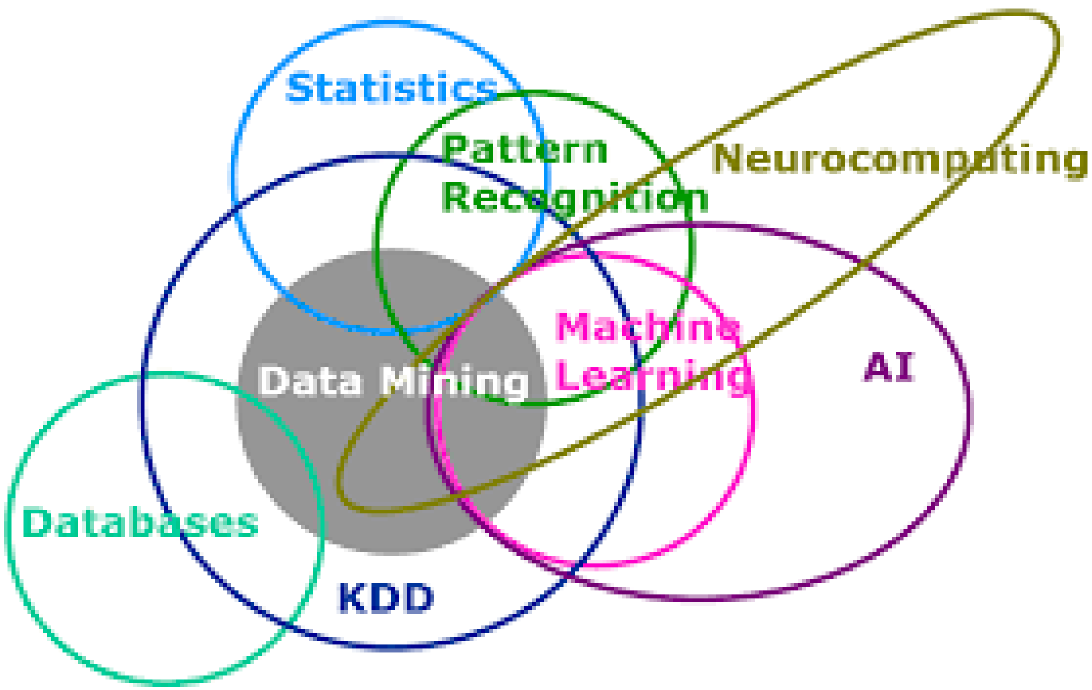
In the name of Allah the most Beneficial ever merciful

بِسۡمِ ٱللَّهِ ٱلرَّحۡمَٰنِ ٱلرَّحِيمِ



Artificial Intelligence (AI) in Software Engineering

ANOVA Table

Copyright © 2020, Dr. Humera Tariq

Department of Computer Science , Univeristy of Karachi (DCS-UBIT)
25th May 2021

Regression Statistics

Week 09 Memory Recall

Problem Statement

A company sets different LOC rates for a particular project in its eight different modules. The accompanying table shows the numbers of LOC and the corresponding rates.

LOC	420	380	350	400	440	380	450	420
Rates (100USD)	5.5	6.0	6.5	6.0	5.0	6.5	4.5	5.0

UNIVERSITY OF KARACHI

Regression Statistics

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.937137027
R Square	0.878225806
Adjusted R Square	0.857930108
Standard Error	12.74227575
Observations	8

Week 10 Agenda: ANOVA

Analysis of Variance Table

ANOVA TABLE Format

The ANOVA (analysis of variance) table splits the sum of squares into its components.

ANOVA

	df	SS	MS	F	Significance F
Regression	1				
Residual	6				
Total	7				

UNIVERSITY OF KARACHI

Total sums of squares

Total sums of squares =

Residual (or error) sum of squares +

Regression (or explained) sum of squares

Thus $\Sigma_i (y_i - ybar)^2 = \Sigma_i (y_i - yhat_i)^2 + \Sigma_i (yhat_i - ybar)^2$

UNIVERSITY OF KARACHI

Today's Agenda

- ✓ Feature…Variable….Factor….component, Vector
- ✓ Why Select, Extract or Rank Feature ??
- ✓ Curse of Dimensionality
- ✓ Weekly Assignment Discussion
- ✓ Strategies for Feature Selection
- ✓ Identify ANOVA as strategy of Feature Selection

UNIVERSITY OF KARACHI

Motivation

We need ANOVA Test in Artificial Intelligence for Feature Selection

Feature…Variable…Factor….Component

Alternate words for feature

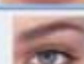
Row/column Vector…..Matrices

"Feature" = a component of data

I	2
you	0
upset	0
unhappy	1
puppy	1
bear	0

⋮

What's is important feature?

	3.29
	-15
	48.3
	25.1
	3.82

⋮

What's is important feature?

Software Defect Prediction Data Analysis | Kaggle

about JM1 Dataset.txt

```
% 7. Attribute Information:
%
%      1. loc             : numeric % McCabe's line count of code
%      2. v(g)            : numeric % McCabe "cyclomatic complexity"
%      3. ev(g)           : numeric % McCabe "essential complexity"
%      4. iv(g)           : numeric % McCabe "design complexity"
%      5. n               : numeric % Halstead total operators + operands
%      6. v               : numeric % Halstead "volume"
%      7. l               : numeric % Halstead "program length"
%      8. d               : numeric % Halstead "difficulty"
%      9. i               : numeric % Halstead "intelligence"
%     10. e               : numeric % Halstead "effort"
%     11. b               : numeric % Halstead
%     12. t               : numeric % Halstead's time estimator
%     13. lOCode          : numeric % Halstead's line count
%     14. lOComment       : numeric % Halstead's count of lines of comments
%     15. lOBlank         : numeric % Halstead's count of blank lines
%     16. lOCodeAndComment: numeric
%     17. uniq_Op         : numeric % unique operators
%     18. uniq_Opnd       : numeric % unique operands
%     19. total_Op        : numeric % total operators
%     20. total_Opnd      : numeric % total operands
%     21: branchCount     : numeric % of the flow graph
%     22. defects         : {false,true} % module has/has not one or more
%                                        % reported defects

% 8. Missing attributes: none

% 9. Class Distribution: the class value (defects) is discrete
%    false: 2106 = 19.35%
%    true:  8779 = 80.65%
```

UNIVERSITY OF KARACHI

about JM1 Dataset.txt

loc	v(g)	ev(g)	iv(g)	n	v	l	d	i	e	b	t	lOCode	lOComme	lOBlank	locCodeA	uniq_Op	uniq_Opn	total_Op	total_Opn	branchCo	defects
1.1	1.4	1.4	1.4	1.3	1.3	1.3	1.3	1.3	1.3	1.3	1.3	2	2	2	2	1.2	1.2	1.2	1.2	1.4	FALSE
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	TRUE
72	7	1	6	198	1134.13	0.05	20.31	55.85	23029.1	0.38	1279.39	51	10	8	1	17	36	112	86	13	TRUE
190	3	1	3	600	4348.76	0.06	17.06	254.87	74202.67	1.45	4122.37	129	29	28	2	17	135	329	271	5	TRUE
37	4	1	4	126	599.12	0.06	17.19	34.86	10297.3	0.2	572.07	28	1	6	0	11	16	76	50	7	TRUE
31	2	1	2	111	582.52	0.08	12.25	47.55	7135.87	0.19	396.44	19	0	5	0	14	24	69	42	3	TRUE
78	9	5	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	17	TRUE
8	1	1	1	16	50.72	0.36	2.8	18.11	142.01	0.02	7.89	5	0	1	0	4	5	9	7	1	TRUE
24	2	1	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	TRUE
143	22	20	10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	43	TRUE
73	10	4	6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	19	TRUE
83	11	10	7	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	21	TRUE
12	3	1	1	37	167.37	0.15	6.87	24.34	1150.68	0.06	63.93	8	0	2	0	11	12	22	15	5	TRUE
48	4	1	4	129	695.61	0.06	17.35	40.1	12067.3	0.23	670.41	29	1	16	0	19	23	87	42	7	TRUE
68	8	1	5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	15	TRUE
138	22	10	8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	43	TRUE
10	1	1	1	9	27	0.5	2	13.5	54	0.01	3	2	0	6	0	4	4	5	4	1	TRUE
250	49	34	16	1469	9673.31	0.01	97	99.72	938311.1	3.22	52128.39	139	92	17	0	32	64	1081	388	97	TRUE
77	8	1	1	284	1160.84	0.02	40.95	28.35	47536.38	0.39	2640.91	59	0	16	0	7	10	167	117	15	TRUE
85	9	1	7	277	1714.58	0.03	32.64	52.53	55961.02	0.57	3108.95	69	0	14	0	26	47	161	118	13	TRUE
110	17	13	8	322	2069.26	0.03	33.41	61.94	69127.22	0.69	3840.4	81	13	14	0	27	59	176	146	33	TRUE
49	6	6	3	171	927.89	0.04	25.33	36.63	23506.58	0.31	1305.92	34	0	13	0	19	24	107	64	11	TRUE
187	35	26	16	526	3296.33	0.02	42.56	77.45	140300	1.1	7794.45	164	1	16	0	21	56	299	227	69	TRUE
27	6	6	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	11	TRUE
38	8	1	3	145	673.36	0.05	20.53	32.8	13824.9	0.22	768.05	29	0	7	0	9	16	72	73	15	TRUE
294	43	33	24	814	5811.59	0.02	40.88	142.15	237606.8	1.94	13200.38	223	41	26	2	28	113	484	330	85	TRUE
29	3	1	3	88	465.12	0.08	12.04	38.63	5599.99	0.16	311.11	21	0	6	0	14	25	45	43	5	TRUE
160	5	4	3	698	4862.12	0.03	33.11	146.86	160969.1	1.62	8942.73	123	11	23	1	22	103	388	310	9	TRUE
94	16	9	5	218	1236.59	0.03	34.52	35.83	42683.63	0.41	2371.31	66	19	6	1	22	29	127	91	31	TRUE
48	3	1	3	157	927.38	0.08	13.09	70.84	12140.27	0.31	674.46	34	1	9	0	16	44	85	72	5	TRUE
14	2	1	2	31	129.27	0.19	5.2	24.86	672.19	0.04	37.34	8	1	3	0	8	10	18	13	3	TRUE
32	6	4	4	116	595	0.06	16.67	35.7	9916.61	0.2	550.92	26	0	4	0	14	21	66	50	11	TRUE
11	1	1	1	9	27	0.5	2	13.5	54	0.01	3	2	0	6	0	4	4	5	4	1	TRUE

UNIVERSITY OF KARACHI

JM1 Data Matrix

10885 Rows x 22 Columns

Dimension d = 22

UNIVERSITY OF KARACHI

ANOVA

Why Select Features ??

