Module 1: Machine Learning – An overview

Lecture 2- Part 2

Theme: *Factors that influence Hypothesis Space Search*

Topic-1: Overfitting in the Search Space

- Training Error:

$$err_{train}(h) = Pr_{x \in D}[h(x) \neq c(x)] = \frac{count\ of\ mismatches}{|D|}$$

$$h_{target} = \underset{h \in H}{\operatorname{argmin}}\{err_{train}(h)\}$$

- True Error:

$$err_{true}(h) = Pr_{x \in S}[h(x) \neq c(x)]$$

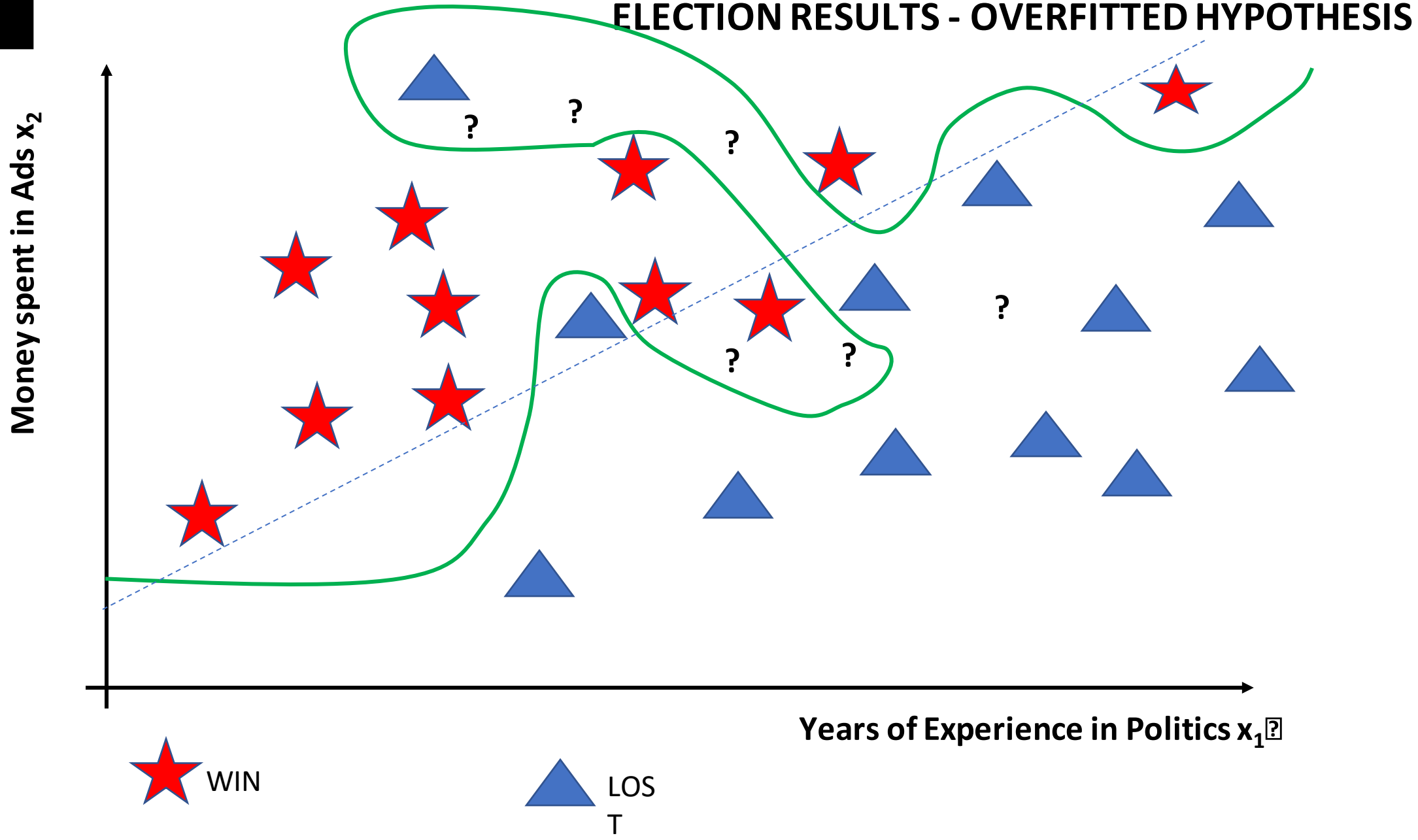
- Overfitting:

