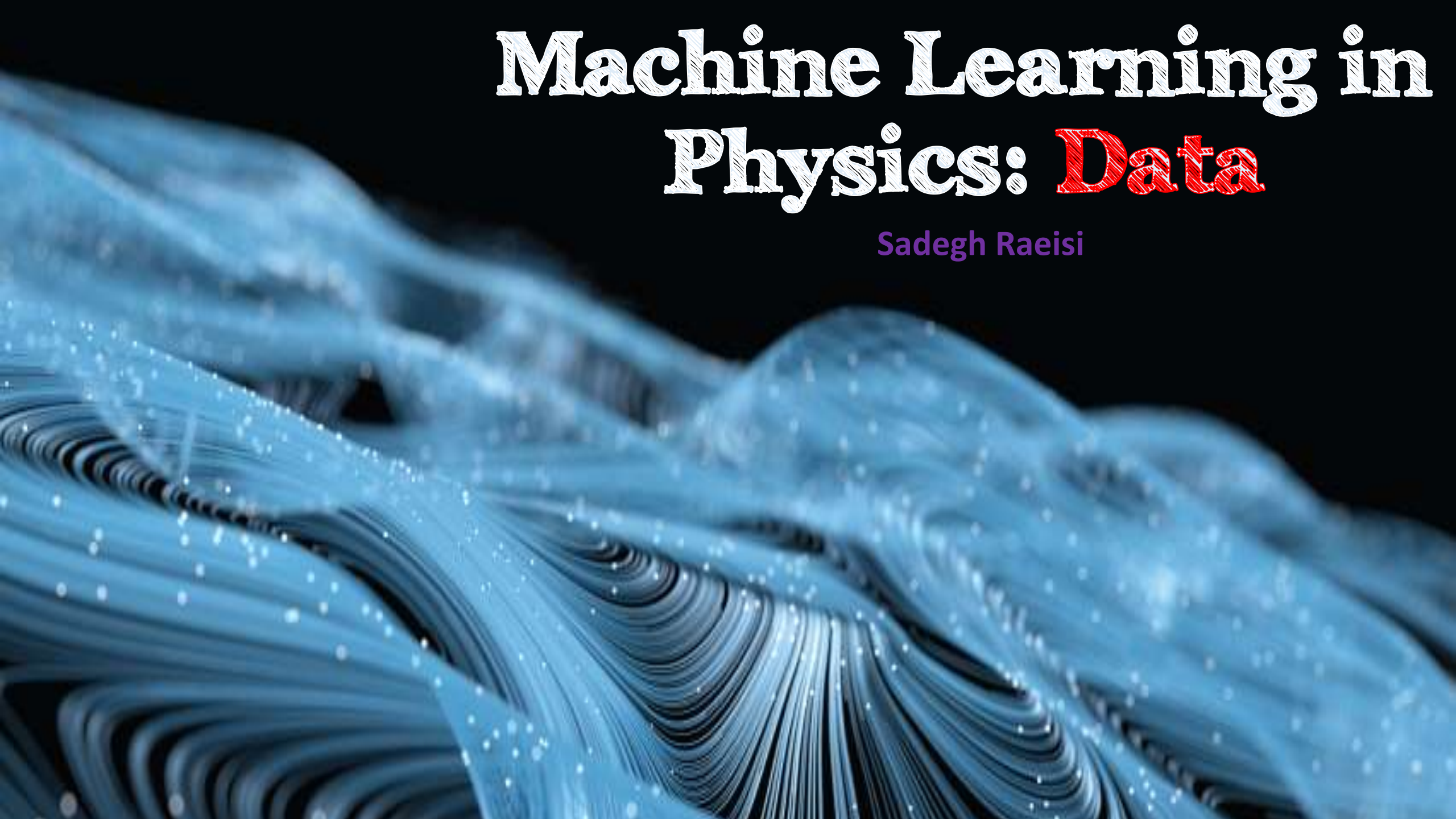
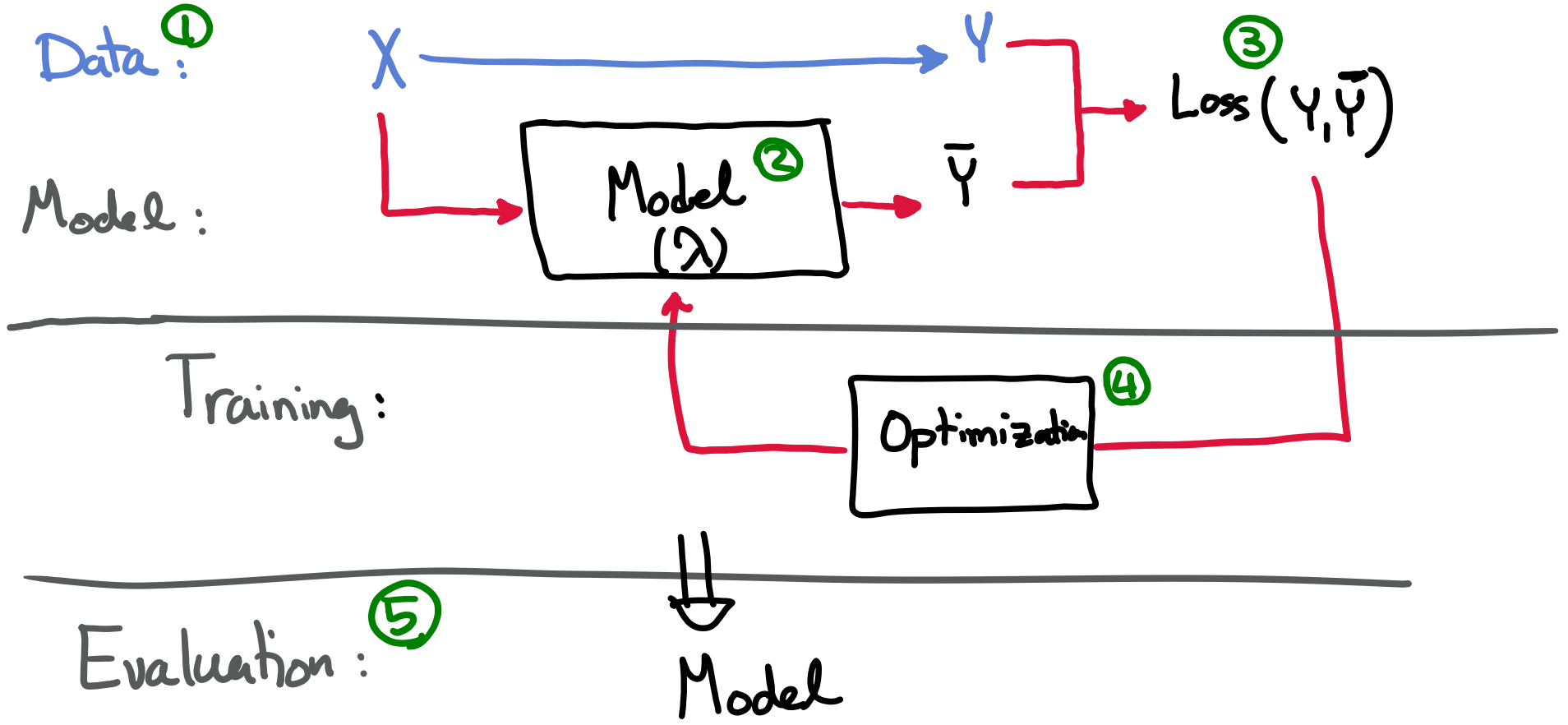


