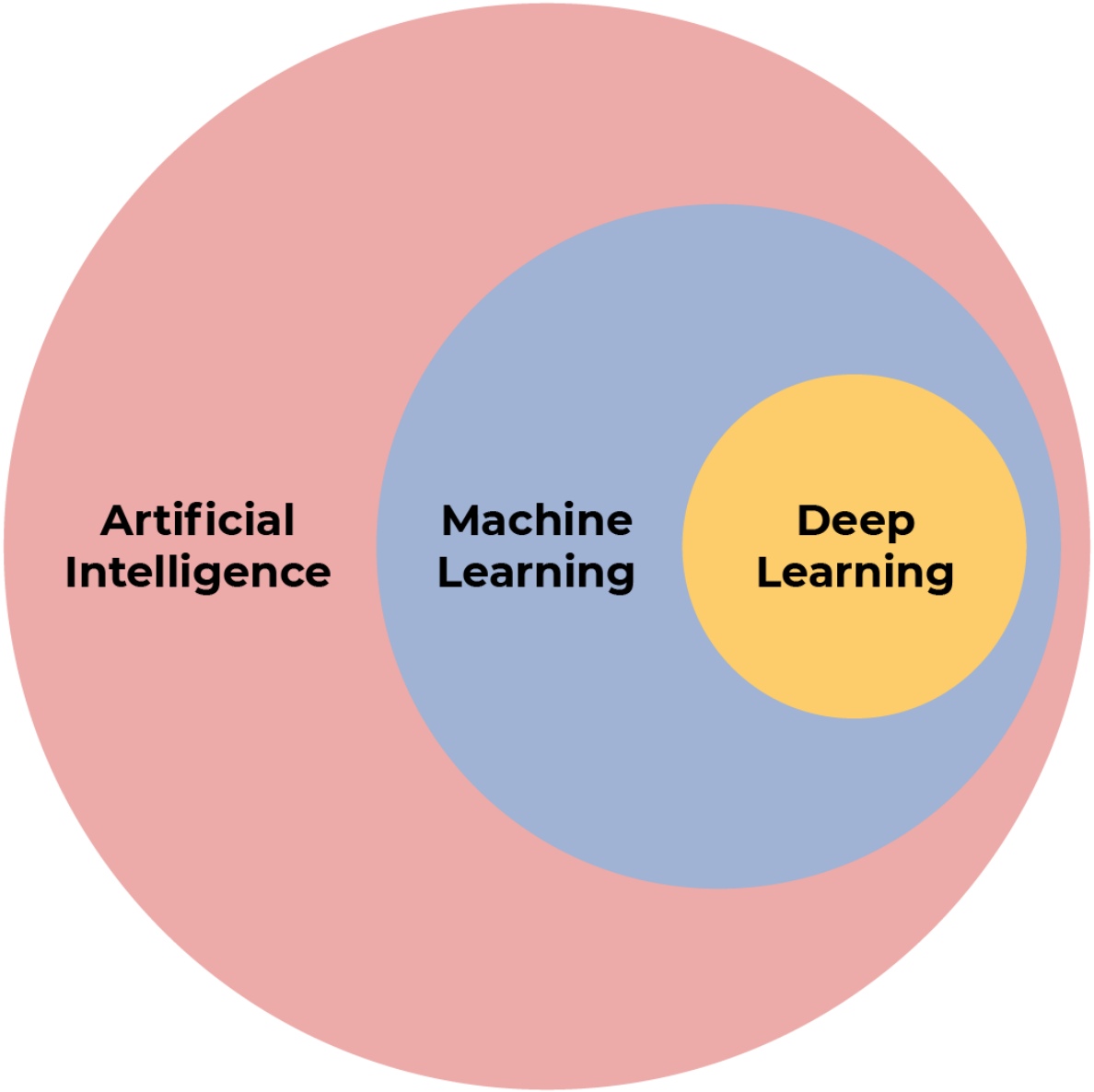
A review on machine learning concepts

Ali Kohan

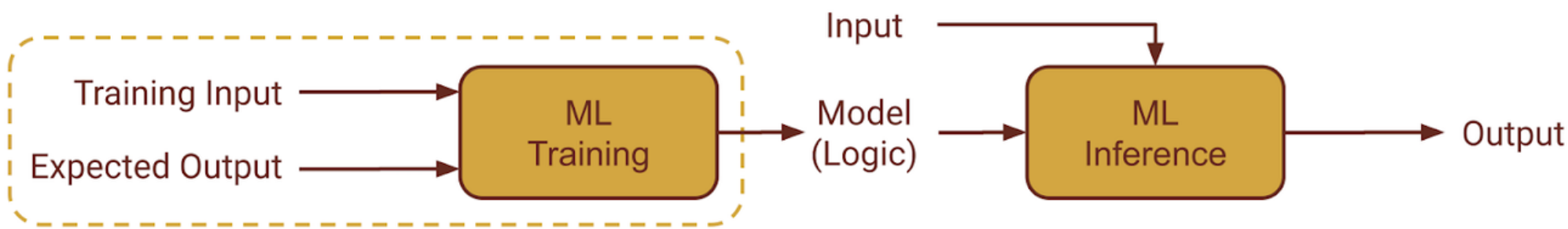
Github.com/alikohan

linkedin.com/in/alikohan/


Artificial Intelligence
Machine Learning
Deep Learning

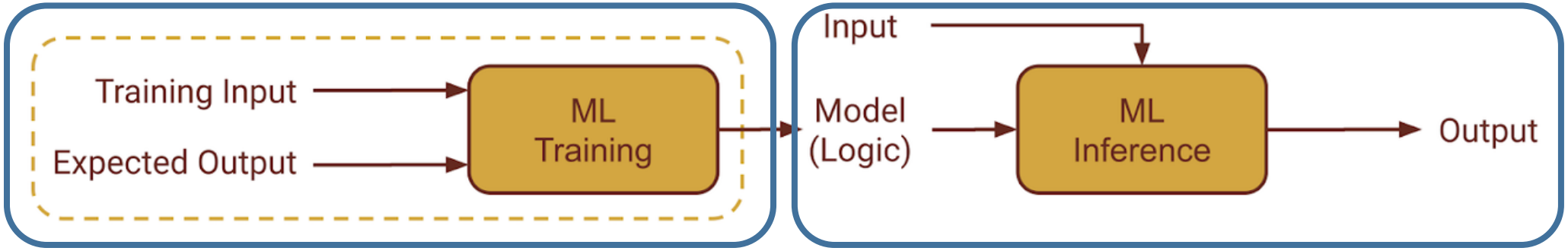
Traditional Programs: Define algo/logic to compute output

