Data Reduction: The idea of data reduction is to obtain the compressed representation of the original data set.

Feature Extraction

$$X=\{\boxed{X_1, X_2, X_3, X_4,}\boxed{C}\} \longrightarrow Y=\{\boxed{Y_1, Y_2, Y_3, Y_4,}\boxed{C}\}$$

DATA REDUCTION IX

Feature Extraction

Data extraction methods discovers new set of reduced features from the given feature space with the objective of improving the end result.

- The two most popular feature extraction methods are:

1. Principal component analysis (PCA)

- PCA is a popular technique used to reduce high dimensional correlated features present in the data set to low dimensional space.
- Low dimensional space is achieved by transforming the variables to a new set of variables from the original feature space, which are known as the principal components.
- PCA has the characteristics that the essence of original features are present in the transformed data.i.e, there is no data loss.
- PCA is forward process.

Factor Analysis (FA)

- Factor analysis is also a dimensionality reduction technique that associates two or more features with one factor (also known as cause).
- FA is a backward process to discover the real causes for the occurrence of variables in the data set.

DATA REDUCTION X

Principal component analysis(PCA)

- The key strategy of PCA is to mine only those features from the given data set that contributes to maximum variance.
- To understand this, let’s consider Figure 25 and Figure 26 that showcase the plot between marks obtained by 8 students in two different subjects (represented by Marks1 and Marks2 respectively).