Machine Learning in Physics: Data

Sadegh Raeisi

Supervised: Ingredients

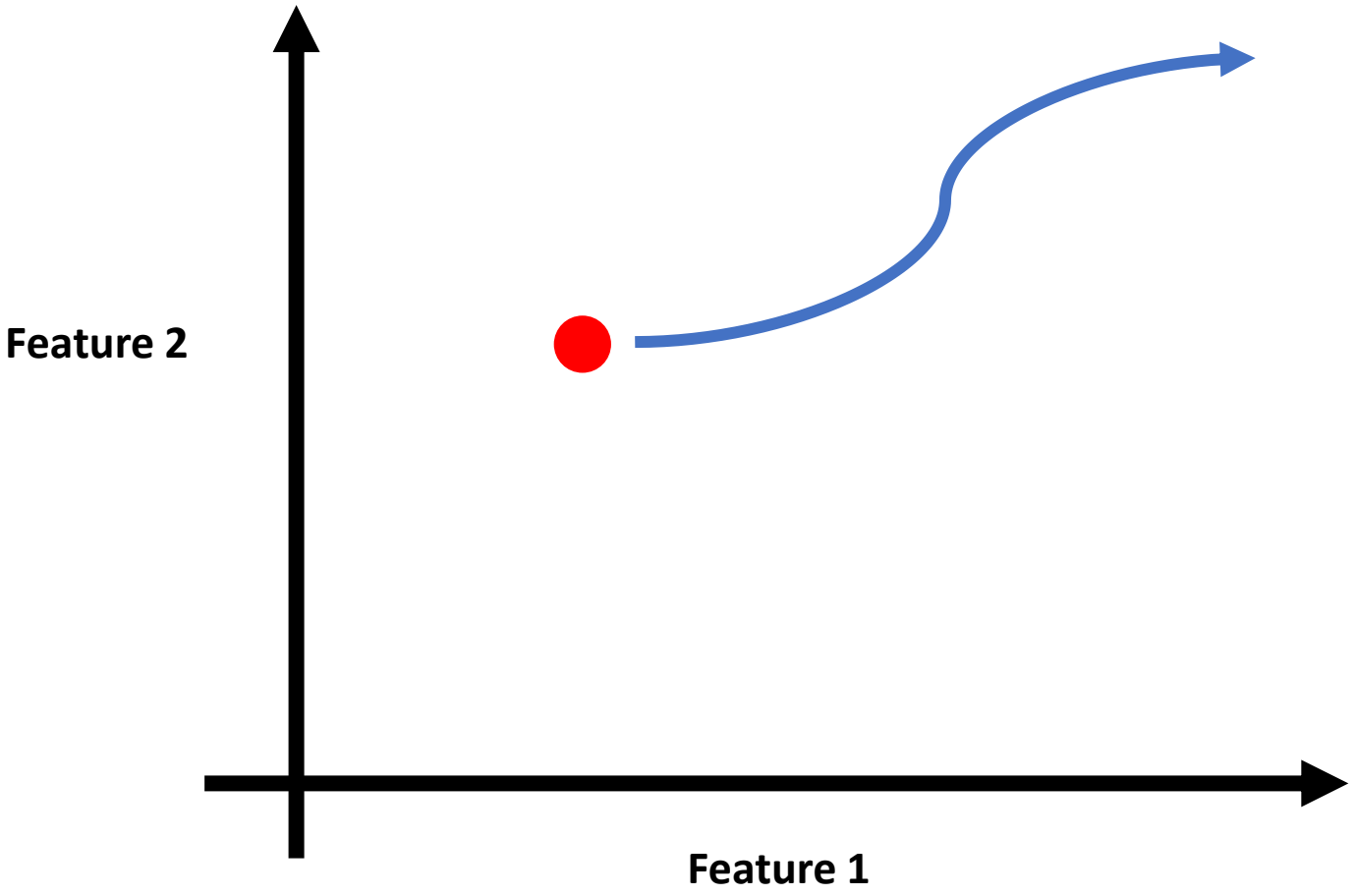


Outline


Notation
Types of Data
Encoding
Transformations and preprocessing

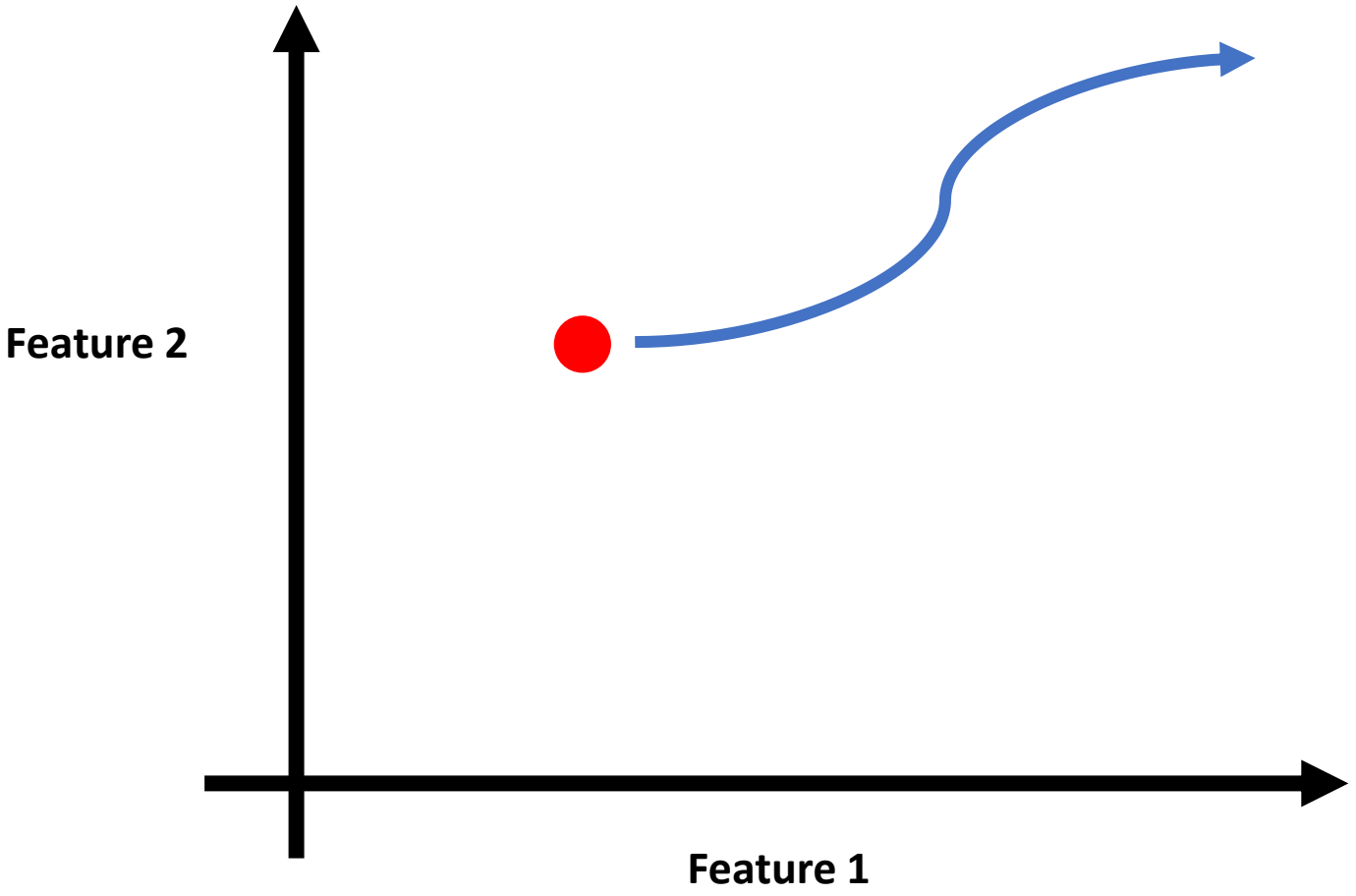
Notation

One sample



Representation of the sample:
(Feature 1, Feature 2)

One sample

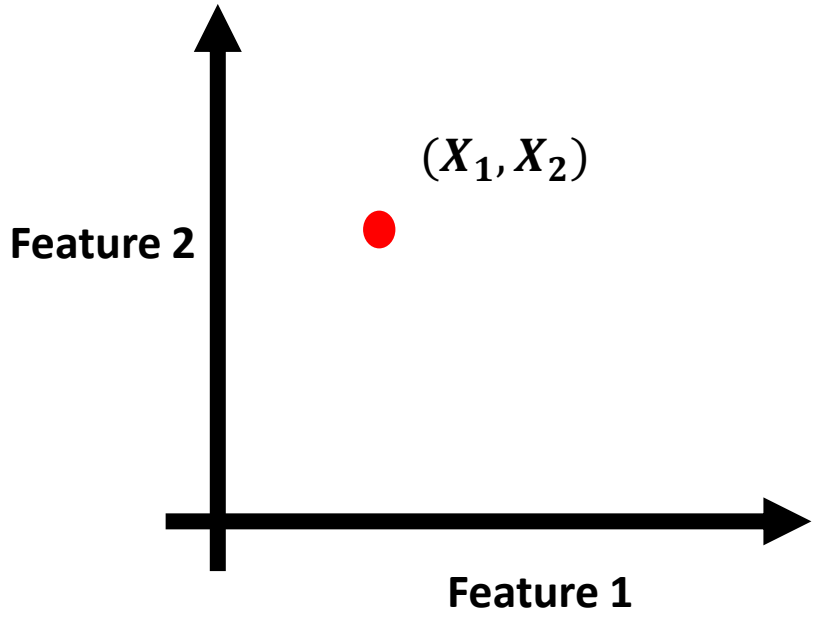


Representation of the sample:
(Feature 1, Feature 2)

Feature 1 => X_1
Feature 2 => X_2

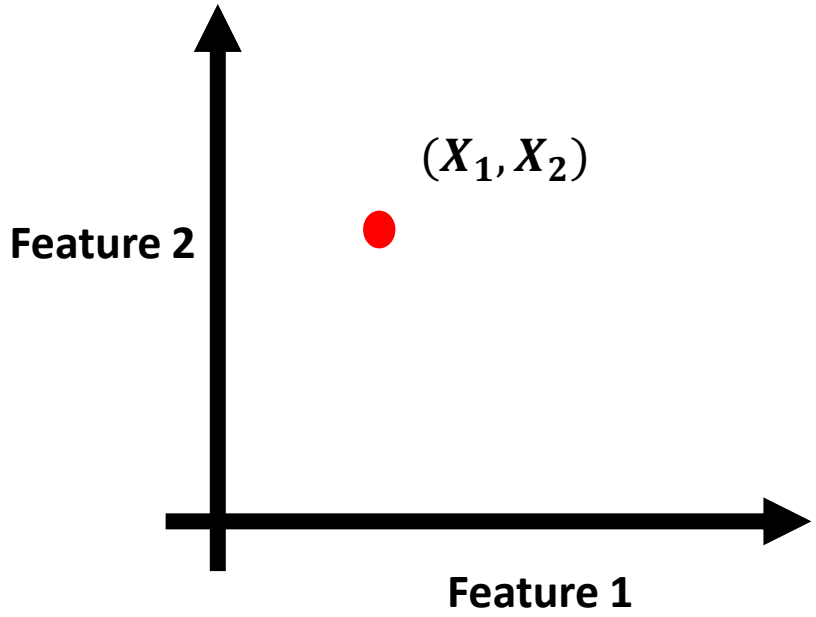
One sample

(X_1, X_2)



One sample

$$\vec{X} = (X_1, X_2, \cdots X_{nf})$$



nf: **Dimension of the space**

Collection of samples

Sample 1: $\vec{X}^1 = (X_1^1, X_2^1, \cdots X_{nf}^1)$

Sample 2: $\vec{X}^2 = (X_1^2, X_2^2, \cdots X_{nf}^2)$

Sample ns: $\vec{X}^{ns} = (X_1^{ns}, X_2^{ns}, \cdots X_{nf}^{ns})$

$$X = \begin{pmatrix} X_1^1, X_2^1, \cdots X_{nf}^1 \\ X_1^2, X_2^2, \cdots X_{nf}^2 \\ \cdots \\ X_1^{ns}, X_2^{ns}, \cdots X_{nf}^{ns} \end{pmatrix}$$

Data : Features

$$\boldsymbol{X}= \begin{pmatrix} X_1^1, X_2^1, \cdots X_{nf}^1 \\ X_1^2, X_2^2, \cdots X_{nf}^2 \\ \cdots \\ X_1^{ns}, X_2^{ns}, \cdots X_{nf}^{ns} \end{pmatrix}$$

nf: **Dimension of the space**
ns **: Number of Samples**

Data : Labels

$$X = \begin{pmatrix} X_1^1, X_2^1, \cdots X_{nf}^1 \\ X_1^2, X_2^2, \cdots X_{nf}^2 \\ \cdots \\ X_1^{ns}, X_2^{ns}, \cdots X_{nf}^{ns} \end{pmatrix} \rightarrow \begin{pmatrix} Y^1 \\ Y^2 \\ \cdots \\ Y^{ns} \end{pmatrix}$$