Why Extract Features??

Why Rank Features ??

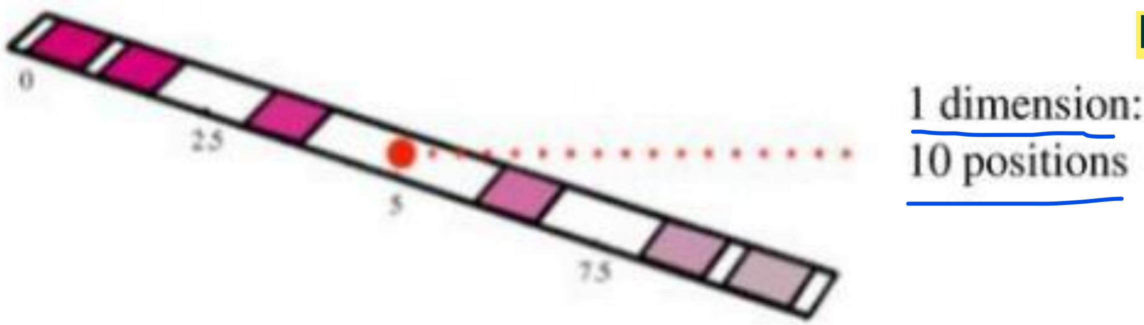
1D-10 Positions, Univariate

With 1D feature space there are only 10 possible positions. Therefore 10 data elements are required to create a representative samples which covers the problem space.

Normalization

Division

Distance

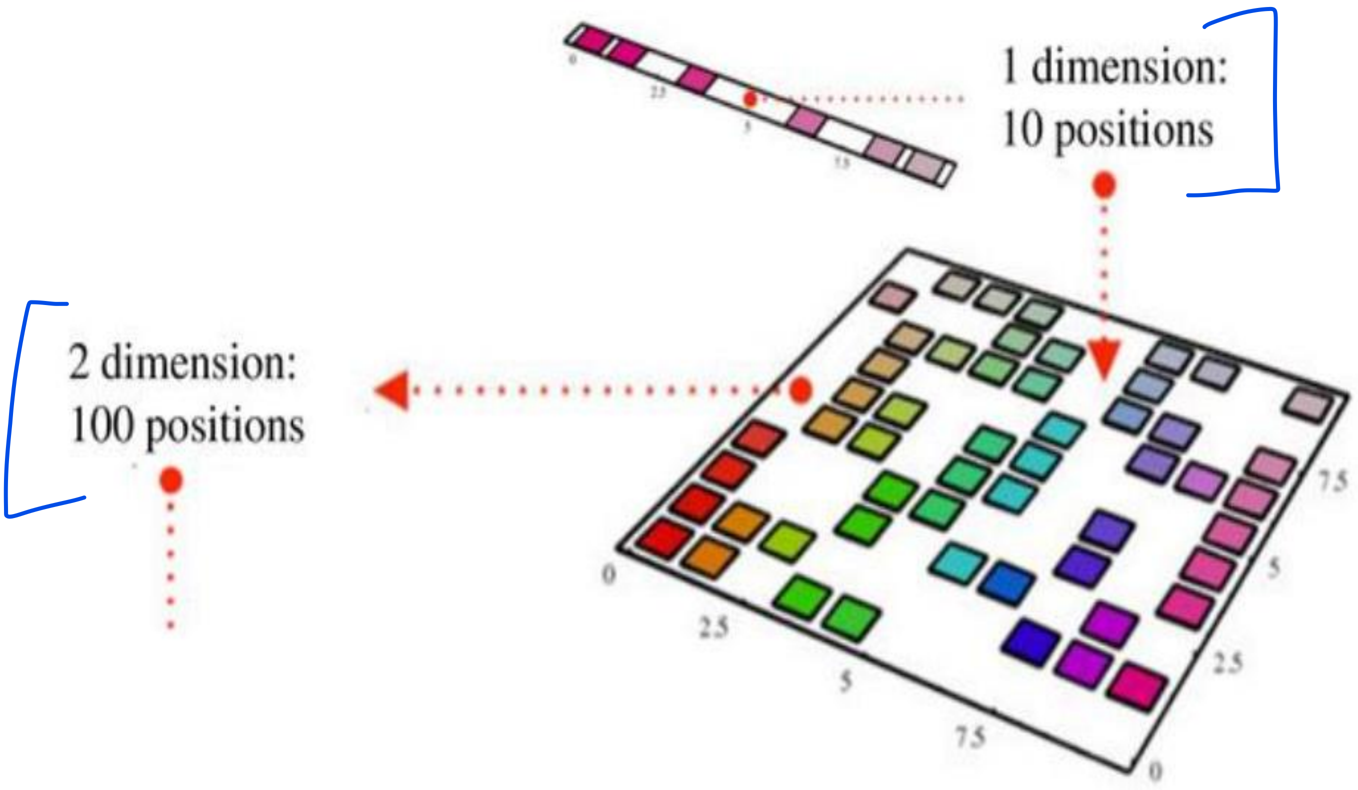


Data/Feature Scale : 0………1 (Normalization)

Data/Feature Scale : 1………10 (un-Normalization)

2D-100 Positions, Multivariate

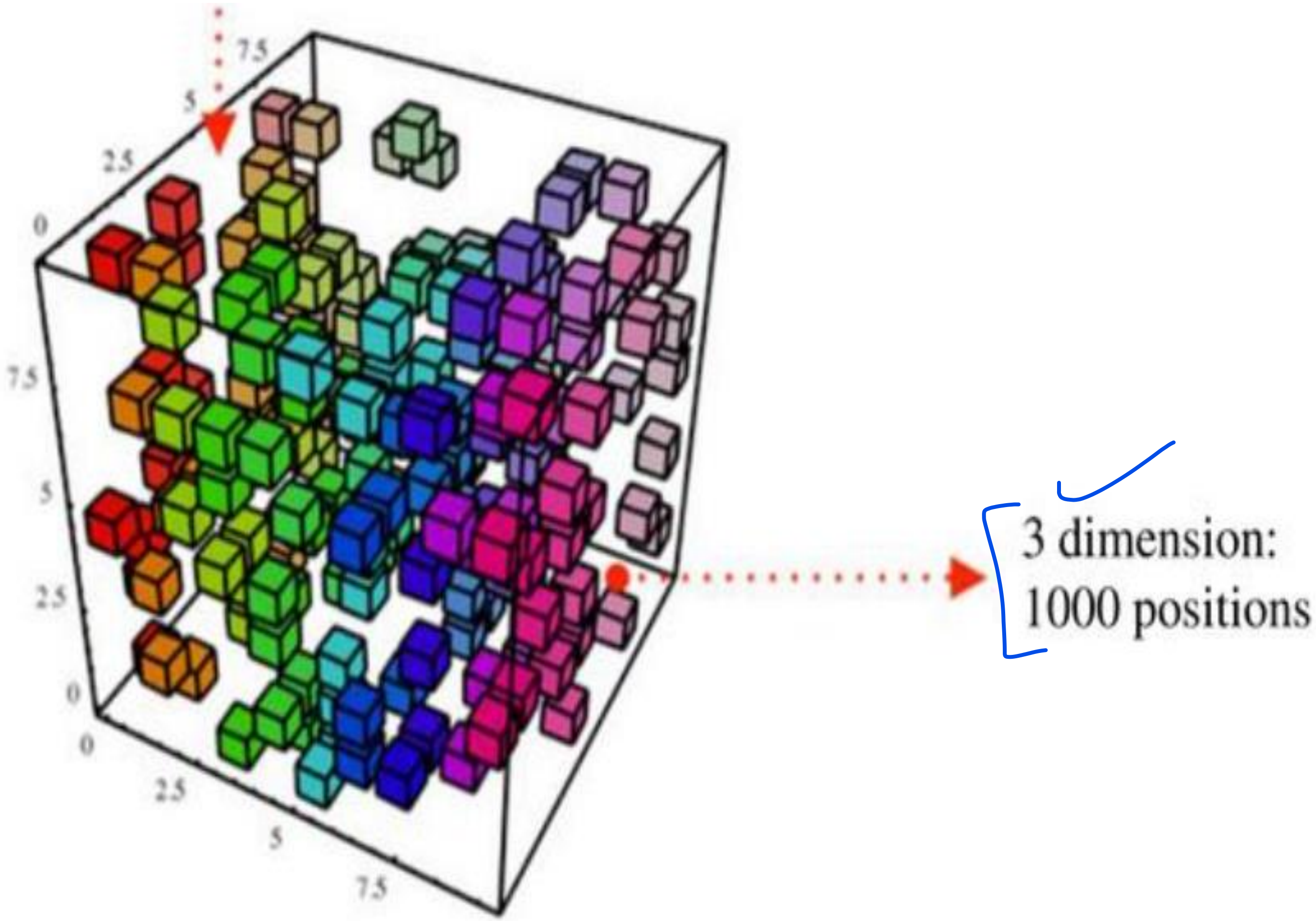
With 2D feature space there are 10^2 = 100 possible positions. Therefore 100 data elements are required to create a representative samples which covers the problem space.



UNIVERSITY OF KARACHI

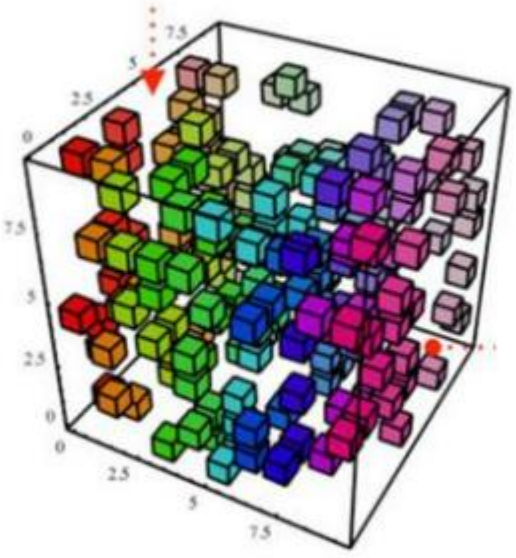
3D-1000 Positions, Multivariate

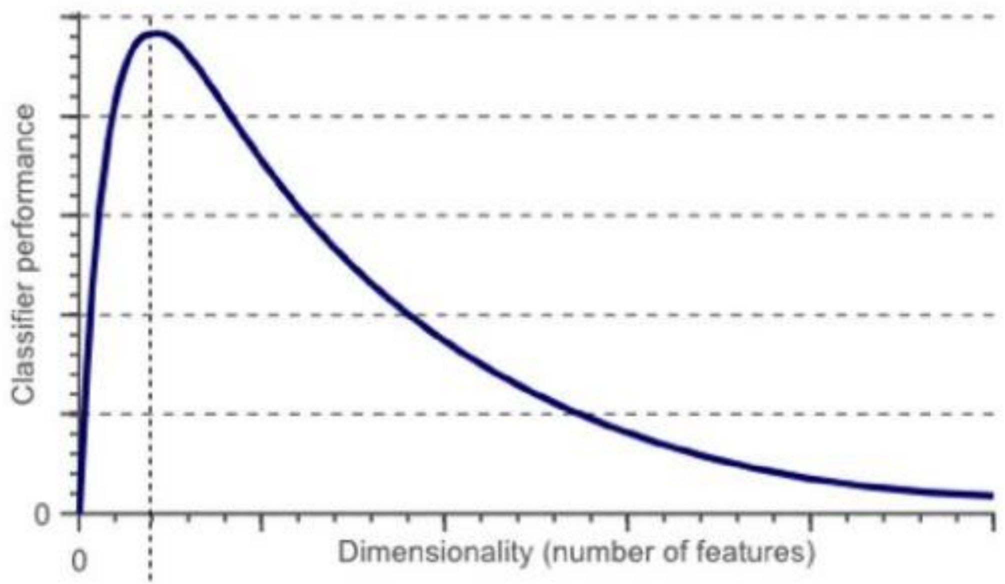
With 3D feature space there are 10^3 = 1000 possible positions. Therefore 1000 data elements are required to create a representative samples which covers the problem space.