No variance in Marks1

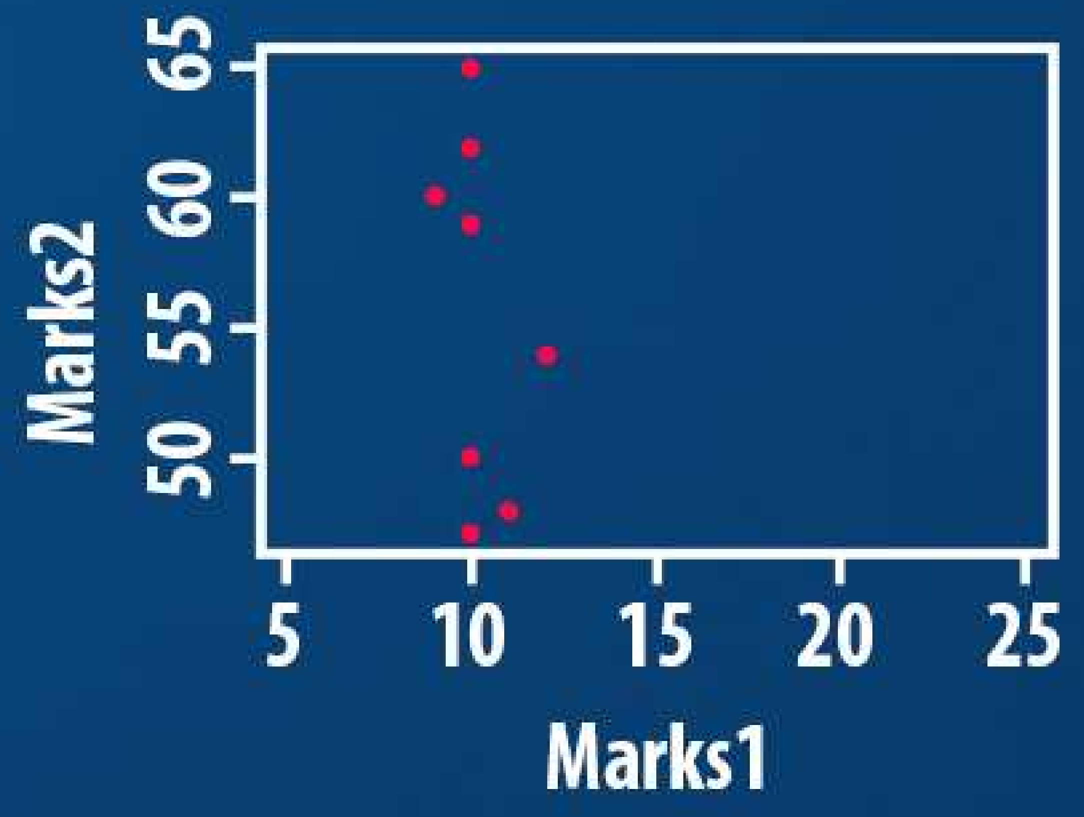


Figure 25: Marks 2 keeps more importance than marks1 for its higher variance

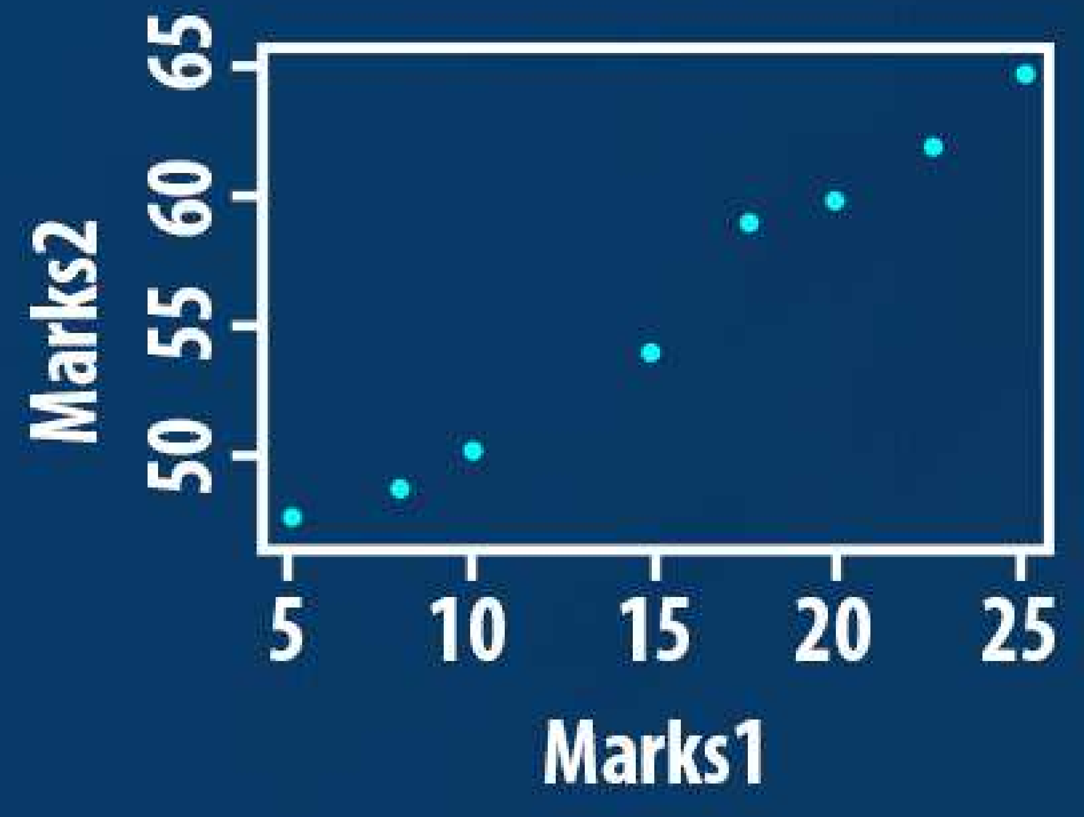


Figure 26: Example of highly correlated features

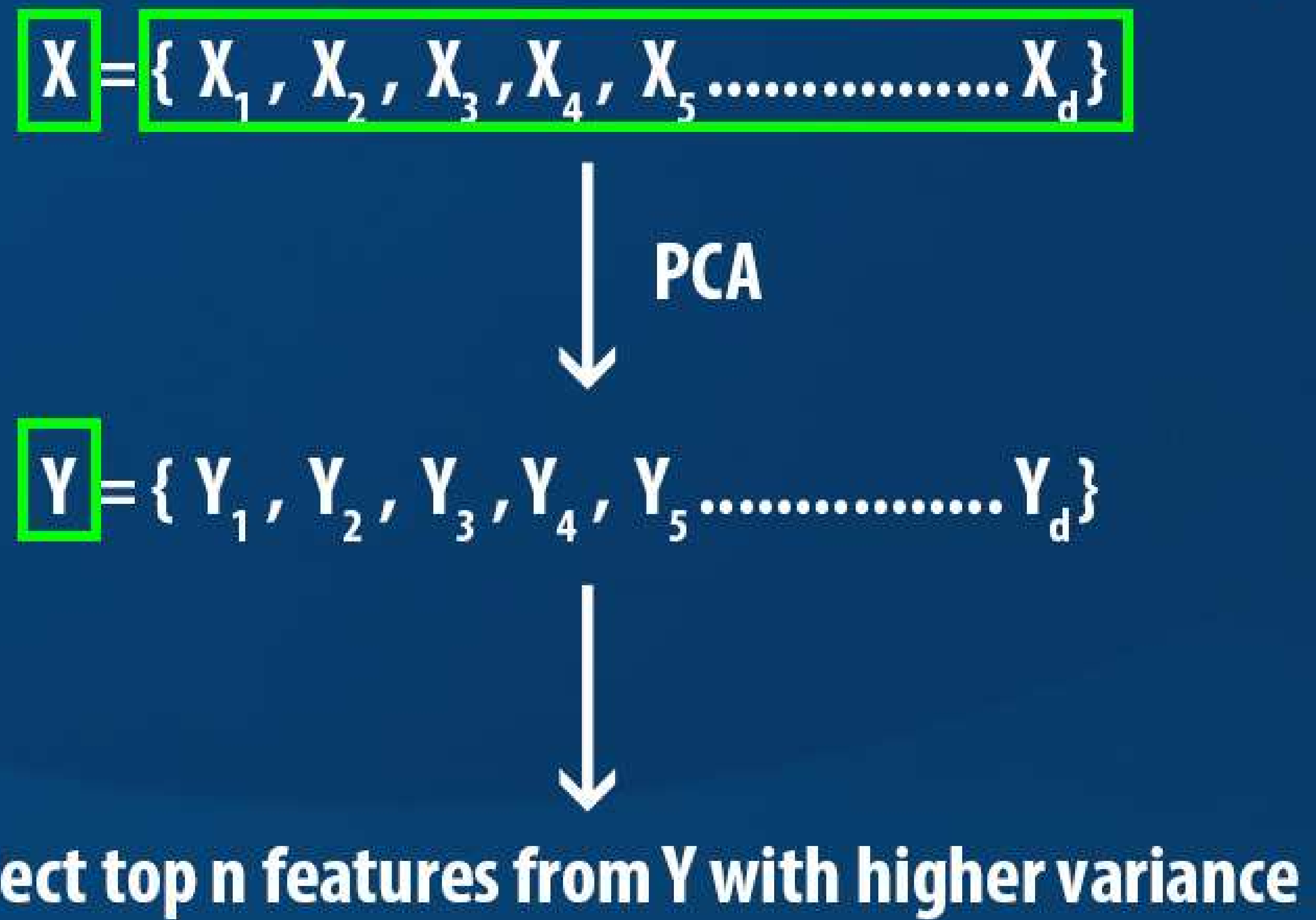
High variance in Marks2

- Consider Figure 26. Suppose our objective is to select only one feature between Marks1 and Marks2.
- It is difficult to decide since both features are giving same variance and are equally important.
- However, PCA can transform the orientation of data in such a way that we can select one feature out of two.

DATA REDUCTION XI

A simple Idea of PCA

- Figure 27 illustrates the simple idea of PCA. Where, set of features X consisting of d dimensions are given in the data set.
- After application of PCA, X is transformed to a new set Y. Where, Y contains d dimensions as given in original data set. However, these d dimensions are new and capture all information as in original data set.
- The features in Y are sorted in descending order of their variance and, top few features are selected.



$$\boxed{X} = \boxed{\{ X_1 , X_2 , X_3 , X_4 , X_5 \ldots\ldots\ldots\ldots\ldots X_d \}}$$

$$\downarrow \text{PCA}$$

$$\boxed{Y} = \{ Y_1 , Y_2 , Y_3 , Y_4 , Y_5 \ldots\ldots\ldots\ldots\ldots Y_d \}$$

$$\downarrow$$

Select top n features from Y with higher variance

Figure 27: PCA process

PCA

$$\begin{tikzcd}
X = \{X_1, X_2, X_3, X_4\} \arrow[d, "PCA"] \\
Y = \{Y_1, Y_2, Y_3, Y_4\} \arrow[d] \\
\boxed{\boxed{Y}_3, \boxed{Y}_2, Y_1, Y_4} \arrow[d, "PCA"] \\
Y_3, Y_2
\end{tikzcd}$$