$$err_{true}(h) \gg err_{train}(h)$$

Overfitting – Case of over-training

- The model performs quite well on training data, but rather poorly on test data
- Overfitting occurs when a model becomes complex due to overtraining.
- Overfitting is **the symptom** indicating *Inductive Learning Hypothesis stands falsified*

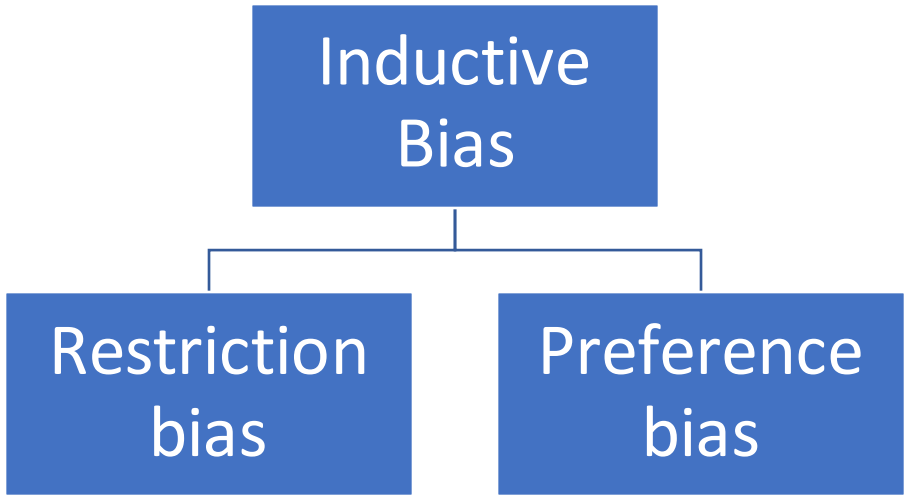
ELECTION RESULTS - OVERFITTED HYPOTHESIS



Optimization Constraints to simplify learning

- Unbiased learning is futile as It leads to an exhaustive search.
- Inductive bias is a *set of assumptions* on *Target Hypothesis* to achieve generalization
- Additional constraints are imposed to guide search and improve generalization

Additional Constraints

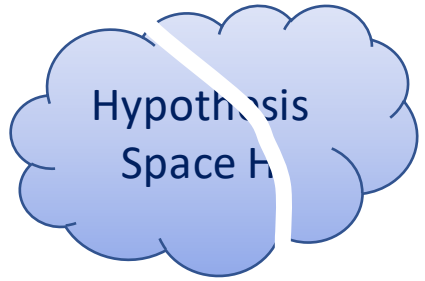


Topic 2: Restriction & Preference Biases

I. Restriction Bias

- Cuts off portions of the Hypothesis space

  Incomplete Hypothesis space
- Forces a less expressive learning model to improve generalization
- Takes Less time to search through the hypothesis space



Restriction Bias

- If hypothesis is a ***Boolean expression***, Allow Only AND terms (conjunctions), ***not*** AND-OR terms (disjunctions of conjunctions)

Go to Cinema?