The Curse of Dimensionality

The exponential growth in the required number of data continues to grow indefinitely.

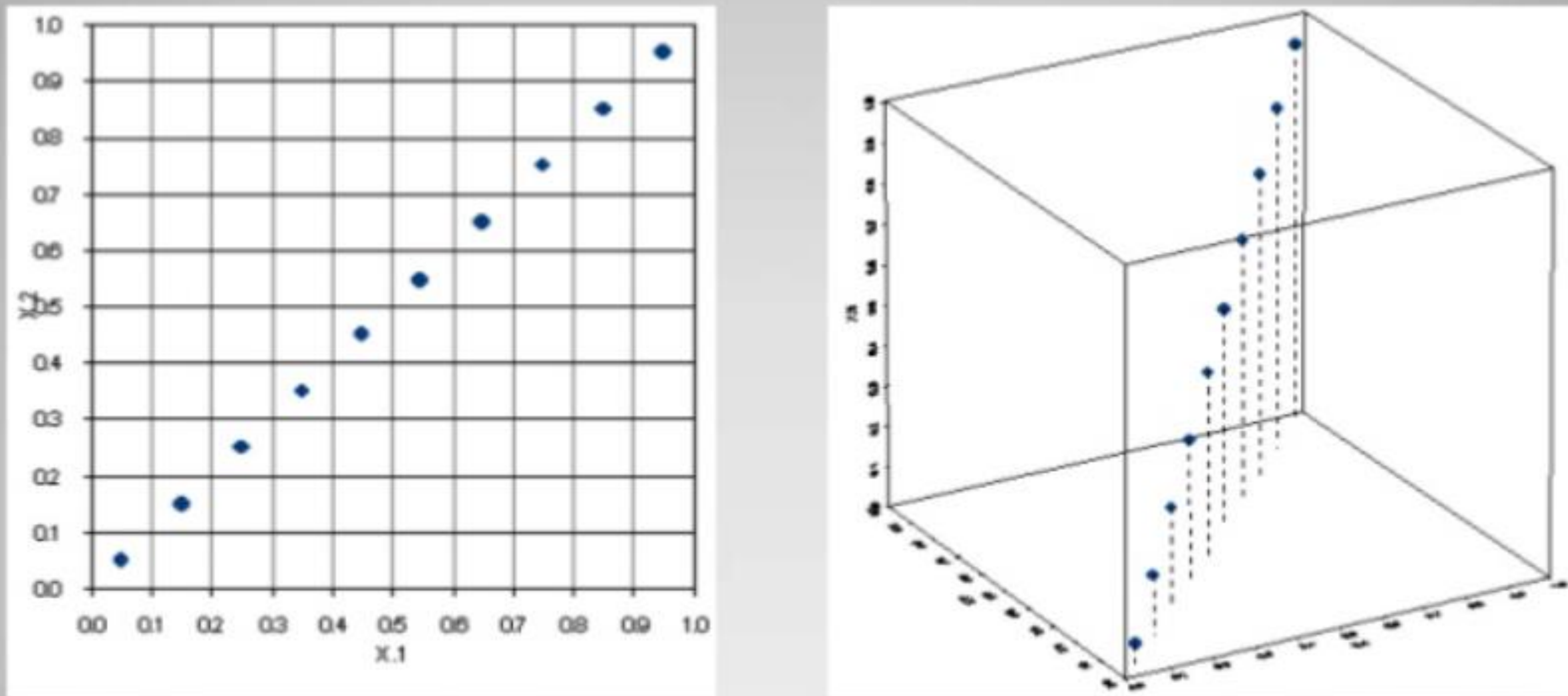




The Curse of Dimensionality

Representation of 10% sample probability space

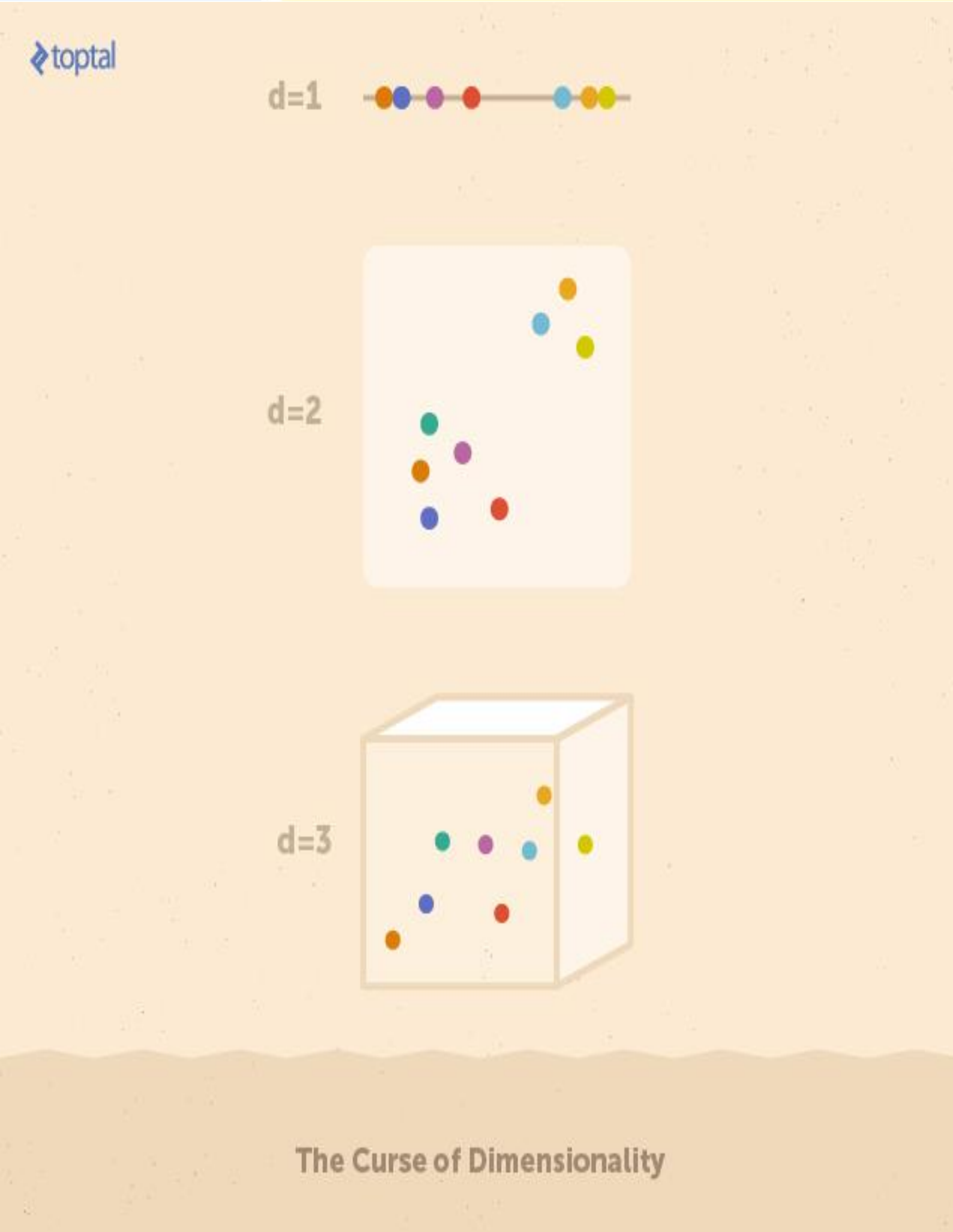
(i) 2-D (ii)3-D

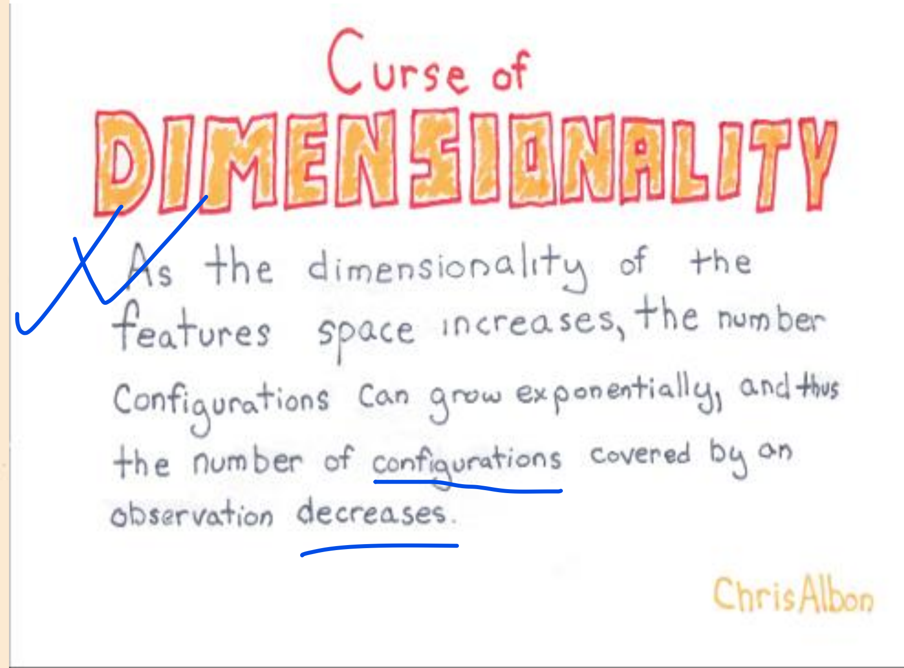


The Number of Points Would Need to Increase Exponentially to Maintain a Given Accuracy.