Machine Learning: Learn model/logic from data

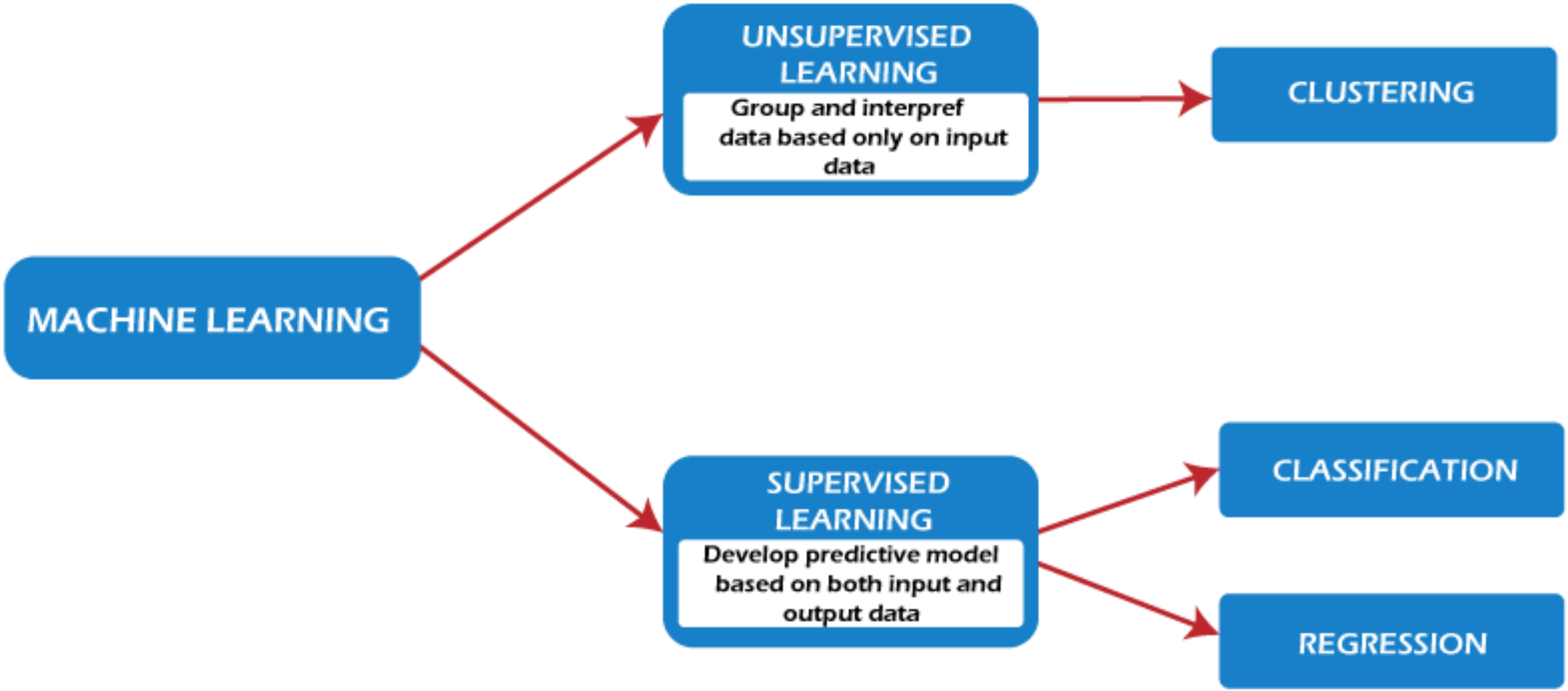


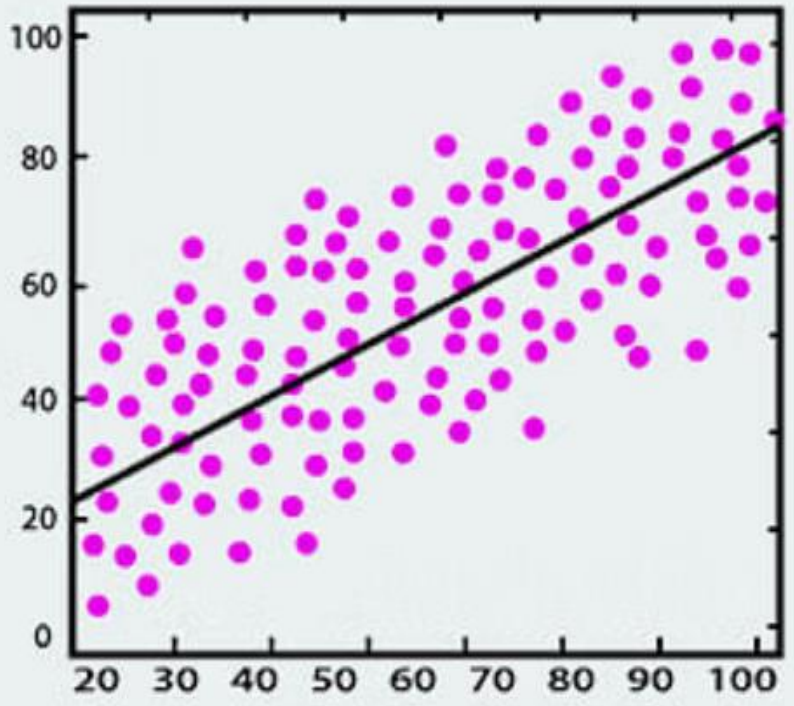
Traditional Programs: Define algo/logic to compute output