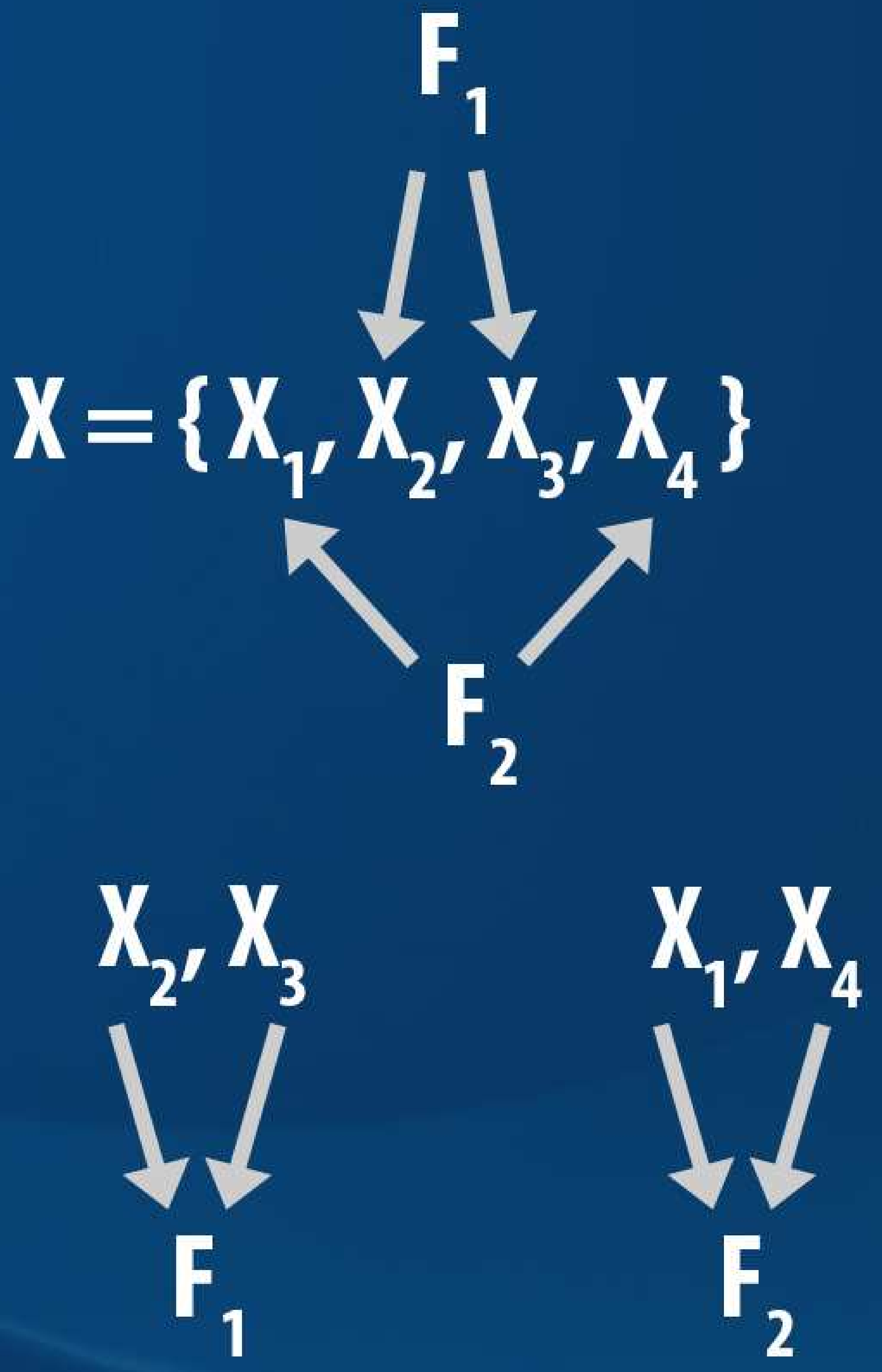
DATA REDUCTION XII

Factor analysis

- A Factor Analysis approaches data reduction in a fundamentally different way. It is a model of the measurement of a latent variable. This latent variable cannot be directly measured with a single variable (think: intelligence, social anxiety, soil health). Instead, it is seen through the relationships it causes in a set of X variables.
- For example, we may not be able to directly measure social anxiety. But we can measure whether social anxiety is high or low with a set of variables like "I am uncomfortable in large groups" and " I get nervous talking with strangers."
  - People with high social anxiety will give similar high responses to these variables because of their high social anxiety.
  - People with low social anxiety will give similar low responses to these variables because of their low social anxiety.
- Key characteristics of Factor analysis are:

1. It identifies correlation between and among variables to bind them into one underlying factor.
2. When factors can be interpreted, new insights are possible.

