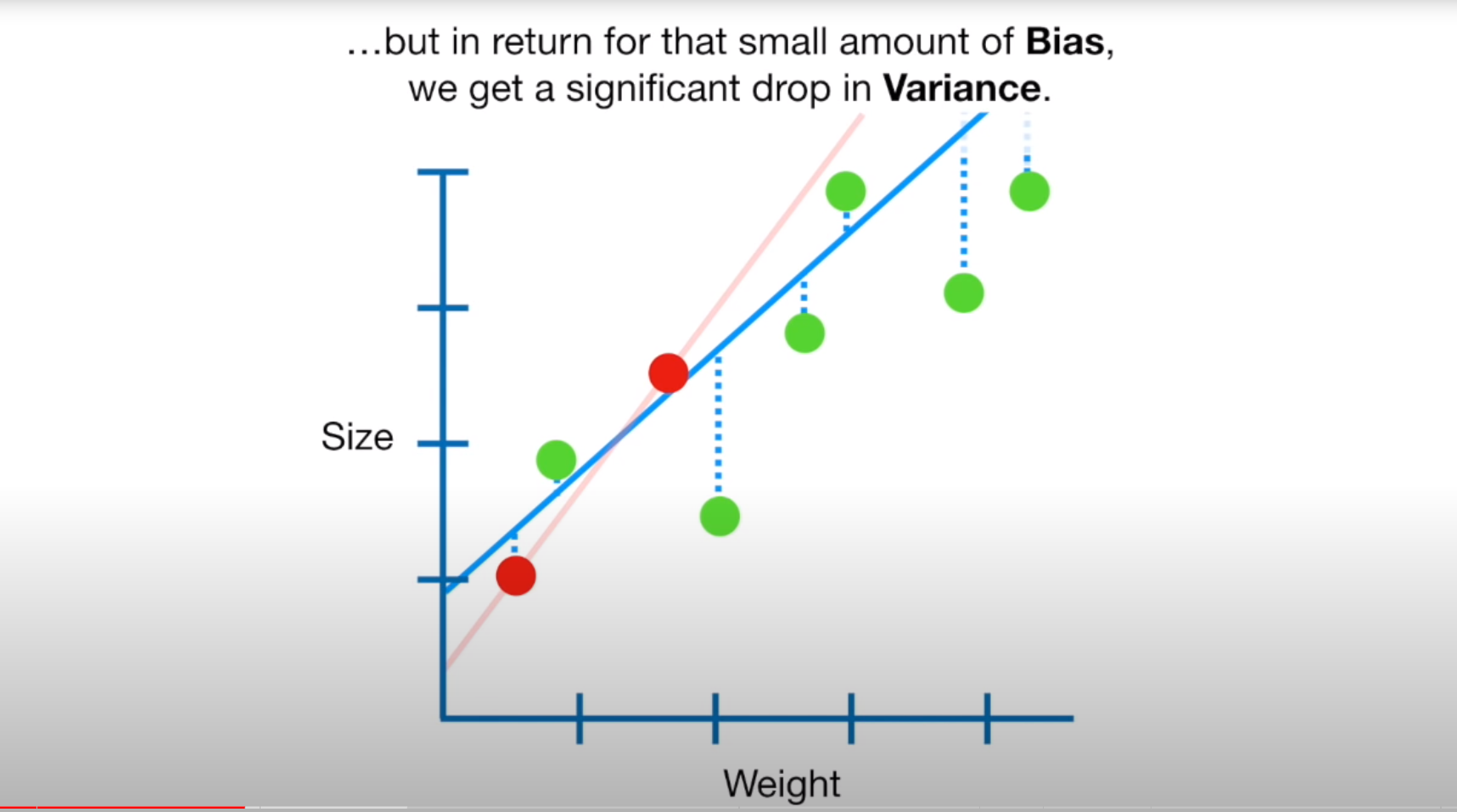
* Bias = error due to model’s assumptions that are made to simplify it
* Variance = amount that the predicted value would change if different training data was used => models that place a higher emphasis on training data will have a higher variance.
* There is an inverse relationship between amount of bias and variance in a given ML model. However, there is ab optimal point which a specific amount of bias and variance results in a minimal amount of total error

