Decision Trees

Introduction and Geometric Intuition

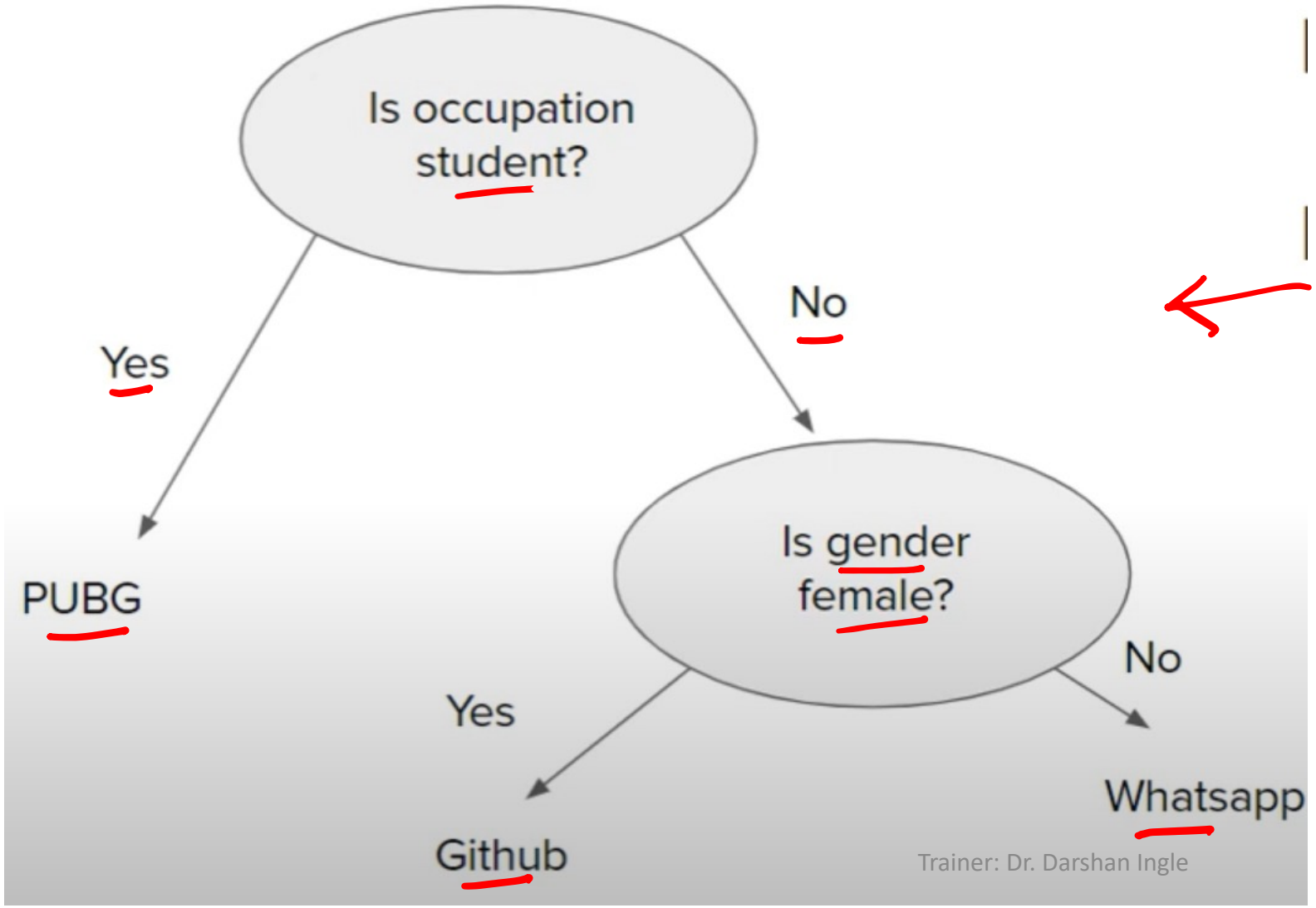
Trainer: Dr Darshan Ingle

Example 1

Gender	Occupation	Suggestion
F	Student	PUBG
F	Programmer	Github
M	Programmer	Whatsapp
F	Programmer	Github
M	Student	PUBG
M	Student	PUBG

```
If occupation==student
    print(PUBG)
Else
    If gender==female
        print(Github)
    Else
        print(Whatsapp)
```

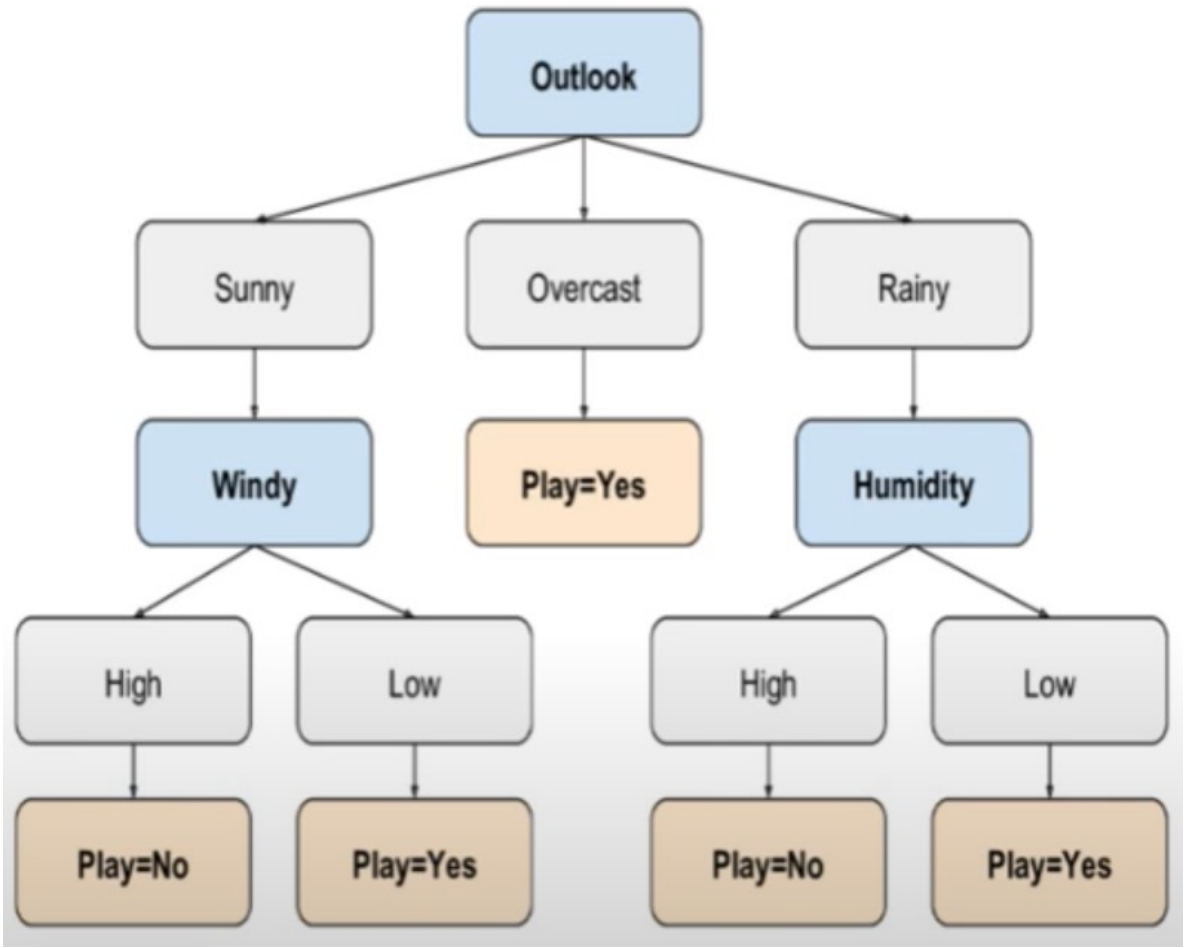
Where is the Tree?

