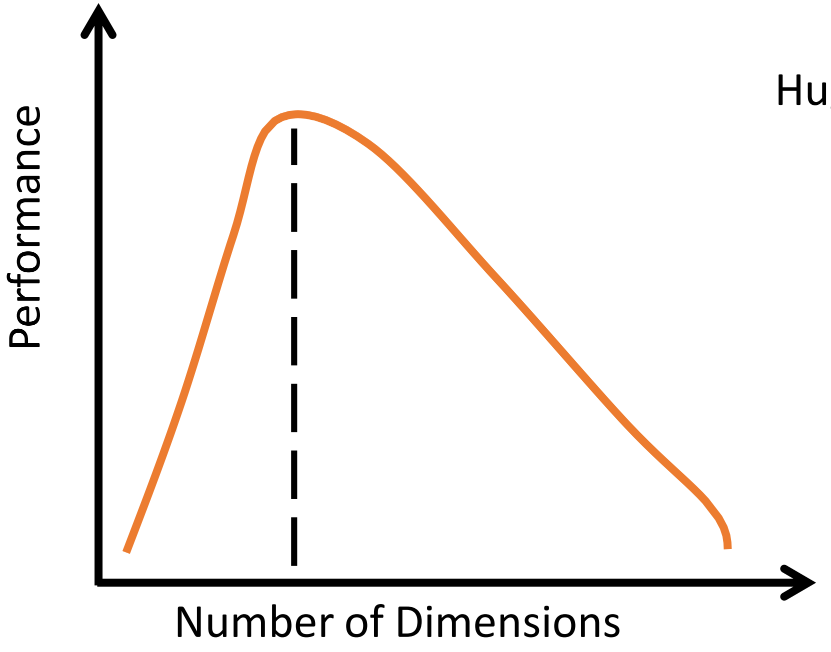
Feature Selection
&
Dimensionality Reduction

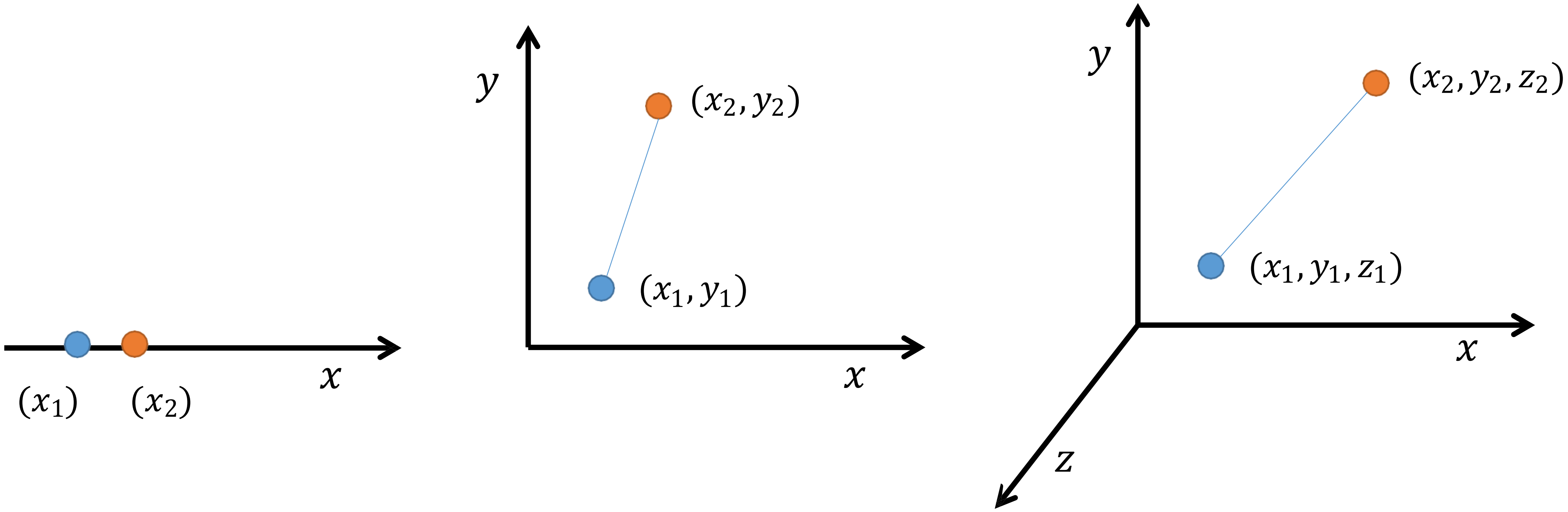
Curse of Dimensionality

What is Curse of Dimensionality?

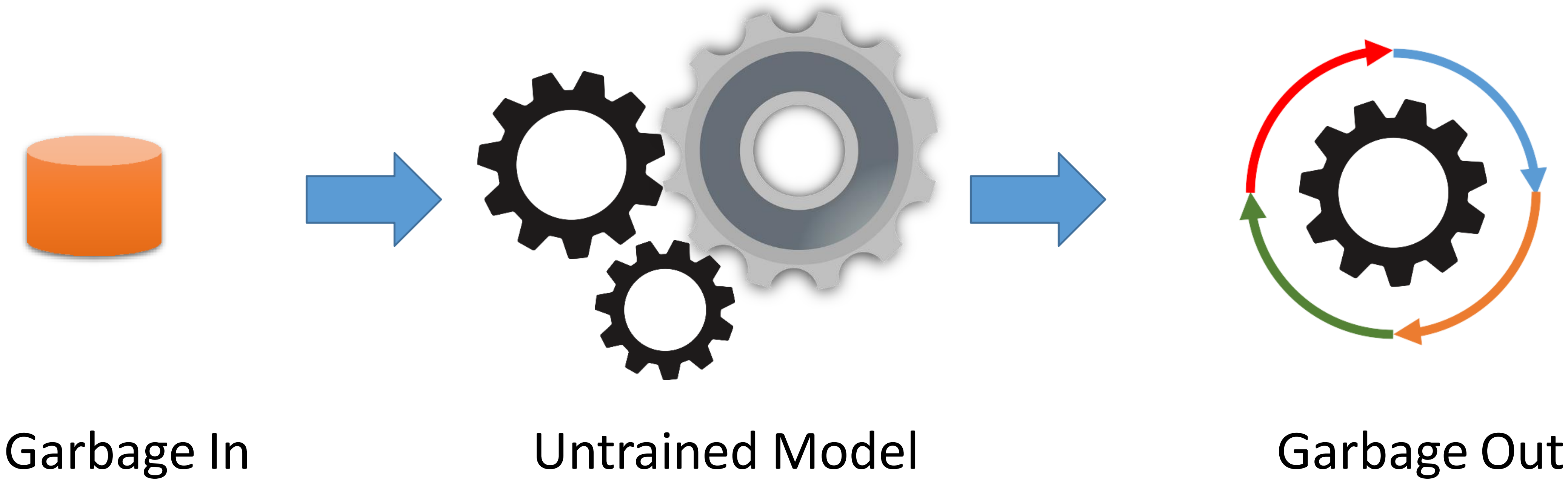


Hughes Phenomenon

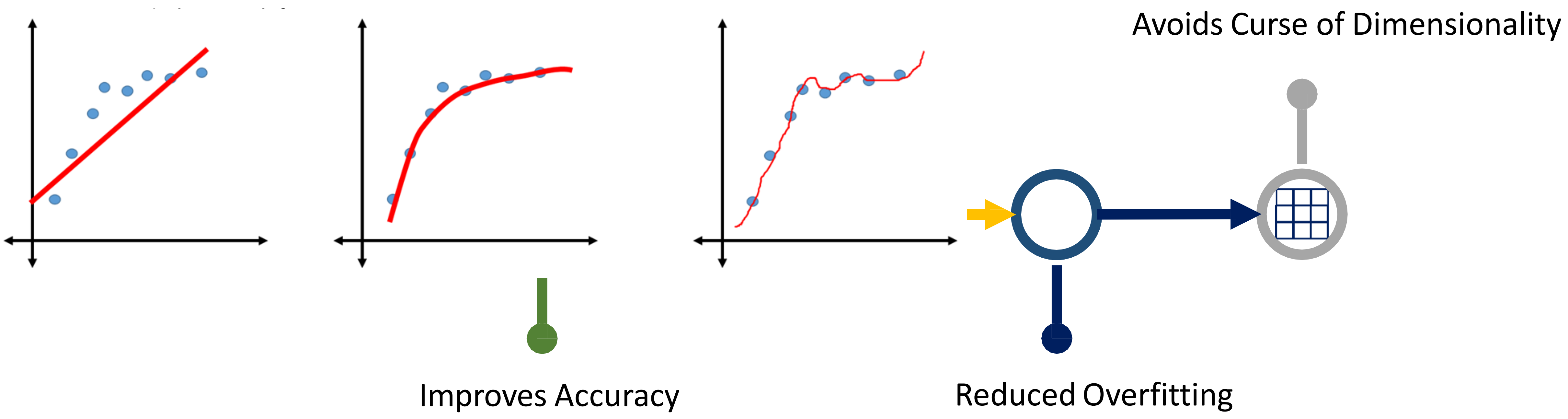
What is Curse of Dimensionality?



Does every feature improve accuracy?



Why to Use Feature Selection?



Univariate Feature Selection

Feature Selection Approaches – Scikit-Learn

- Variance Threshold
- Univariate
  - SelectKBest
  - SelectPercentile
  - SelectFpr
  - SelectFdr
  - SelectFwe
  - Generic Univariate
- Recursive Elimination
- Select From Model
  - L1 Based
  - Tree Based
- Pipeline