nf: **Dimension of the space**
ns **: Number of Samples**

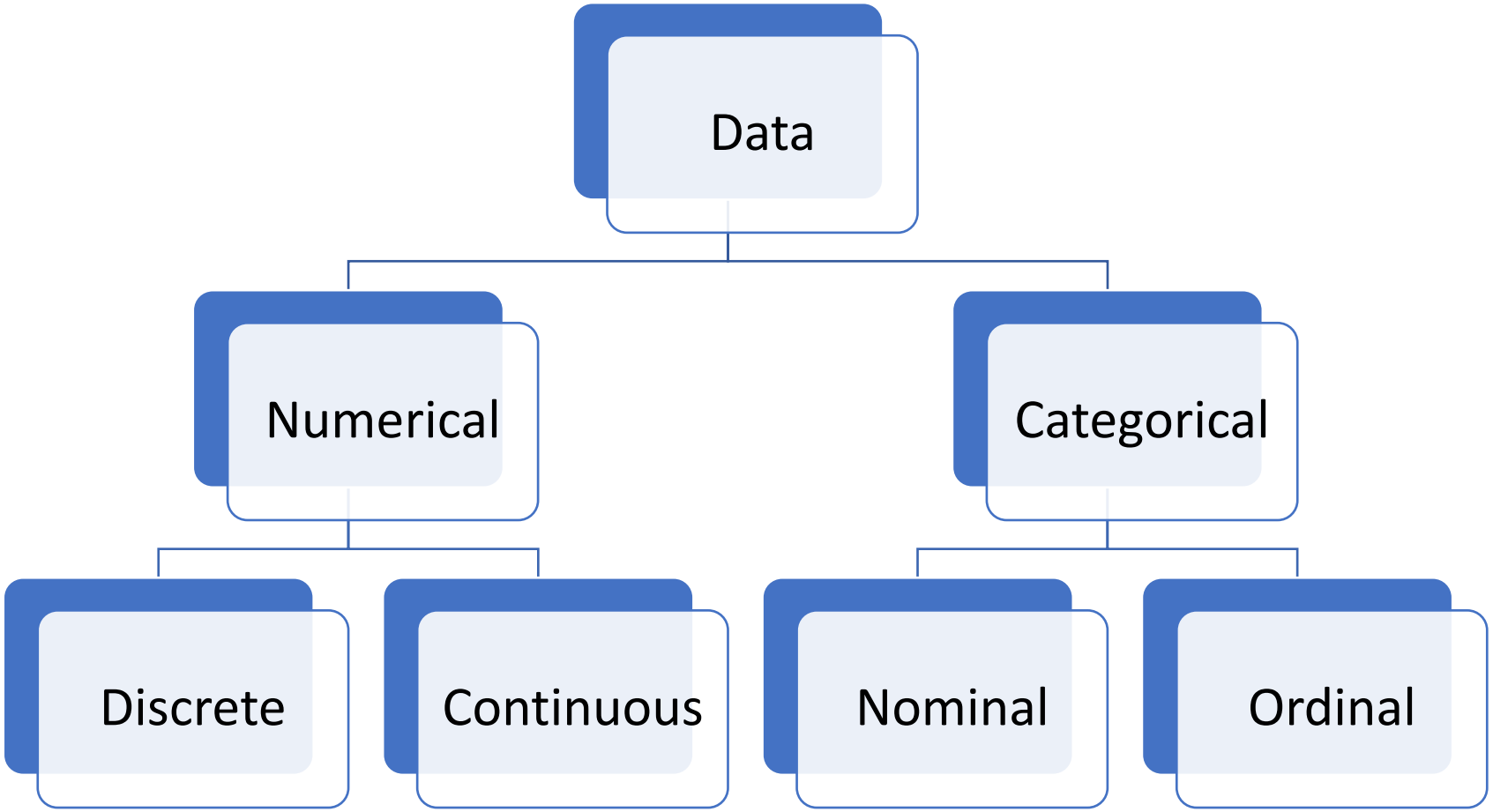
Data : Labels

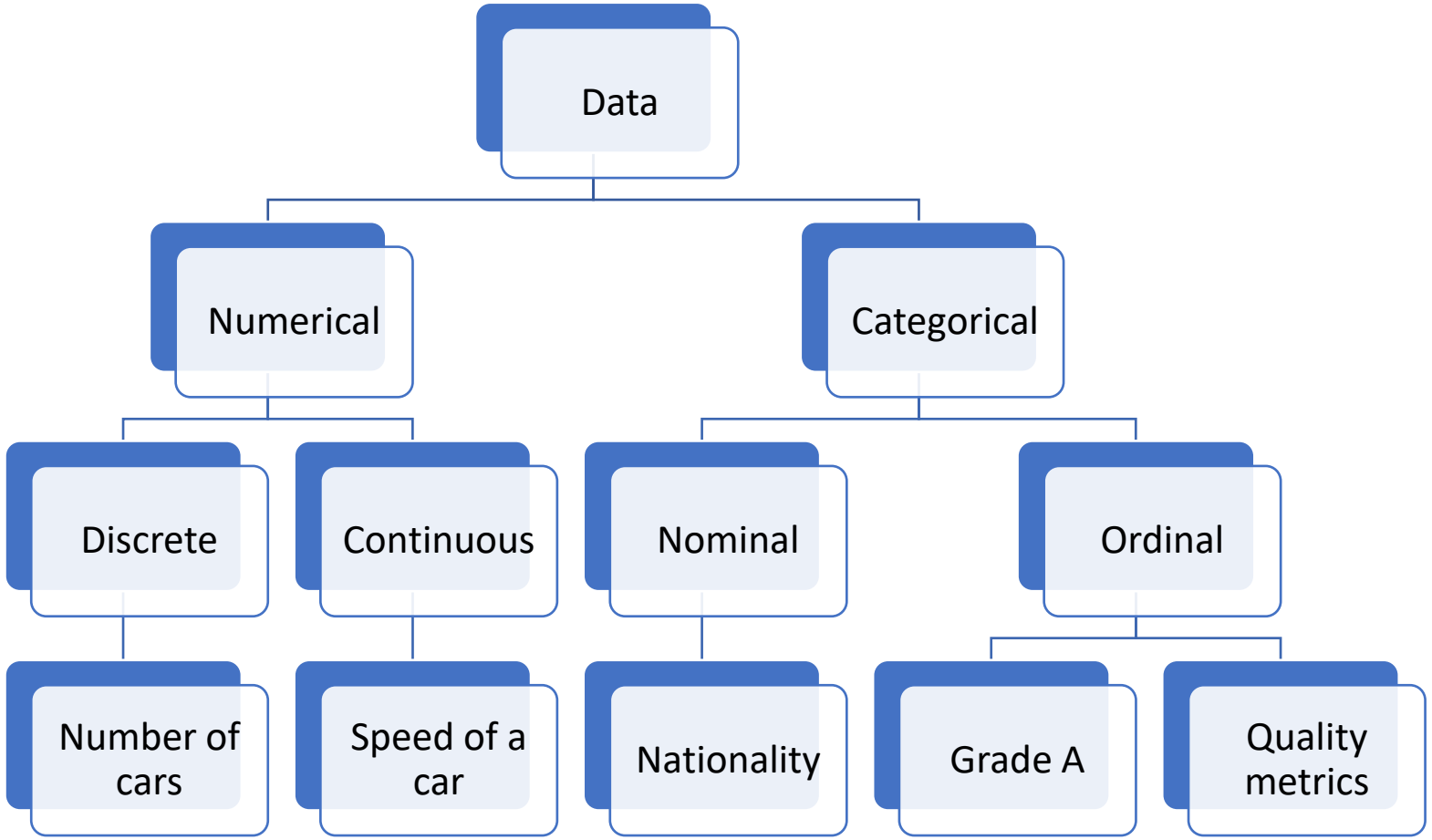
$$\boldsymbol{X}=\begin{pmatrix} X_1^1, X_2^1, \cdots X_{nf}^1 \\ X_1^2, X_2^2, \cdots X_{nf}^2 \\ \cdots \\ X_1^{ns}, X_2^{ns}, \cdots X_{nf}^{ns} \end{pmatrix} \qquad \boldsymbol{Y}=\begin{pmatrix} Y^1 \\ Y^2 \\ \cdots \\ Y^{ns} \end{pmatrix}$$

nf: Dimension of the space
ns : Number of Samples

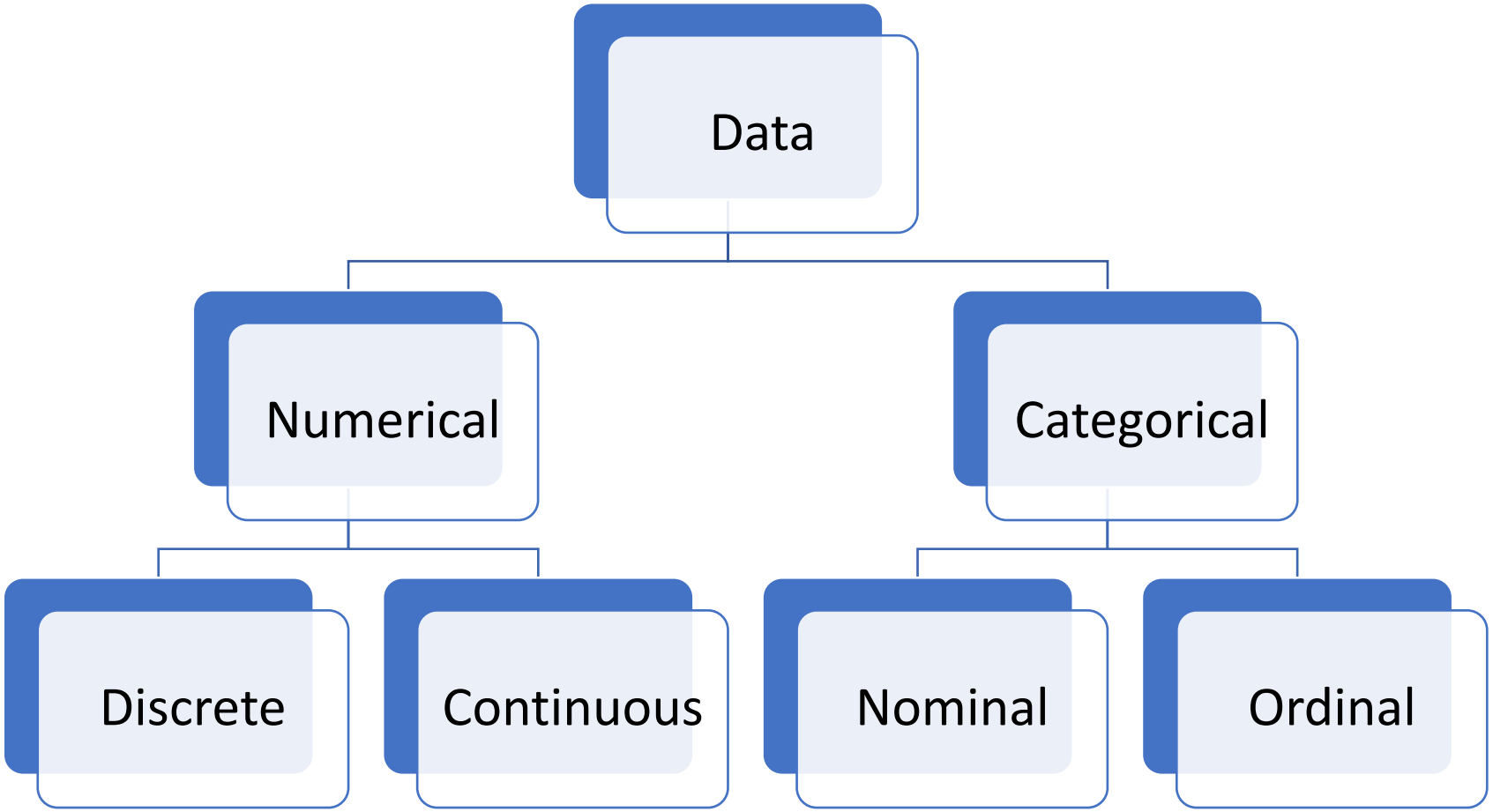
Types of Data

There are more than one categorization …


Data
Numerical
Categorical
Discrete
Continuous
Nominal
Ordinal


Data
Numerical
Categorical
Discrete
Continuous
Nominal
Ordinal
Number of cars
Speed of a car
Nationality
Grade A
Quality metrics

Encoding of Data


Data
Numerical
Categorical
Discrete
Continuous
Nominal
Ordinal

Exercise

How would you encode categorical types of data?