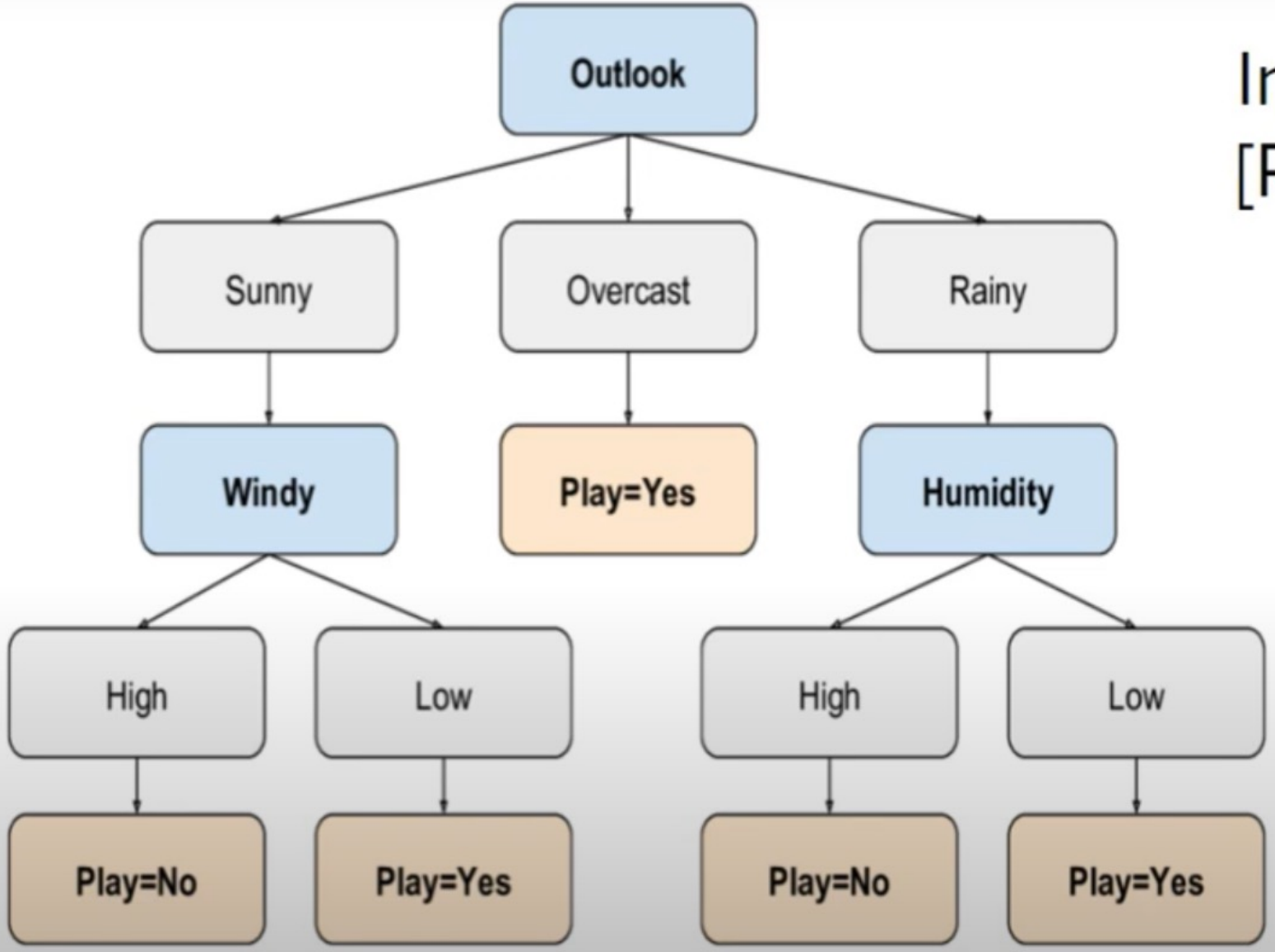



```
If occupation==student
    print(PUBG)
Else
    If gender==female
        print(Github)
    Else
        print(Whatsapp)
```

Example 2

Day	Outlook	Temp	Humid	Wind	Play?
1	Sunny	Hot	High	Weak	No
2	Sunny	Hot	High	Strong	No
3	Overcast	Hot	High	Weak	Yes
4	Rain	Mild	High	Weak	Yes
5	Rain	Cool	Normal	Weak	Yes
6	Rain	Cool	Normal	Strong	No
7	Overcast	Cool	Normal	Strong	Yes
8	Sunny	Mild	High	Weak	No
9	Sunny	Cool	Normal	Weak	Yes
10	Rain	Mild	Normal	Weak	Yes
11	Sunny	Mild	Normal	Strong	Yes
12	Overcast	Mild	High	Strong	Yes
13	Overcast	Hot	Normal	Weak	Yes
14	Rain	Mild	High	Strong	No