10^n samples would be required for a *n*-dimension problem.

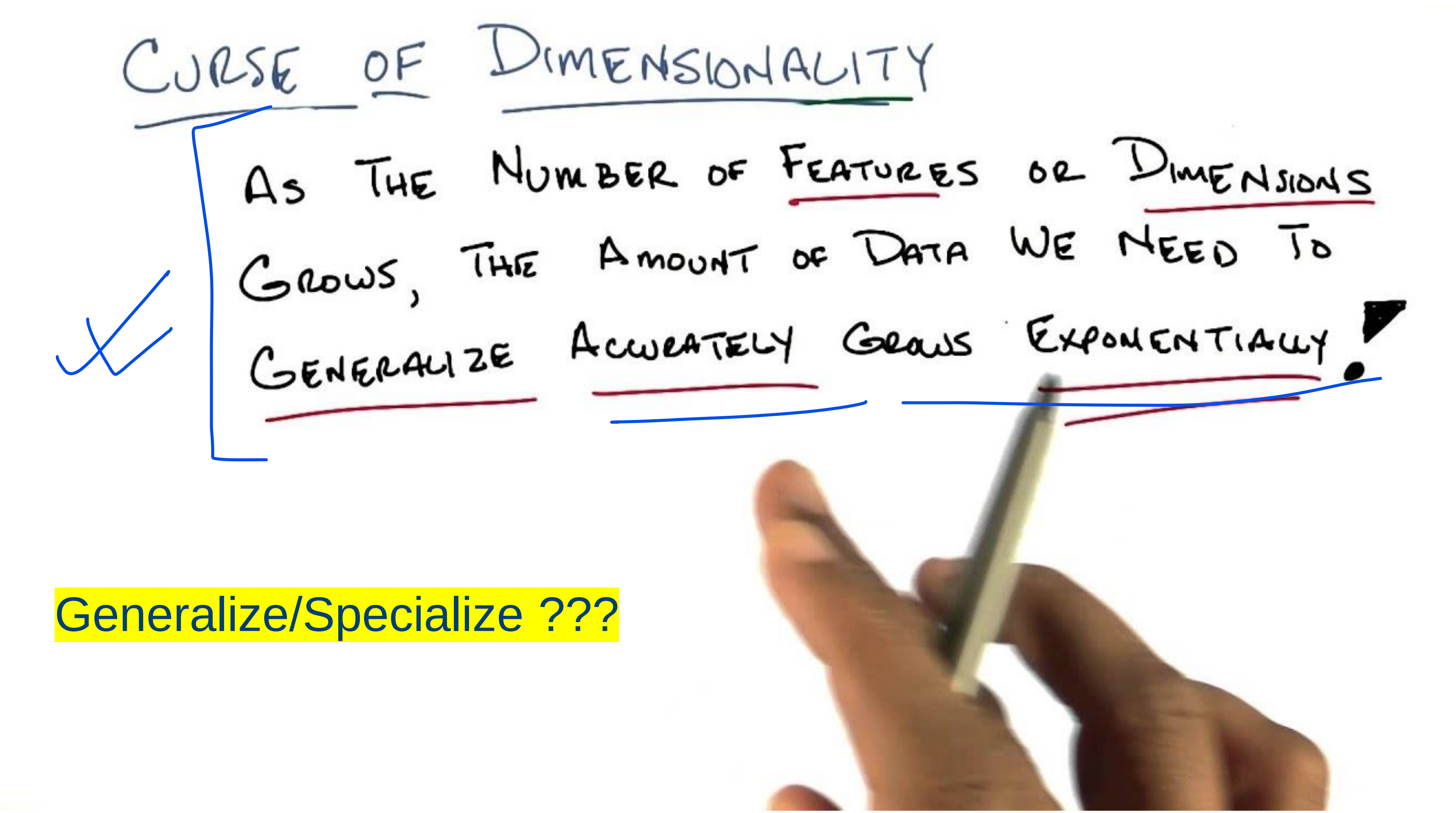
The Curse of Dimensionality





UNIVERSITY OF KARACHI

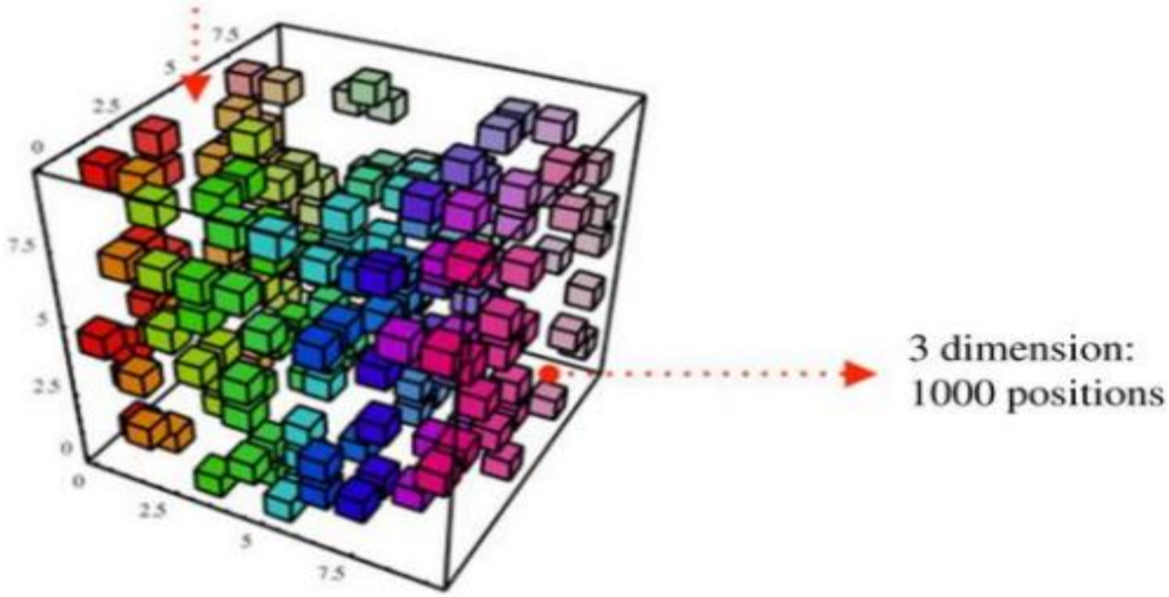
How should Model Behave??



Generalize/Specialize ???

UNIVERSITY OF KARACHI

The Curse of Dimensionality



This mean higher the dimension,(less/more) space the data occupies in the whole space.

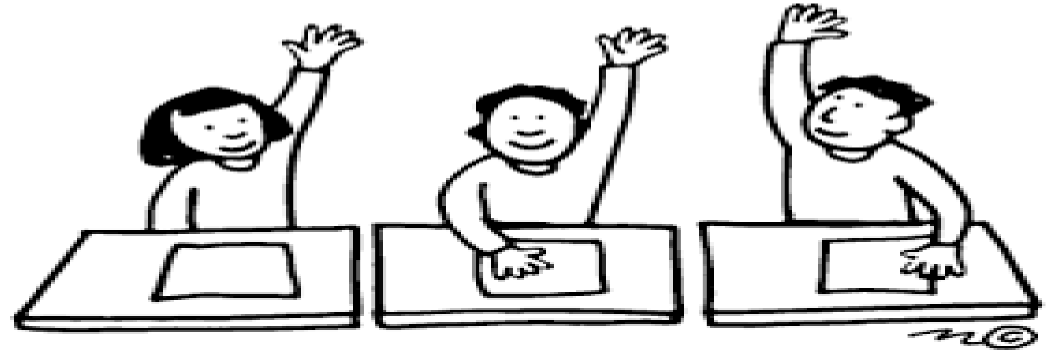
Sparse Data

Row	Feature with Sparse Data	Feature with Missing Data
1	0	null
2	1	4
3	0	3
4	0	null

As the data becomes sparse, the new data is likely to be (further/closer) from train data, requiring much more work to be done for prediction.

UNIVERSITY OF KARACHI

Class Participation and Homework
25-31 May 2021



I am giving you 5 pictures based on artificial intelligence concepts.

Take Print of each picture, Explore and write at-least 15 technical/AI relevant points that shows your understanding.

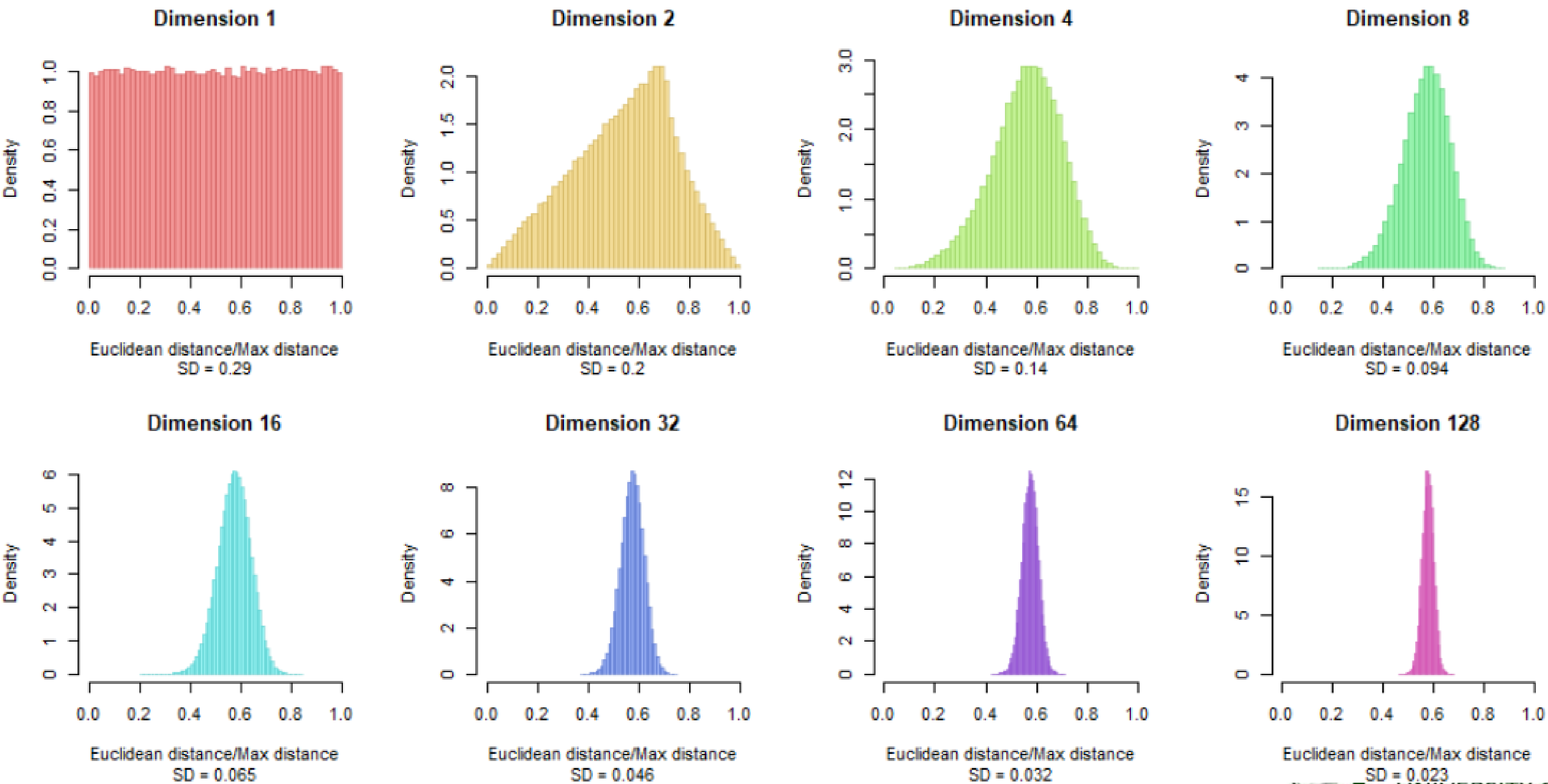
Format: Mainly Handwritten

Bonus:
- Attach/support Lab work
- Relevant Toy Example
- Relevant Mathematical Formulas
- Relevant Table