Processing the Data

Missing data

Missing data

$$\boldsymbol{X}= \begin{pmatrix} X_1^1, X_2^1, \cdots X_{nf}^1 \\ X_1^2, \boldsymbol{NA}, \cdots X_{nf}^2 \\ \cdots \\ X_1^{ns}, X_2^{ns}, \cdots X_{nf}^{ns} \end{pmatrix}$$

Missing data

$$\boldsymbol{X}= \begin{pmatrix} X_1^1, X_2^1, \cdots X_{nf}^1 \\ X_1^2, \boldsymbol{NA}, \cdots X_{nf}^2 \\ \cdots \\ X_1^{ns}, X_2^{ns}, \cdots X_{nf}^{ns} \end{pmatrix}$$

What can we do about the missing data?

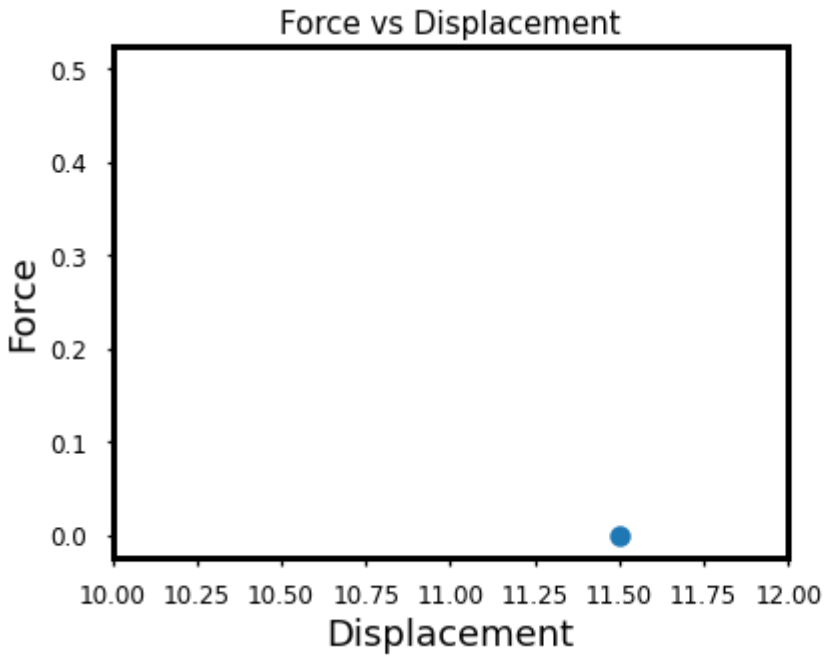
Missing data

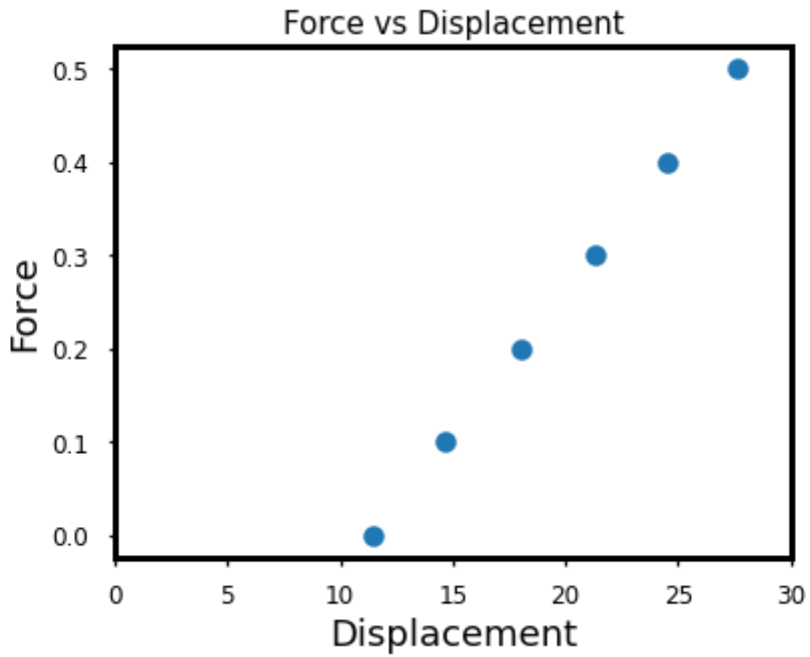
$$X = \begin{pmatrix} X_1^1, X_2^1, \cdots X_{nf}^1 \\ X_1^2, NA, \cdots X_{nf}^2 \\ \cdots \\ X_1^{ns}, X_2^{ns}, \cdots X_{nf}^{ns} \end{pmatrix}$$

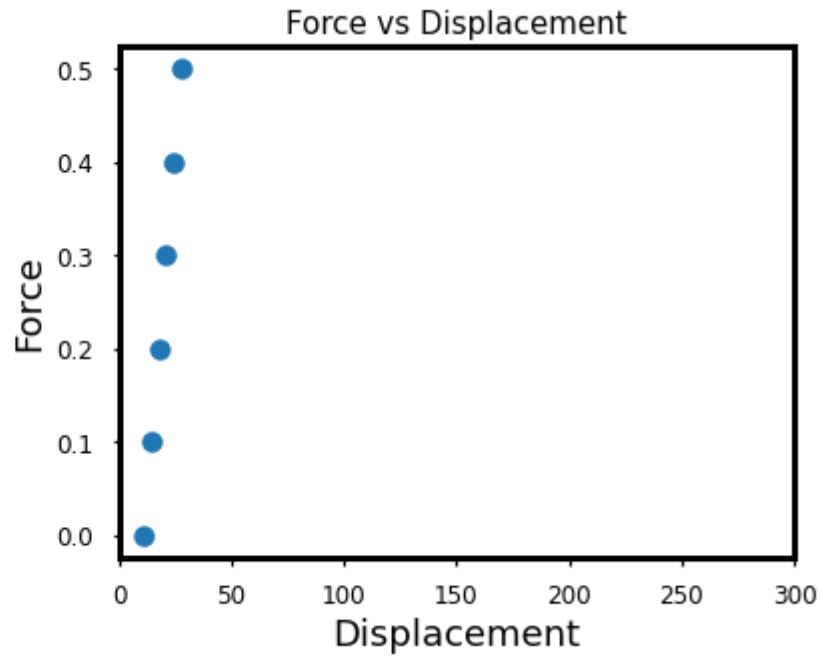
1. **Get rid of the sample**
   - **What's the disadvantage?**
2. **Get rid of the feature**
   - **When does it make sense to do this?**
3. **Assign a value to it?**
   - **How?**