Machine Learning: Learn model/logic from data



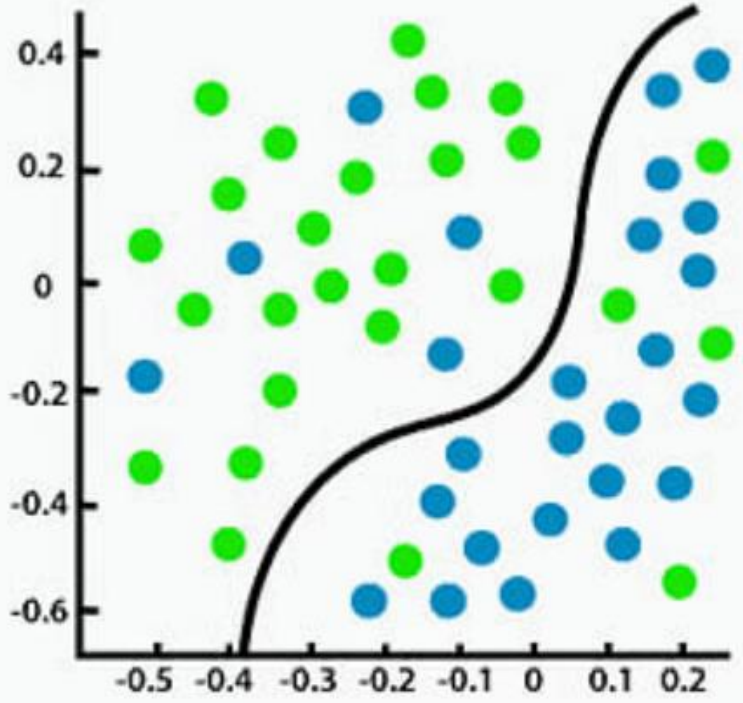
Training time **Inference time**


MACHINE LEARNING
UNSUPERVISED LEARNING
Group and interpref data based only on input data
CLUSTERING
SUPERVISED LEARNING
Develop predictive model based on both input and output data
CLASSIFICATION
REGRESSION



Regression

versus



Classification

Preliminary: Classification

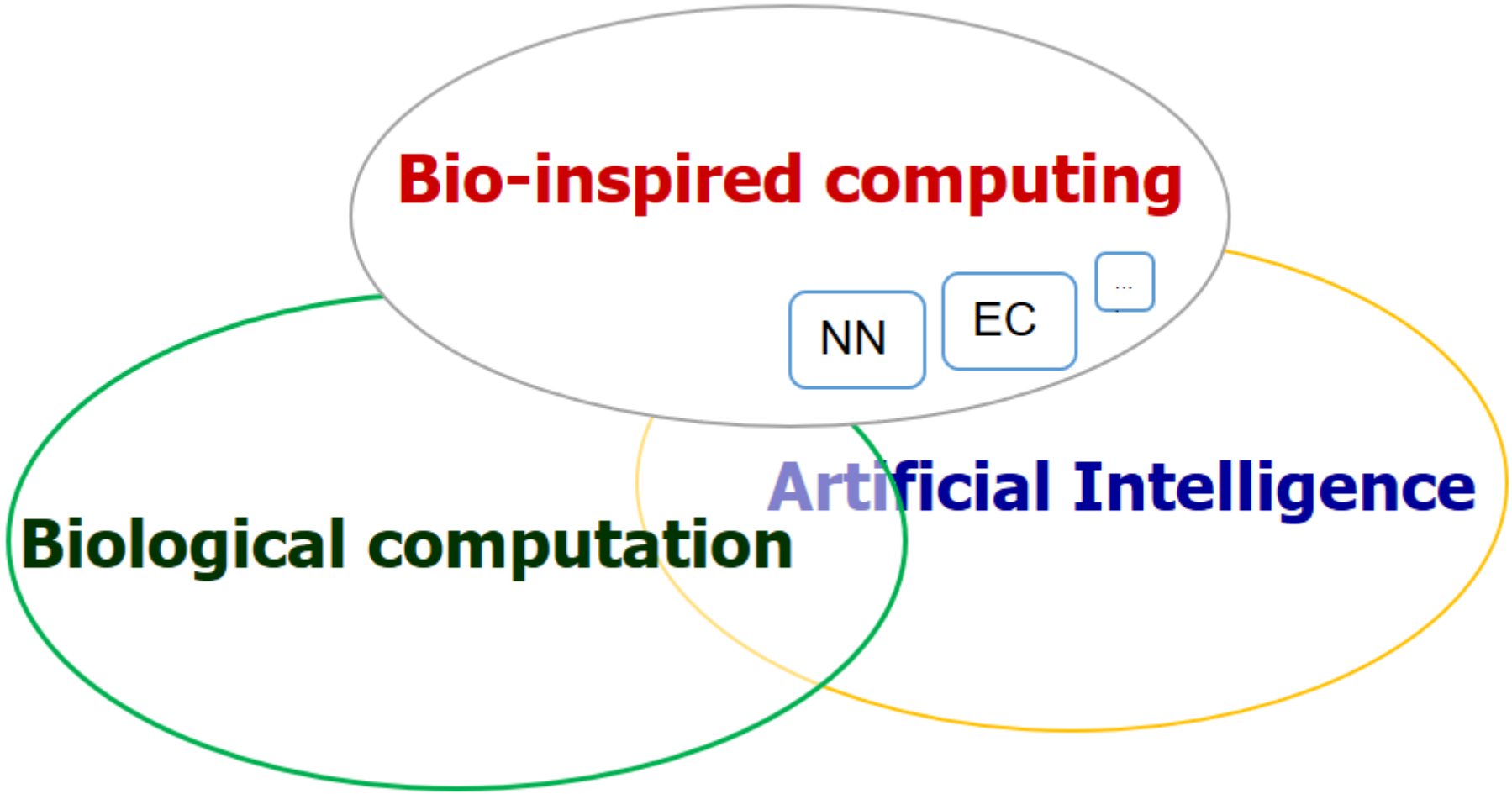
Column Name	Description
Email ID	Unique identifier for each email. Not used in classification but helps track emails.
Subject Length (words)	Number of words in the subject. Spam emails often have short, catchy subject lines.
Contains "Free" (Yes/No)	Indicates if the word "Free" is in the email. Spam frequently uses such words to grab attention.
Number of Links	Counts the number of links in the email. Spam usually contains multiple links to suspicious sites.
Has Attachment (Yes/No)	Indicates if the email has an attachment. Spam often includes attachments that may be malicious.
Class (Spam/Not Spam)	Target label that indicates whether the email is classified as spam or not spam. This is the prediction goal for the model.