12

Density plots and Dimension curse

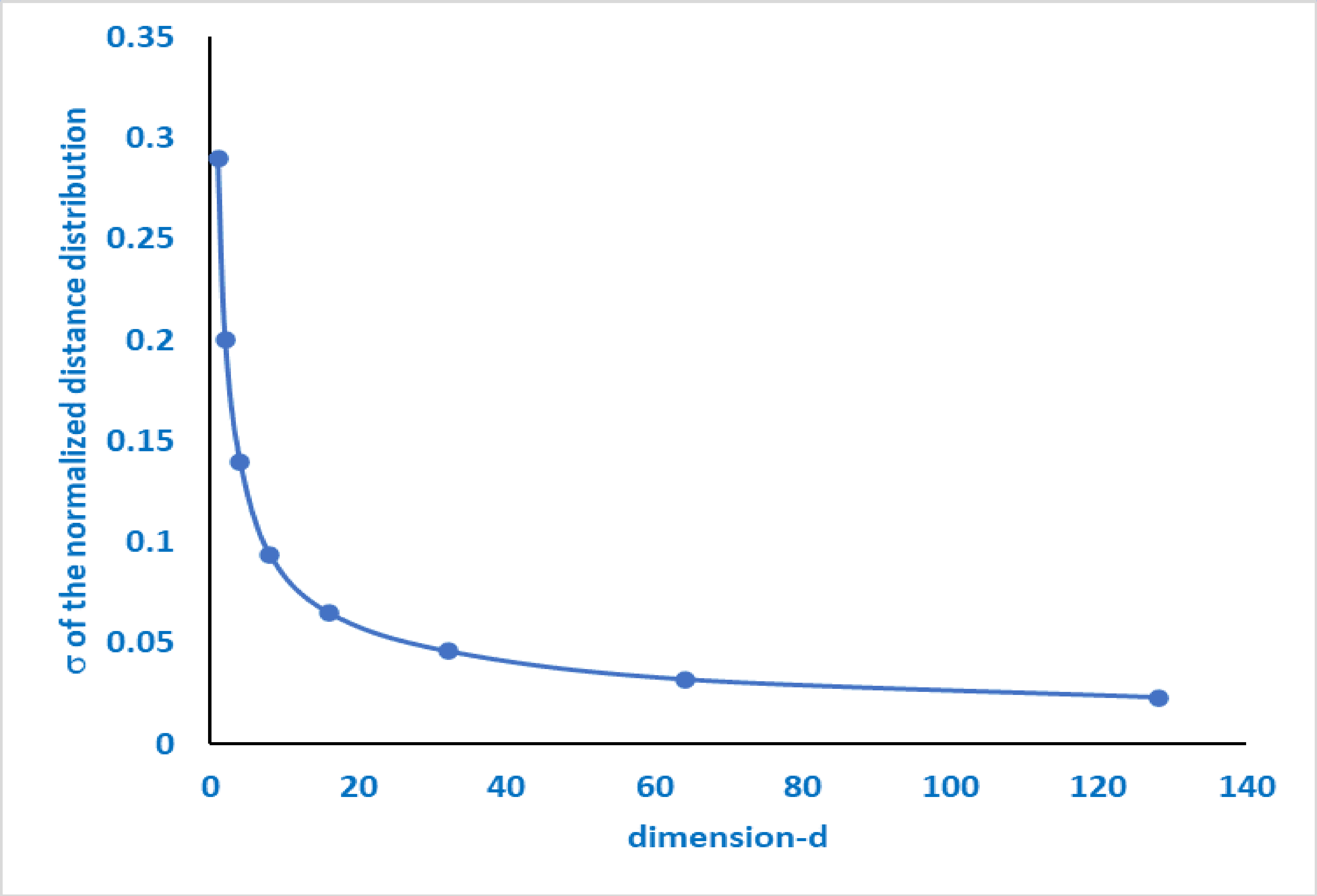
As the number of dimensions increases, we see that the spread of the frequency plot decreases indicating that distances between different samples or points tend towards a single value as the dimension increases.



What is Curse of Dimensionality in Machine Learning? (mygreatlearning.com)