Scale of the data

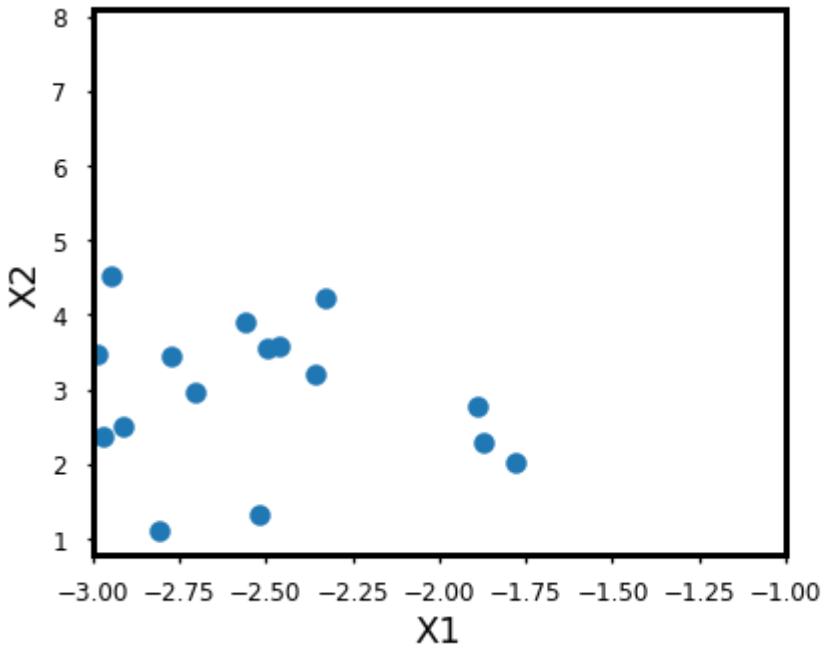
Scale of the data for regression

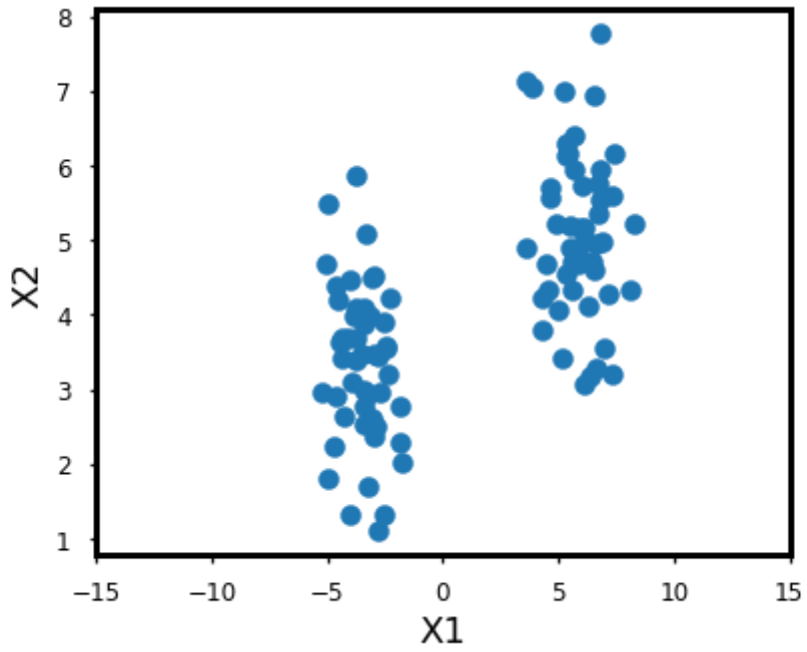


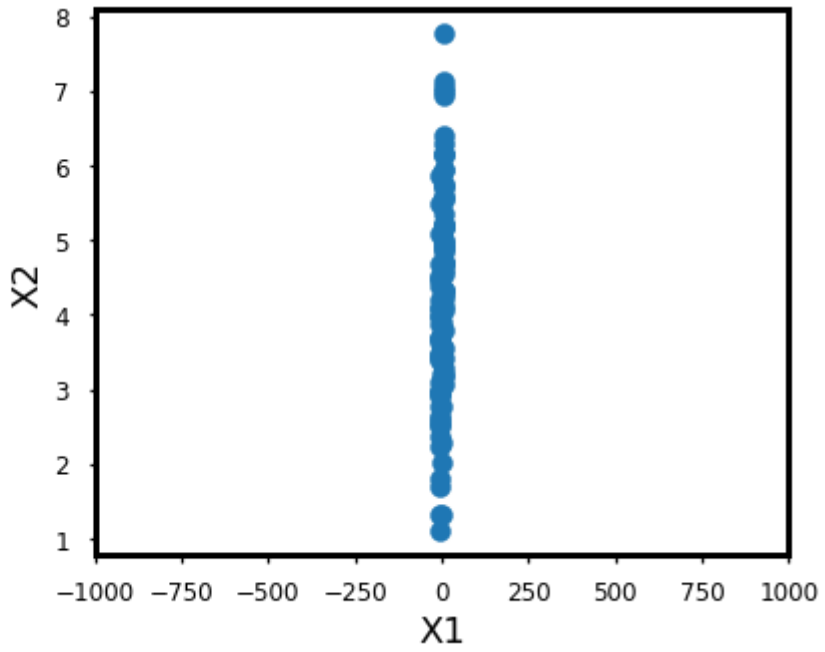




Scale of the data: Clustering

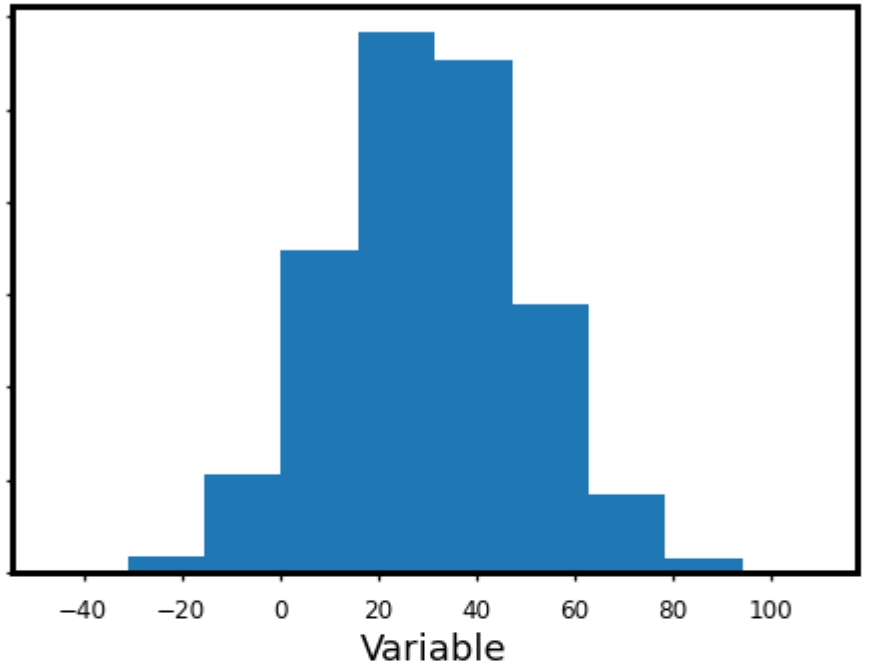


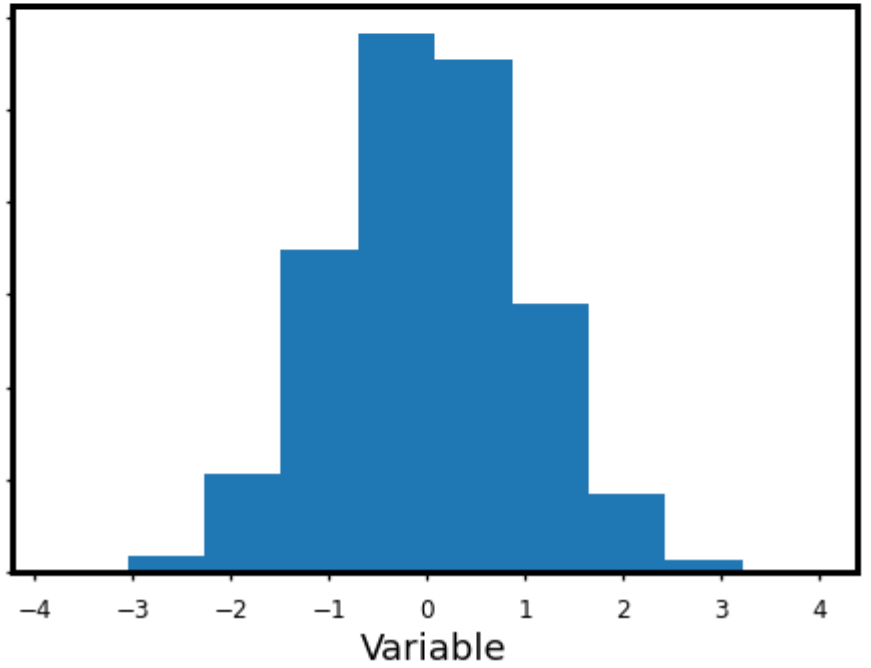




Re-scaling the data

1. **Physical Scale**
2. **Scale on which it is changing**





Questions

- What other ways are there to scale the data? What are they good for?
- Does scaling the data affect optimization/training models?

Data Reduction

Why

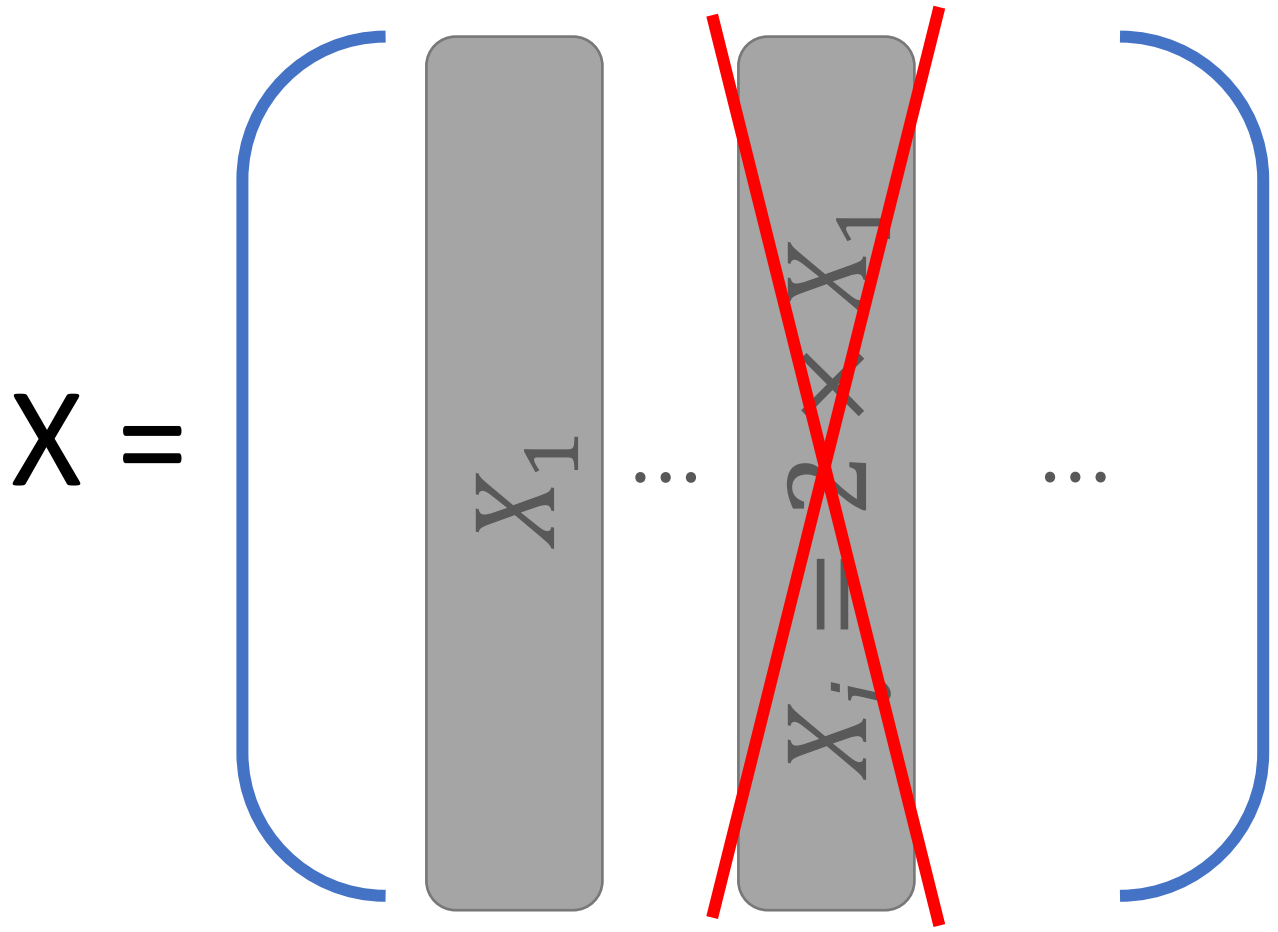
- Visualization
- Computational efficiency
- Curse of dimensionality