Steps For Univariate Feature Selection

Step 1 – Get all Independent Features

Step 2 – Apply relevant statistical method

Step 3 – Get P-Value and compare with the significance level

Step 4 – Select the feature if $P < \alpha$

Step 1 – Get all Independent Features

x_1

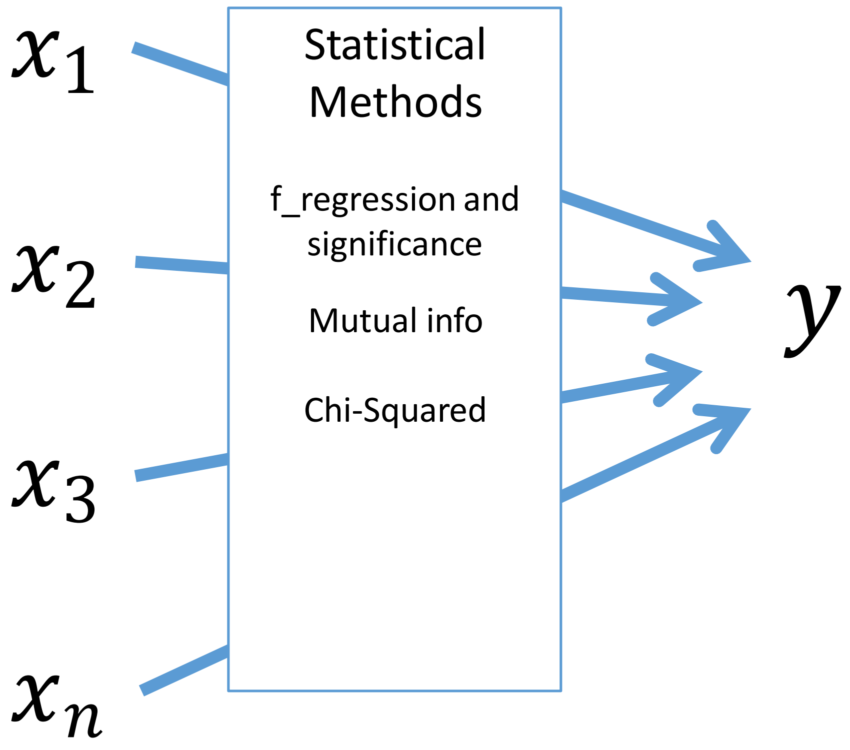
x_2

x_3

x_n

y

Step 2 – Apply relevant statistical method



$H_0 \rightarrow$ The feature has no impact on predictor

$H_a \rightarrow$ The feature has significant impact on predictor

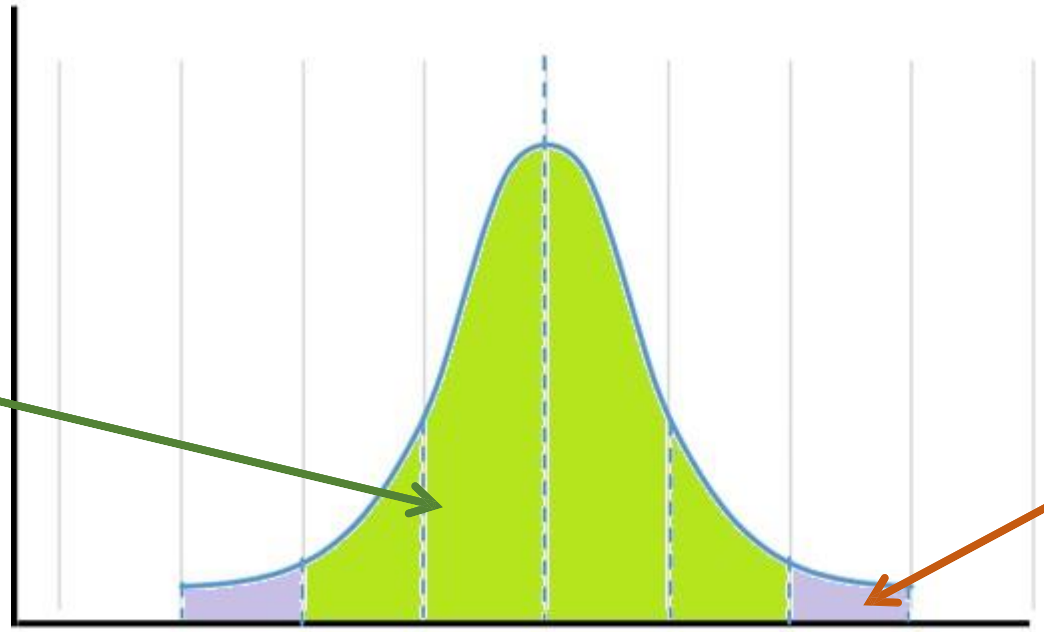
Statistical Significance

H_0

We expect this to happen.

It happened.

Nothing changes.
Null Hypothesis is true.
Status quo remains.



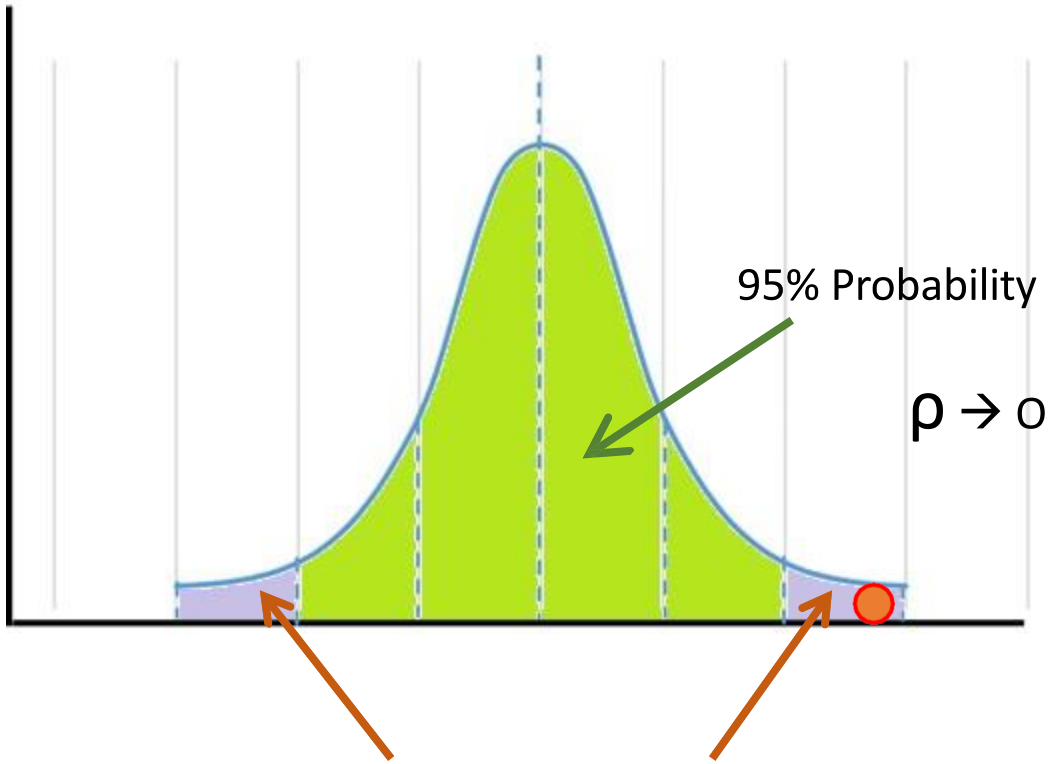
H_a

We don't expect this.

Still it happened.

