# Live Session | Choosing a Good Model

Tamojit Maiti
Masters in Applied Statistics and Operations Research, ISI Kolkata
Data Scientist at Sixt R&D, previously at Rapido & AB InBev

# Agenda

- Algorithms and Models
  - Distinction
  - Selecting Algorithms
  - Selecting Models
- Tradeoffs
  - Complexity vs Interpretability
  - Bias &/vs Variance
- Eliminating Overfitting
  - Cross-Validation
  - Feature Selection
  - Regularization
- Mini-Quiz
  - 10 Qs
- Playground
  - Feature Selection using Ridge and Lasso Regression in Colab

# Algorithms & Models

# Algorithms & Models | Distinction

- Model is a function that relates the inputs (also known as features) to the outputs (also known as the labels)
- Algorithm is the set of rules that are used to create this function
- A single algorithm can create infinitely many models, but all of them will belong to the same general family.
- The first step involves selecting the most suitable algorithm
- The next step involves searching for the best model of all models generated by the algorithm
- The best model needs to to further modified so that it is generalizable

# Algorithms & Models | Selecting Algorithms

- Class of problems
  - Supervised
  - Unsupervised

- Type of Data available
  - Structured
  - Unstructured

- How frequently do you need output
  - Batch Processing
  - Stream Processing

- What is the purpose
  - Prediction
  - Inference

# Algorithms & Models | Selecting Models

After we are done selecting a subset of class of algorithms, we focus on finding a good model amongst all models within that subset

- Predictability How closely does my model reproduce reality?
- Interpretability How well can I explain what is going on in the model?
- Bias How closely is my model expected to reproduce reality on average?
- Variance How consistently is my model expected to reproduce reality on average?

Model Selection has the objective of making the model more generalizable while maintaining high predictability, and reducing bias and variance.

# Tradeoffs

# Tradeoffs| Complexity vs Interpretability

- Complexity of a model is the number of parameters that can be independently changed

$y = w_0$ vs $y = u_o + u_1 x + u_2 x^2 + u_3 x^3$

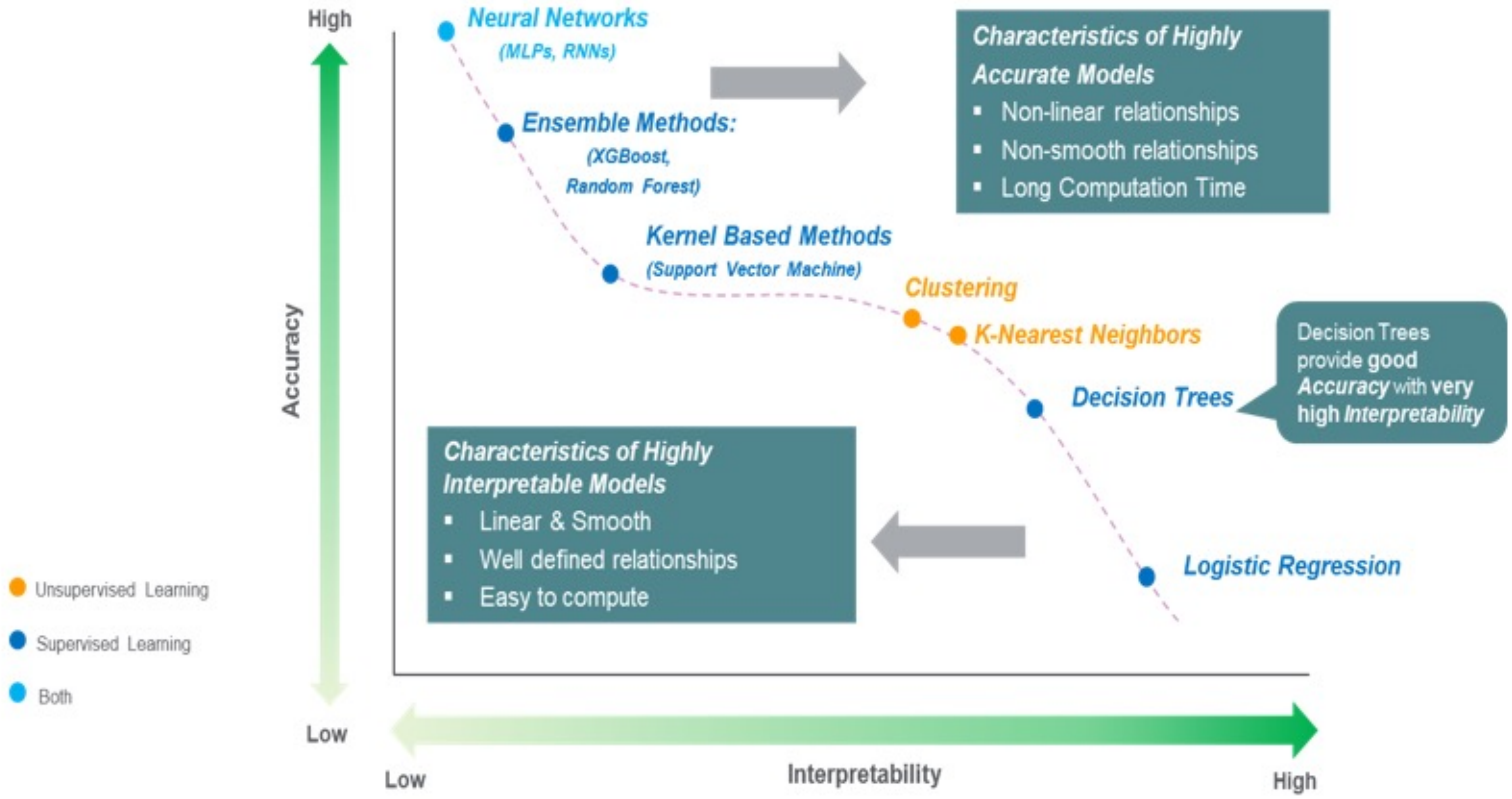
- We want our models to be as simple as possible without sacrificing predictability