Techniques

- Feature selection

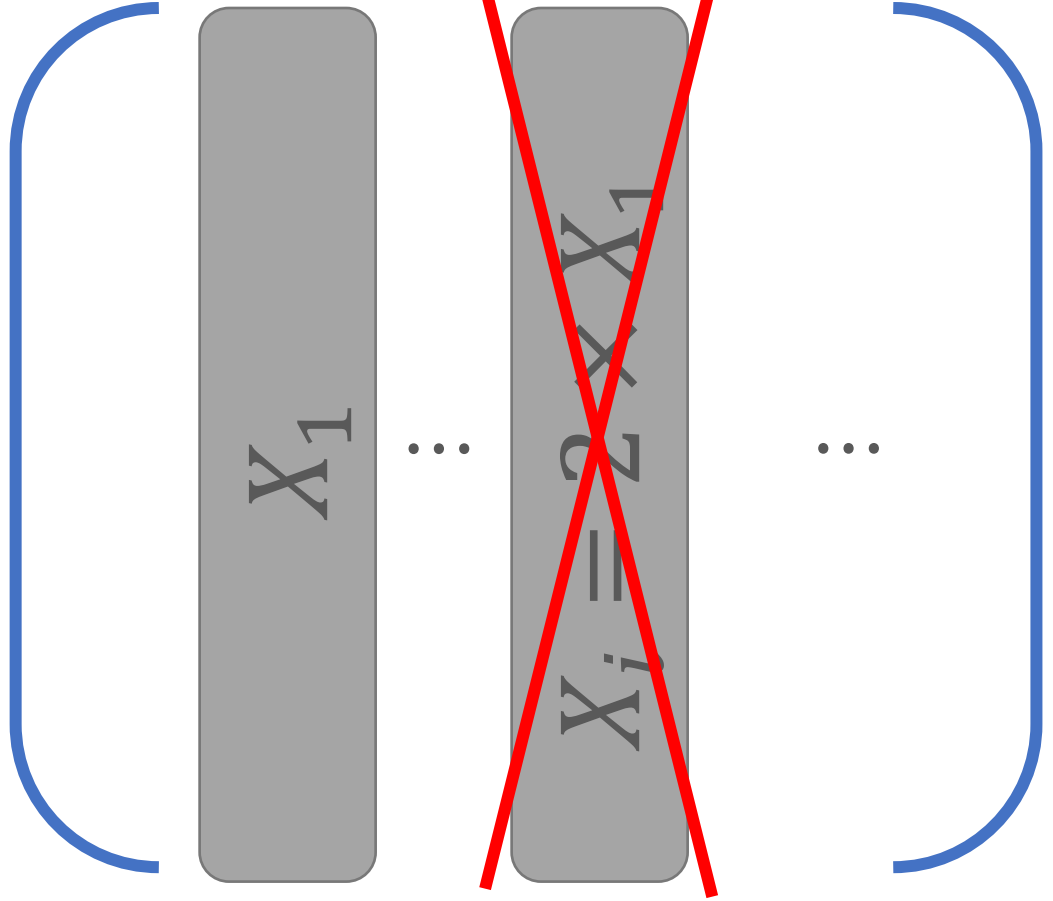
- Feature transformation
  - Linear transformation
  - Manifold learning

Feature Selection



When?

Feature Selection

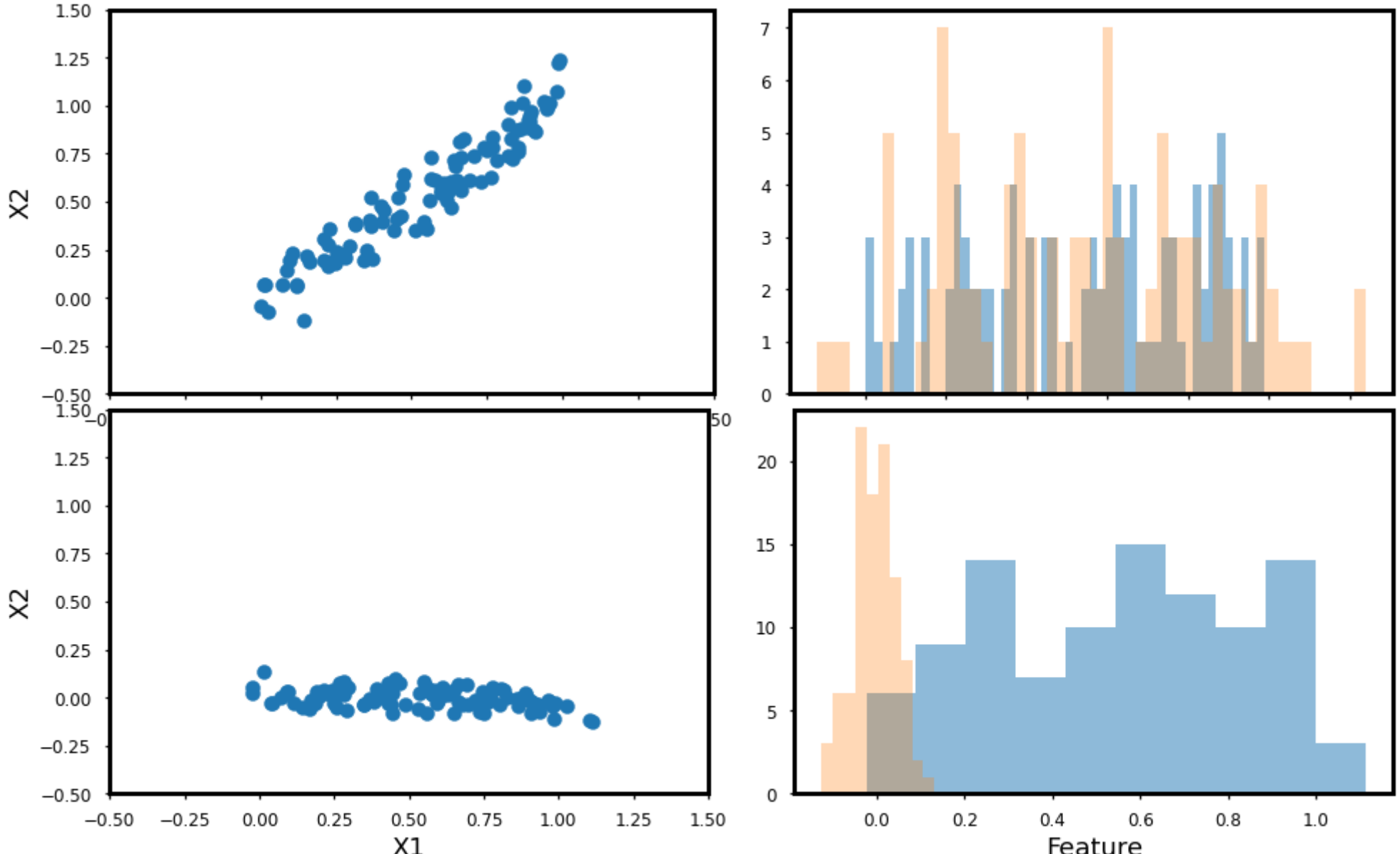


How to:

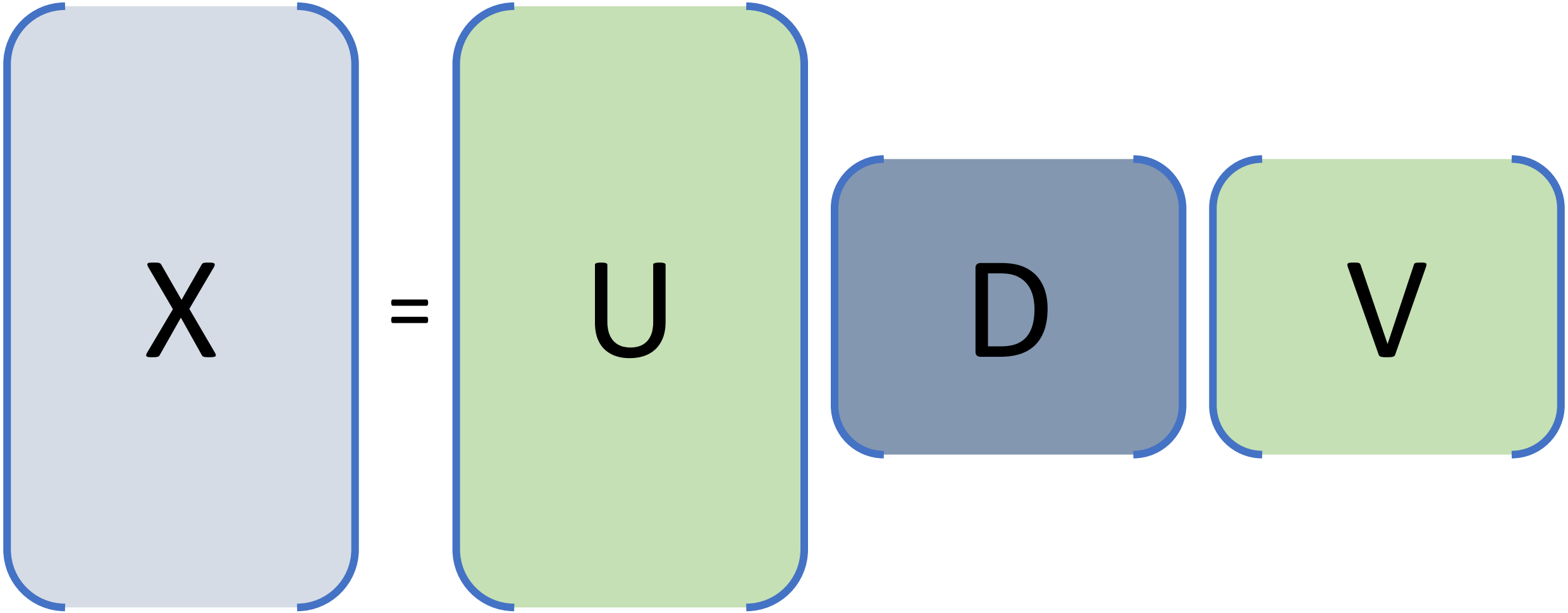
- Un-Supervised
  - Variance
  - Importance/significance