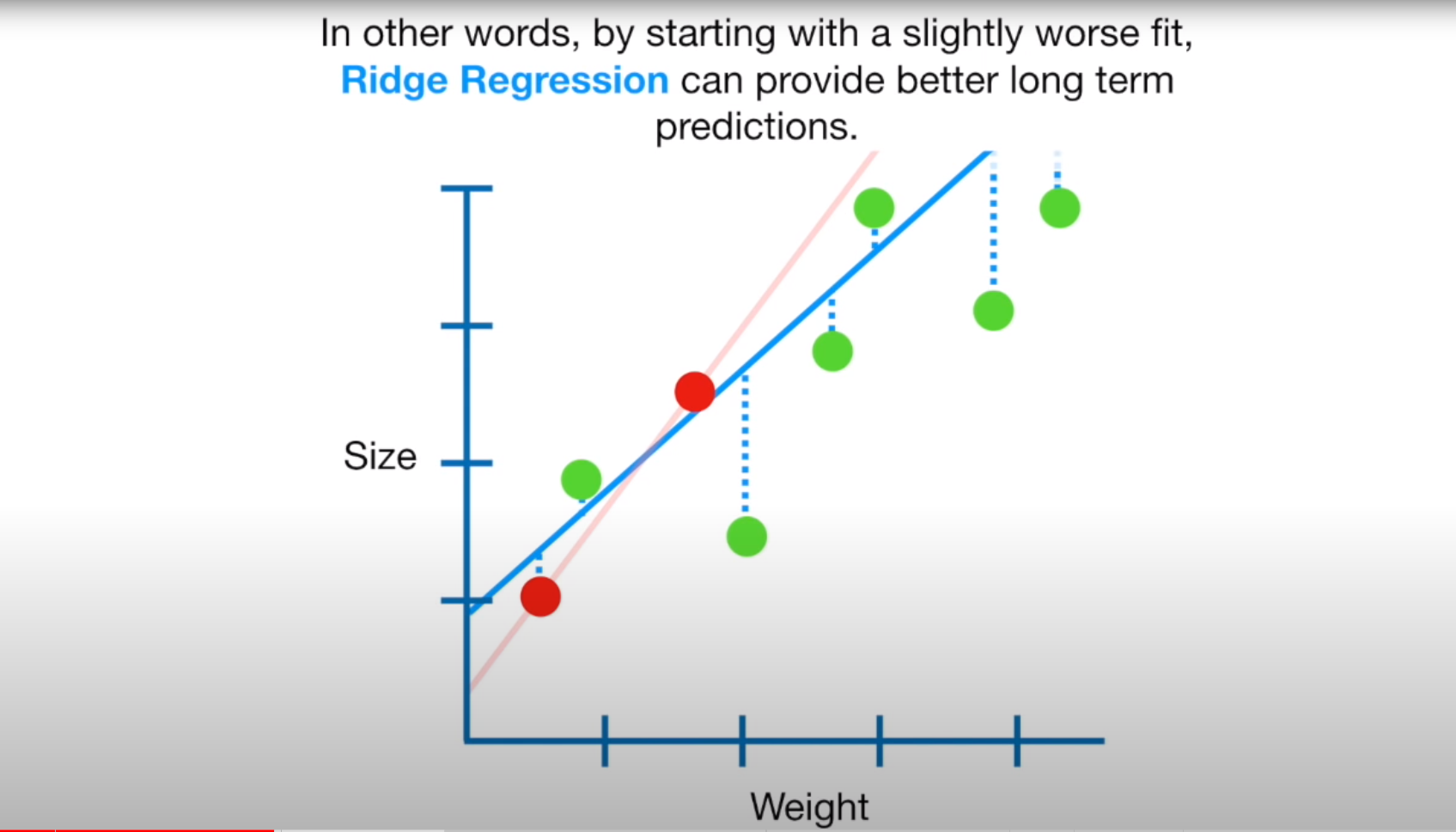
c/ Regularization:

* L1 & L2 to reduce the over fitting of training data
* L2 = ridge regression = minimize the sum of squared residual + lambda \* slope squared
* L1 = Lasso regression = minimize the sum of squared residual + lambda \* absolute slope
* L2 is less robust but has a stable solution, always 1 solution
* L1 is more robust but has unstable solution, can possibly has multiple solution



d/ Cross Validation

* Technique used to assess how well a model performs on a new independent dataset
* Ex: Split dataset to training, validation, testing.