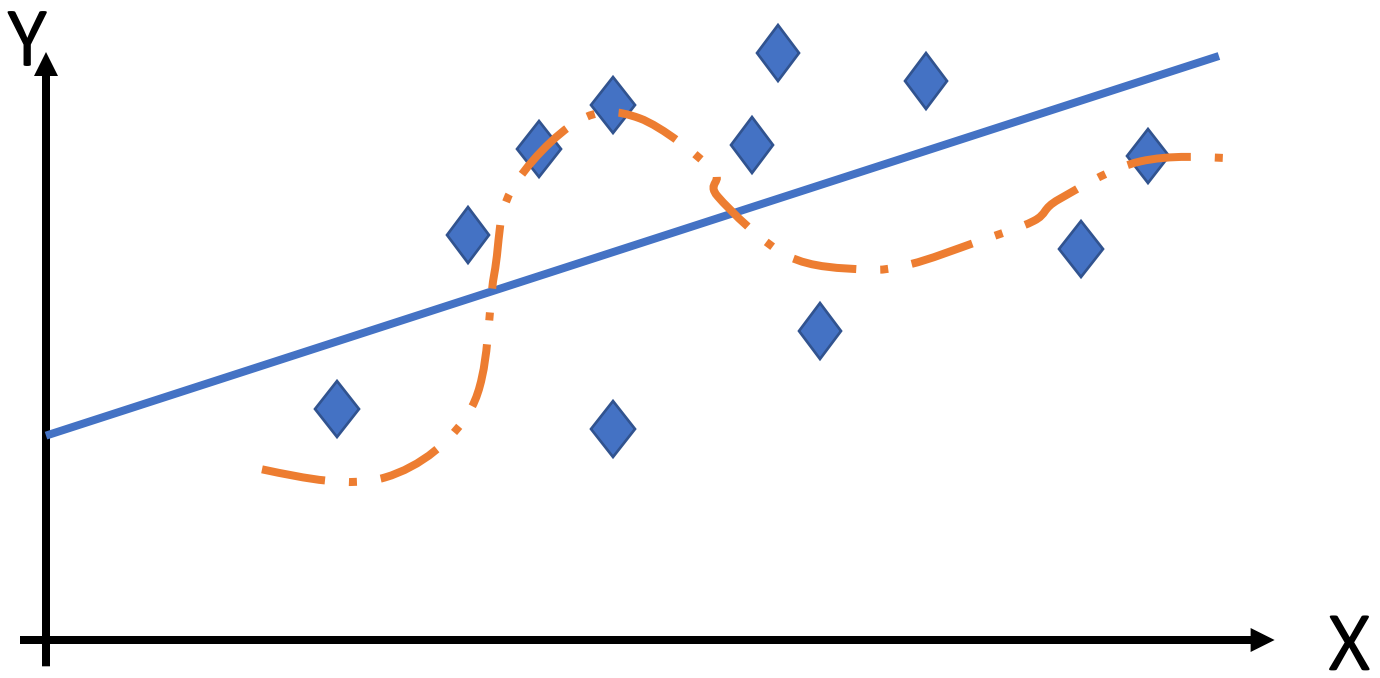
Weekend ?	Pending work?	New Movie?	Friend Available?	See Movie?
Y	N	Y	Y	Y
N	N	Y	Y	Y
N	Y	Y	Y	N

- Not allowed: *If Weekend OR New Movie THEN See movie*

RESTRICTION BIAS

- If hypothesis is ***a curve*** that separates or fits data***:***

  Allow only first order linear x terms, do not allow higher order terms such as : x^2, $x_1 * x_2$ etc. in the equation of curve:



II. Preference bias

Principle of OCCAM's Razor* or Parsimony: ***Given two models that can solve a problem, Prefer the simpler one***