- Interpretability of a model is the ease at which a human can understand the reasoning behind model outputs

- We additionally want our models to be interpretable in critical applications

- No absolute unit exists to quantify interpretability

# Tradeoffs| Complexity vs Interpretability



# Tradeoffs| Bias & Variance

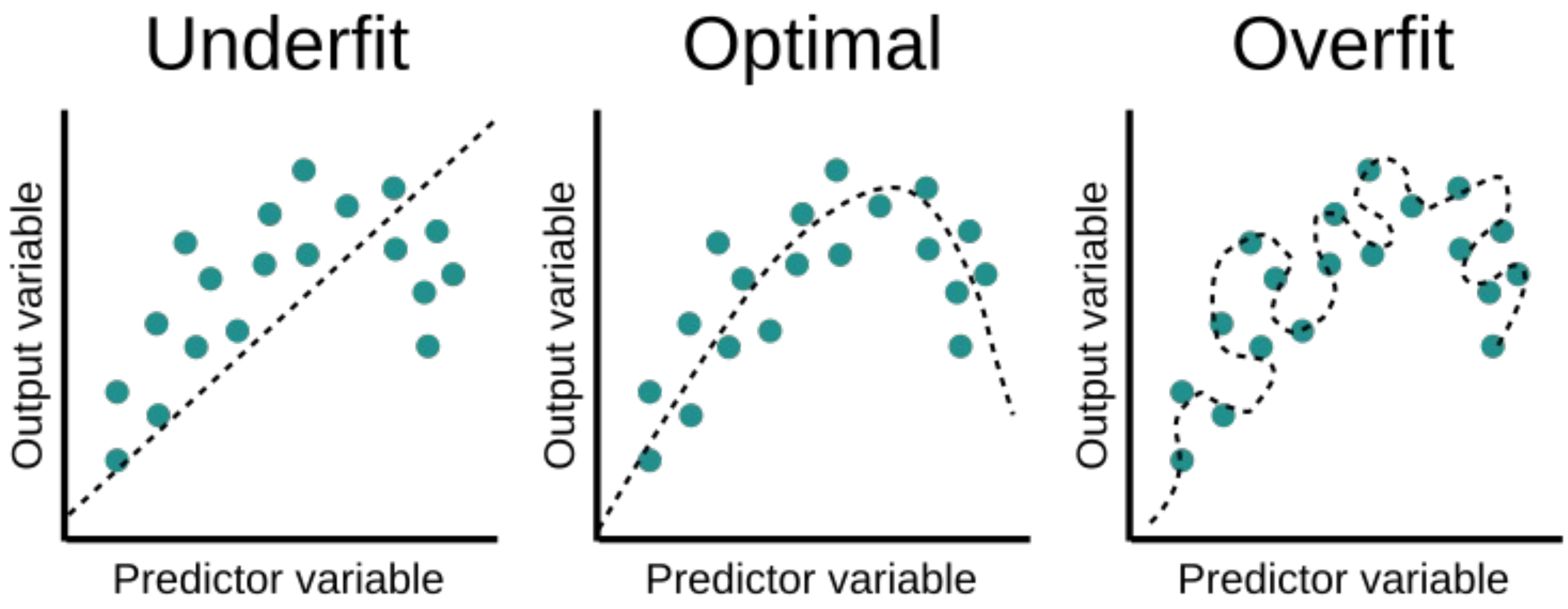
We want to approximate $y = g\,(X)$ by $y \approx \hat{f}(X) + \,\epsilon$

