

Step 2-B:

Pre-processing Data

After this video you will be able to..

- Identify some problems with real-world data
- Describe what is needed to transform raw data to data that can be used for analysis

Big Data Engineering

Computational Big Data Science

ACQUIRE

PREPARE

ANALYZE

REPORT

ACT

Step 2-A: Explore

Step 2-B: Pre-process

Clean



Transform



Real-world data is messy!

Data Quality Issues

Inconsistent values

Duplicate records

Missing values

Invalid data

Outliers

Addressing Data Quality Issues

Addressing Data Quality Issues

Remove data with
missing values

Addressing Data Quality Issues

Remove data with
missing values

Merge duplicate records

Addressing Data Quality Issues

Remove data with
missing values

Generate best estimate
for invalid values

Merge duplicate records

Addressing Data Quality Issues

Remove data with
missing values

Generate best estimate
for invalid values

Merge duplicate records

Remove outliers

Addressing Data Quality Issues

Remove data with
missing values

Generate best estimate
for invalid values

Merge duplicate records

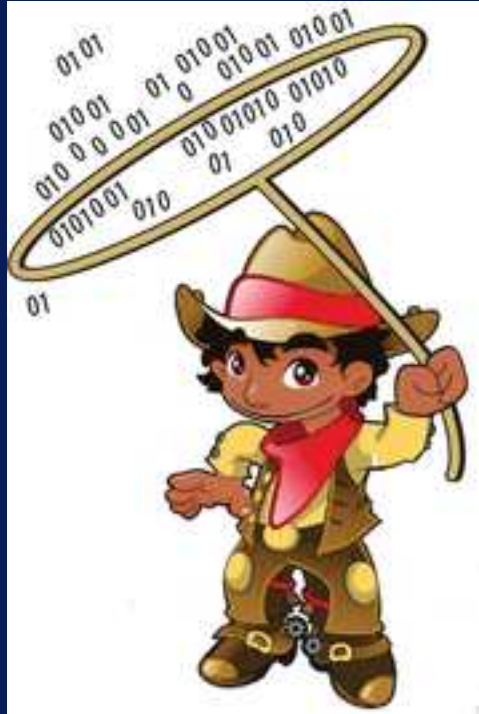
Remove outliers

*Domain
Knowledge*

Getting Data in Shape

**Data
Munging**

**Data
Preprocessing**



**Data
Wrangling**

Data Munging

*Dimensionality
Reduction*

*Data
Manipulation*

Transformation

*Feature
Selection*

Scaling

Scaling

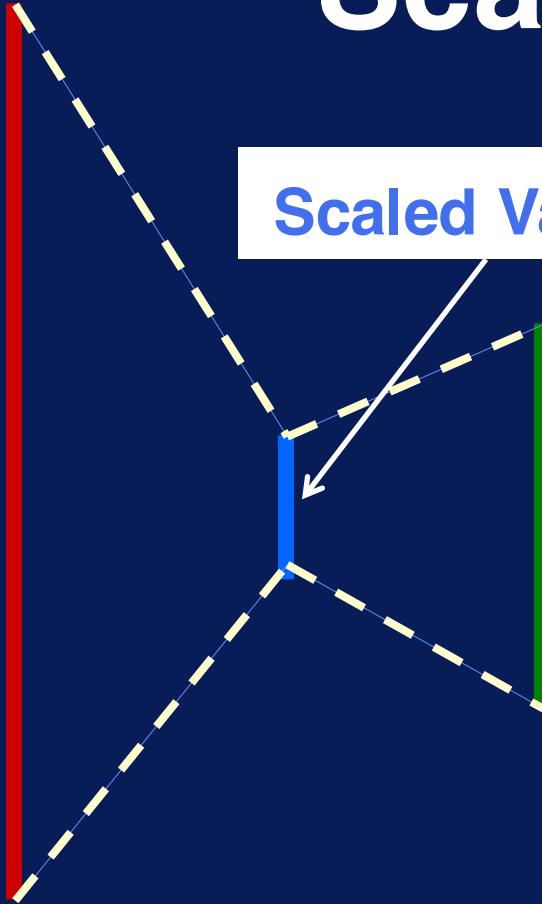


Weight

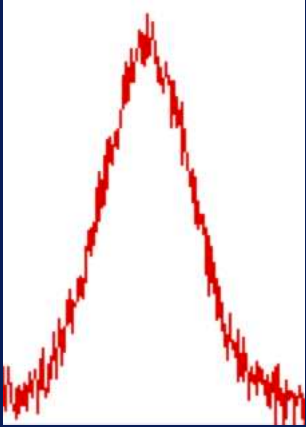
Scaled Values



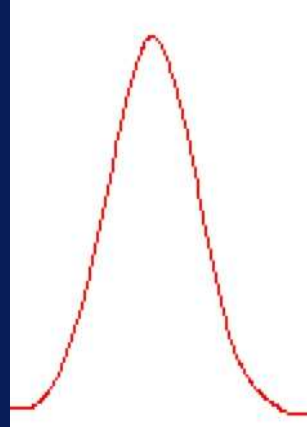
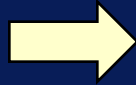
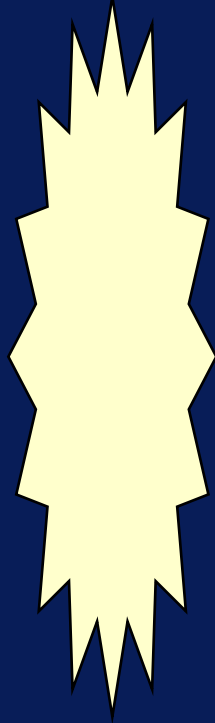
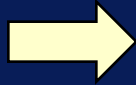
Height