Email ID	Subject Length	Contains "Free"	Number of Links	Has Attachment	Class (Spam/Not Spam)
1	5	Yes	3	Yes	Spam
2	12	No	1	No	Not Spam
3	8	Yes	0	No	Spam
4	20	No	2	Yes	Not Spam
5	6	Yes	5	Yes	Spam
…	…	…	…	…	…

Preliminary: Classification

Column Name	Description
Email ID	Unique identifier for each email. Not used in classification but helps track emails.
Subject Length (words)	Number of words in the subject. Spam emails often have short, catchy subject lines.
Contains "Free" (Yes/No)	Indicates if the word "Free" is in the email. Spam frequently uses such words to grab attention.
Number of Links	Counts the number of links in the email. Spam usually contains multiple links to suspicious sites.
Has Attachment (Yes/No)	Indicates if the email has an attachment. Spam often includes attachments that may be malicious.
Class (Spam/Not Spam)	Target label that indicates whether the email is classified as spam or not spam. This is the prediction goal for the model.

Email ID	Subject Length	Contains "Free"	Number of Links	Has Attachment	Class (Spam/Not Spam)
1	5	Yes	3	Yes	Spam
2	12	No	1	No	Not Spam
3	8	Yes	0	No	Spam
4	20	No	2	Yes	Not Spam
5	6	Yes	5	Yes	Spam
…	…	…	…	…	…