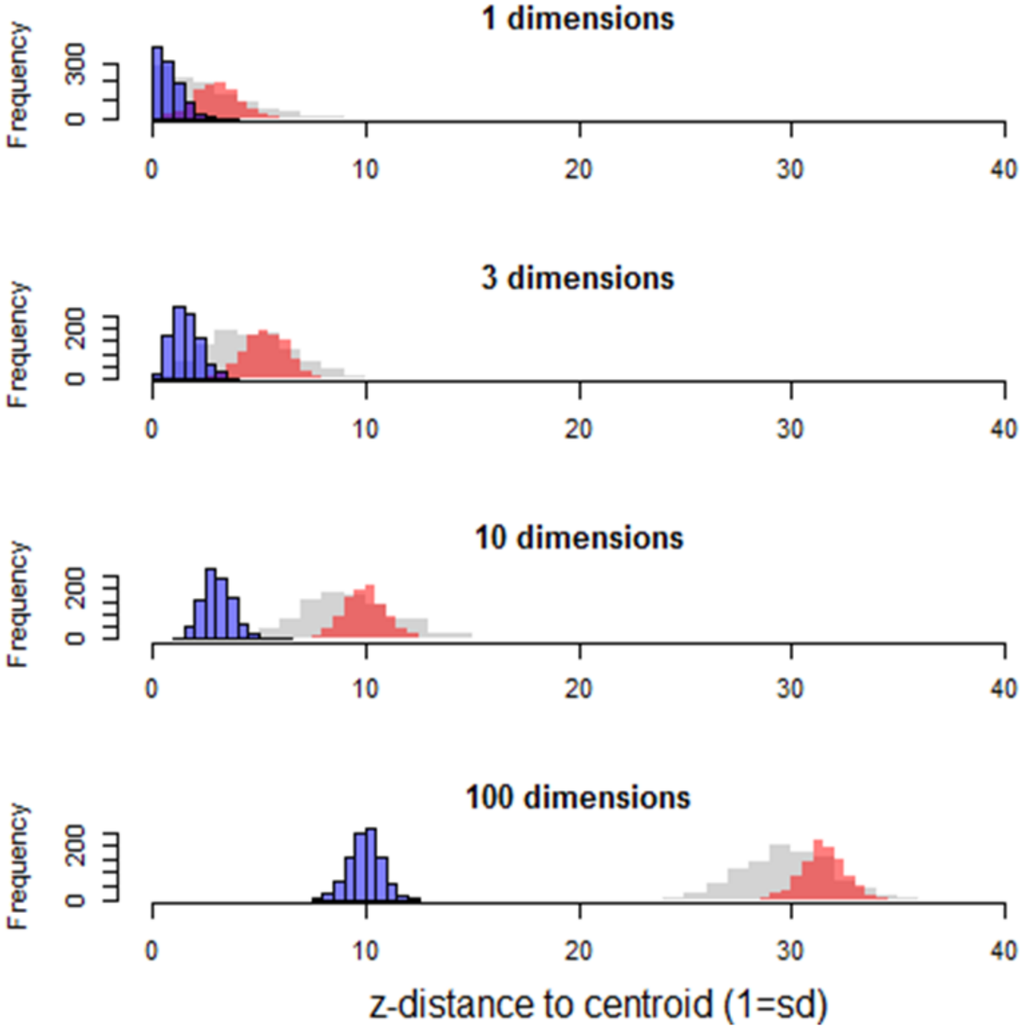
UNIVERSITY OF KARACHI

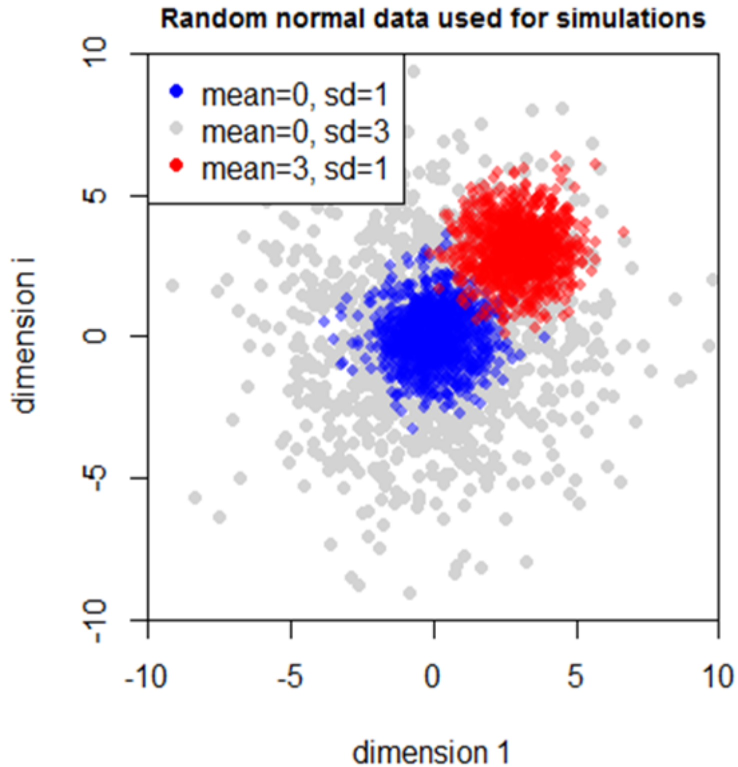
Standard Deviation and Dimension Curse



UNIVERSITY OF KARACHI

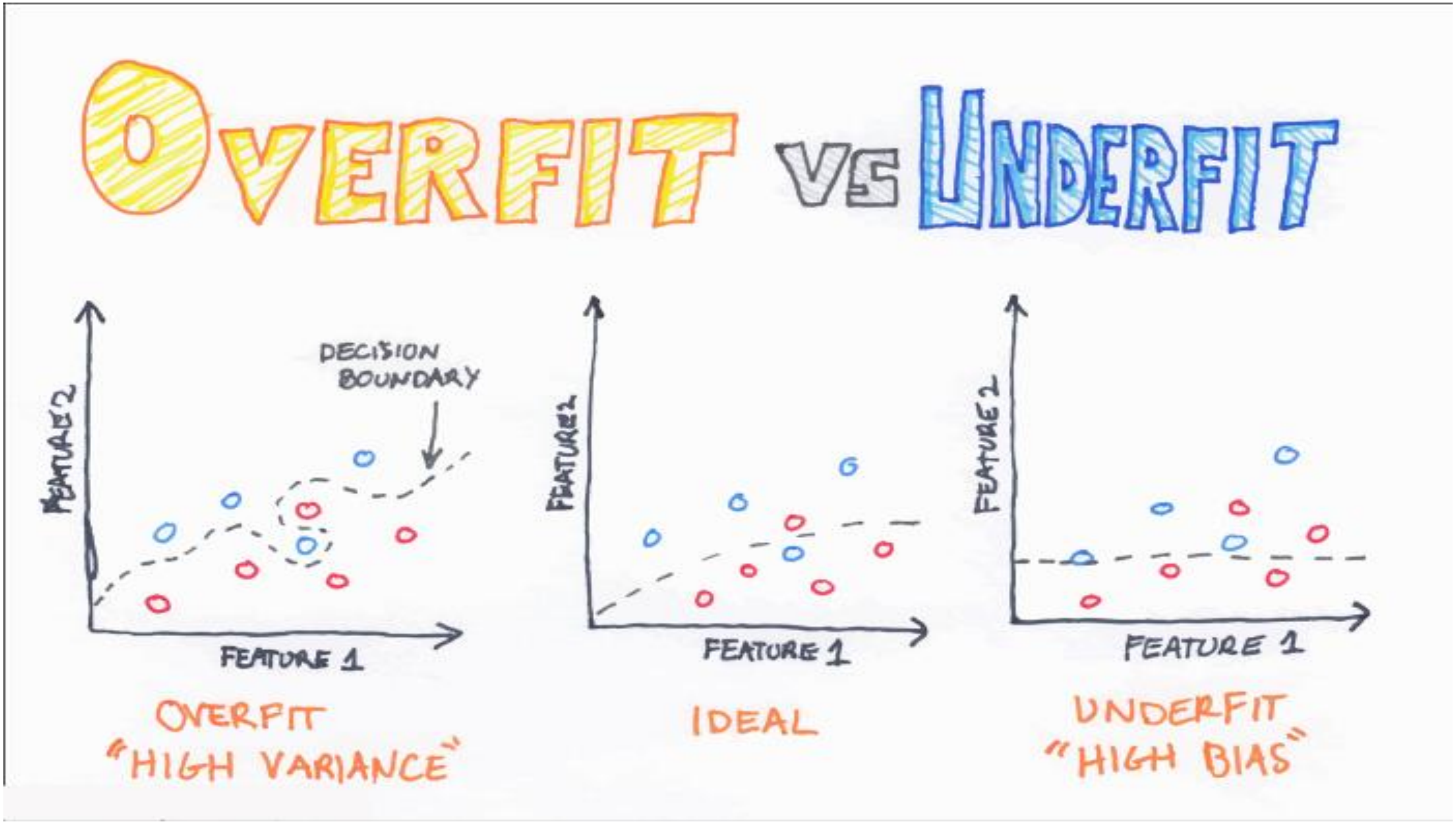
Higher dimension is Good/bad for Prediction ??



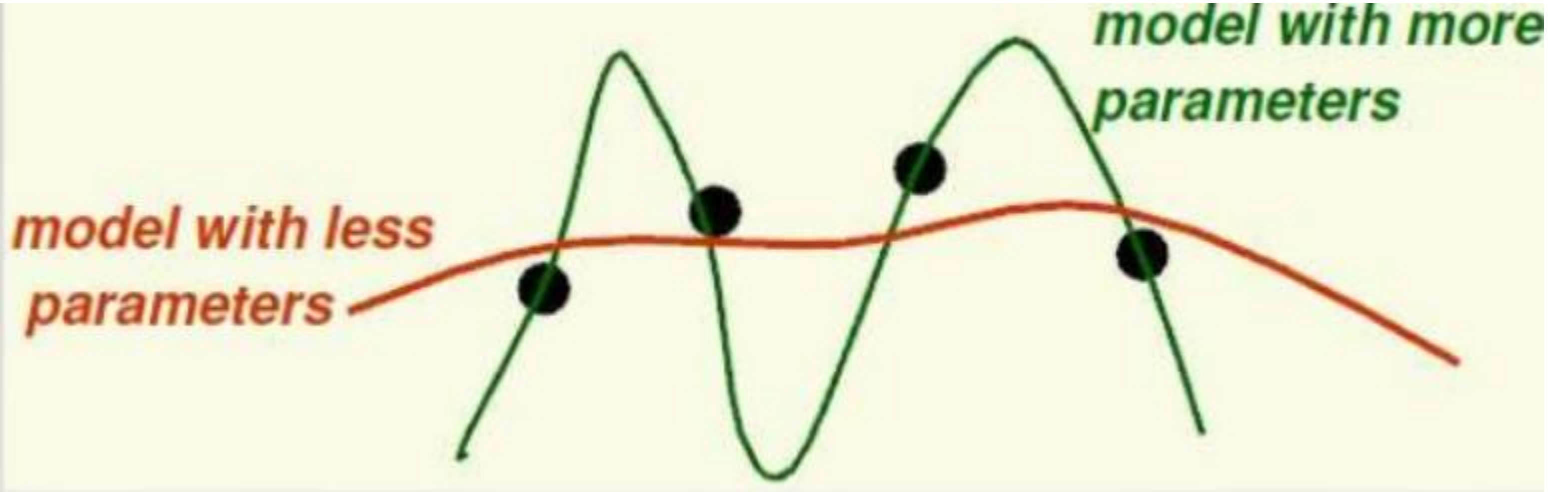


UNIVERSITY OF KARACHI

What : Classification Or Regression ??



Model : Linear vs. Non-Linear



We are still on chasing :

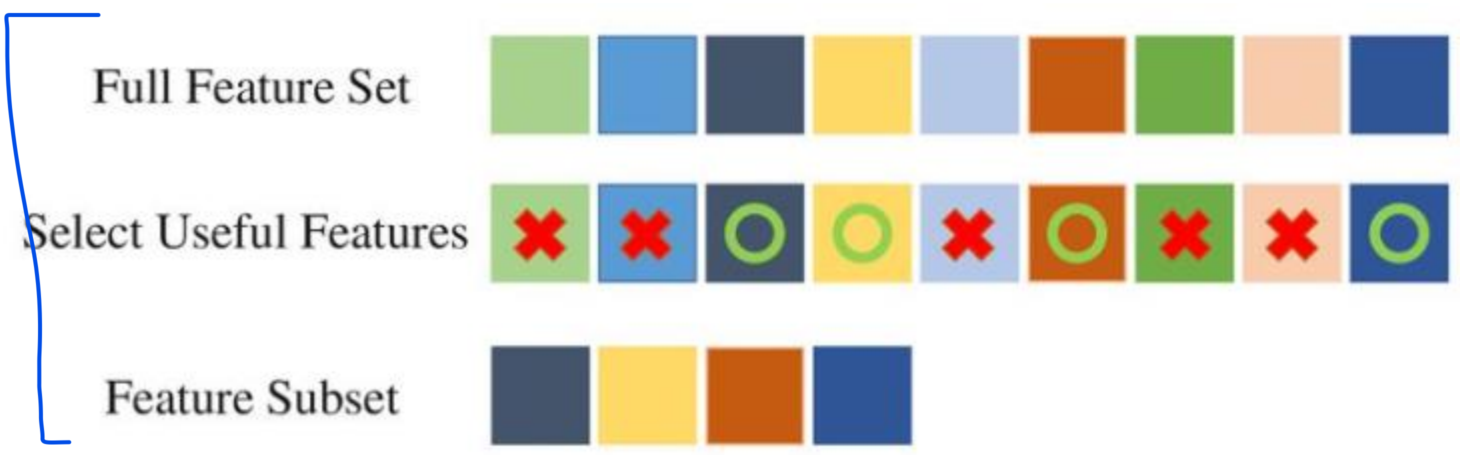
Why ANOVA ???