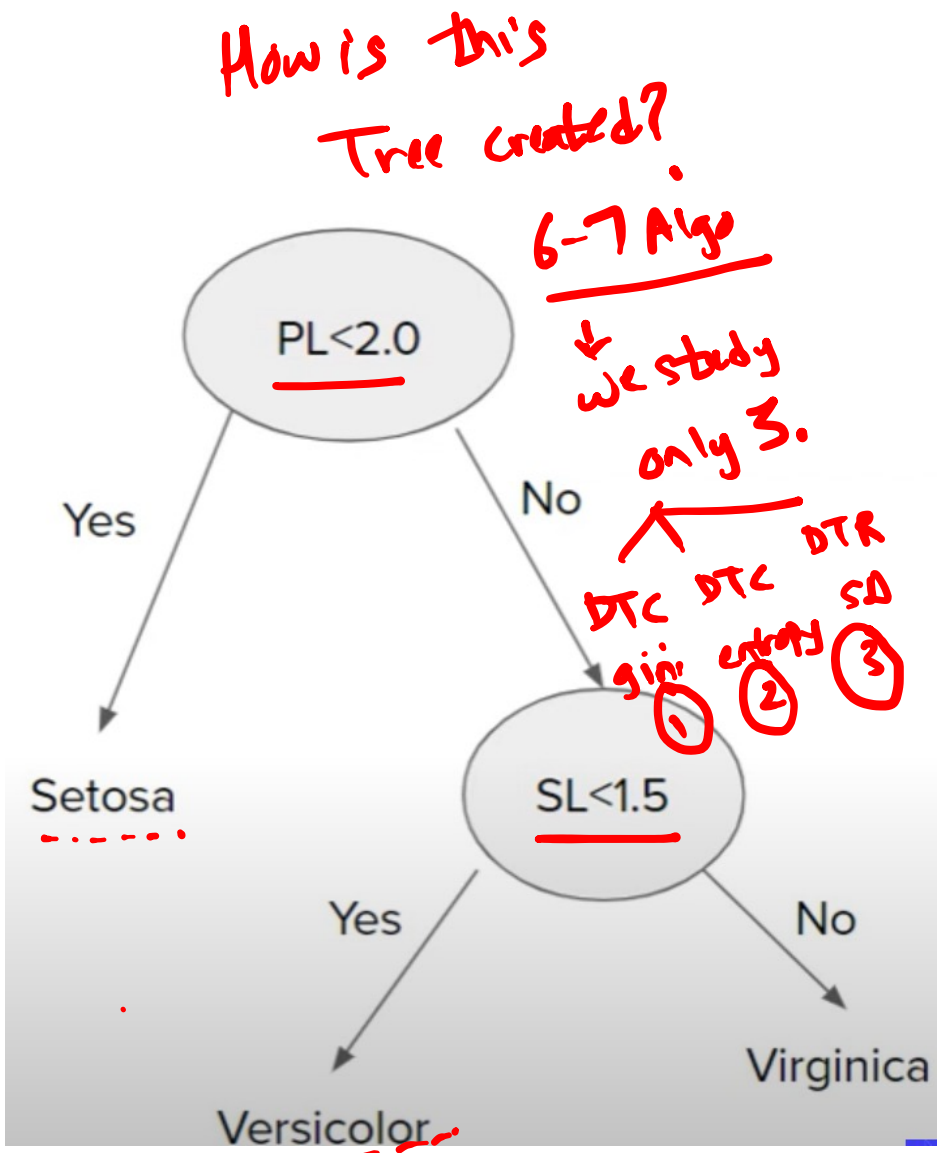


Input query point:
[Rainy, Mild, High, Strong]

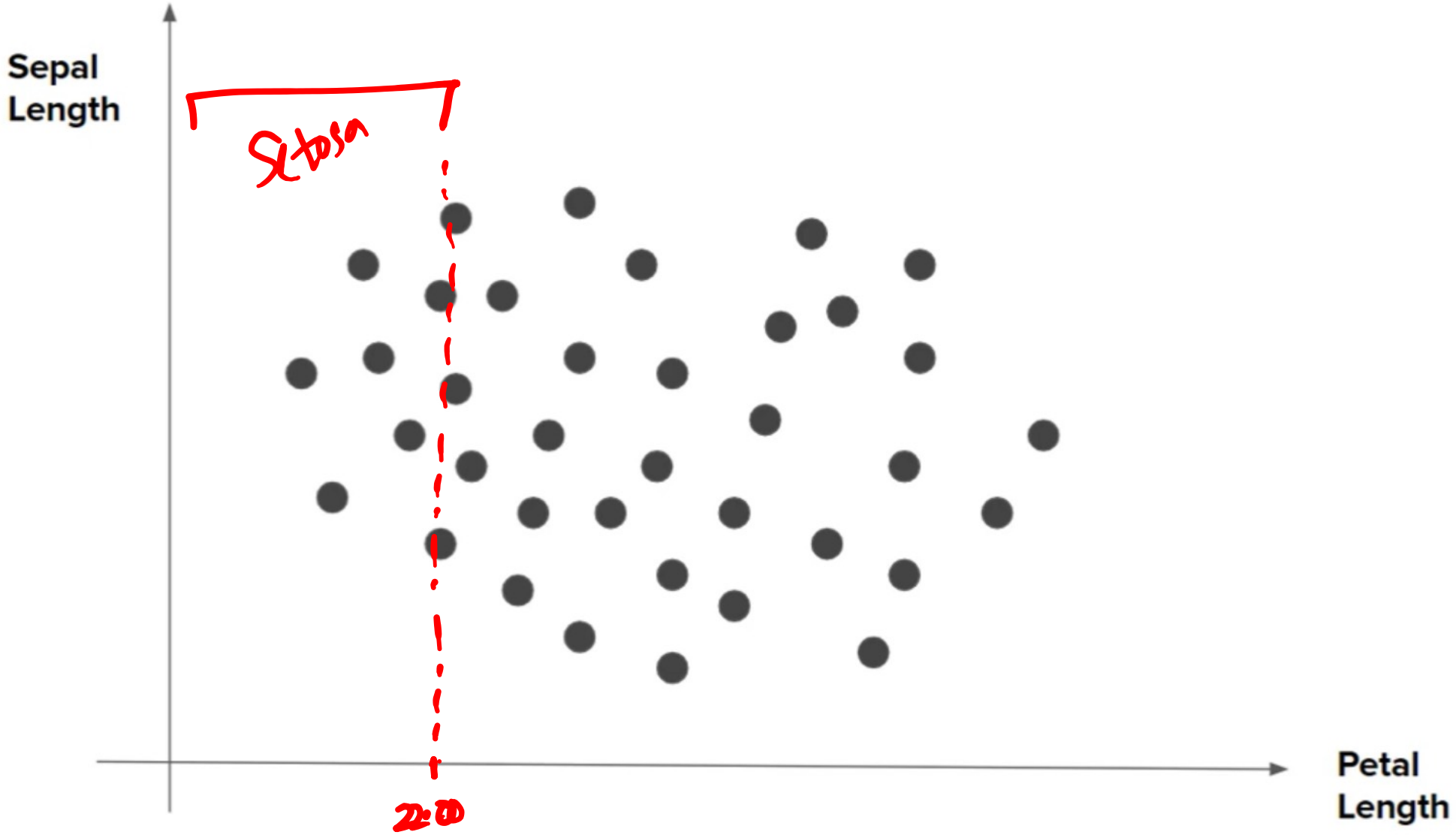
What if we have numerical data?