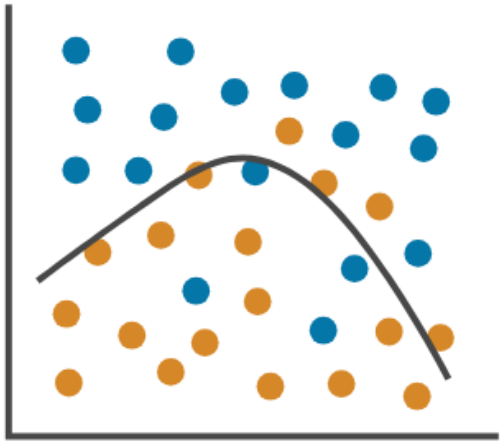

Bio-inspired computing
NN
EC
...
Artificial Intelligence
Biological computation

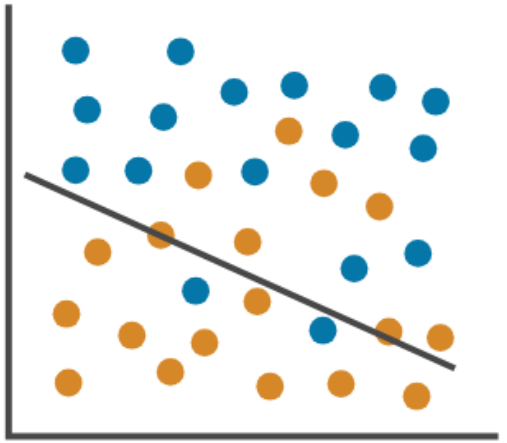
Overfitting

Right Fit

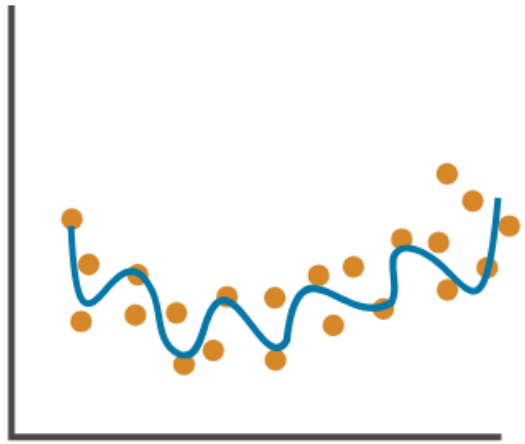