Null Hypothesis rejected.
Status quo or claim is rejected

Important terms – Statistical Significance



ρ → Observed or seen Probability of the sample

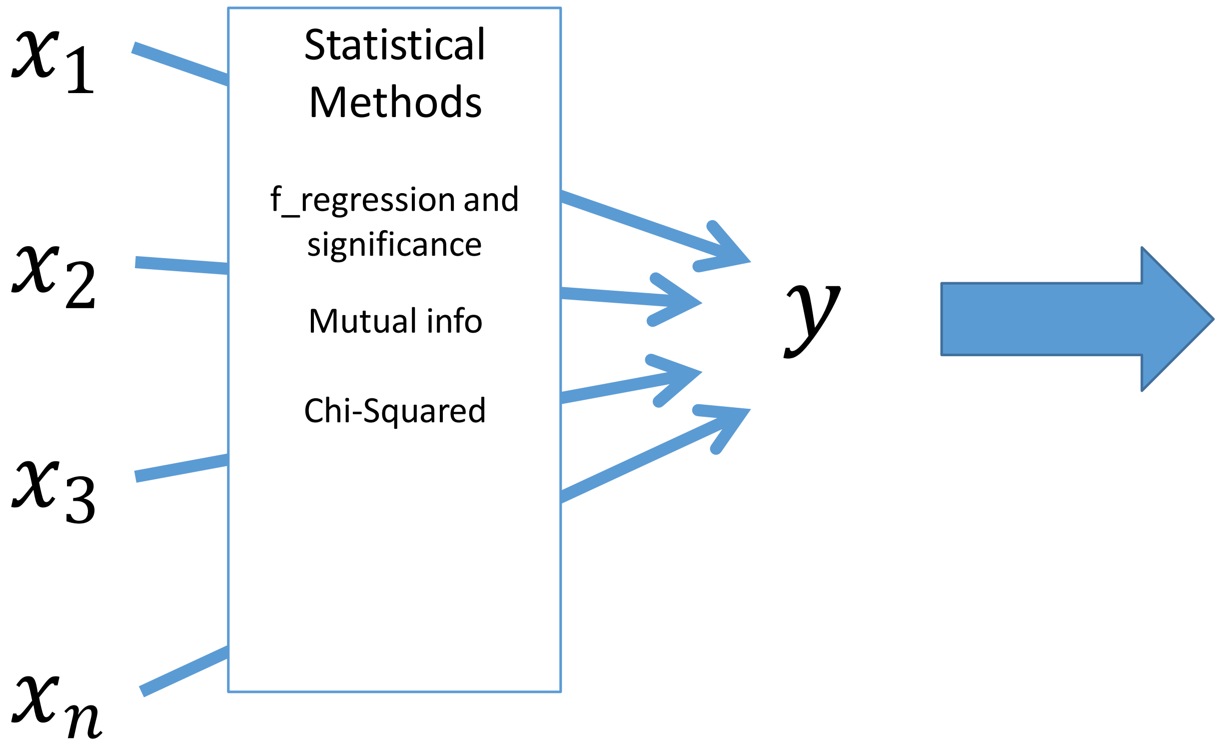
Rejection Region

Probability of rejection region → α

$\rho < \alpha$

Reject Null Hypothesis

Step 3 – Get P-Value and compare with the significance level



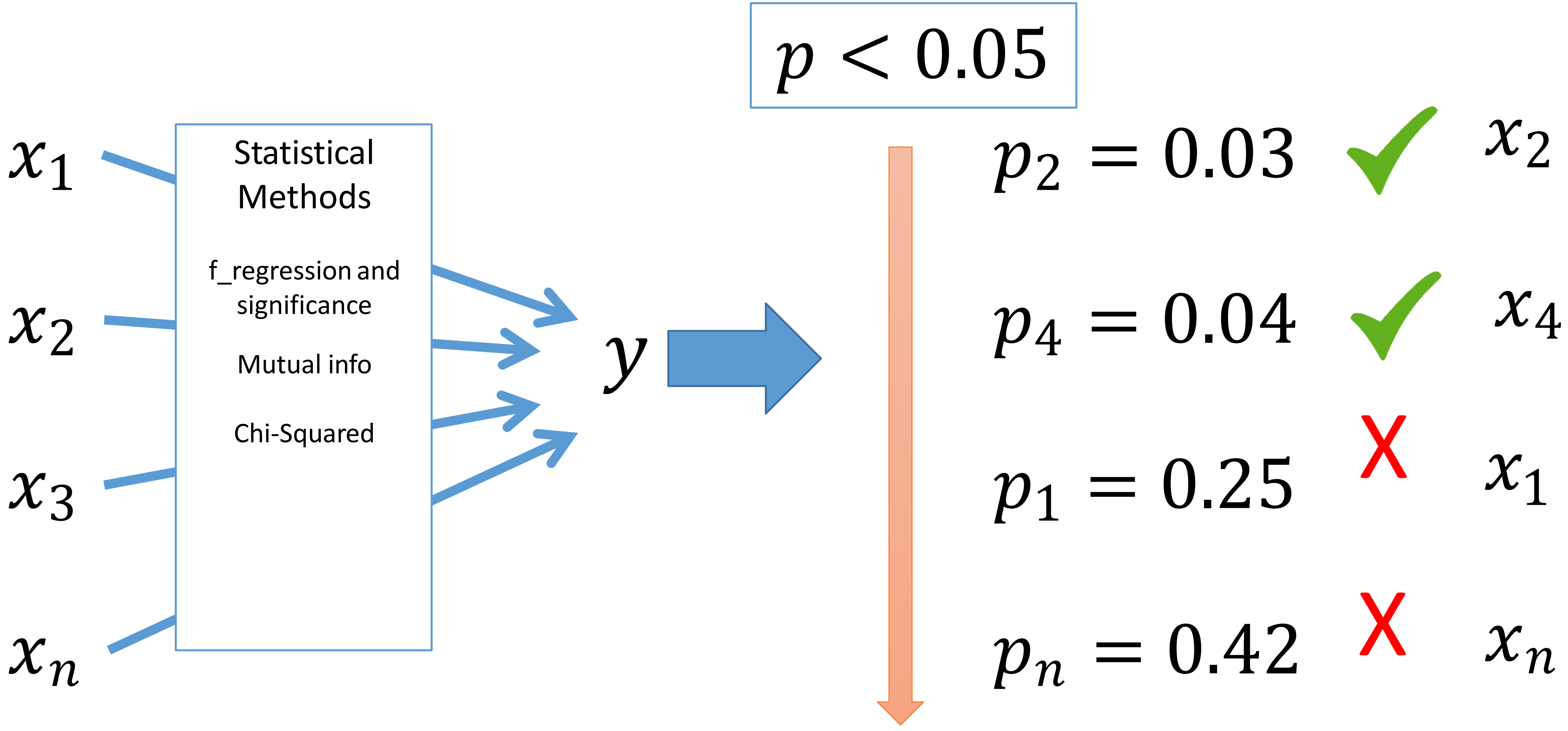
$p_1 < \alpha\,?$

$p_2 < \alpha\,?$

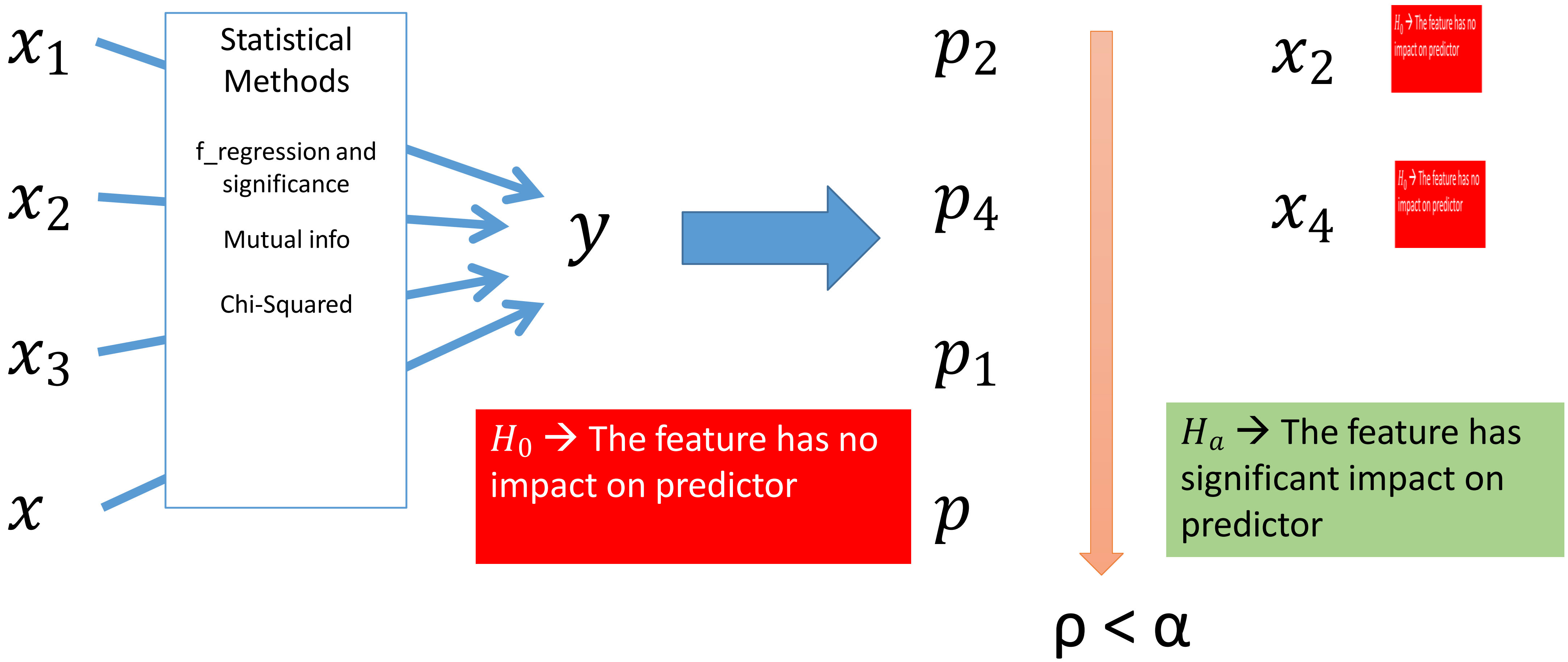
$p_3 < \alpha\,?$

$p_n < \alpha\,?$

Step 4 – Select the feature if P < α

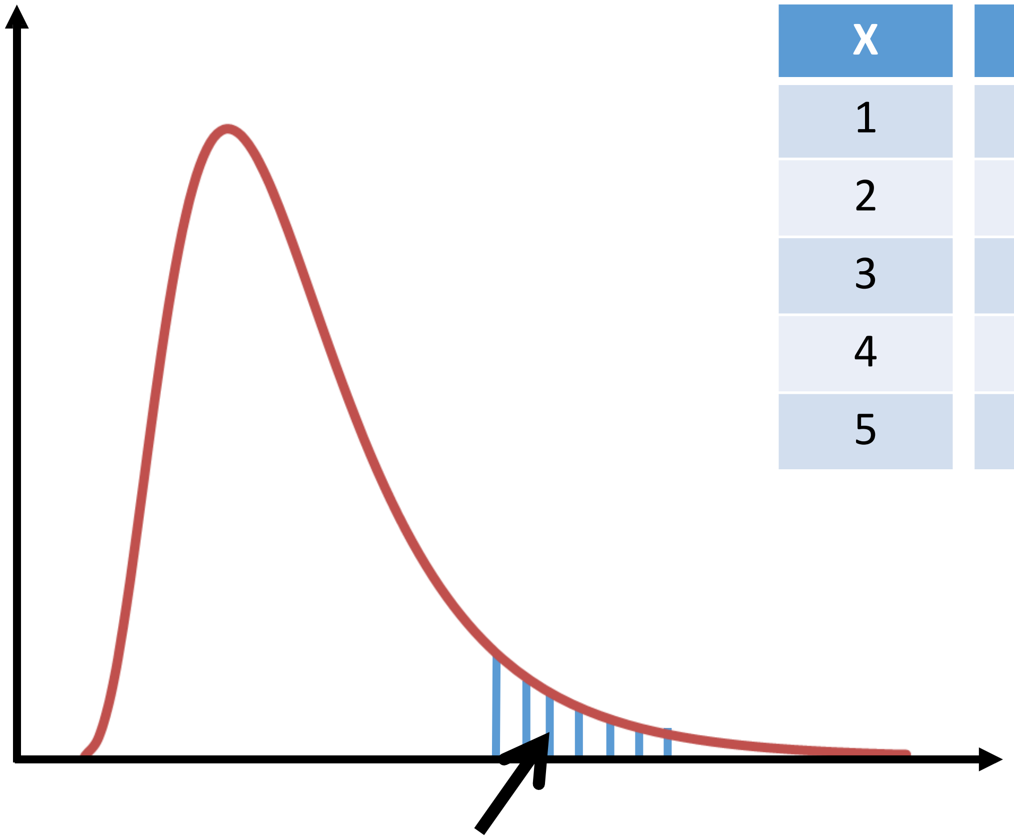


Step 4 – Select the feature if P < α



F-Distribution

F-Distribution



Probability of rejection region → α

X	Y
1	2
2	4
3	5
4	8
5	10

$$\text{F-Score} = \frac{R^2}{1-R^2} * \frac{df_2}{df_1}$$

R = Correlation Coefficient

$$R = \frac{\sum(x-\bar{x}) * (y-\bar{y})}{\sigma_x \sigma_y}$$

df_2 = Degrees of freedom within the group

df_1 = Degrees of freedom between the groups

F-Test for Target variables

$$y \ \rightarrow Continuous \ \rightarrow f\_regression$$

$$y \ \rightarrow Categorical \ \rightarrow f\_classif$$

Chi-Square

Chi Squared Test of Independence

- Developed by Karl Pearson
- Evaluates the relationship when the target variable is categorical
- Steps to Evaluate the Independence
  - Define Hypothesis – Null and Alternate
  - Define Alpha
  - Calculate the Degrees of Freedom
  - State Decision Rule
  - Calculate Test Statistics
  - Results
  - Conclusion

Chi Square Test of Independence

Flight Status	Weather
Delayed	Rainy
Delayed	Rainy
Delayed	Rainy
Ontime	Rainy
Delayed	Rainy
Ontime	Sunny
Delayed	Rainy
Delayed	Rainy
Ontime	Sunny
Delayed	Rainy
Delayed	Overcast
Delayed	Overcast
Delayed	Overcast
Delayed	Overcast
Delayed	Overcast

Step 1

Null Hypothesis – There is no relationship between Flight Status and Weather

Alternate Hypothesis – There is relationship between Flight Status and Weather

Step 2

alpha, $\alpha = 0.05$

Chi Square Test of Independence

	Rainy	Sunny	Overcast	
Delayed	36	16	13	**65**
On time	11	84	40	**135**
	47	**100**	**53**	

Step 3

Calculate the degrees of freedom

Total df = (no of Rows – 1) * (no of Columns – 1)

= (2 – 1) * (3 – 1)

= **2**

Chi Square Test of Independence

Step 4

State Decision Rule using chi square degrees of freedom table

df = 2

alpha, α = 0.05

Table 3-1 Critical Values of the χ^2 Distribution

	P						
df	**0.995**	**0.975**	**0.9**	**0.5**	**0.1**	**0.05**	**0.025**
1	.000	.000	0.016	0.455	2.706	3.841	5.024
2	0.010	0.051	0.211	1.386	4.605	5.991	7.378
3	0.072	0.216	0.584	2.366	6.251	7.815	9.348
4	0.207	0.484	1.064	3.357	7.779	9.488	11.143
5	0.412	0.831	1.610	4.351	9.236	11.070	12.832
6	0.676	1.237	2.204	5.348	10.645	12.592	14.449
7	0.989	1.690	2.833	6.346	12.017	14.067	16.013
8	1.344	2.180	3.490	7.344	13.362	15.507	17.535
9	1.735	2.700	4.168	8.343	14.684	16.919	19.023