+ training => build model

+ validation => tune hyper parameters

+ testing => evaluate final model

e/ Evaluation Metrics:

* R Squared =
* Adjusted R Squared =
* Mean Absolute Error =
* Mean Squared Error =

f/ Metrics for classification model:

* True positive = outcome where the model correctly predict positive class
* True negative = outcome where the model correctly predicts the negative class
* False positive (type 1 error) = outcome where the model incorrectly predict the positive class
* False negative (type 2 error) = outcome where the model incorrectly predict the negative class
* Accuracy = fraction of predictions that a model got right
* Precision = What proportion of positive identifications was actually correct?
* F1 score = a measurement of test’s accuracy. Max of 1 min of 0. It measure of the preciseness and robustness of a model
* AUC-ROC curve = a performance measurement for classification problem that tells us how much a model is capable of distinguishing between classes. Higher AUC = model is more accurate.

g/ Dimension Reduction:

* Def = process of reducing the number of features in a dataset. Important when you want to reduced variance in your model (overfitting)
* Popular = PCD = principal component analysis => To reduce memory + speed up algorithm + data visualization

h/ Data Wrangling:

* The process of cleaning + transforming raw data into moe usable state.
* Steps:

+ Checking for outlier + possibly removing them

+ imputation of missing data

+ Encoding categorical data

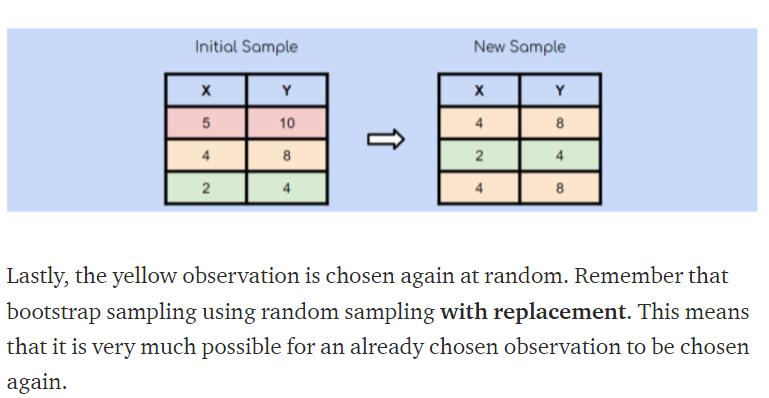
+ Normalization + standardizing data

+ Feature engineering

+ Dealing wit imbalances in data by under or over sampling dataset

i/ Bootstrap Sampling

* Def = resampling method that used random sampling with replacement



k/ Ensemble Learning, Bagging, Boosting:

* Ensemble Learning = multiple ML algo are used in conjunction to allow higher predictive performance
* Bagging = boostrap aggregating = multiple models of same learning algo are trained with bootstrapped samples of the original dataset
* Boosting = variation of bagging = a variation of bagging where each individual model is built sequentially, iterating over the previous one. Specifically, any data points that are falsely classified by the previous model is emphasized in the following model.

**Machine Learning Full Course**

**1/ Introduction:**

* AI = technique which enables machine to mimic human behavior
* ML = Subset of AI technique which use statistical methods to enable machine to improve with experience
* DL = subset of ML which make the computation of multi-layer neural network feasible.
* Supervised Learning:

+ 3 popular algo: Linear Regression, Random Forest, SVM

* Unsupervised Learning:

+ Clustering

+ Apriori Algo

+ K means

+ Hierachical Clustering

* RL: Environment + learning agent

2/