Feature Selection

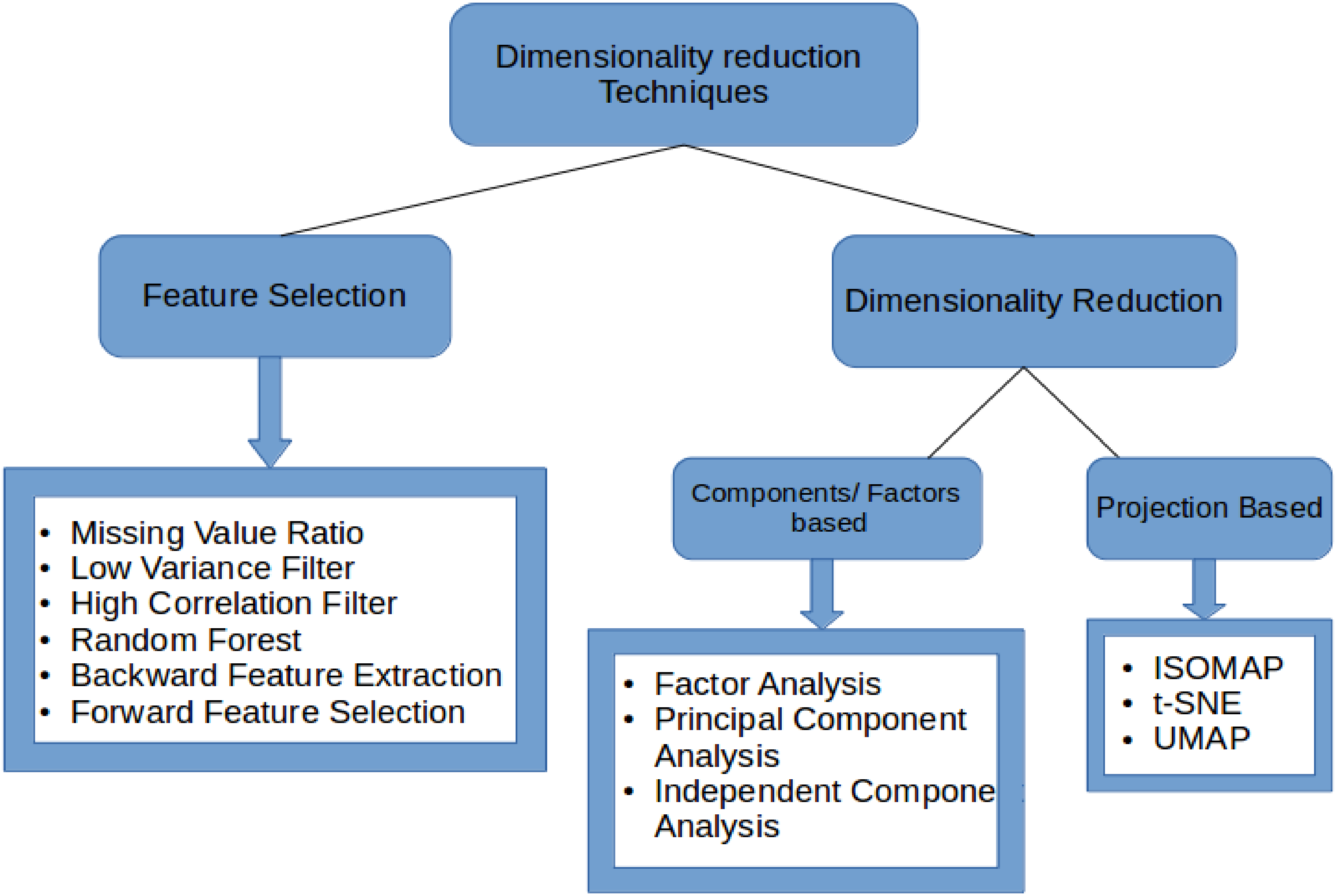
Feature Selection

- ✓ The focus of feature selection is to select a nice subset from the input data.

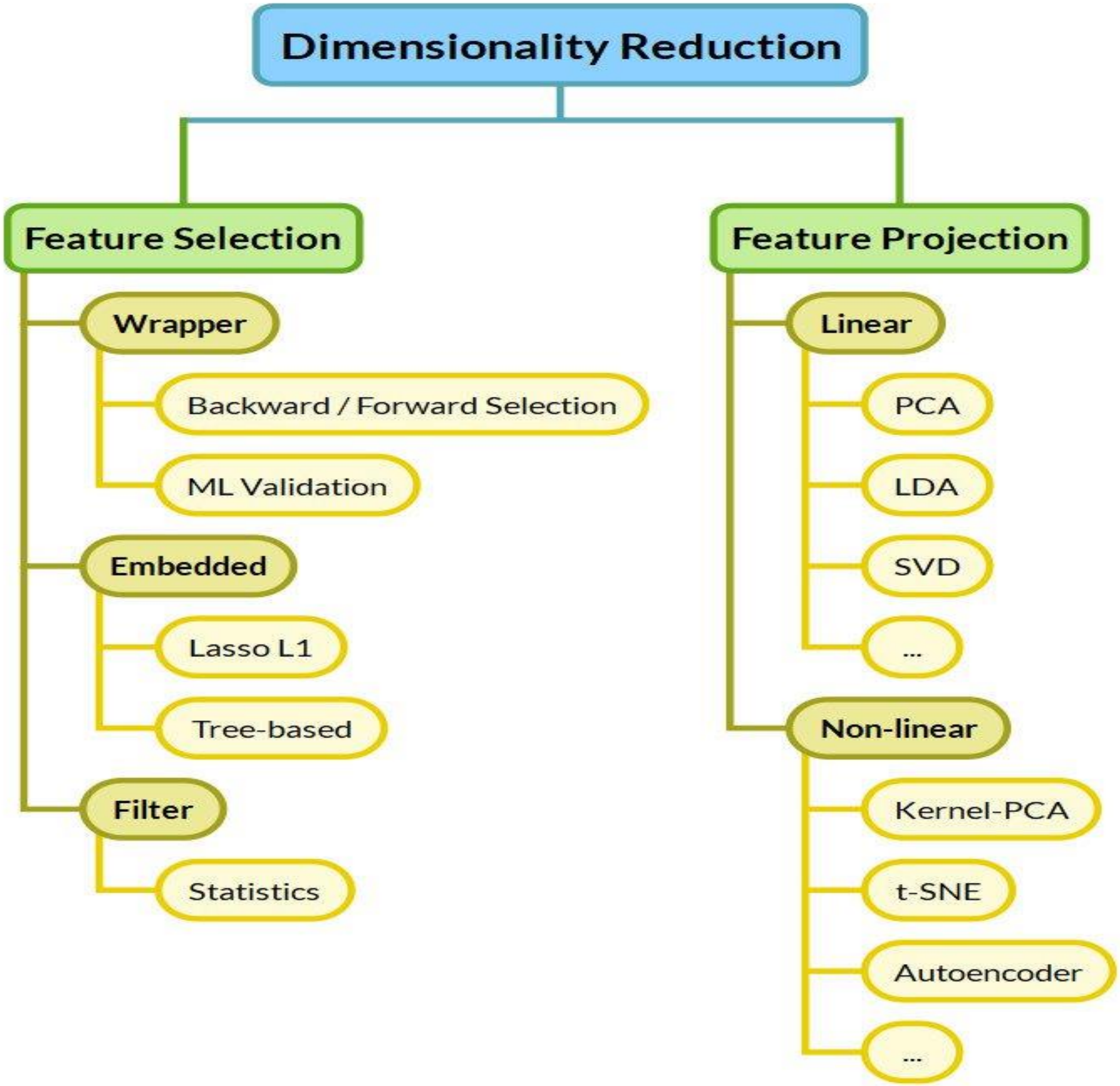
- ✓ It can make nice predictive accuracy while reducing noise or irrelevant features.



Feature Selection Strategies-1



Feature Selection Strategies-2



UNIVERSITY OF KARACHI

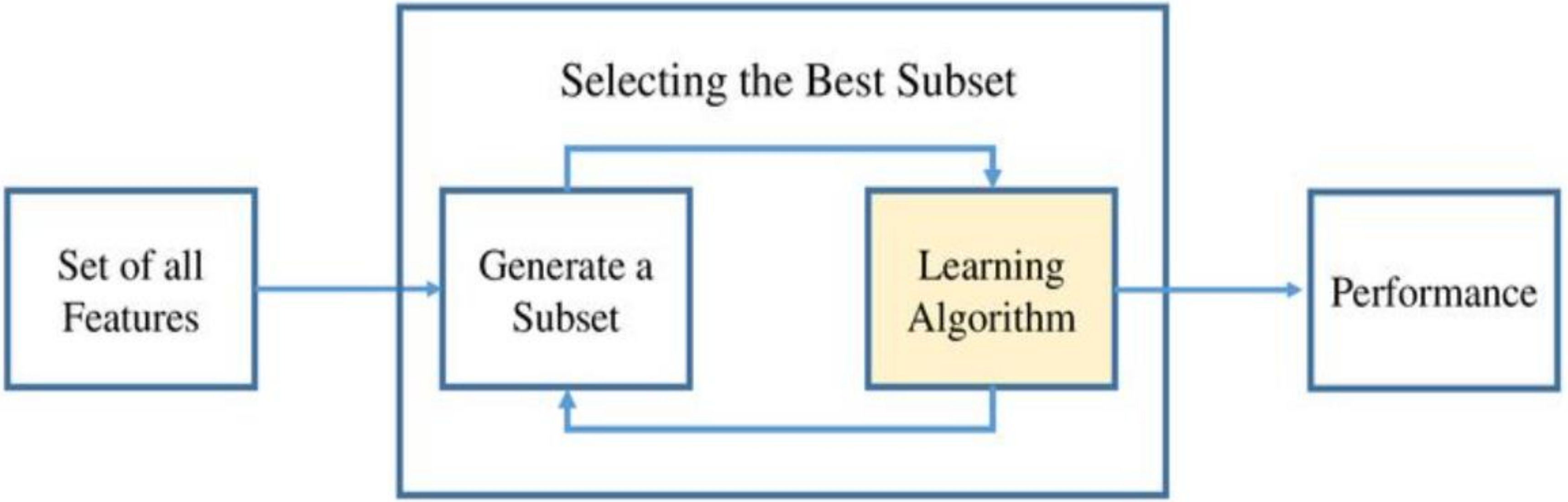
We are still on chasing :

Why ANOVA ???

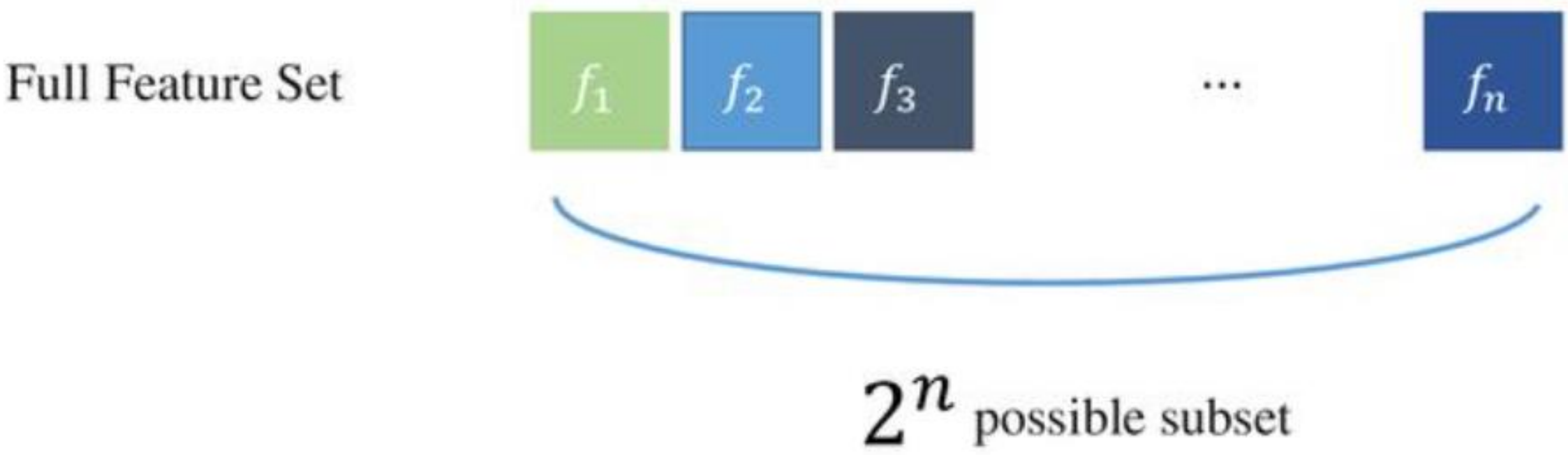
Feature Selection → Wrapper strategy

Wrapper Concept

- ✓ Feature set search component first generate a subset of features
- ✓ Learning Algorithm acts as a black box to evaluate the quality of these subsets/folds based on learning performance.
- ✓ The whole Process works iteratively until:
  - The best learning is achieved .
  - The desired number of selected feature is obtained.