Reject the Null Hypothesis if the X square value is greater than 5.991

Chi Squared Test of Independence

Actual

	Rainy	Sunny	Overcast	
Delayed	36	16	13	**65**
On time	11	84	40	**135**
	47	**100**	**53**	

Step 5 Calculate Test Statistics

$$f = \frac{f_c * f_r}{n}$$

Expected

	Rainy	Sunny	Overcast	
Delayed	15			
On time				

Expected (Delayed, Rainy)

= (65*47)/200

= 15.275 ~ 15

Chi Squared Test of Independence

Actual

	Rainy	Sunny	Overcast	
Delayed	36	16	13	**65**
On time	11	84	40	**135**
	47	**100**	**53**	

Expected

	Rainy	Sunny	Overcast	
Delayed	15	33		
On time				

Step 5 Calculate Test Statistics

$$f = \frac{f_c * f_r}{n}$$

Expected (Delayed, Sunny)

= (65*100)/200

= 32.5 ~ 33

Chi Squared Test of Independence

Step 5 Calculate Test Statistics

Actual

	Rainy	Sunny	Overcast	
Delayed	36	16	13	**65**
On time	11	84	40	**135**
	47	**100**	**53**	

$$f = \frac{f_c * f_r}{n}$$

Expected

	Rainy	Sunny	Overcast	
Delayed	15	33		
On time	32			
	47			

Expected (OnTime, Rainy)

= (135*47)/200

= 31.725 ~ 32

Chi Squared Test of Independence

Actual

	Rainy	Sunny	Overcast	
Delayed	36	16	13	**65**
On time	11	84	40	**135**
	47	**100**	**53**	

Step 6 Calculate Results

$$Score = \frac{(f_o - f_e)^2}{f_e}$$

Expected

	Rainy	Sunny	Overcast	
Delayed	15	33	17	**65**
On time	32	67	36	**135**
	47	**100**	**53**	

	Rainy	Sunny	Overcast
Delayed	29.4	8.76	0.94
On time	13.78	4.31	0.44

Chi Squared Test of Independence

Flight Status	Weather
Delayed	Rainy
Ontime	Rainy
Delayed	Rainy
Ontime	Sunny
Delayed	Overcast
Delayed	Overcast

Step 6 Calculate Results

$$Score = \frac{(f_o - f_e)^2}{f_e}$$

$$Total\ Score = 57.64$$

	Rainy	Sunny	Overcast
Delayed	29.4	8.76	0.94
On time	13.78	4.31	0.44

Chi Squared Test of Independence

Step 7 Conclusion

Chi squared (57.64) > 5.991

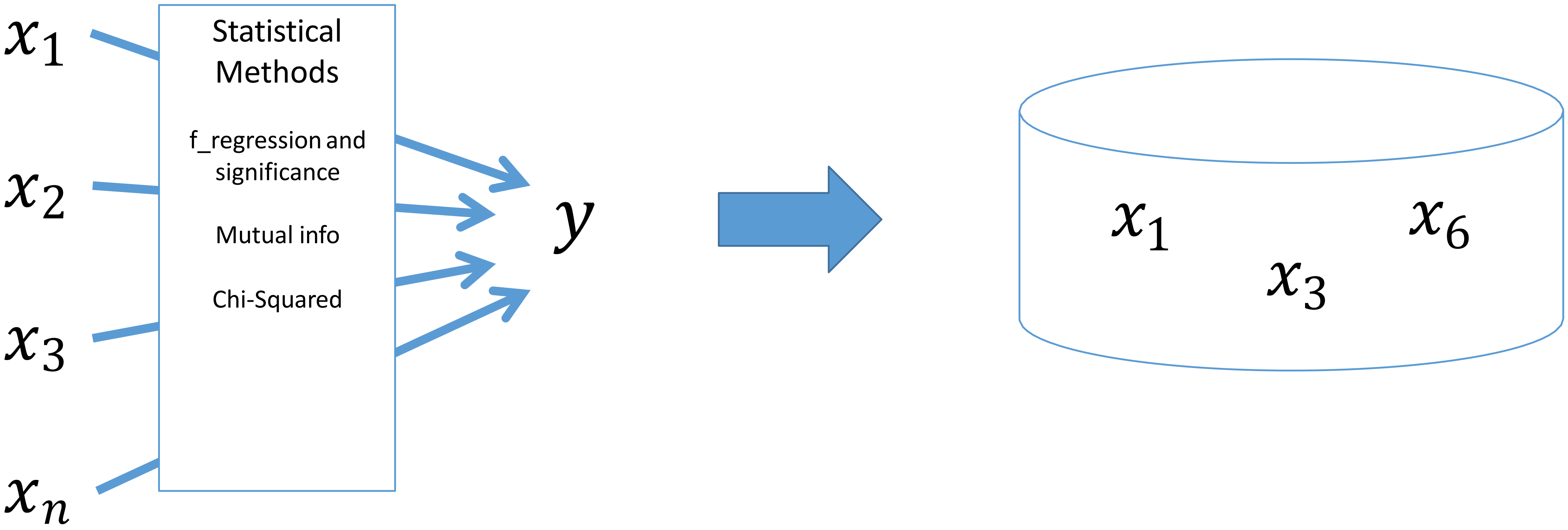
Reject the Null Hypothesis.

Table 3-1 **Critical Values of the χ^2 Distribution**

	P						
df	**0.995**	**0.975**	**0.9**	**0.5**	**0.1**	**0.05**	**0.025**
1	.000	.000	0.016	0.455	2.706	3.841	5.024
2	0.010	0.051	0.211	1.386	4.605	5.991	7.378
3	0.072	0.216	0.584	2.366	6.251	7.815	9.348
4	0.207	0.484	1.064	3.357	7.779	9.488	11.143
5	0.412	0.831	1.610	4.351	9.236	11.070	12.832
6	0.676	1.237	2.204	5.348	10.645	12.592	14.449
7	0.989	1.690	2.833	6.346	12.017	14.067	16.013
8	1.344	2.180	3.490	7.344	13.362	15.507	17.535
9	1.735	2.700	4.168	8.343	14.684	16.919	19.023

The weather and Flight Status are correlated.

Selection Transforms



Most common Feature Selection Transforms

GenericUnivariateSelect

mode

score_func

param

SelectKBest

SelectPercentile

score_func

f_regression
f_classif
chi2

K

Number of
top features

Integer

Percentile

Top percentile of
features

Float

Most common Feature Selection Transforms

GenericUnivariateSelect

mode	=	K_best	K_best	K_best
score_func	=	f_regression	f_classif	chi2
param	=	**K=10**	**K=10**	**K=10**

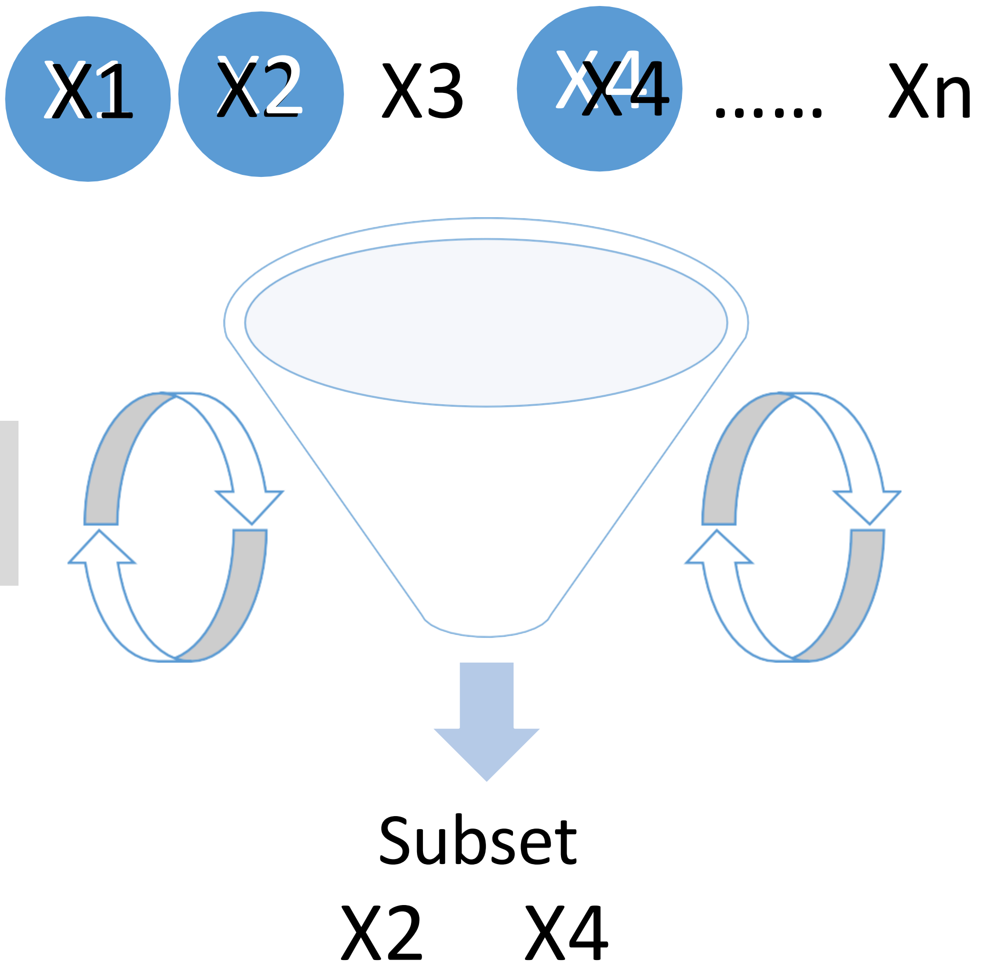
Most common Feature Selection Transforms

GenericUnivariateSelect

mode	=	percentile	percentile	percentile
score_func	=	f_regression	f_classif	chi2
param	=	**Percentile=20.0**	**Percentile=20.0**	**Percentile=20.0**

Recursive Feature Elimination

Recursive Feature Elimination



Learning Algorithm

Classification
Regression

Performance

Feature Importance
Coefficients

Criteria for Feature Selection or Elimination

- Coefficients or weights
- Feature Importance

Rank Ordered Features for elimination

x_1

x_2

x_3

.

.

.

x_n

Recursive Feature Elimination

Input Features

x_1

x_2

x_3

.

.

.

x_n

Coefficients or weights

$$y = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_n x_n$$

Learning Algorithm

Feature Importance

$$Entropy = -1 * \sum_{i=1}^{n} p_i \log_2 p_i \qquad Gini = 1 - \sum_{i=1}^{n} p_i^2$$

Rank Ordered Features

x_2

x_3

x_7

Principal Component Analysis

What is a Principal Component?