Underfitting

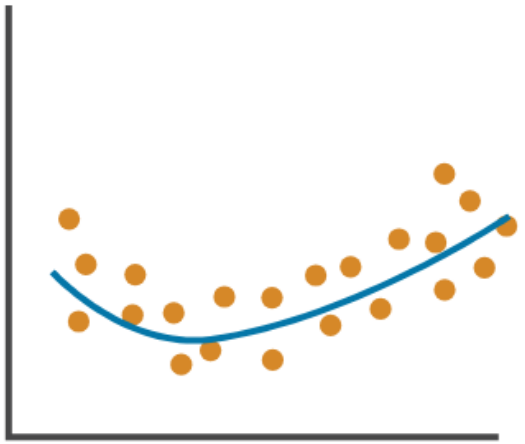
Classification

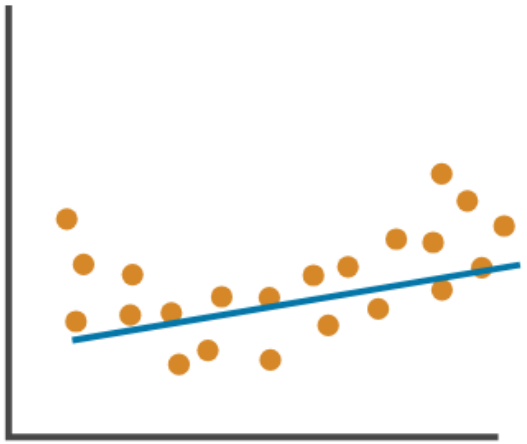


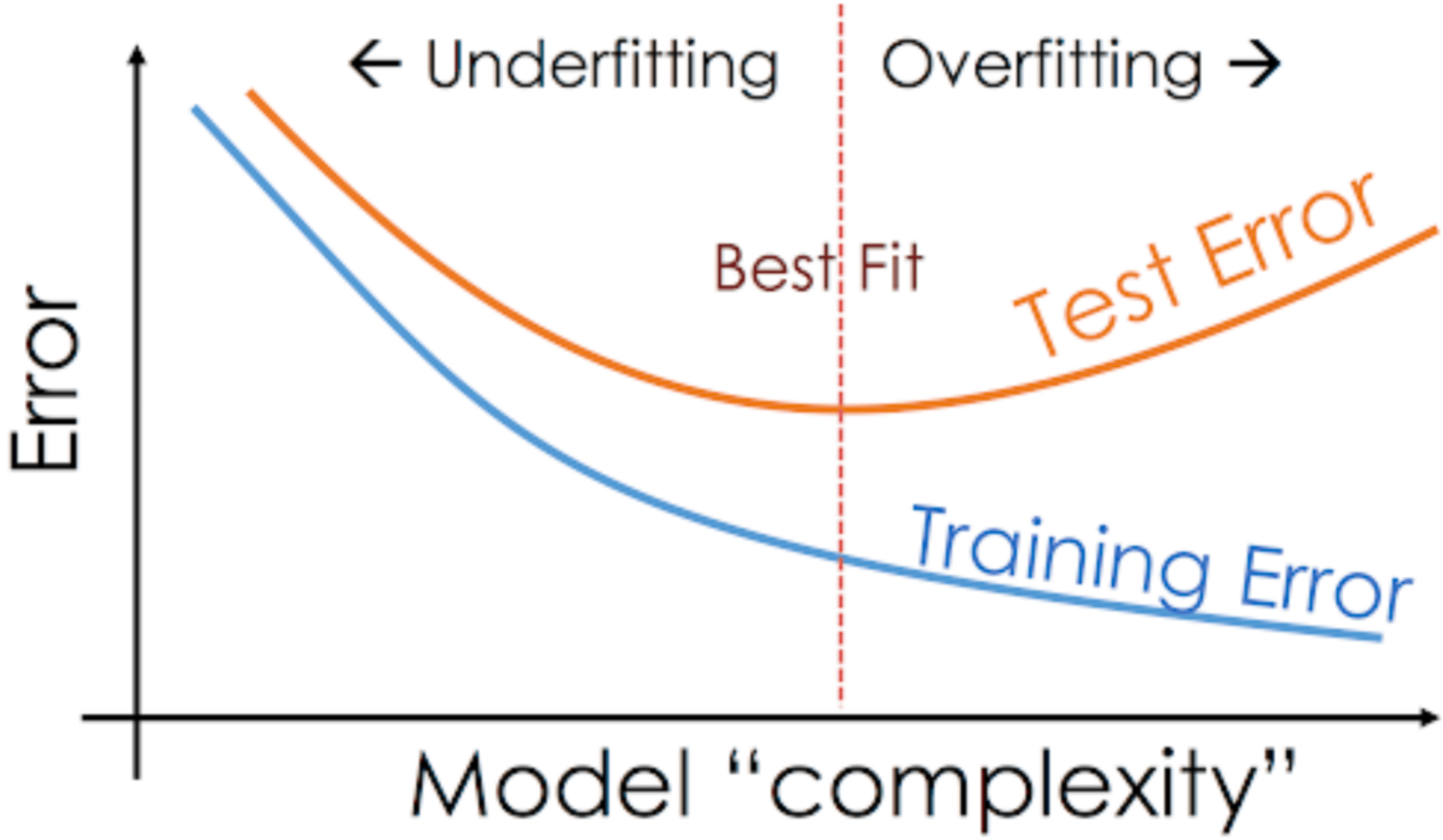


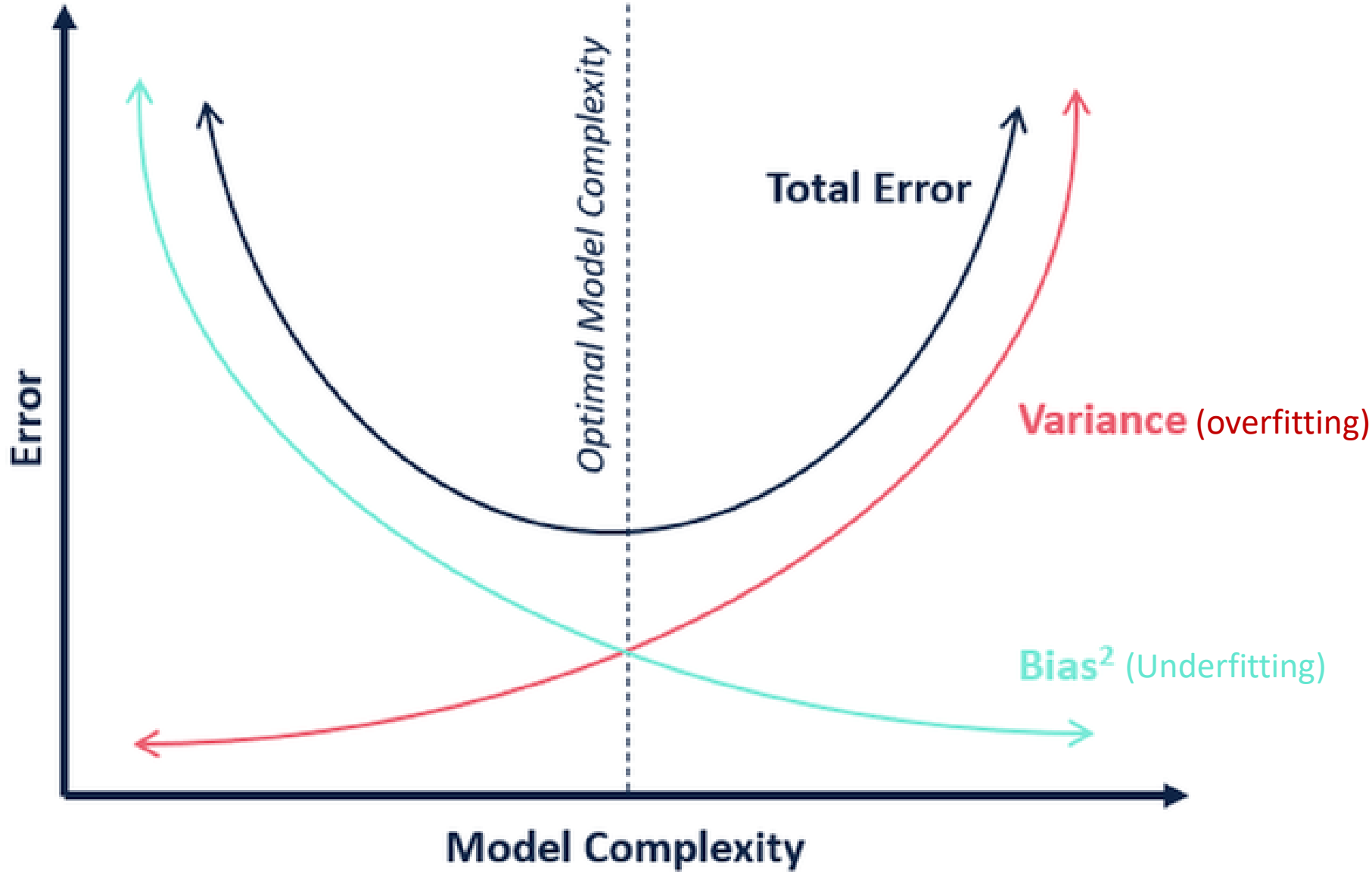
Regression