- supervised

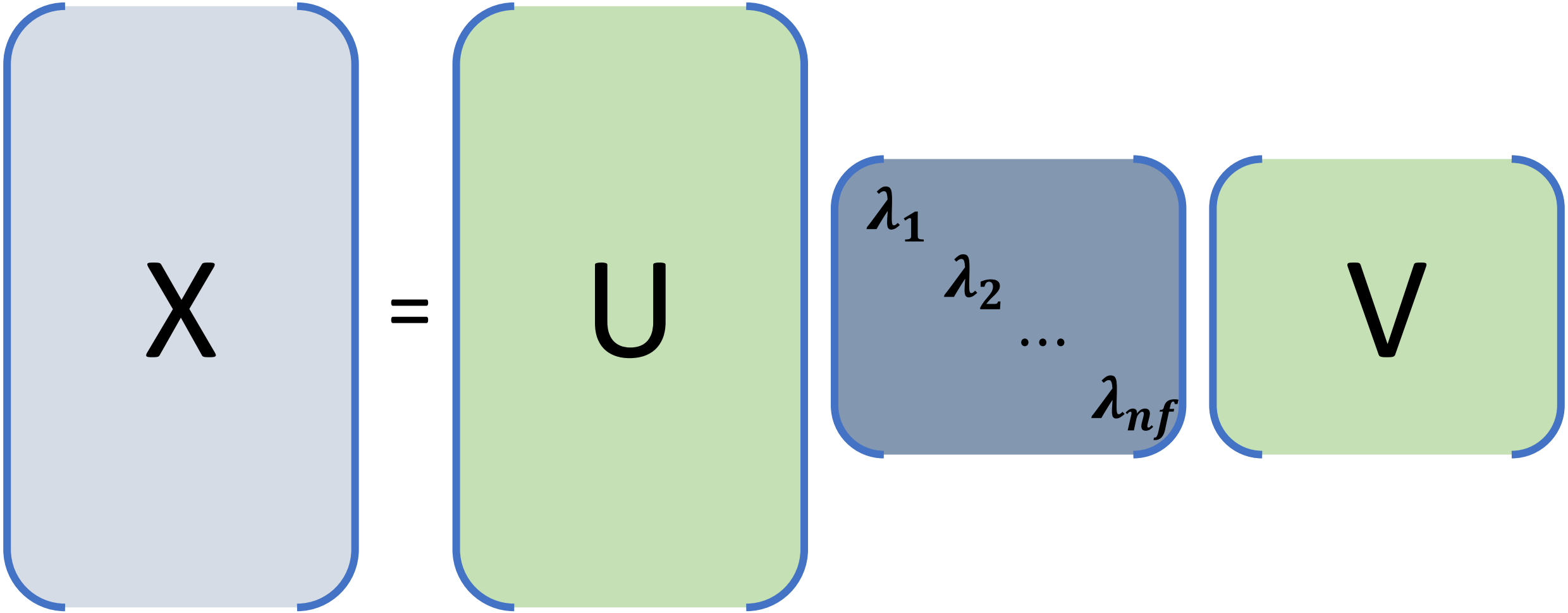
Feature Transformation



Feature Transformation: PCA



Feature Transformation: PCA



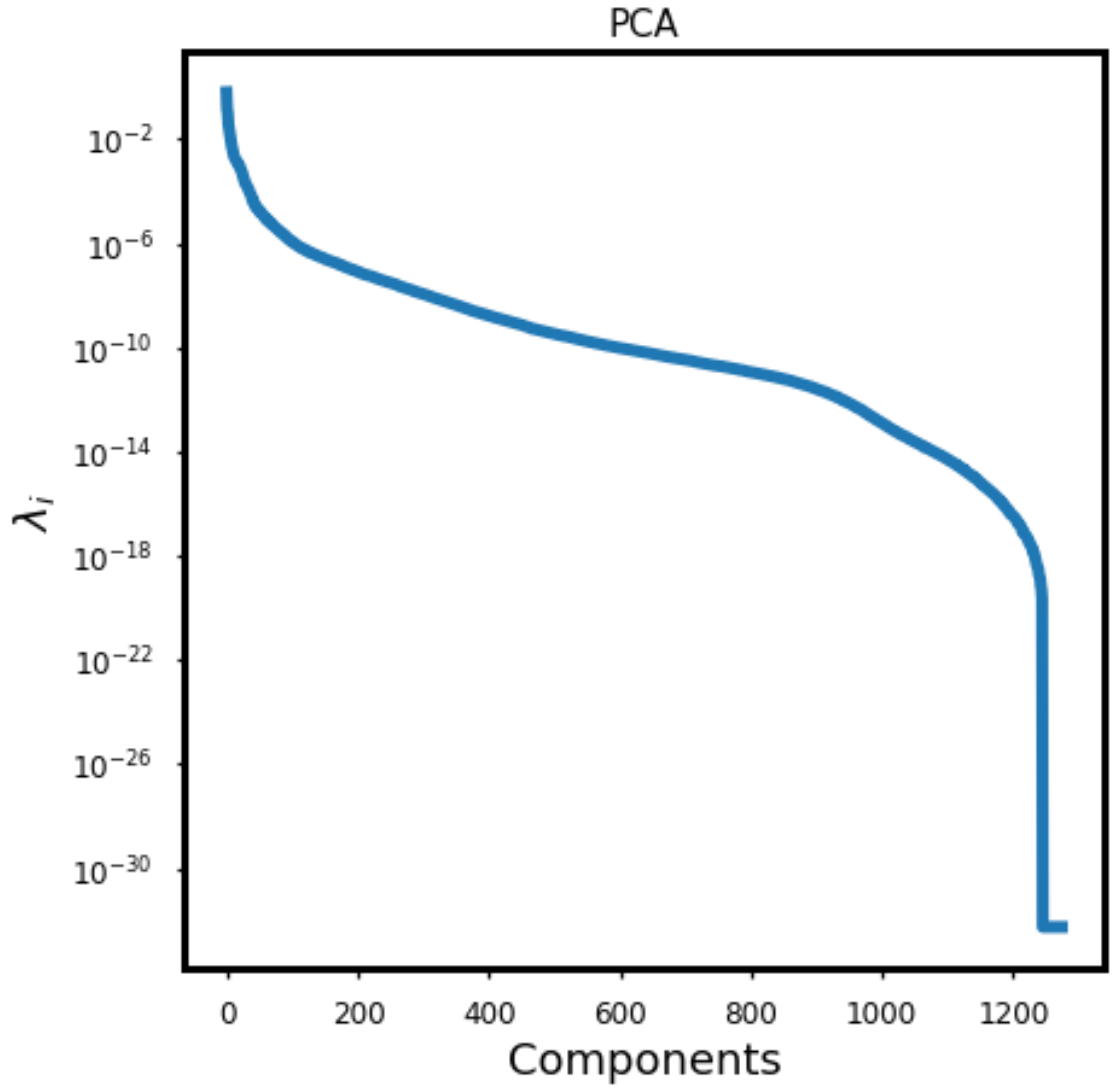
Feature Transformation: PCA

λ_1

λ_2

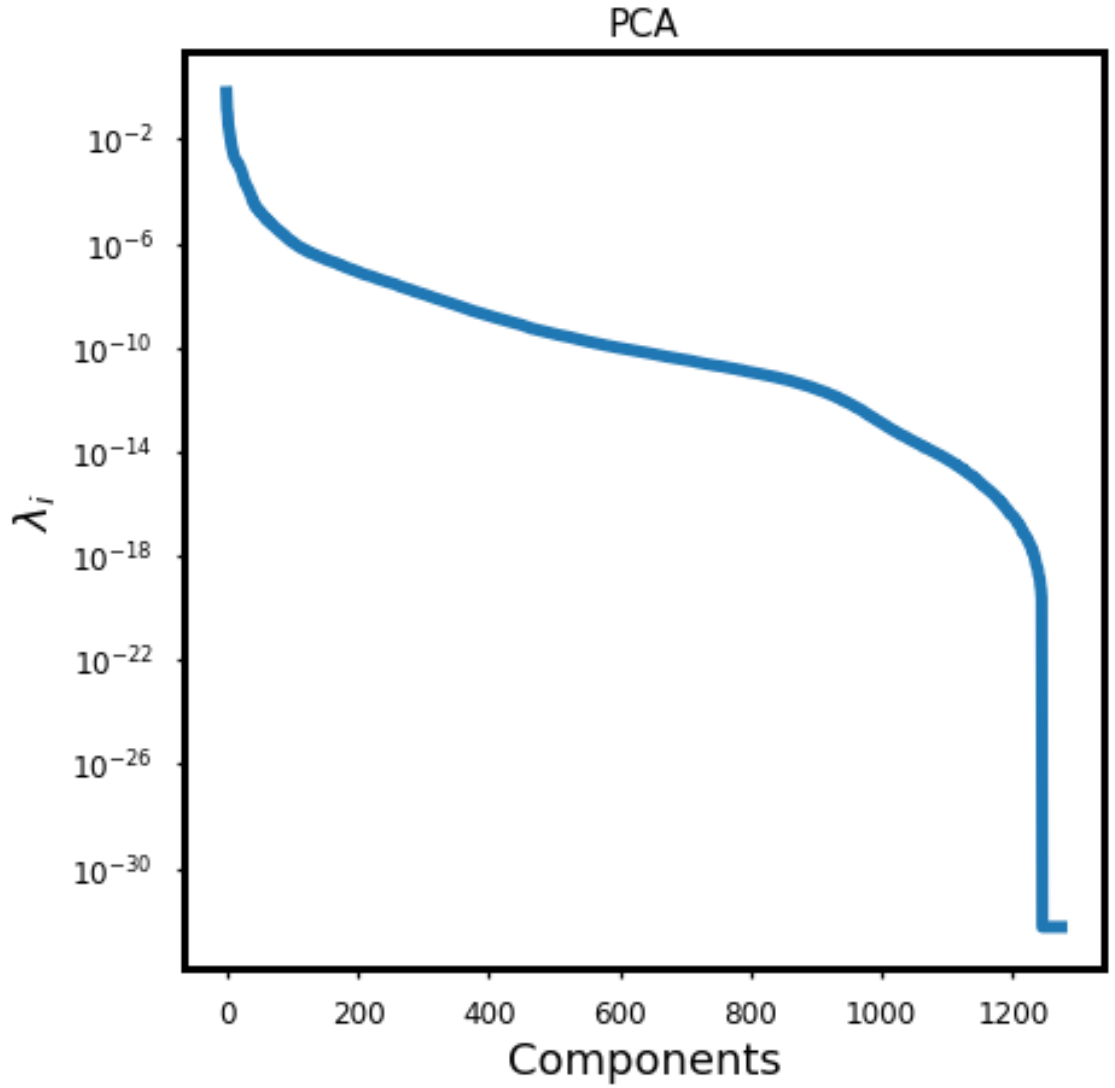
...