- Creates a new set of coordinates for the data
- Reveals the internal structure of the data that best explains the variance in data
- Reduces the dimensionality of the multivariate dataset

Predict the demand for bikes

https://www.capitalbikeshare.com

How Capital Bikeshare Works

Unlock

Pick up a bike at one of hundreds of stations around the metro DC area. See bike availability on the System Map or mobile app.

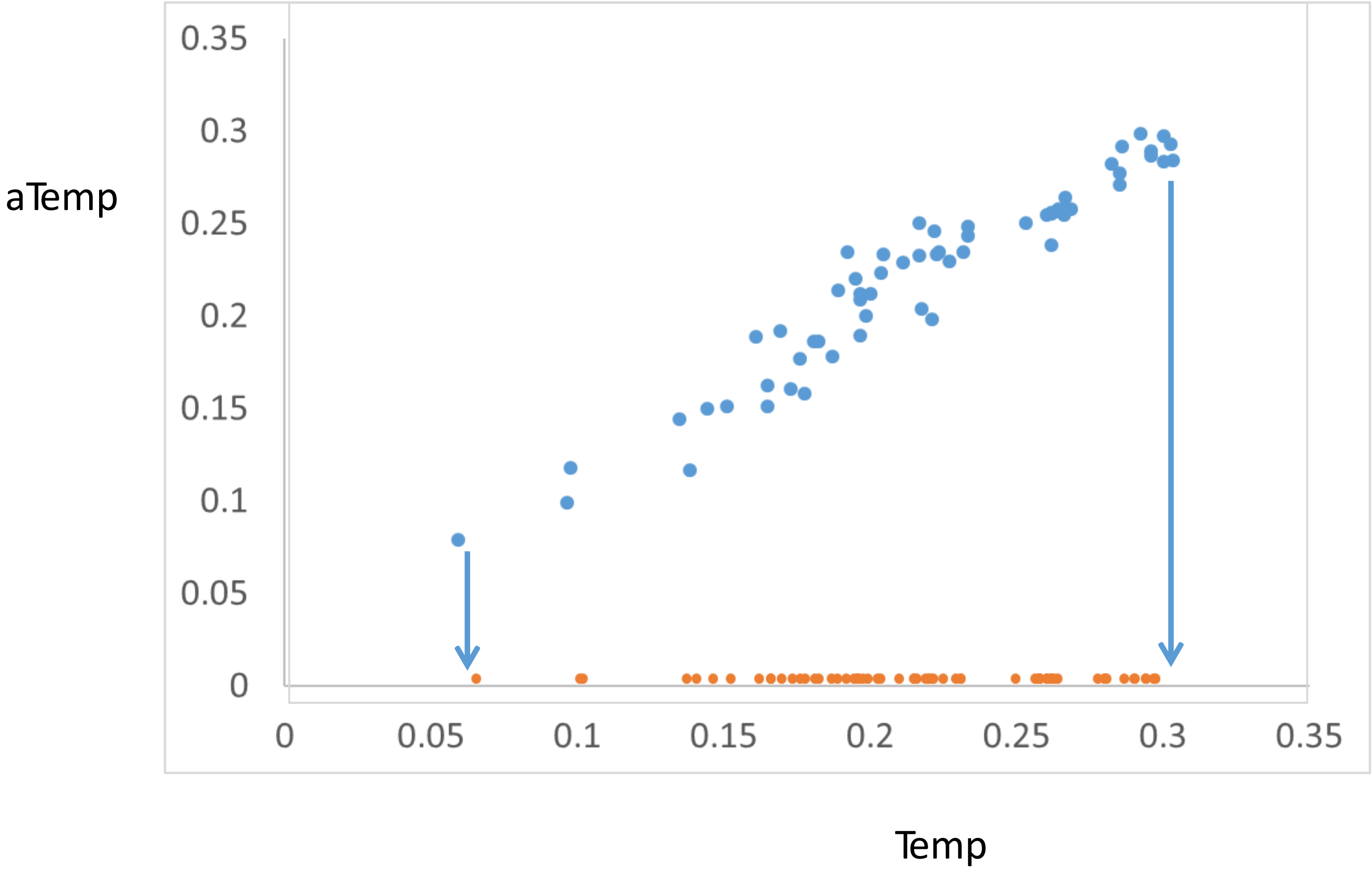
Ride

Take as many short rides as you want while your pass is active. Passes and memberships include unlimited classic bike trips under 30 minutes.

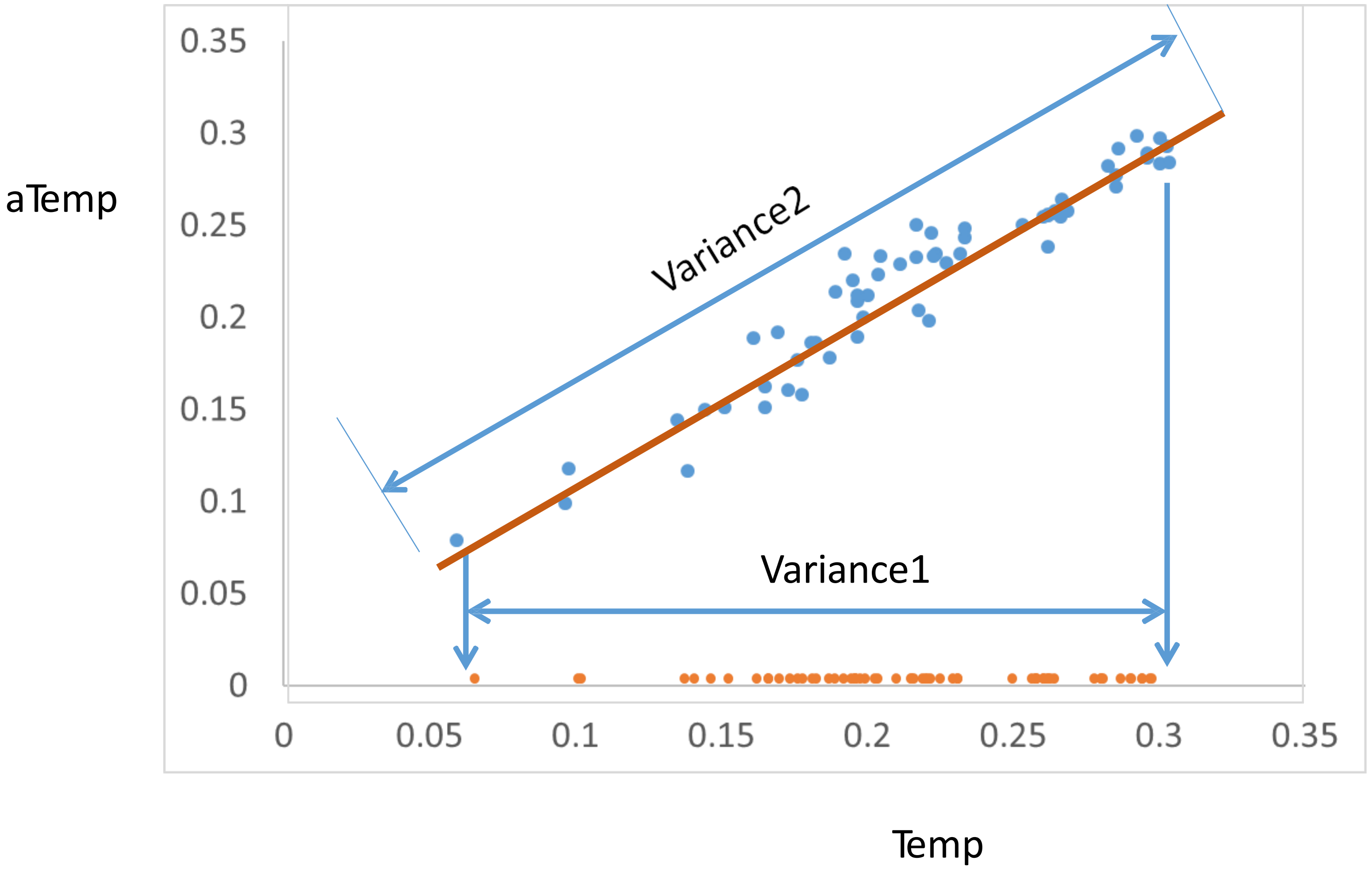
Return

End a ride by returning your bike to any station. Push your bike firmly into an empty dock and wait for the green light to make sure it's locked.

