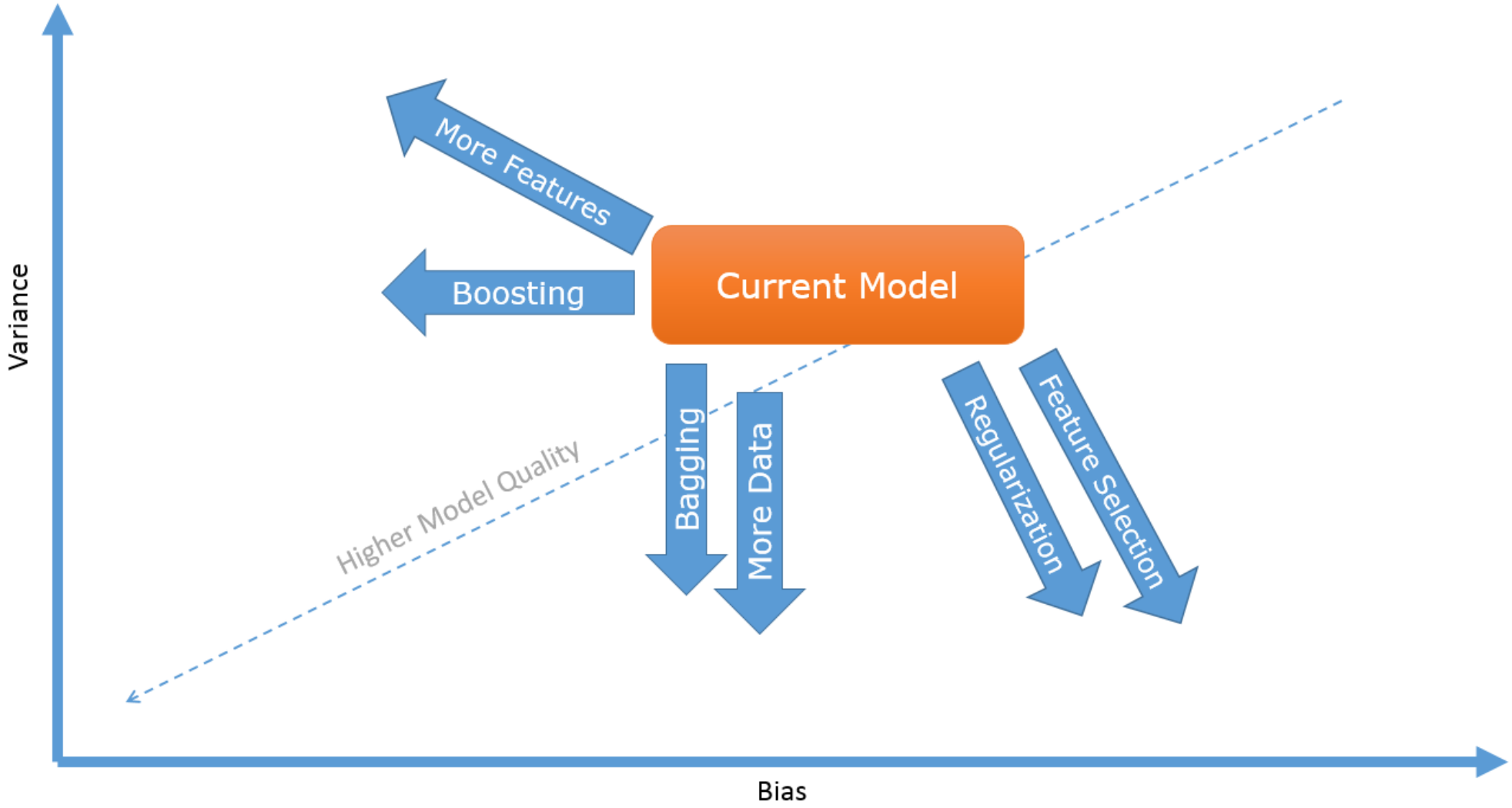
Improving Machine Learning Models

Improving Machine Learning Models



Bias and variance

- Bias and variance are the two components of imprecision in predictive models
- In general there is a trade-off between them
- So normally reducing one tends to increase the other.
- Bias in predictive models means that model is not capturing all the signal it could.
- Bias is also known as under-fitting.
- High variance models tend to perform very well on some data points and really bad on others.
- This is also known as over-fitting
- It means that model is too flexible for the amount of training data you have and ends up picking up noise in addition to the signal, learning random patterns that happen by chance and do not generalize beyond your training data.

Detect Bias / Variance

- High Variance:
  - If your model is performing really well on the training set, but much poorer on the hold-out set
- High Bias:
  - If your model is performing poorly on both training and test data sets, it is suffering from high bias.