Sub-set Selection in wrapper strategy is challenging



Unfortunately, If we have *n* features, the number of possible subsets is 2 *to the power n*. It is impossible for us to enumerate each of these possible subsets and check which good it is. Therefore, Wrapper methods usually uses the Heuristic Search Algorithm or Sequential Selection Algorithm to obtain the final subset within a reasonable time.

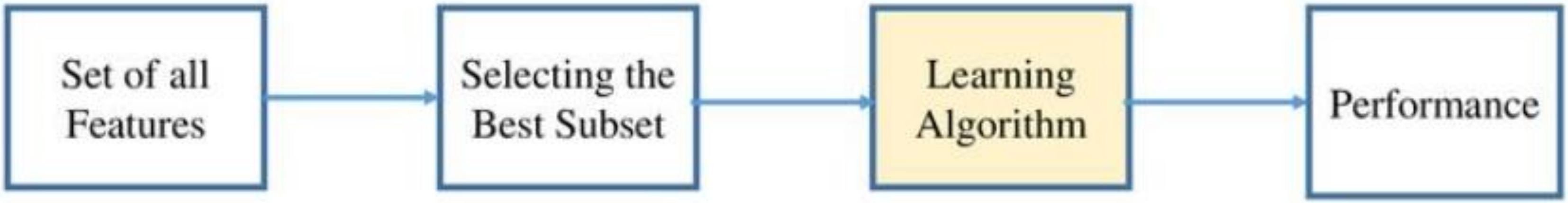
UNIVERSITY OF KARACHI

We are still on chasing :

Why ANOVA ???

Feature Selection → Filter strategy

Filter Concept

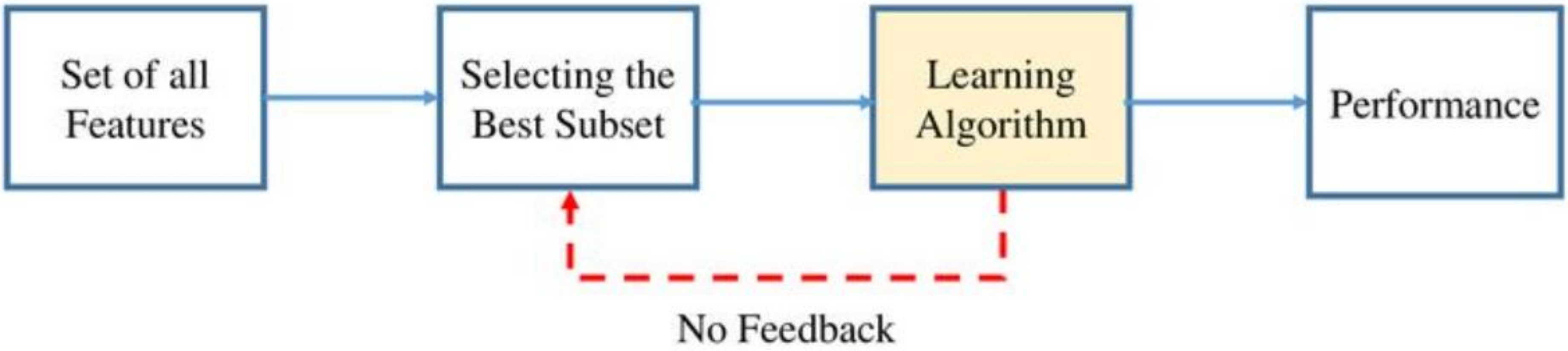


- ✓ Filter methods are independent of any learning algorithm.
- ✓ They rely on statistical measure about data to evaluate performance of each feature.
- ✓ They are more computationally efficient than wrapper methods.

UNIVERSITY OF KARACHI

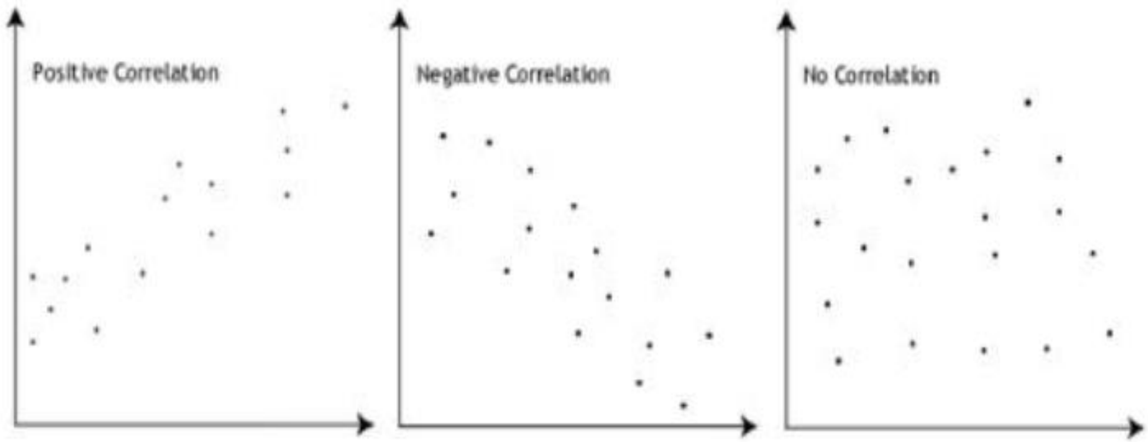
Sub-set Selection in Filter strategy is challenging

- ✓ Due to lack of learning algorithm guidance/ feed back in feature selection phase, the selected features may not be optimal for target learning algorithms



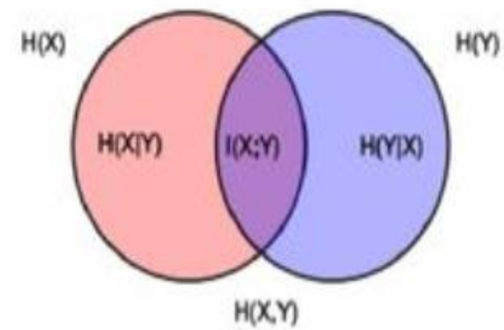
UNIVERSITY OF KARACHI

We are Close to our Target → Why ANOVA ?



Correlation criteria

$$I(X;Y) = \sum_{y \in Y} \sum_{x \in X} p(x,y) \log \left(\frac{p(x,y)}{p(x)\,p(y)} \right)$$



Mutual Information

Simplest Filter: Pearson Correlation Coefficient

One of the simplest criteria is the Pearson correlation coefficient defined as (1). Where x_i is the i_{th} variable, Y is the class labels, $cov()$ is the covariance and $var()$ the variance. Correlation ranking can only detect linear dependencies between variable and target.

- ✓ Numerical vs. categorical variable
- ✓ Regression vs. class label prediction
- ✓ Variance
- ✓ Co-Variance
- ✓ Correlation ranking → Detecting Linear Dependencies

UNIVERSITY OF KARACHI

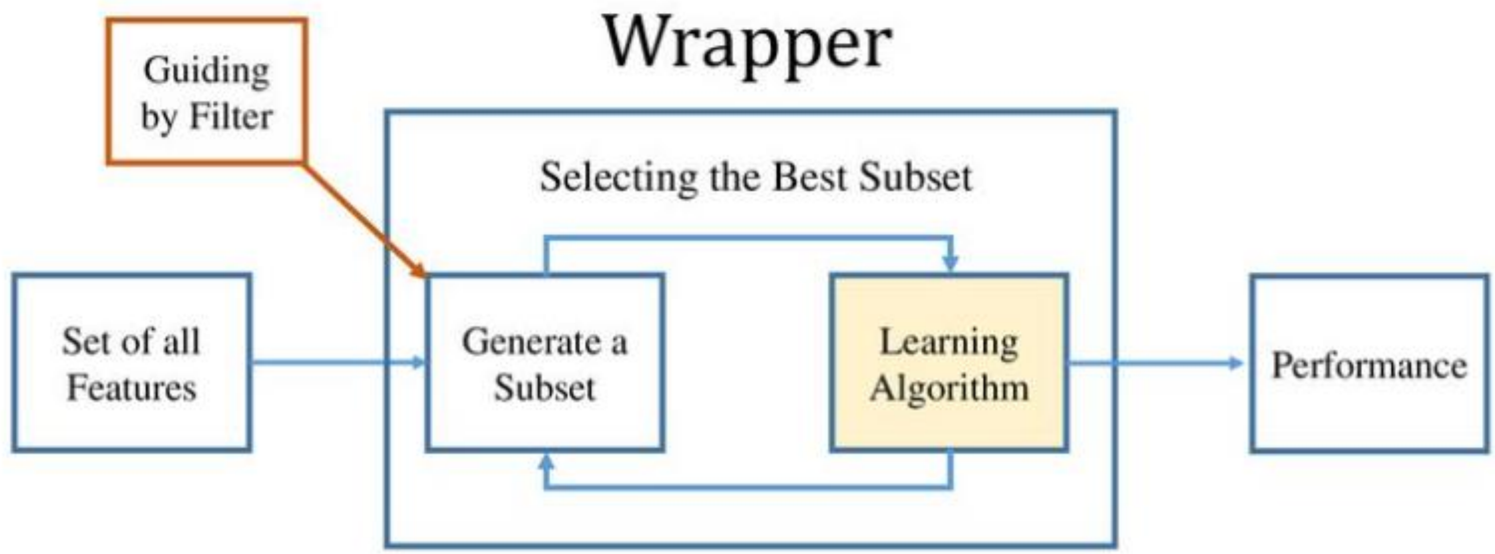
Pearson Correlation Formula

$$R(i) = \frac{cov(x_i, Y)}{\sqrt{var(x_i) * var(Y)}}$$

One of the simplest criteria is the Pearson correlation coefficient defined as (1). Where x_i is the i_{th} variable, Y is the class labels, $cov()$ is the covariance and $var()$ the variance. Correlation ranking can only detect linear dependencies between variable and target.

UNIVERSITY OF KARACHI

Recent Research: Wrapper + Filter



Recently research, It is effective to apply the Filter method when using the Wrapper methods. We can use the filter method when the Wrapper method is initialization phase or reproduction phase. It allows the wrapper to focus on promising features and increase the performance.