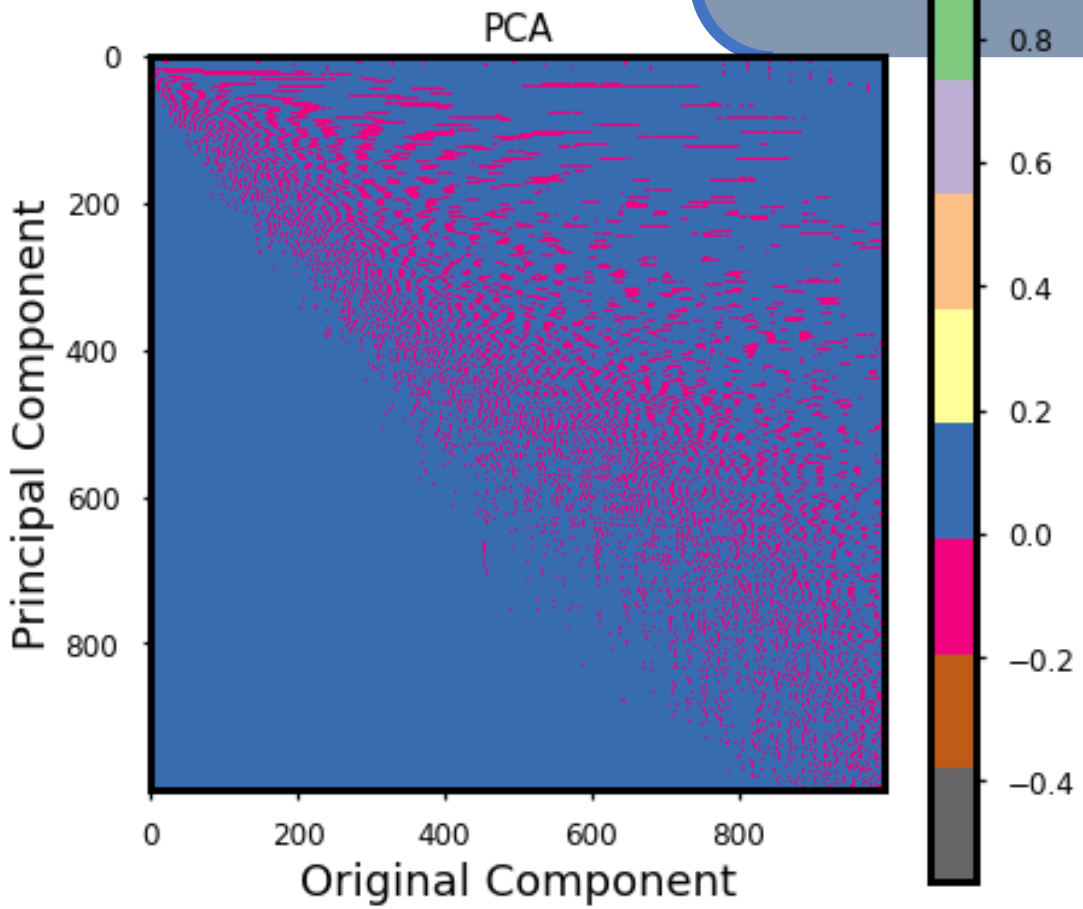
λ_{nf}



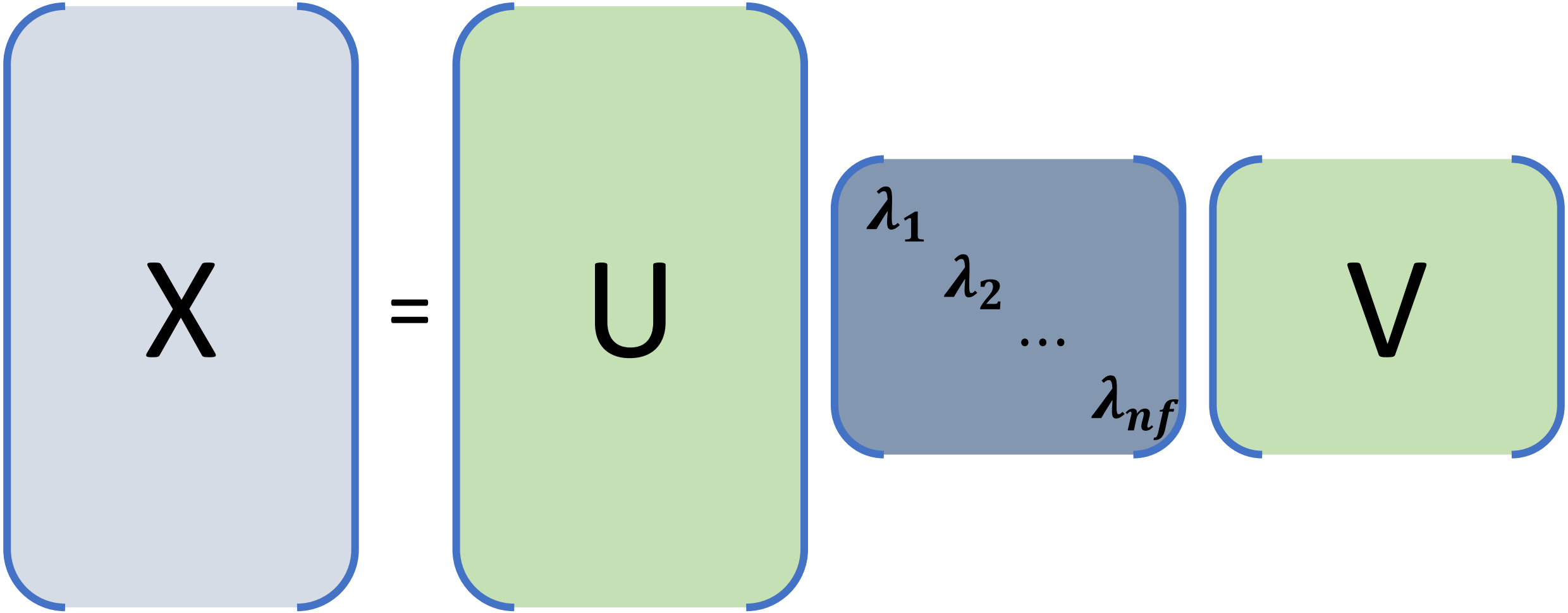
Feature Transformation: PCA

λ_1 λ_2 … λ_{nf}

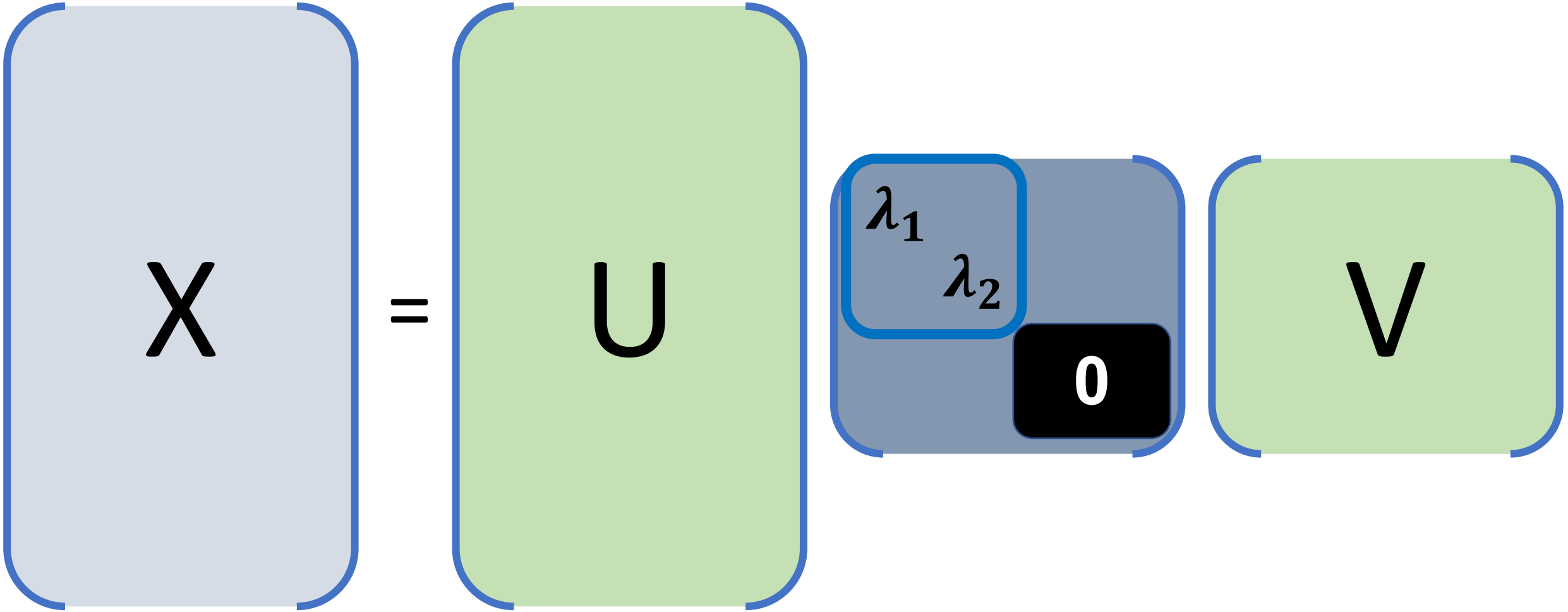




Feature Transformation: PCA



Feature Transformation: PCA



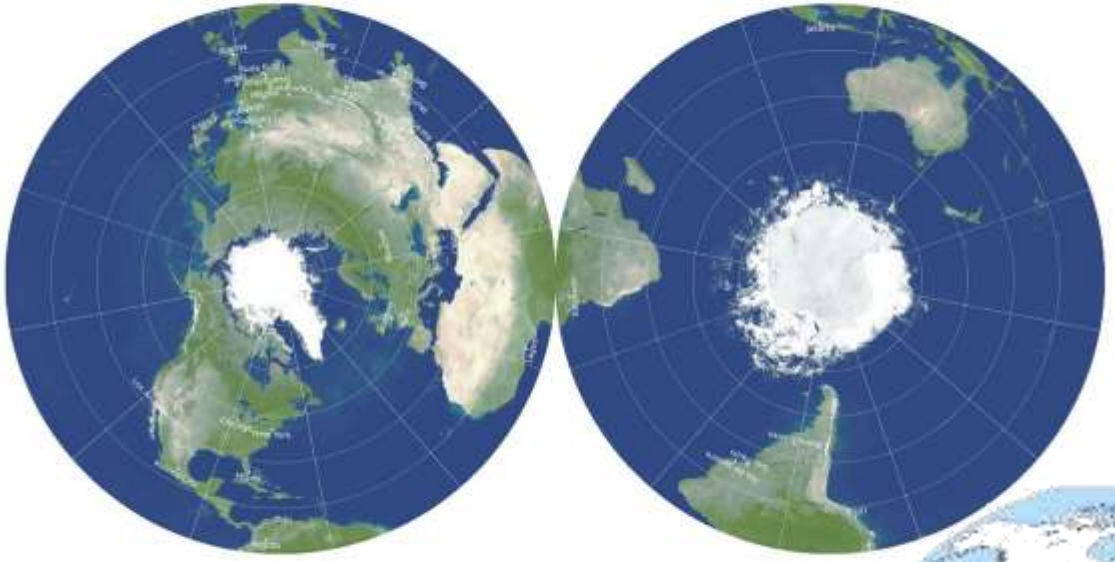
Feature Transformation: PCA

Non-linear transformations: Manifold learning

In three dimension

2D reduction

Non-linear transformations: Manifold learning



2D reduction

Summary

- Notation
- Types of Data
- Encoding
- Transformations and preprocessing
  - Missing data
  - Scaling the data
  - Data reduction

Supervised: Ingredients