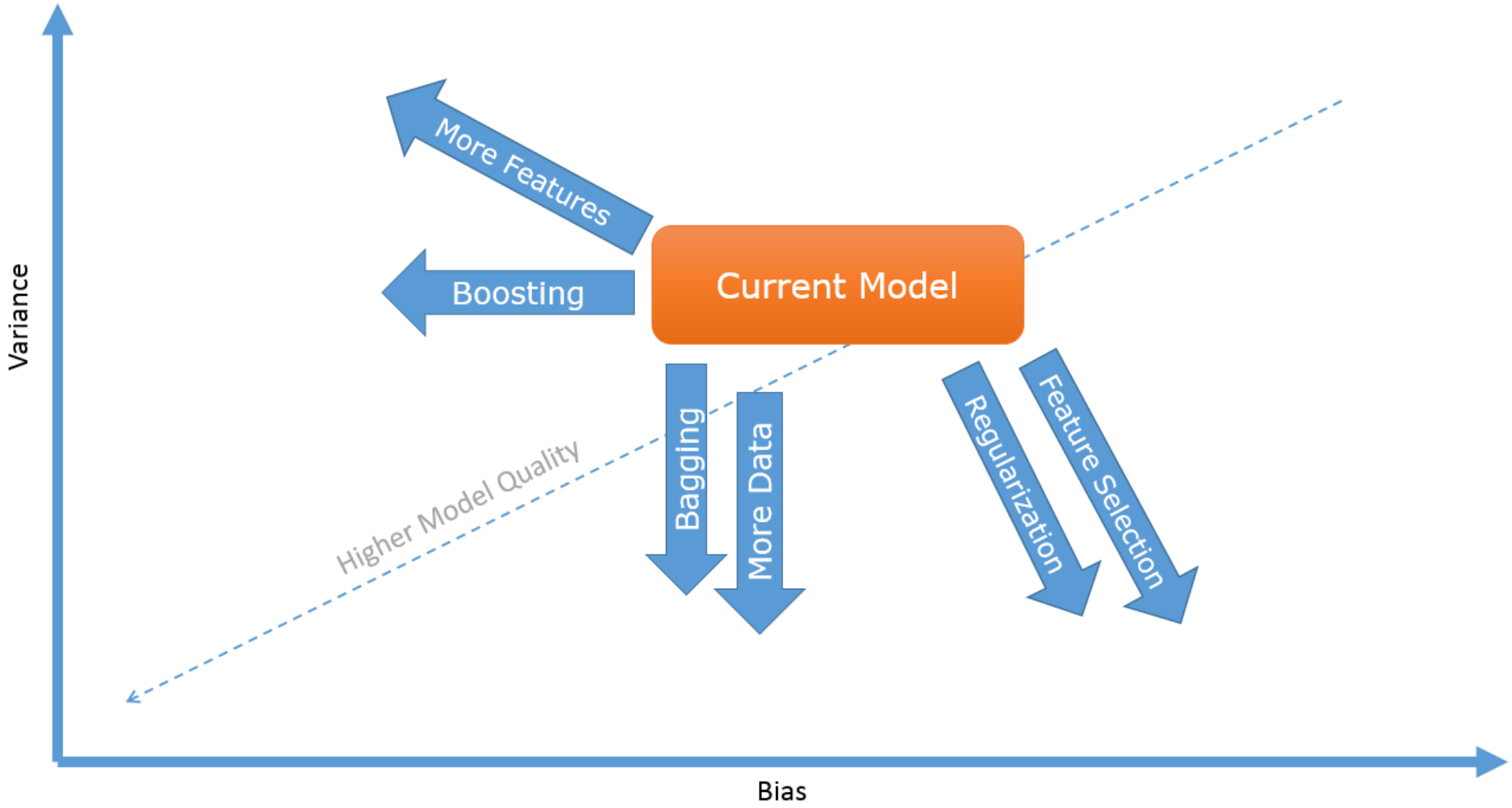
Techniques to bring your model where you want it to be

- Depending on the performance of model and whether it is suffering more from high bias or high variance, you can resort to one or more of these seven techniques to bring your model where you want it to be:
  - Add More Data
  - Add More Features
  - Do Feature Selection
  - Use Regularization
  - Bagging
  - Boosting
  - Use a more different class of models

Which method is to use when?

Technique	When to use?
Add More Data	High Variance / Overfitting
Add More Features	High Bias / Underfitting
Do Feature Selection	High Variance / Overfitting
Use Regularization	High Variance / Overfitting
Bagging	High Variance / Overfitting
Boosting	High Bias / Underfitting
Use a more different model	If our data-set is suitable to some different kind of model, then chose that model.

Lets see this diagram again



Thanks