x-axis X_1 y-axis X_2 y

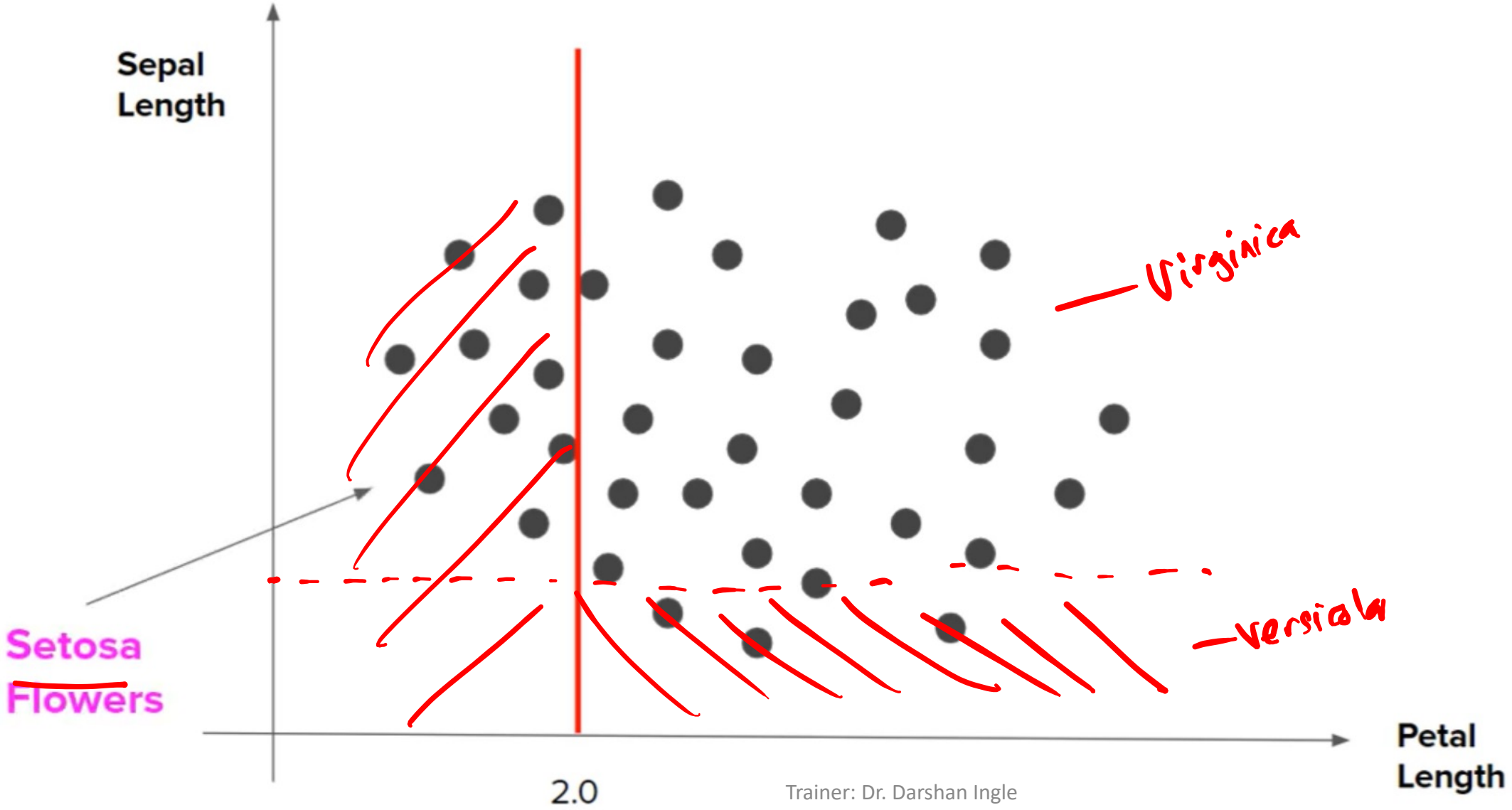
Petal Length	Sepal Length	Type
1.34	0.34	Setosa
3.45	1.45	Versicolor
1.69	0.98	Setosa
2.56	1.79	Virginica
3.00	1.13	Versicolor
1.3	0.88	Setosa