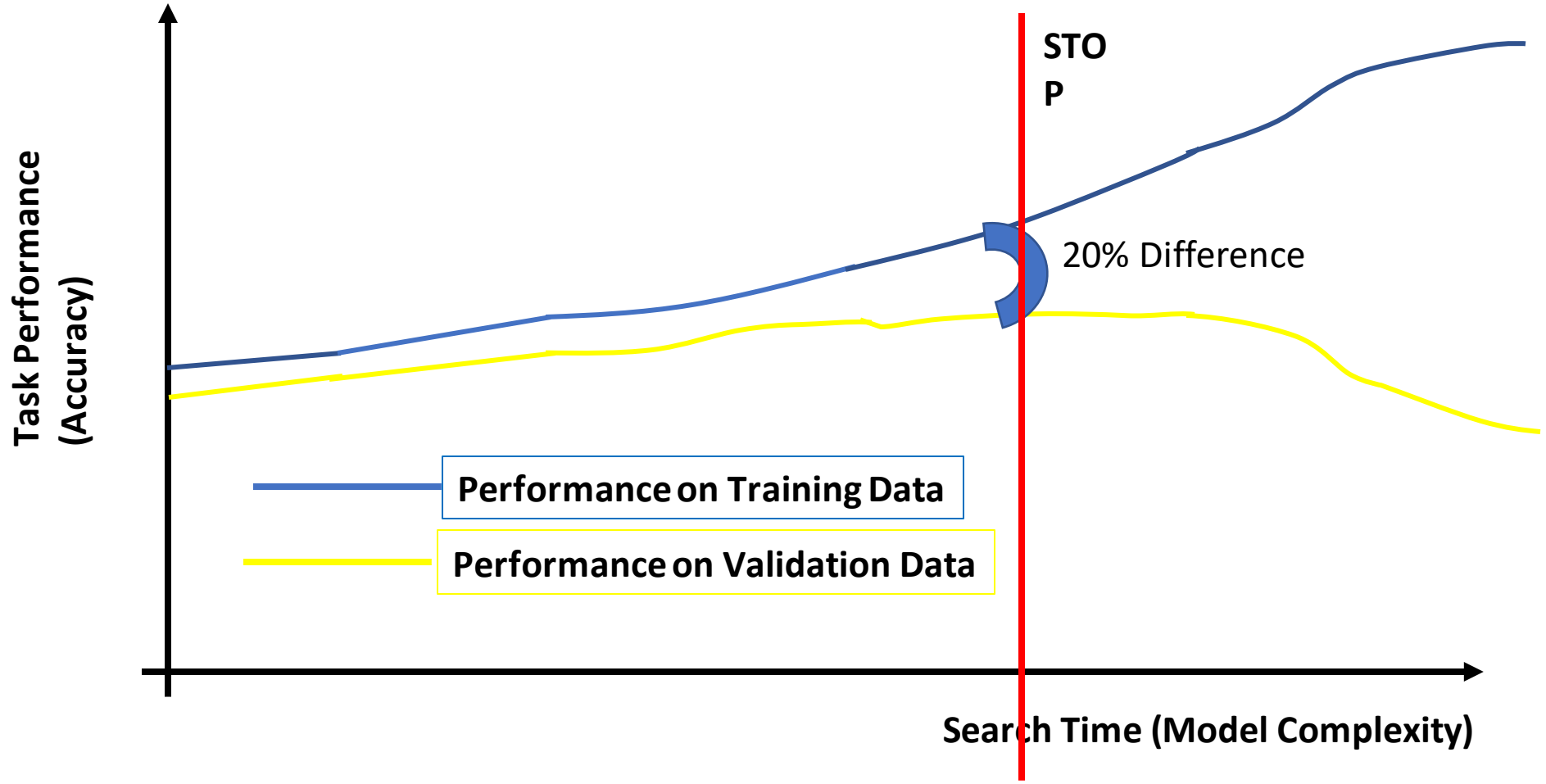
**Occam is a place, Scottish logician Sir William Hamilton of Occam gave the principle in the 14th Century!*

- If Hypothesis is a Decision Tree: Search greedily towards immediate rewarding paths
- If Hypothesis is a Boolean Function: Prefer lesser attributes and lesser number of terms
- If Hypothesis is a Decision Tree: Prefer shorter trees over long trees

Preference Bias - Regularization

- If Hypothesis is curve: $w_0+w_1x_1+w_2x_2=0$, *Prefer lower values of weights* especially for unimportant attributes
- The weight parameters are normally adjusted by optimizing the Objective Function (OF) during the Hypothesis space search process: *e.g.* Minimize{MSE} / Maximize{Accuracy}
- Now, OF is augmented with a regularization term to prohibit excessive weights, or even make them vanish!
- This process is called Regularization

Detecting and Arresting Overfitting during Hypothesis Space Search



Topic 3- Underfitting & Data Considerations

Case of Underfitting

- Underfitting happens when the model is too simple
- It gives high training error as it is unable to extract mappings / concepts
- It gives a constant average error over training data called **Bias error**
- It also gives high true error on test data due to lack of generalization
- It can be tackled by increasing training time, adding more effective features and reducing regularization

Factor C: Data considerations for Improving Search Quality

Recall that Training Data D is but a subset of Sample Space S.
Training data considerations:

- Is Training Data sufficient?