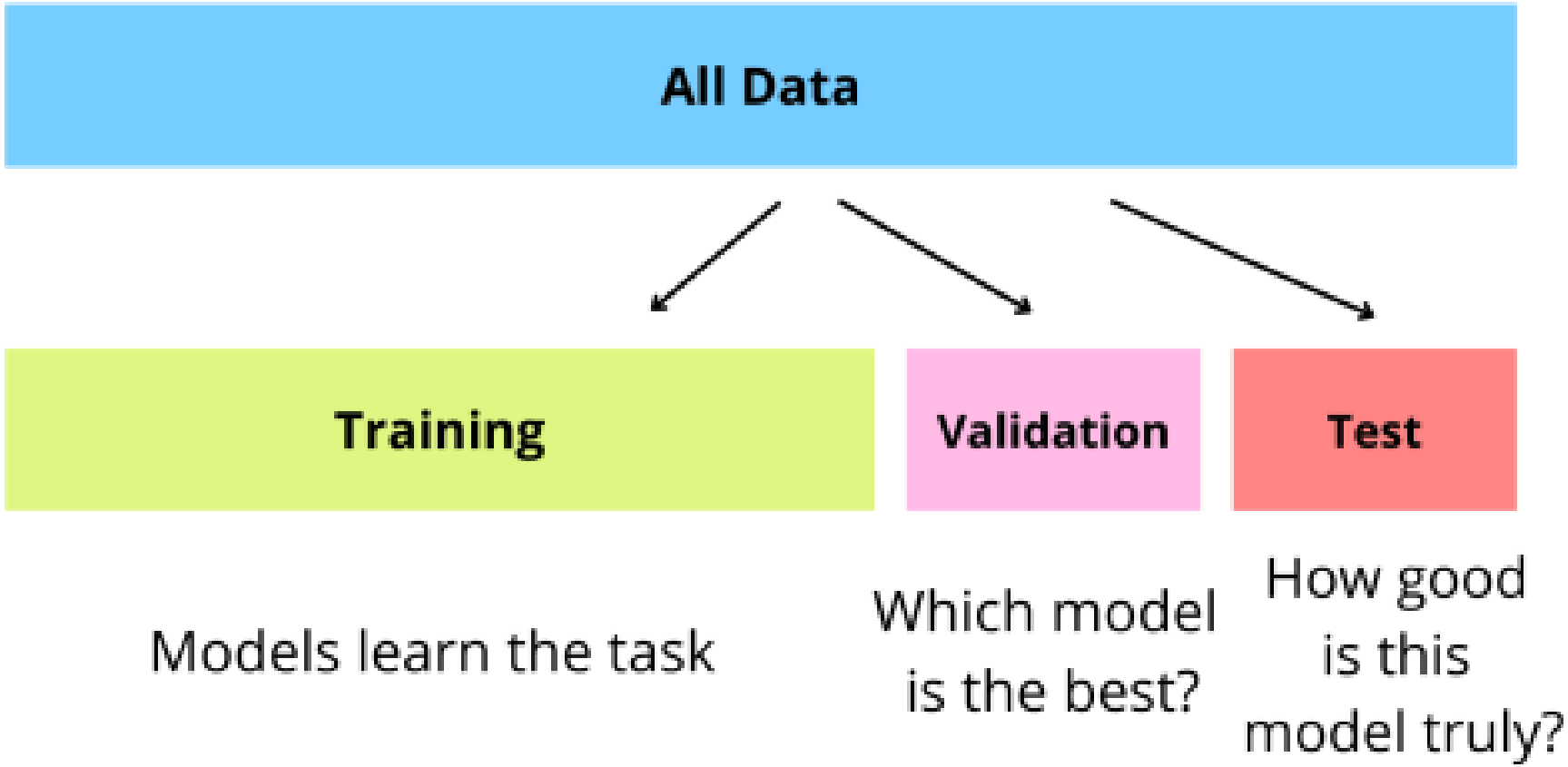


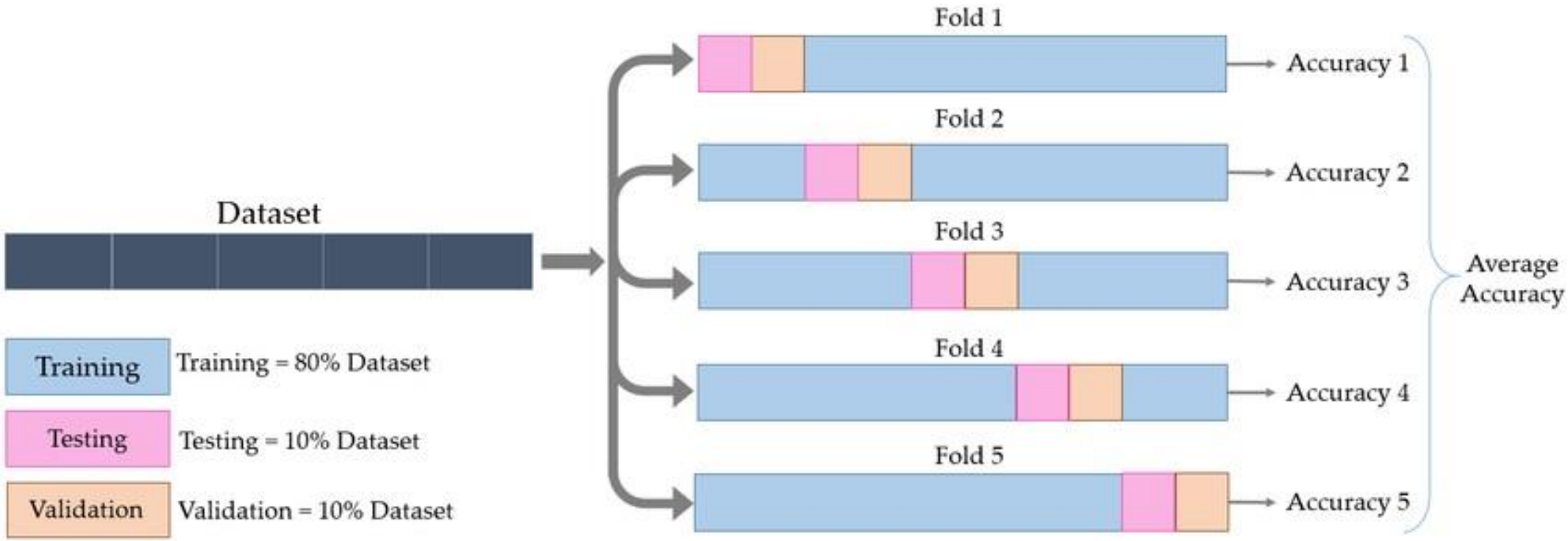




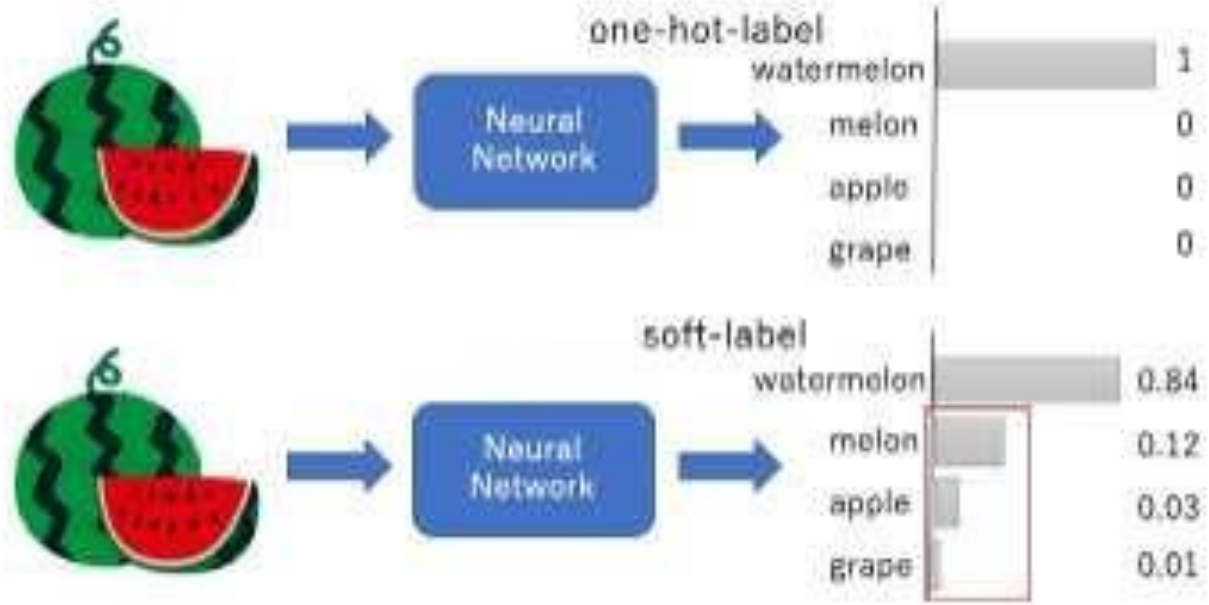

← Underfitting
Overfitting →
Best Fit
Test Error
Training Error
Error
Model "complexity"