Actual Temperature Vs Feels Like

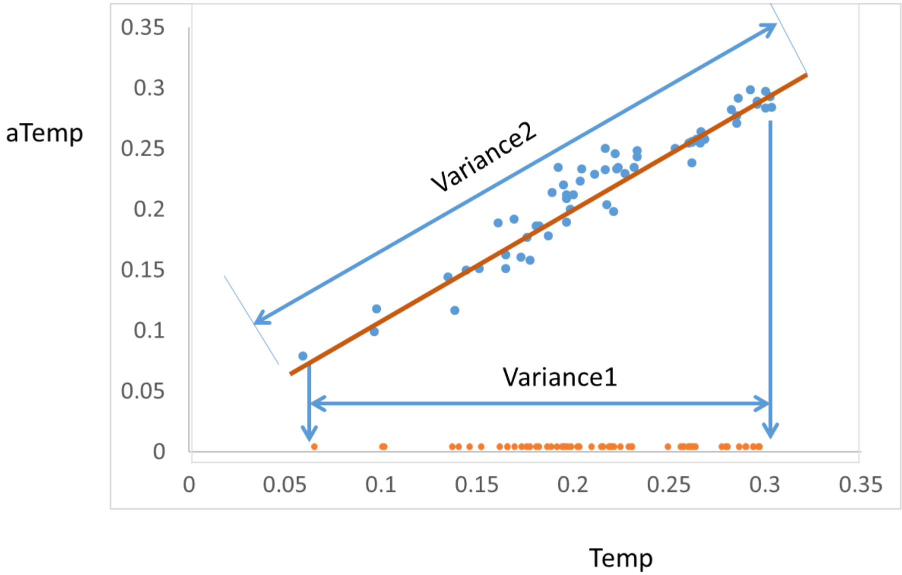


Actual Temperature Vs Feels Like

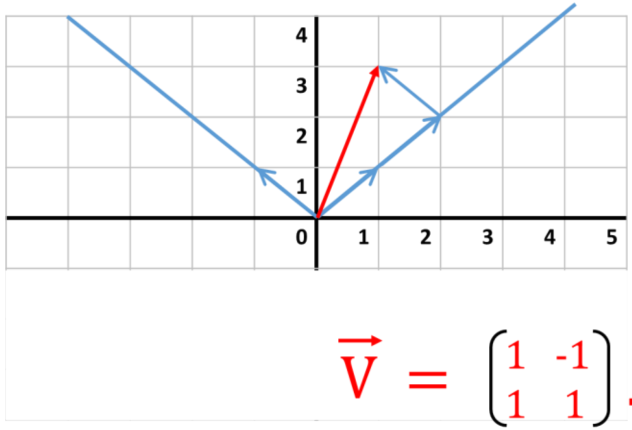


Important concepts to know for PCA

- Variance and covariance among variables



- Change of Basis using matrix transformation



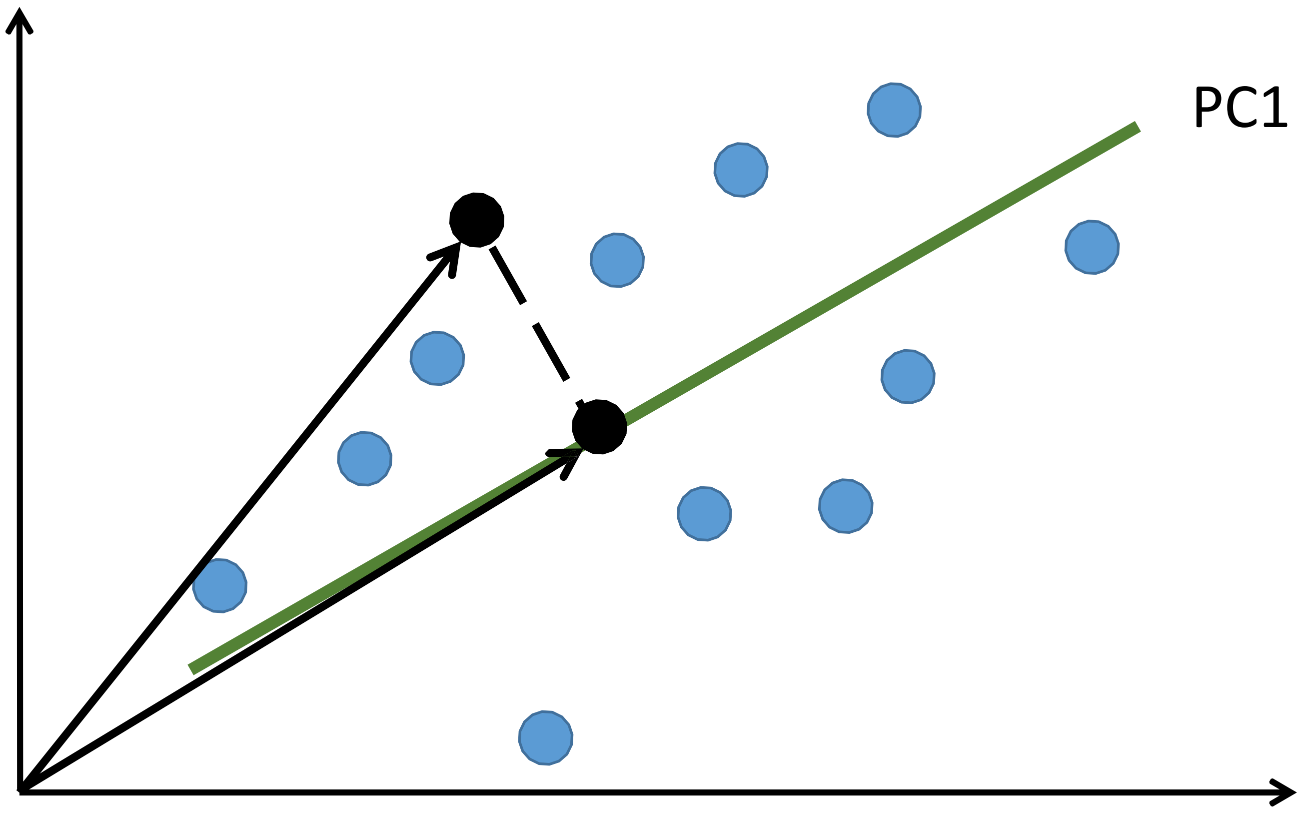
$$b_1 = \begin{pmatrix} 1 \\ 1 \end{pmatrix} \quad b_2 = \begin{pmatrix} -1 \\ 1 \end{pmatrix}$$

$$\vec{W} = \begin{pmatrix} 2 \\ 1 \end{pmatrix} = 2b_1 + b_2$$

$$\vec{V} = \begin{pmatrix} 1 & -1 \\ 1 & 1 \end{pmatrix} . \vec{W}$$

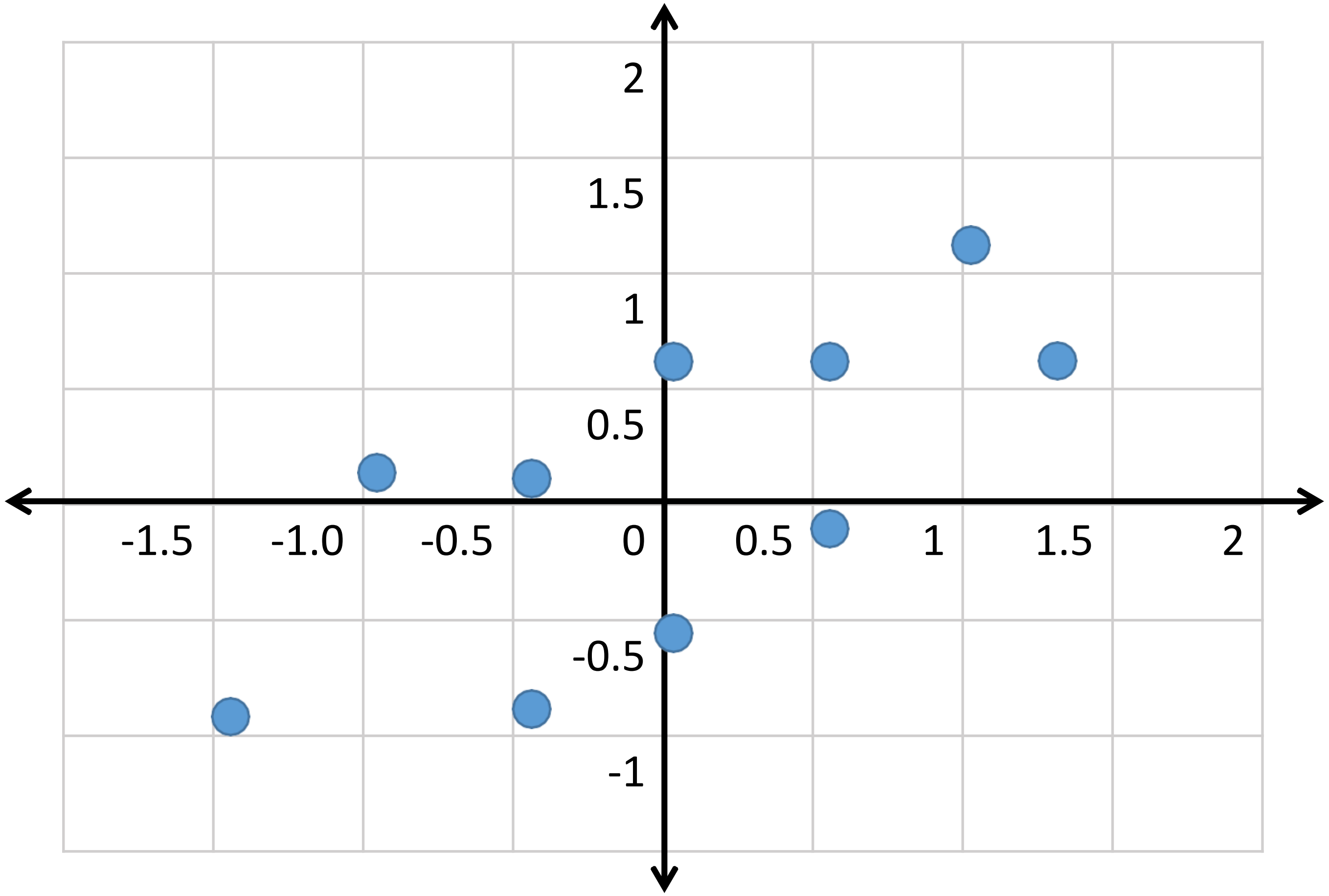
Matrix Transformation of $\vec{W}$

Projection of Data on new axis



Steps in Creating the Principal Components

Step 1 – Center the Data

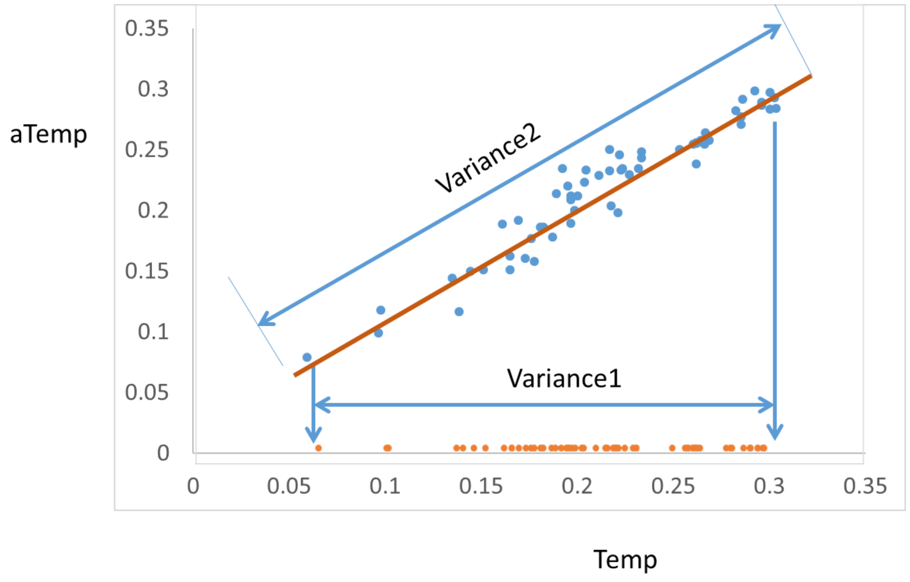


X1	X2
-1.475	-0.955
-0.975	0.045
-0.475	-0.955
-0.475	0.045
0.025	-0.655
0.025	0.545
0.525	-0.205
0.525	0.545
1.025	1.045
1.275	0.545

Steps in Creating the Principal Components

Step 1 – Center the Data

Step 2 – Create Variance-Covariance Matrix



Covariance Matrix

	Height X	Weight Y	$X - \bar{X}$	$Y - \bar{Y}$	$(X - \bar{X}) * (Y - \bar{Y})$
	160	130	-15.625	-40.625	634.7656
	170	150	-5.625	-20.625	116.0156
	165	145	-10.625	-25.625	272.2656
	180	190	4.375	19.375	84.76563
	175	175	-0.625	4.375	-2.73438
	190	210	14.375	39.375	566.0156
	185	180	9.375	9.375	87.89063
	180	185	4.375	14.375	62.89063
Mean	**175.625**	**170.625**			**1821.875**
Std Dev	**10.155**	**25.651**			

	x	y
x	Variance(x)	Covariance(x, y)
y	Covariance (y, x)	Variance (y)