Else learner may underfit and classify all examples similarly

- *Is data noisy (contains measurement, transmission and input errors or has missing values)?*

Else may overfit noisy data and misclassify valid new data!

- *Is data complexity manageable?*

Else it may be difficult to arrange for data

Data considerations

- *Does it represent all kinds of variety in original sample population?*

Else trained *model may not be able to correctly classify new instances which was not represented !*

- *Does it have a balance of all kinds of response (say 'yes' and 'no')*

Else trained model may be biased towards one class, and not be able to classify the less represented class!

Challenges in Preparing Data…..

1. **Domain understanding** – Adopt interdisciplinary approach
   Medical Diagnostics, E-Commerce, Stock analysis
2. **Data Collection:** Can collect only samples from population of real data
3. May need to collect from multiple sources – **Data integration**
   - Published databases
   - Collect field data
   - IOT devices
   - Social media
   - Crowd sourcing

Preparing Data…

4. May need to remove noise or errors – **Data Cleaning**

5. May need to fill up missing values in table – **Data completion**

6. Pre-processing data

- Standardise & Normalize numerical data
- Convert text words to root form
- Find statistical properties of data

7. Feature selection/elimination

Tackling Bias towards a response class

- Part of the training data is reserved for validation / testing a learner.

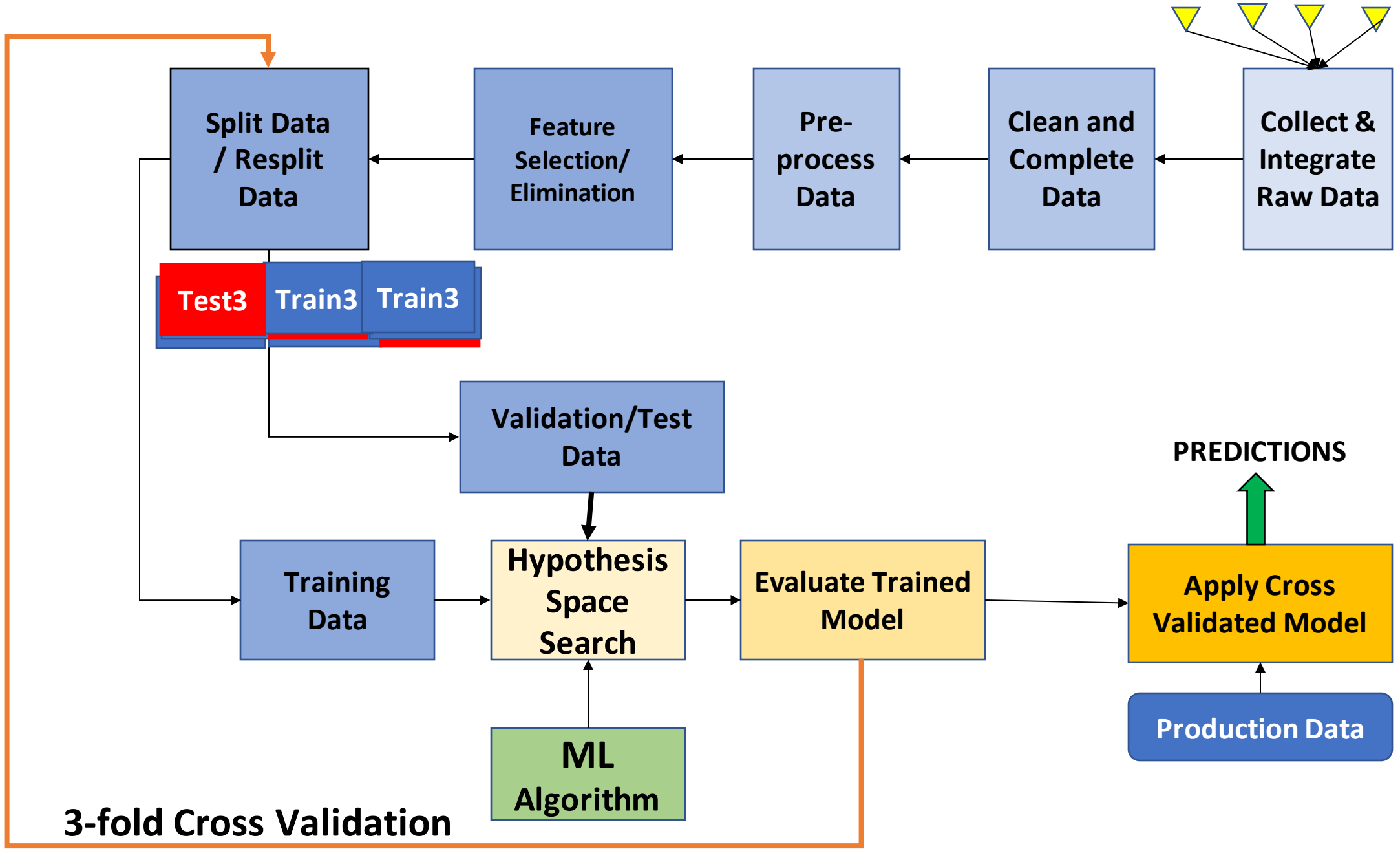
Illustration of biased training / testing:

Training Data For Predicting Recruitment Results

Highest Qualification	Years of Experience	Level of Coding Skill	Reference	Interview Perf.	Result
UG	5	Good	Excellent	Good	**Accept**
PG	3	Med	Good	Bad	**Reject**
UG	6	Poor	Medium	Excellent	**Accept**
UG	2	Good	Excellent	Good	**Accept**
…	…	…	…		…

Testing Data

Highest Qualification	Years of Experience	Level of Coding Skill	Reference	Interview Perf	Result
PG	5	Bad	Med	Good	**???/Reject**
UG	3	Med	Good	Med	**???/Reject**


Collect & Integrate Raw Data
Clean and Complete Data
Pre-process Data
Feature Selection/ Elimination
Split Data / Resplit Data
Test3
Train3
Train3
Validation/Test Data
Training Data
Hypothesis Space Search
Evaluate Trained Model
Apply Cross Validated Model
PREDICTIONS
Production Data
ML Algorithm
3-fold Cross Validation

Learning Points…

- ***Overfitting*** **occurs when a model performs well on training data but poorly on test data (high true error)**
- **To avoid overfitting in supervised learning, In addition to assuming an Inductive Bias on the target hypothesis, additional constraints are imposed on learning, that is – while searching through Hypothesis Space**
  - ***Restriction bias*** **limits expressive power of hypothesis space**
  - ***Preference bias*** **favours simpler models in accordance with Occam's Principle of Parsimony. Preference bias can also prefer searching certain portions of the search space.**
  - ***Regularization – a preference bias,*** **can be achieved in many ways during the Hypothesis Space search process, such as (i) by adding a regularization term in OF (ii) by Stopping the search process early.**

Learning Points

- Underfitting can be detected easily during training and avoided by increasing training time, adding features and reducing regularization.
- To make search effective, Data must be sufficient, have enough variety to represent all possible cases, remain free of bias and have noisy data purged
- Data must be pre-processed to improve its quality. This includes data collection, integration, cleaning, filling missing values, and feature selection/elimination
- Cross-validation ensures that all parts of available data are utilized for training and testing. It removes bias towards a class.

Learning Points

- **Thus, we conclude that there are three factors that influence Hypothesis Search Space:**

1. **Overfitting**
2. **Underfitting**
3. **Data quality**

See Transcript for Link for self-assessment:

"If there is anything worse than knowing too little, it is knowing too much"

Enjoy learning!!