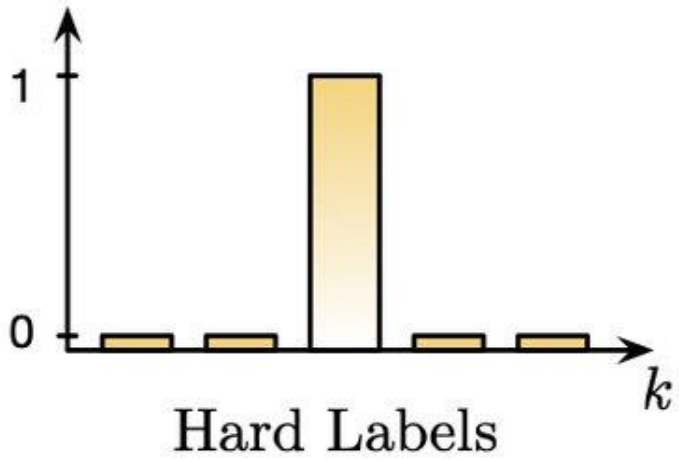

Error
Optimal Model Complexity
Total Error
Variance (overfitting)
Bias^2 (Underfitting)
Model Complexity

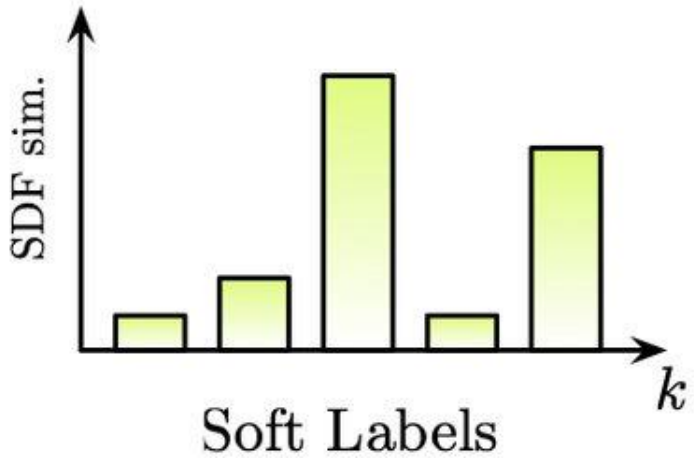

All Data
Training
Validation
Test
Models learn the task
Which model is the best?
How good is this model truly?


Dataset
Training
Training = 80% Dataset
Testing
Testing = 10% Dataset
Validation
Validation = 10% Dataset
Fold 1
Accuracy 1
Fold 2
Accuracy 2
Fold 3
Accuracy 3
Fold 4
Accuracy 4
Fold 5
Accuracy 5
Average Accuracy

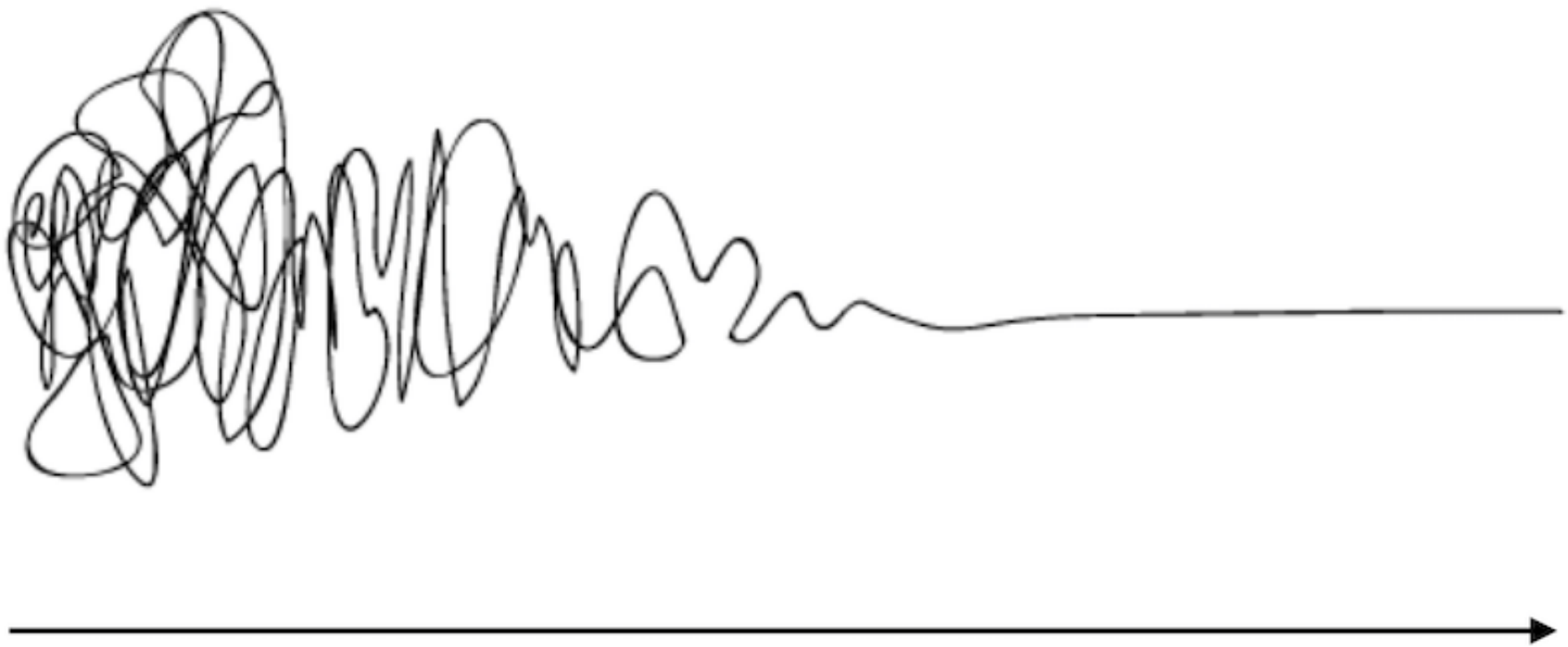
Soft label vs hard label







Explore **Exploit**



Higher uncertainty Lower uncertainty

Explore

Exploit