Variance – Covariance Matrix

$$\text{Covariance, } S_{xy}^2 = \frac{\sum (x - \bar{x}) * (y - \bar{y})}{(N - 1)}$$

$$\text{Covariance, } S_{xy}^2 = \frac{\sum (x - \bar{x}) * (y - \bar{y})}{(N - 1)}$$

	X1	X2
X1	Variance(x)	Covariance(x, y)
X2	Covariance (y, x)	Variance (y)

$$\begin{array}{c c} & \begin{array}{cc} X1 & X2 \end{array} \\ \begin{array}{c} X1 \\ X2 \end{array} & \begin{pmatrix} 0.76 & 0.42 \\ 0.42 & 0.48 \end{pmatrix} \end{array}$$

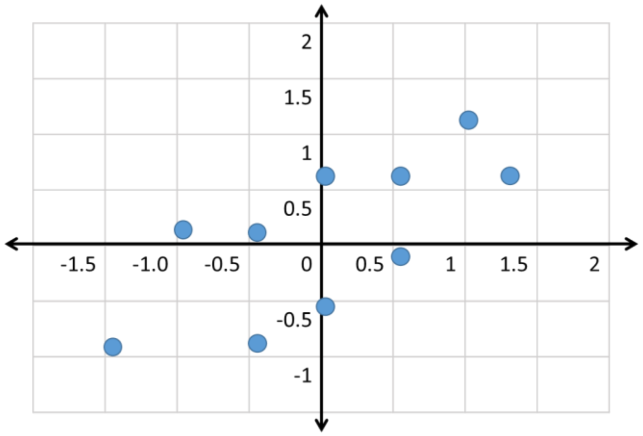
X1	X2
-1.475	-0.955
-0.975	0.045
-0.475	-0.955
-0.475	0.045
0.025	-0.655
0.025	0.545
0.525	-0.205
0.525	0.545
1.025	1.045
1.275	0.545

Steps in Creating the Principal Components

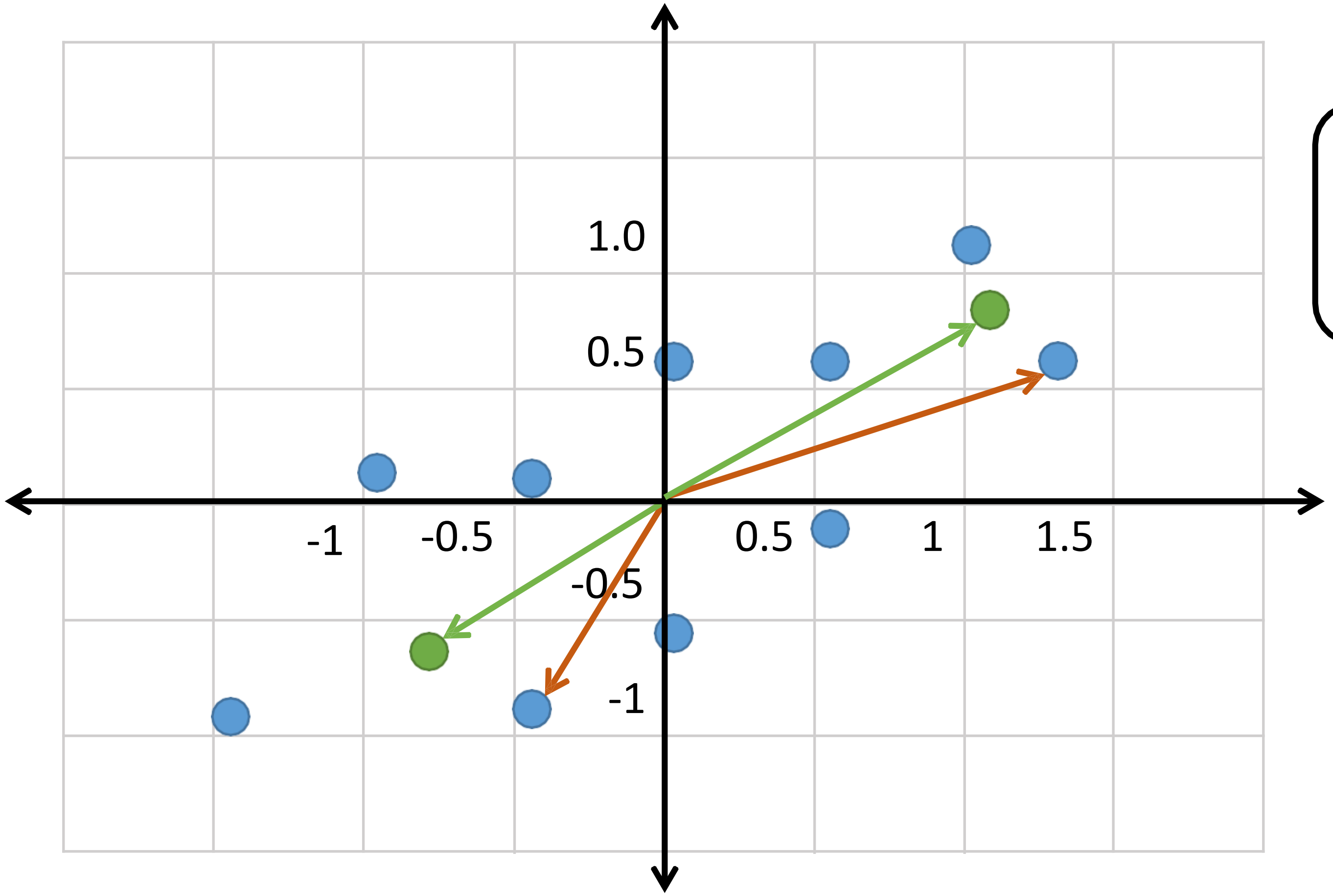
Step 1 – Center the Data

Step 2 – Create Variance-Covariance Matrix

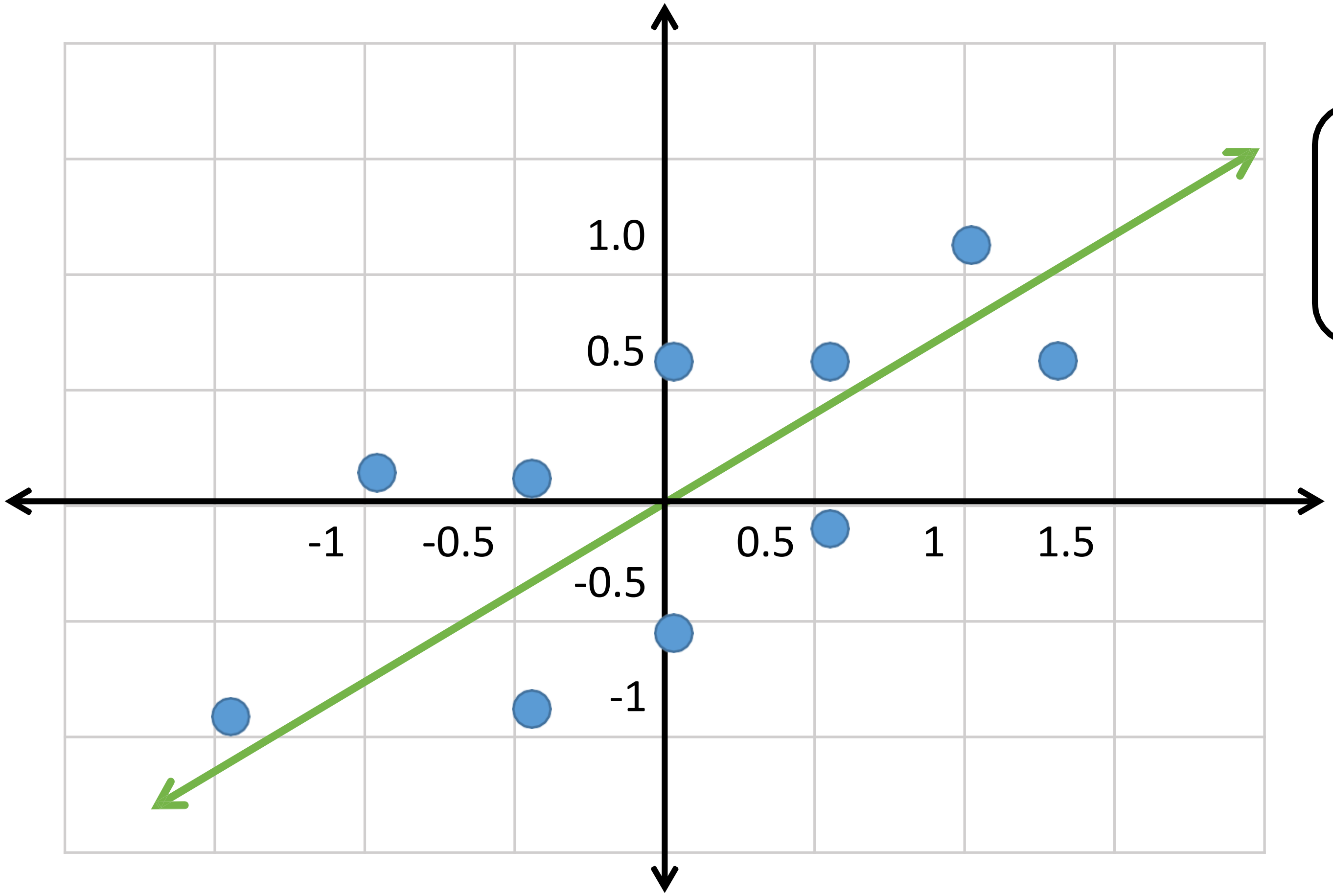
Step 3 – Project Vectors towards variance