We can only hope to approximate the true function, never find it. This introduces several errors in our model.

The error that our model $\hat{f}$ makes can be divided into three groups

- Bias is the average difference between model predictions and the actual labels
- Variance is the amount of change in the predictions with changes in the training data
- Irreducible Error is the error that cannot be reduced, usually a consequence of noise in the data generating process

# Tradeoffs| Bias vs Variance

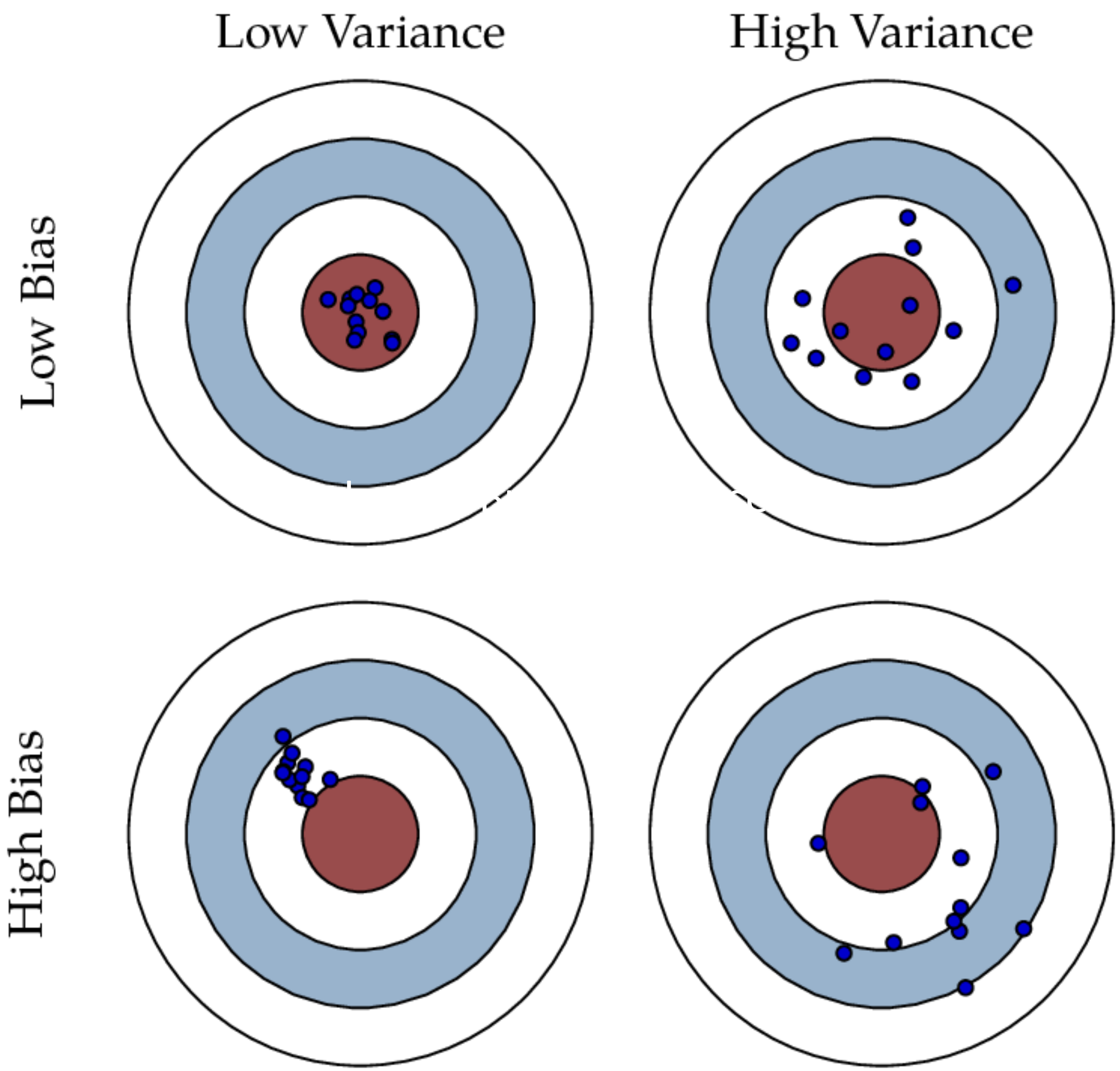



High Bias
Low Variance

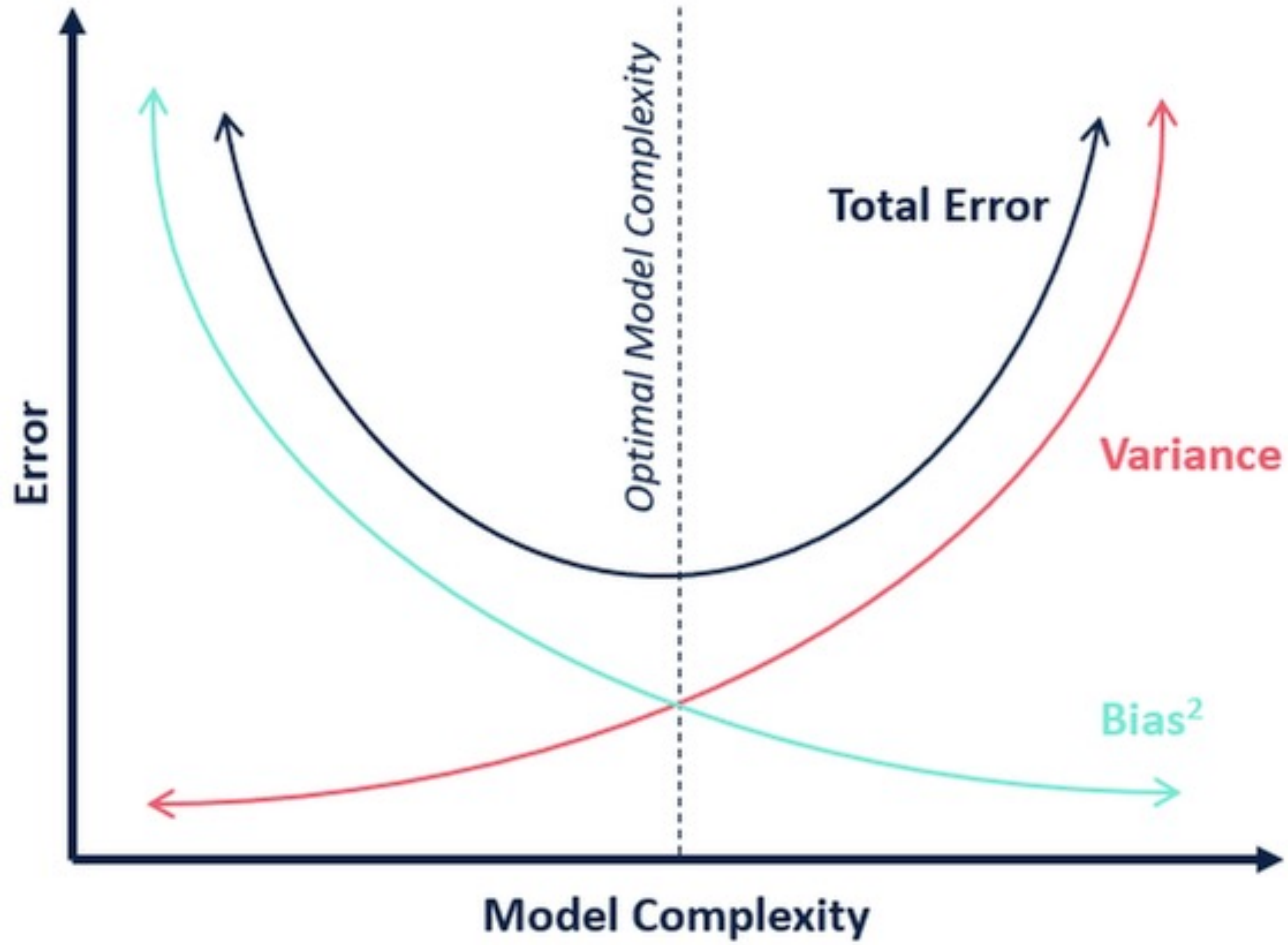
Medium Bias
Medium Variance

Low Bias
High Variance

# Tradeoffs| Bias & Variance



# Tradeoffs| Bias & Variance



# Overcoming Overfitting

# Overcoming Overfitting | Cross Validation

- Technique to extend the validity of any model to unseen data
- Tells us the average performance of the model under all possible circumstances
- Usually, two techniques are used
  - K-Fold CV
  - Leave One Out CV