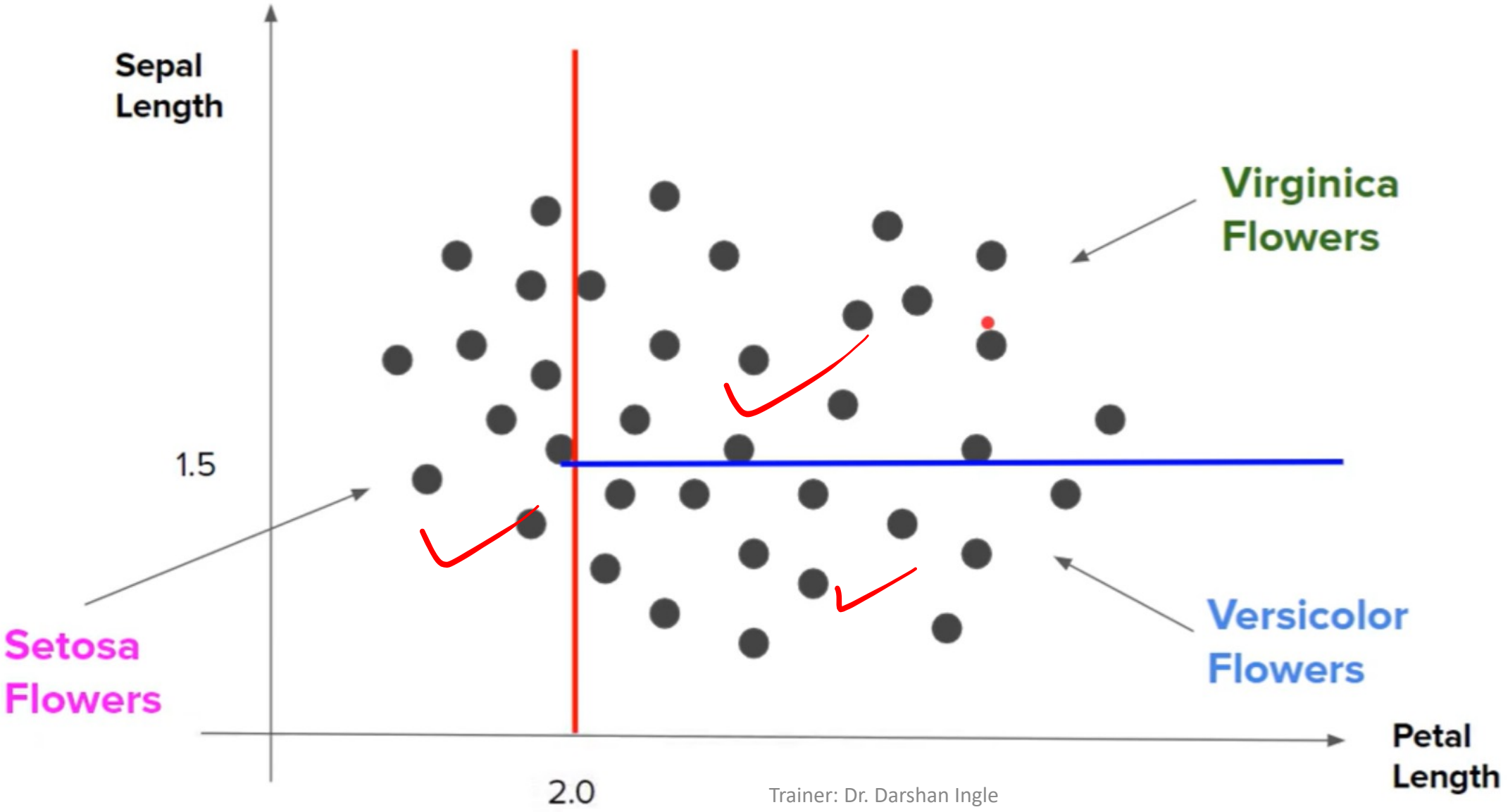
Geometric Intuition



Geometric Intuition



Geometric Intuition



Pseudo code

- Begin with your training dataset, which should have some feature variables and classification or regression output.
- Determine the “best feature” in the dataset to split the data on; more on how we define “best feature” later
- Split the data into subsets that contain the correct values for this best feature. This splitting basically defines a node on the tree i.e each node is a splitting point based on a certain feature from our data.
- Recursively generate new tree nodes by using the subset of data created from step 3.

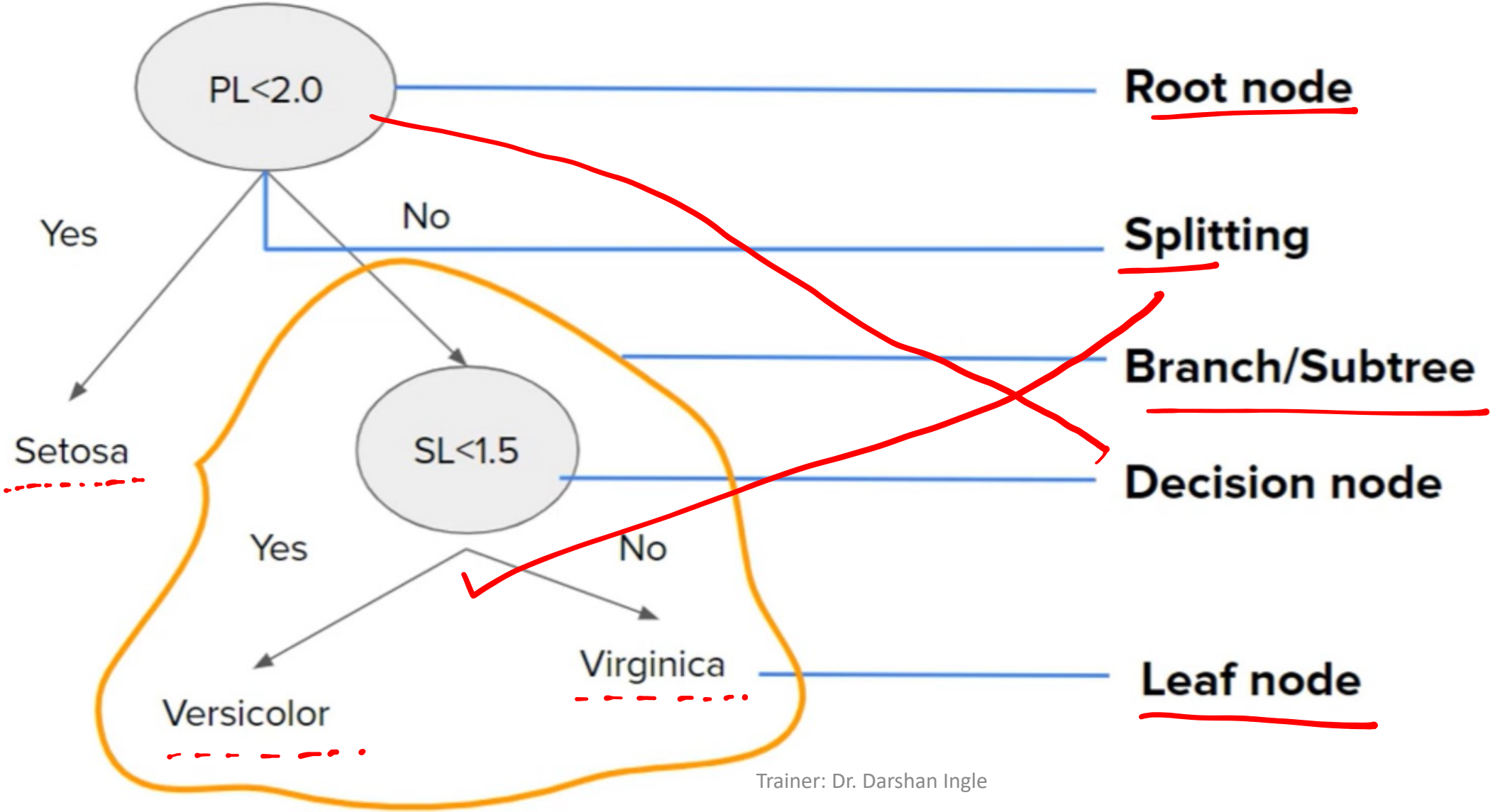
Conclusion

Programatically speaking, Decision trees are nothing but a giant structure of nested if-else condition

Mathematically speaking, Decision trees use hyperplanes which run parallel to any one of the axes to cut your coordinate system into hyper cuboids

Hyperplane → 3D AD ....

Terminology



Some unanswered questions

How to decide which column should be considered as root node?

How to select subsequent decision nodes?

How to decide splitting criteria in case of numerical columns?

PL<2.0

SL<1.5

Advantages

Intuitive and easy to understand

Minimal data preparation is required

The cost of using the tree for inference is **logarithmic** in the number of data points used to train the tree