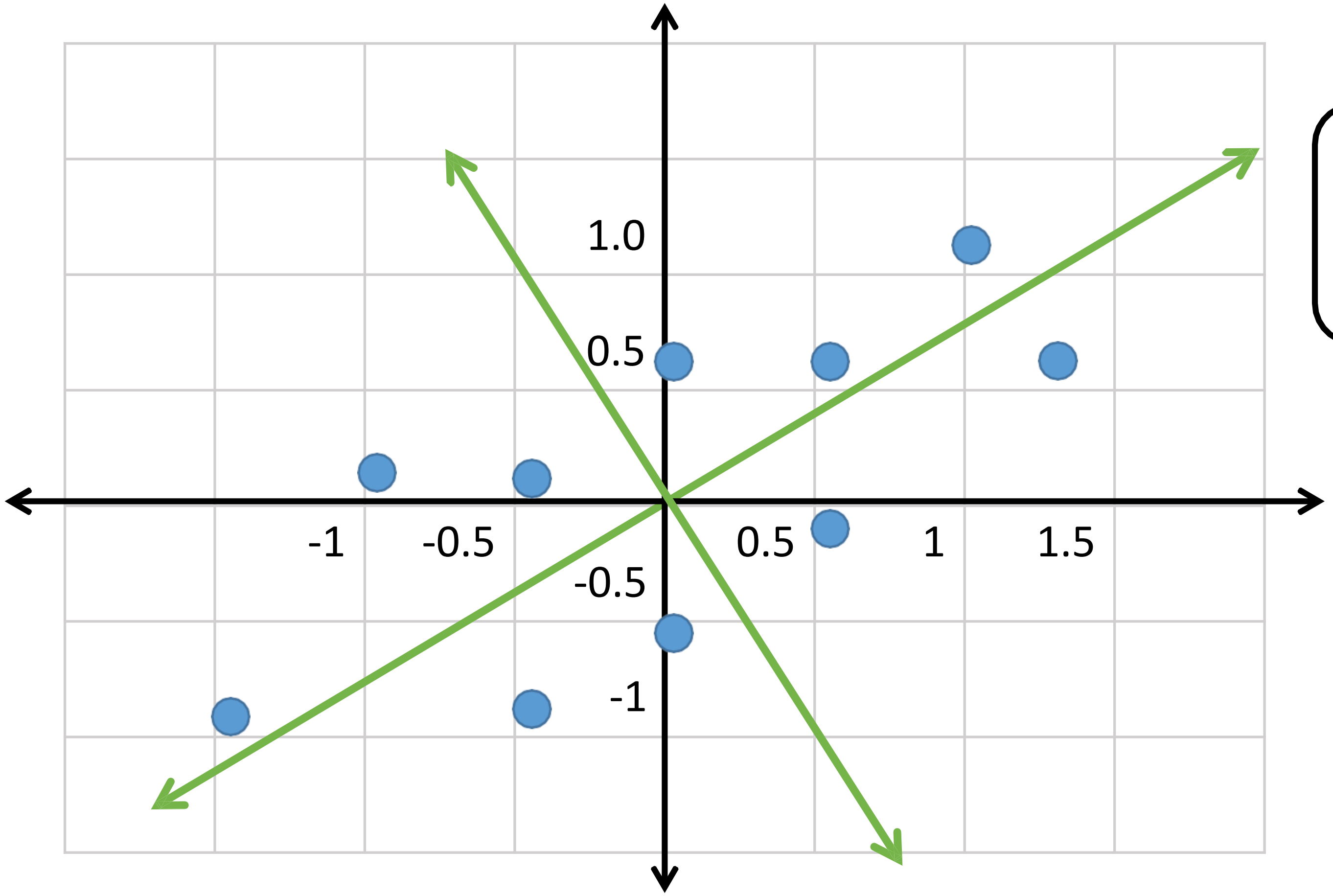
$$
\begin{array}{c|cc}
 & X1 & X2 \\
\hline
X1 & 0.76 & 0.42 \\
X2 & 0.42 & 0.48
\end{array}
$$



$$\begin{pmatrix} 0.76 & 0.42 \\ 0.42 & 0.48 \end{pmatrix} \begin{pmatrix} 1.275 \\ 0.545 \end{pmatrix} = \begin{pmatrix} 1.19 \\ 0.79 \end{pmatrix}$$



$$\begin{pmatrix} 0.76 & 0.42 \\ 0.42 & 0.48 \end{pmatrix}$$



$$\begin{pmatrix} 0.76 & 0.42 \\ 0.42 & 0.48 \end{pmatrix}$$

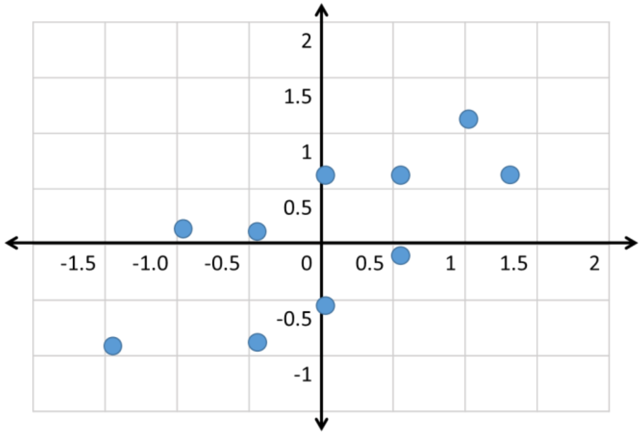
Steps in Creating the Principal Components

Step 1 – Center the Data

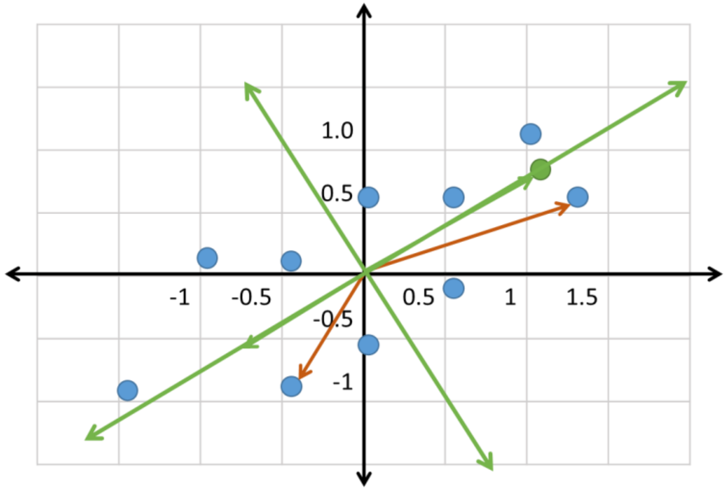
Step 2 – Create Variance-Covariance Matrix

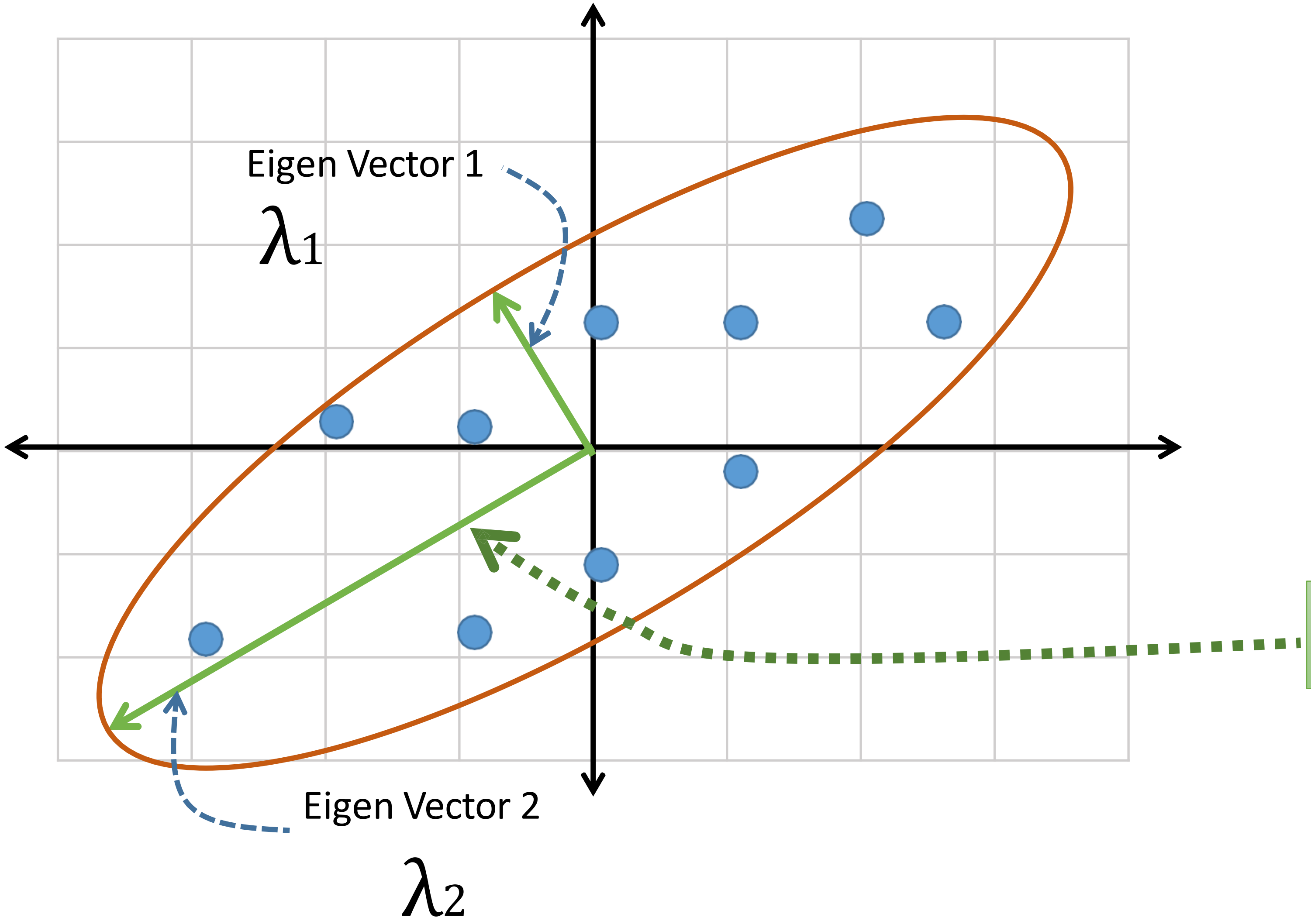
Step 3 – Project Vectors towards variance

Step 4 – Find Eigen Vectors and Eigen Values



$$
\begin{array}{c c}
 & \begin{array}{cc} X1 & X2 \end{array} \\
\begin{array}{c} X1 \\ X2 \end{array} & \begin{pmatrix} 0.76 & 0.42 \\ 0.42 & 0.48 \end{pmatrix}
\end{array}
$$





$$T.\vec{V} - \lambda.\vec{V} = 0$$

$$\lambda_2 > \lambda_1$$

Principal Component

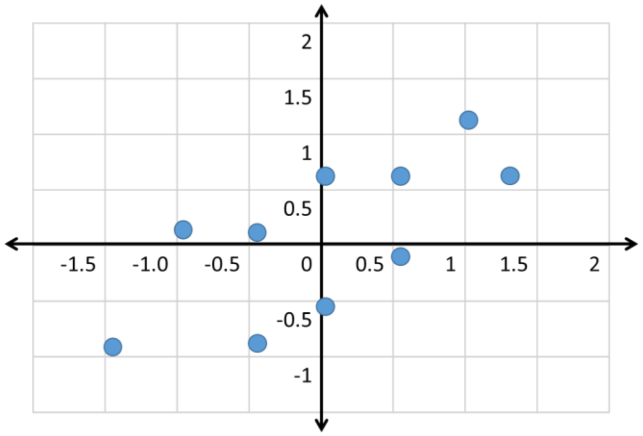
Steps in Creating the Principal Components

Step 1 – Center the Data

Step 2 – Create Variance-Covariance Matrix

Step 3 – Project Vectors towards variance

Step 4 – Find Eigen Vectors and Eigen Values



$$\begin{array}{cc} & \begin{array}{cc} X1 & X2 \end{array} \\ \begin{array}{c} X1 \\ X2 \end{array} & \begin{pmatrix} 0.76 & 0.42 \\ 0.42 & 0.48 \end{pmatrix} \end{array}$$