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| Fold 1 | TEST 1 | TR | TR | TR | TR | TEST | $M_1$ |
| Fold 2 | TR | TEST 2 | TR | TR | TR | TEST | $M_2$ |
| Fold 3 | TR | TR | TEST 3 | TR | TR | TEST | $M_3$ |
| Fold 4 | TR | TR | TR | TEST 4 | TR | TEST | $M_4$ |
| Fold 5 | TR | TR | TR | TR | TEST 5 | TEST | $M_5$ |

# Overcoming Overfitting | Feature Selection

- Technique to only select features relevant to the task at hand
- Reduces model complexity if done right
    - Filter Methods use
    - Wrapper Methods
    - Embedded Methods

# Overcoming Overfitting | Feature Selection | Filter

- They are based on the singular and isolated relationship between the predictor and the response variable

- These can be applied to any model

- They are computationally very fast

- They have statistical foundations

  - Correlation Value
  - Coefficient p-values
  - VIF

# Overcoming Overfitting | Feature Selection | Wrapper

- They are based on fitting models using not all but a subset of features

- These can also be applied to any model

- They are computationally intensive, and mostly impossible if number of features exceed 15

- They operate in a greedy fashion, hence have no statistical foundations, and are often found to be weaker than other feature selection methods

  - Forward Selection
  - Backward Selection
  - Stepwise Selection

# Overcoming Overfitting | Feature Selection | Embedded

- They are based on exploiting the fitted model to select features
- These can be applied to a certain family of models
- They are computationally moderate, feasible even for ~100 features
- They have statistical foundations, rooted in biostatistics
    - Entropy based feature importance
    - MSE based feature importance

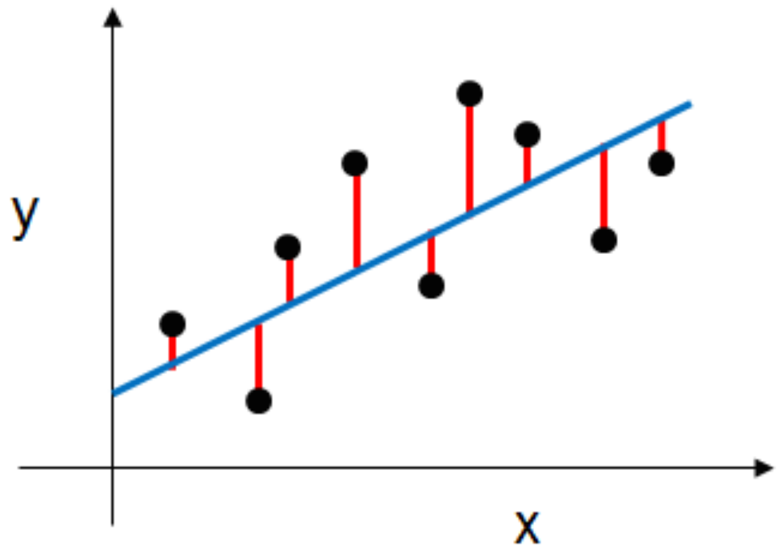
# Overcoming Overfitting | Regularization

- In a regression problem, given a scatter of black points, we want to find the blue line that minimizes the sum of distances of the line to all the black points

- The blue line is represented by a linear combination of all the features

$$y = \beta_1 x_1 + \beta_2 x_2 + + \cdots + \beta_p x_p$$

- The $\beta$ values tell how much the response variable will change with each unit increase in that predictor variable



# Overcoming Overfitting | Regularization

- We can reduce model complexity by trying to impose a constraint on all the values of $\beta$

- The loss function is modified accordingly as

$$Loss_{new} = Loss_{old} + Constraint_{\beta}$$

- For Ridge:
  Loss function $MSE + \lambda \sum_j \beta_j^2$

- For Lasso:
  Loss function $MSE + \lambda \sum_j |\beta_j|$

# Playground

- Colab Notebook for Ridge and Lasso Regression
- Colab Notebook for Bias – Variance Tradeoff

# Q & A