DATA REDUCTION



F_1

$X = \{ X_1, X_2, X_3, X_4 \}$

F_2

X_2, X_3

X_1, X_4

F_1

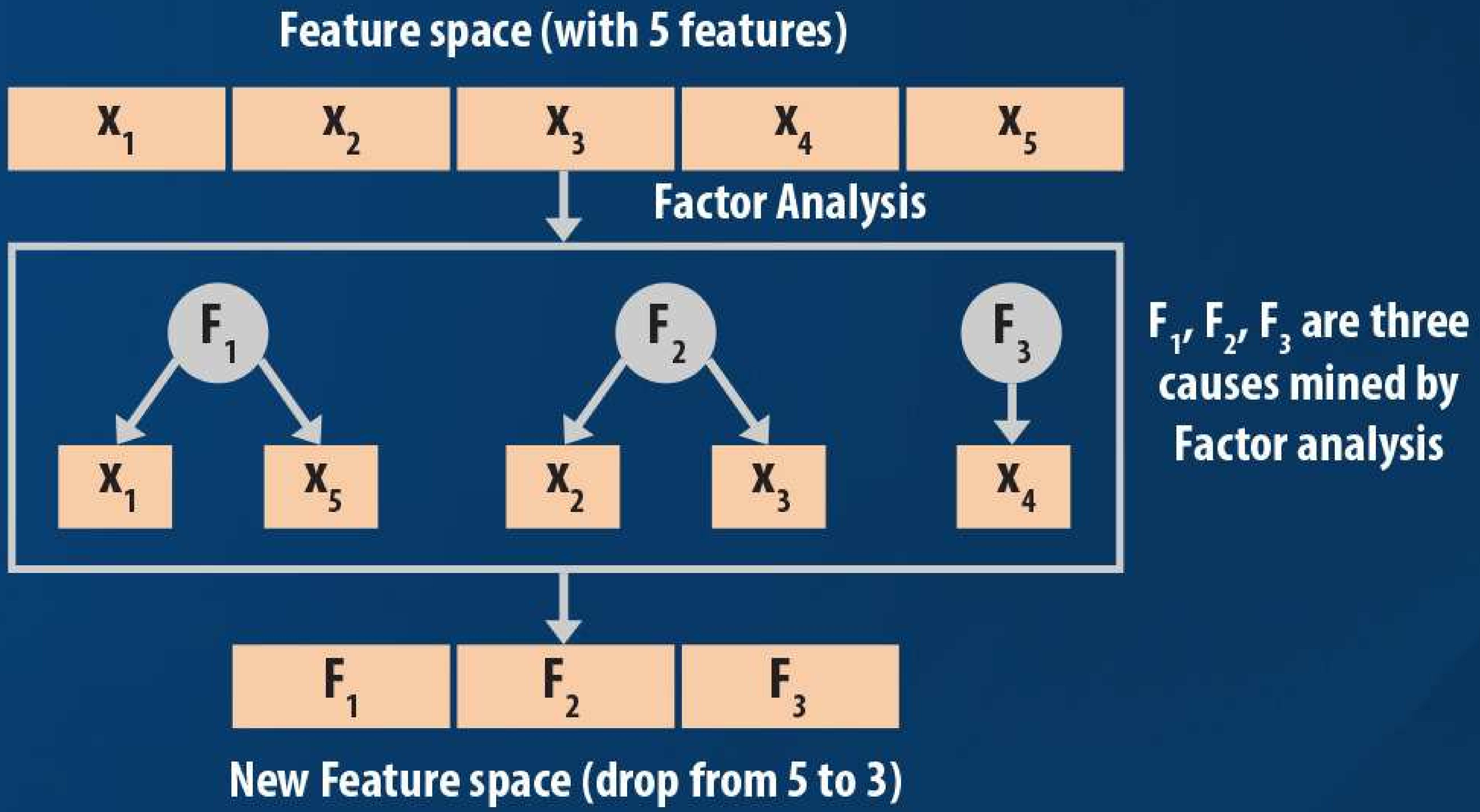
F_2

DATA REDUCTION XIII

Factor Analysis - Simple Idea

Consider Figure 28 where, feature space X is given with five features namely, x_1, x_2, x_3, x_4 and x_5. After factor analysis technique applied on X, we receive that features x_1, x_5 are clubbed in one cause, i.e., f_1. Whereas, f_2 is revealed as a common cause for features x_2 and x_3.

The feature x_4 is discovered as an independent variable.



Feature space (with 5 features)
x_1
x_2
x_3
x_4
x_5
Factor Analysis
F_1
F_2
F_3
x_1
x_5
x_2
x_3
x_4
F_1, F_2, F_3 are three causes mined by Factor analysis
F_1
F_2
F_3
New Feature space (drop from 5 to 3)

Figure 28: Hypothetical example of Factor analysis

In this example, we started with five features but using factor analysis technique, we reduce the features by three.