Disadvantages

Overfitting

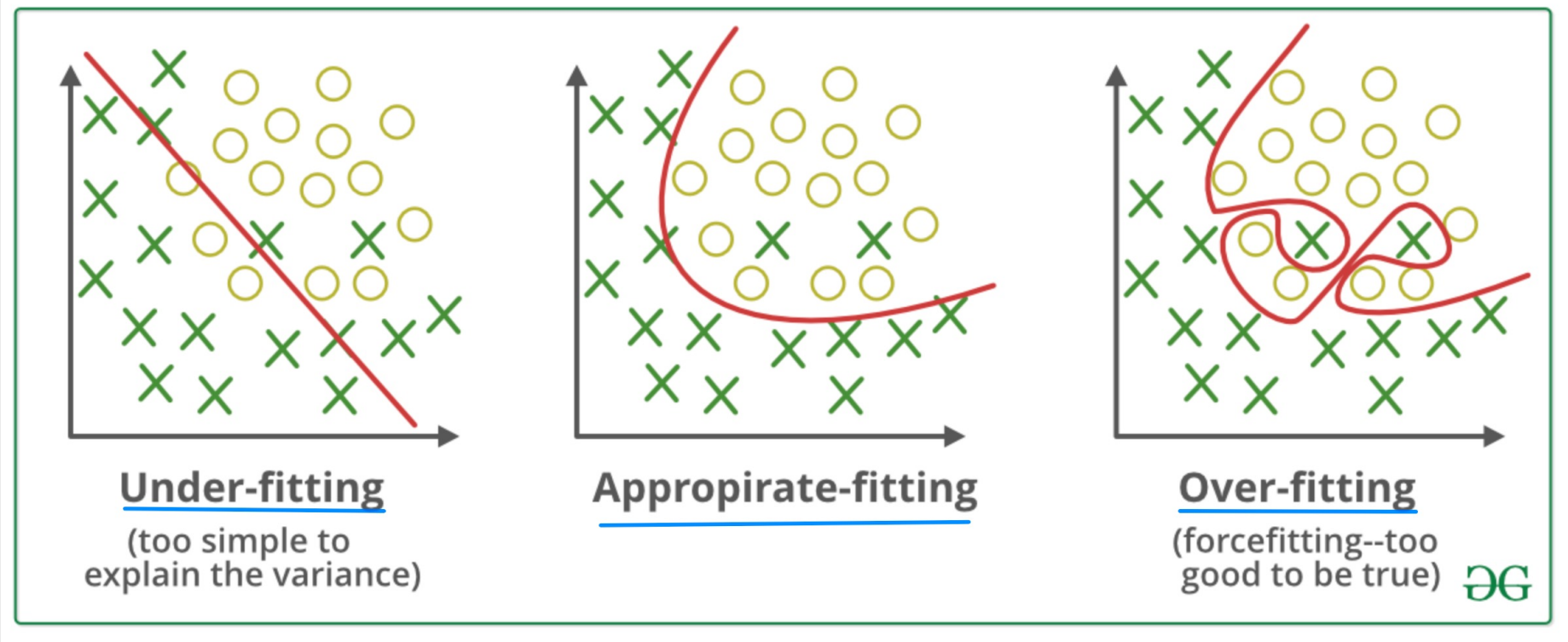
Prone to errors for imbalanced datasets

Overfitting :

Inbox (29,7 | (99+) Feed | Google Cale | Post Attend | My Meeting | decision tre | 1*HXQJ8Eb | precision ac | overfitting u | overfitting_

media.geeksforgeeks.org/wp-content/cdn-uploads/20190523171258/overfitting_2.png

Classification

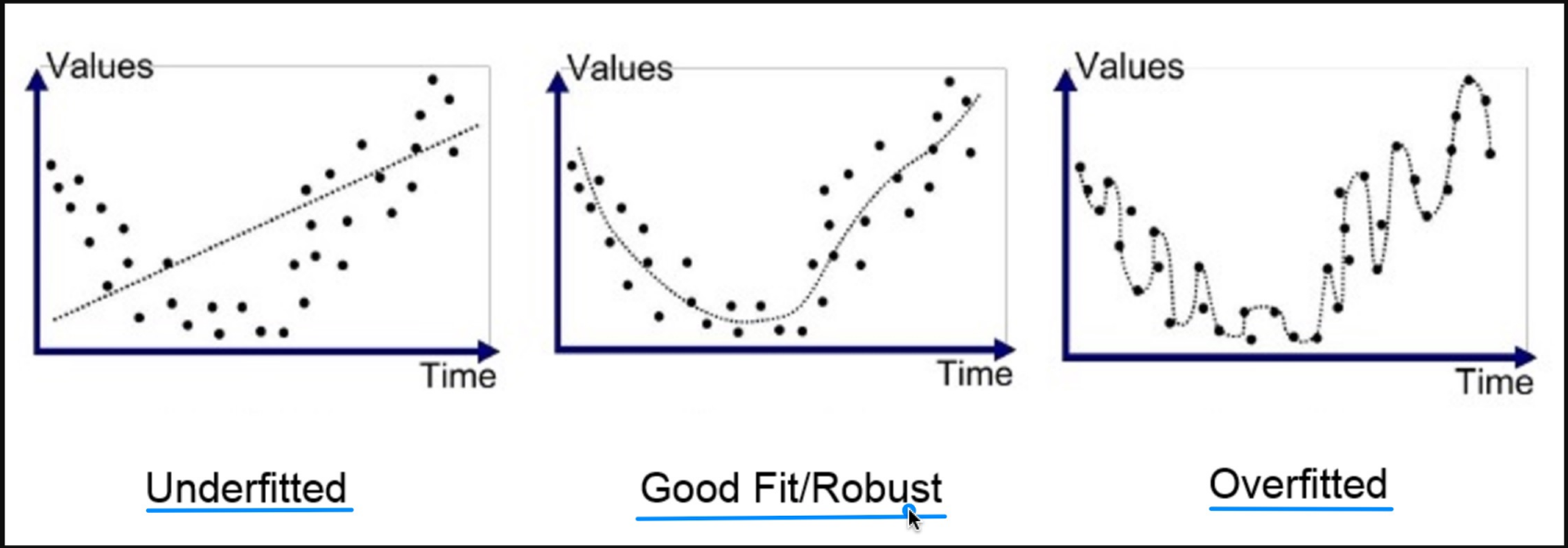


You are screen sharing | Stop Share

Inbox (29,784) | (99+) Feed | Li | Google Calend | Post Attendee | My Meetings | decision tree | 1*HXQJ8Eb8v | precision accu | 1*_7OPgojau8

miro.medium.com/max/1125/1*_7OPgojau8hkiPUiHoGK_w.png

Regression:



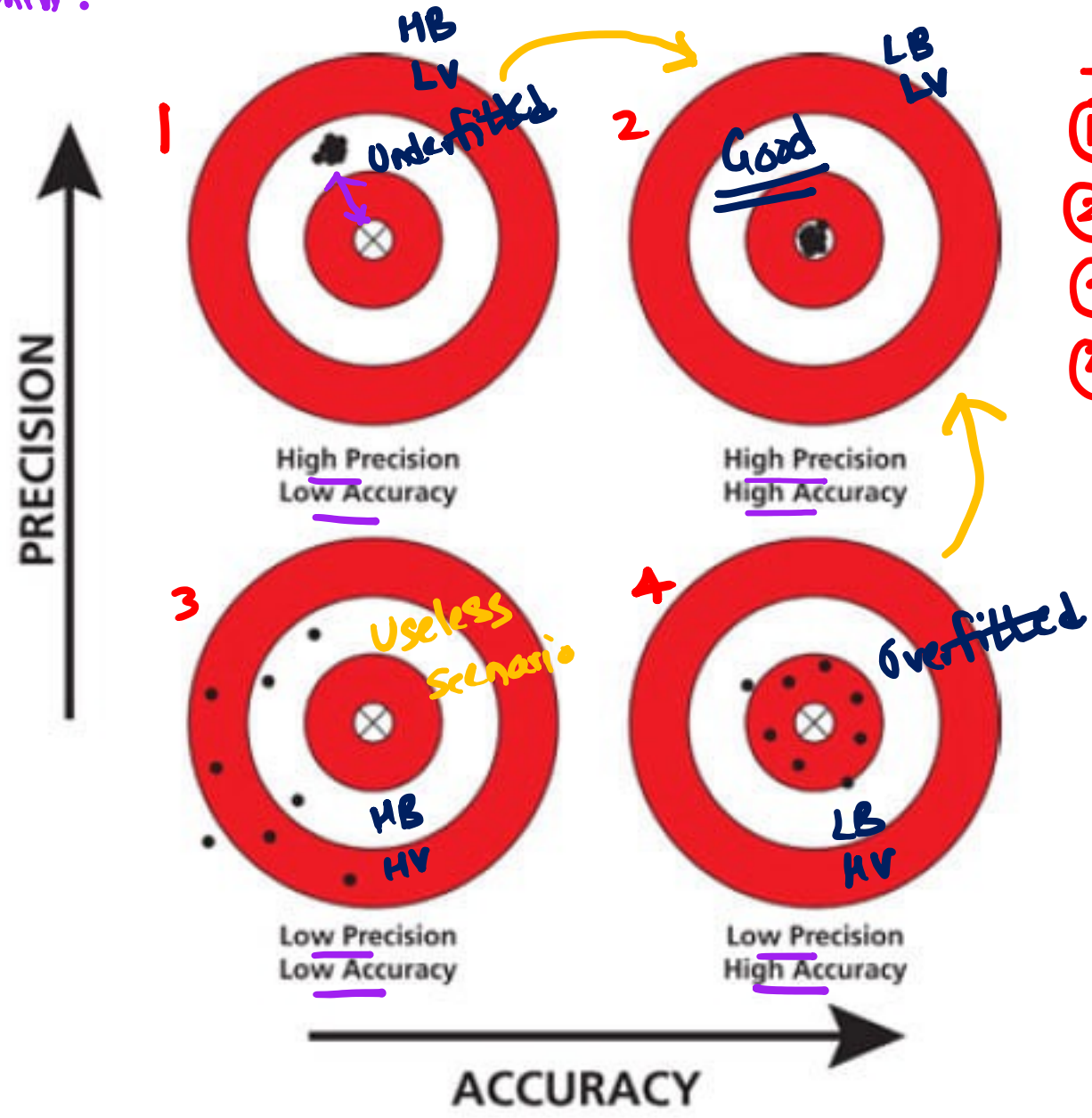
You are screen sharing | Stop Share

Precision: Spread is less/min.

Accuracy: Close to target / less error.

Bias: error

Variance: Spread of data



Archery: 20 arrows

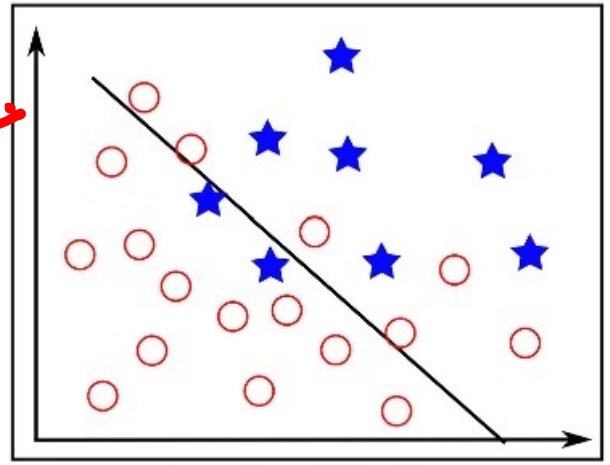
(1) Arjun (Mahabharat)^2

(2) Arjun with squint eyes !

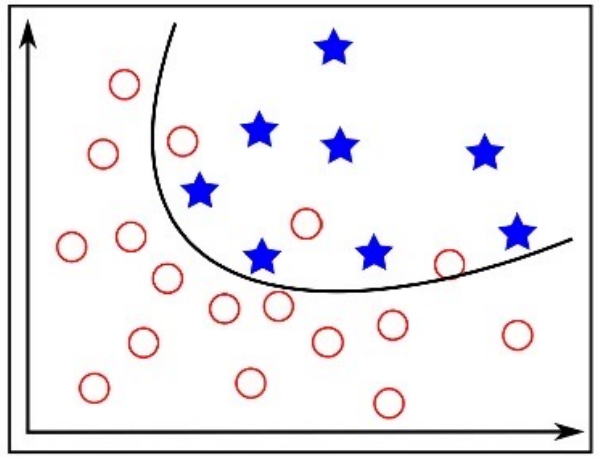
(3) Darshan +

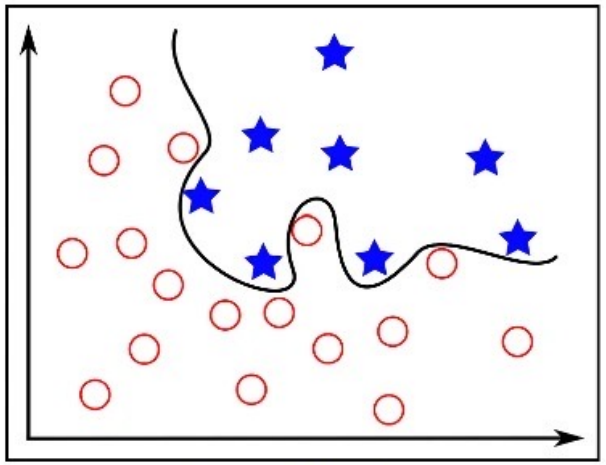
(4) Darshan with squint eyes 3

Classification:



Underfitting

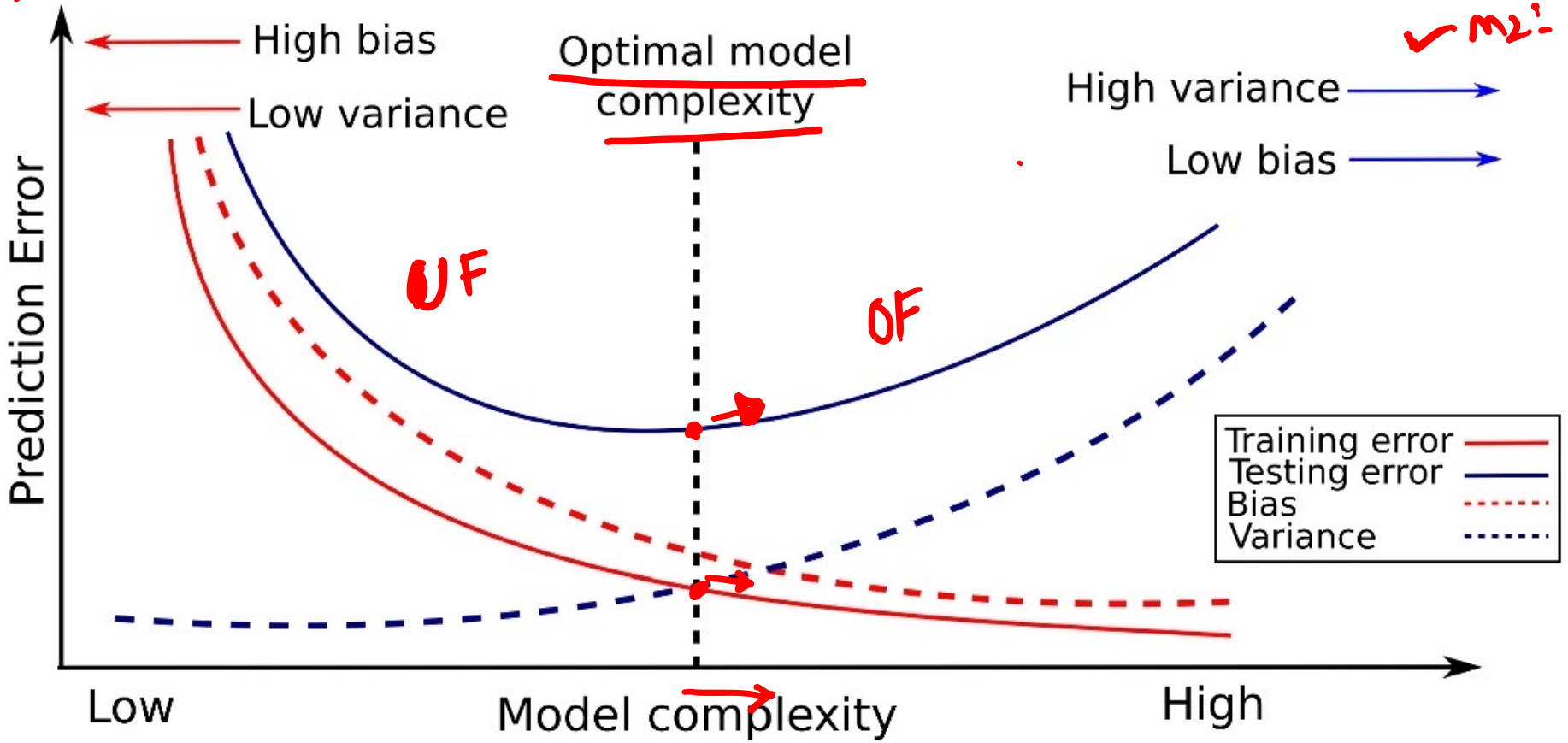




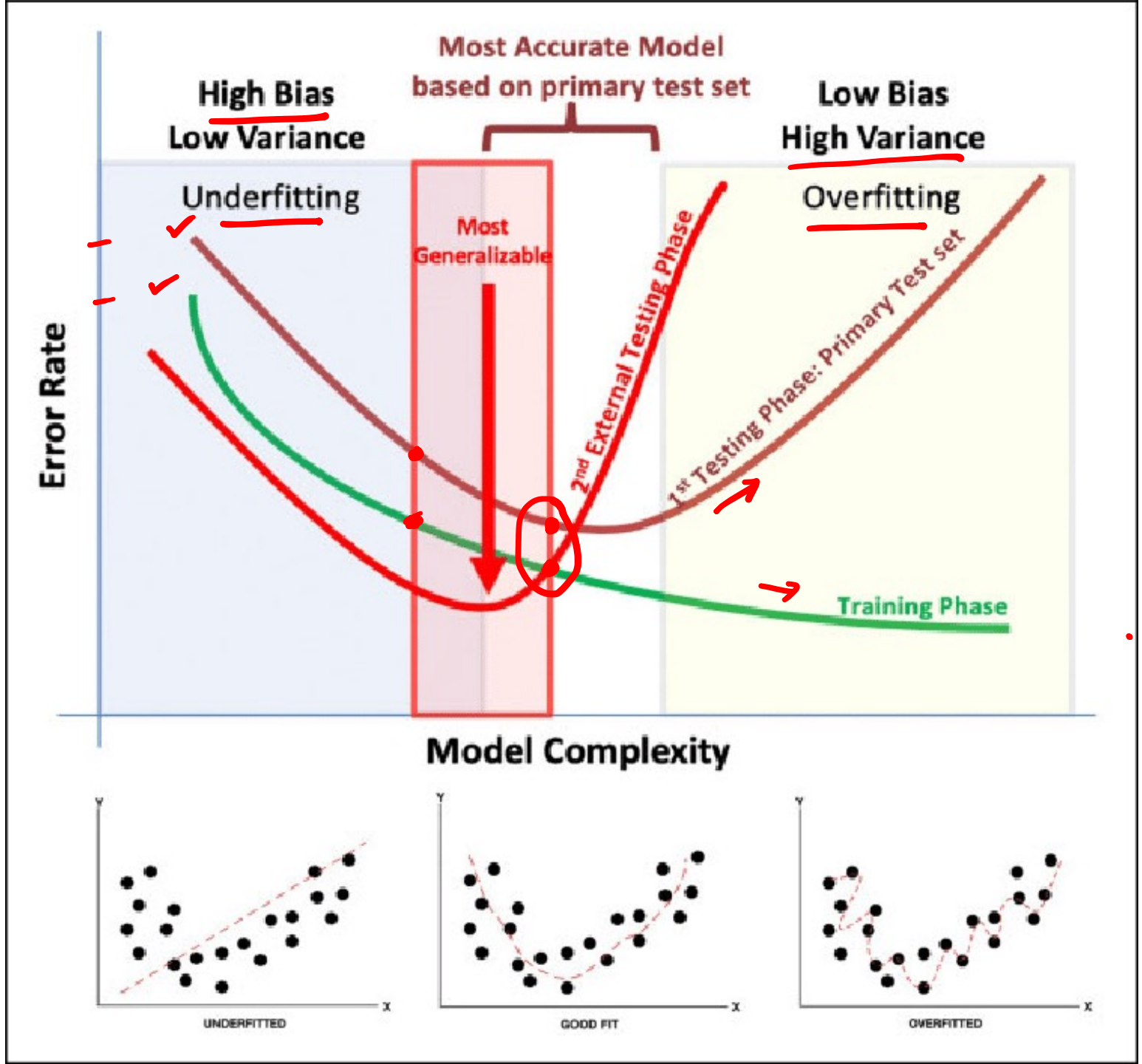
Overfitting

Acc. Tr. | Acc Test
M1: 90 | 85
M2: 89 | 88.5

fit(X_tr, y_tr)
pred(X_test)



Regression:


Most Accurate Model
based on primary test set
High Bias
Low Variance
Underfitting
Most
Generalizable
Low Bias
High Variance
Overfitting
2nd External Testing Phase
1st Testing Phase: Primary Test set
Training Phase
Error Rate
Model Complexity
UNDERFITTED
GOOD FIT
OVERFITTED

CART - Classification and Regression Trees

The logic of decision trees can also be applied to regression problems, hence the name CART

A fun example

https://en.akinator.com/

Thank You