Transformation

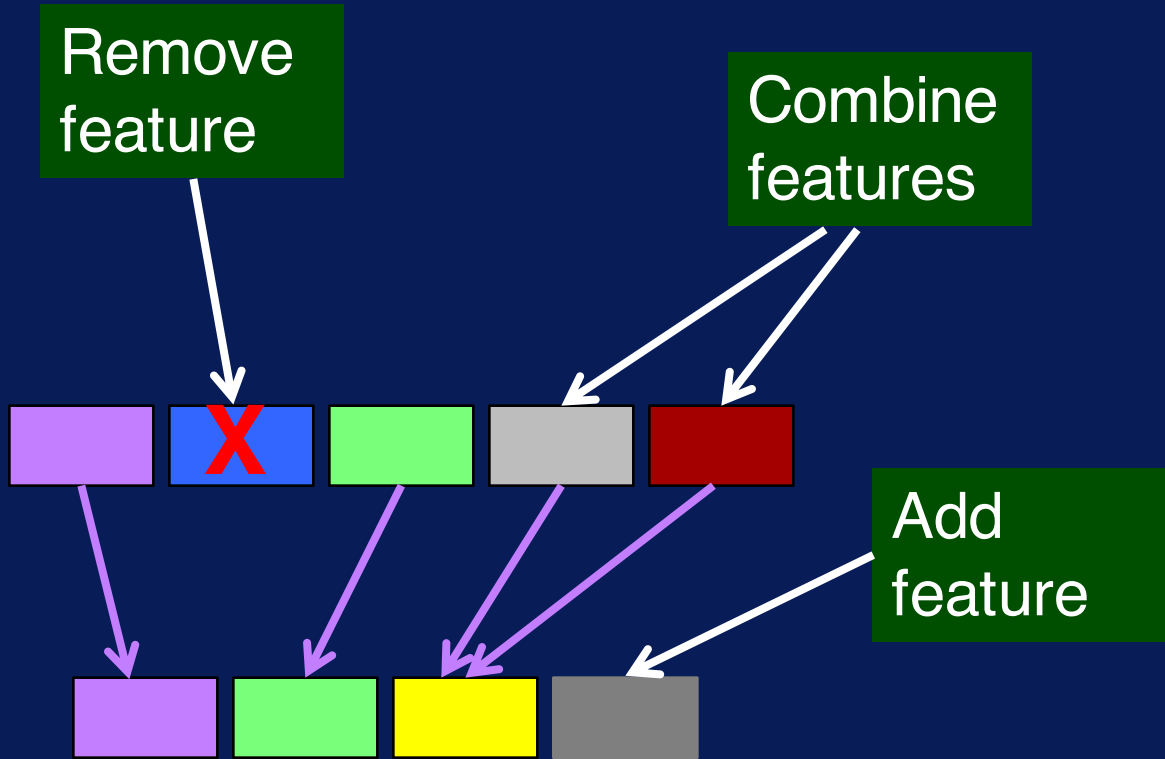


**Original
Data**

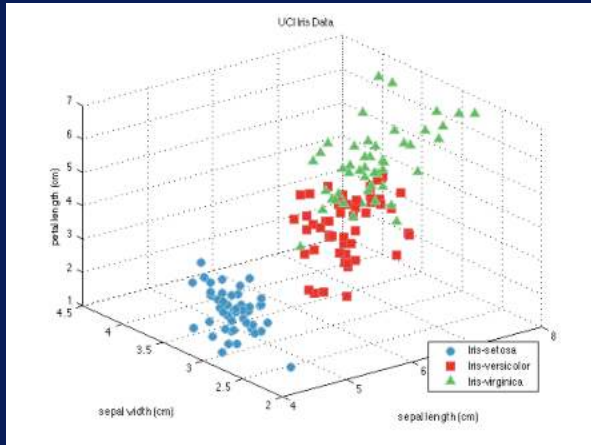


**Transformed
Data**

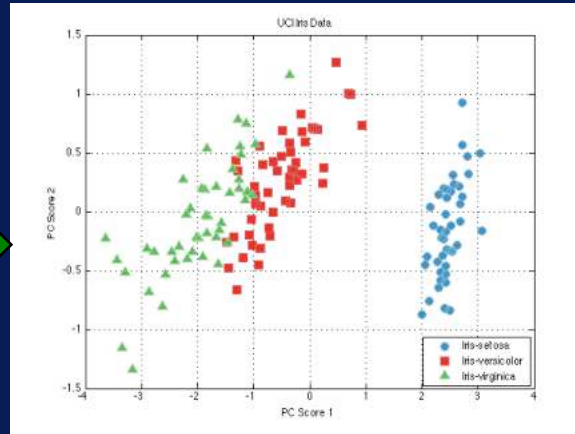
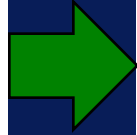
Feature Selection



Dimensionality Reduction

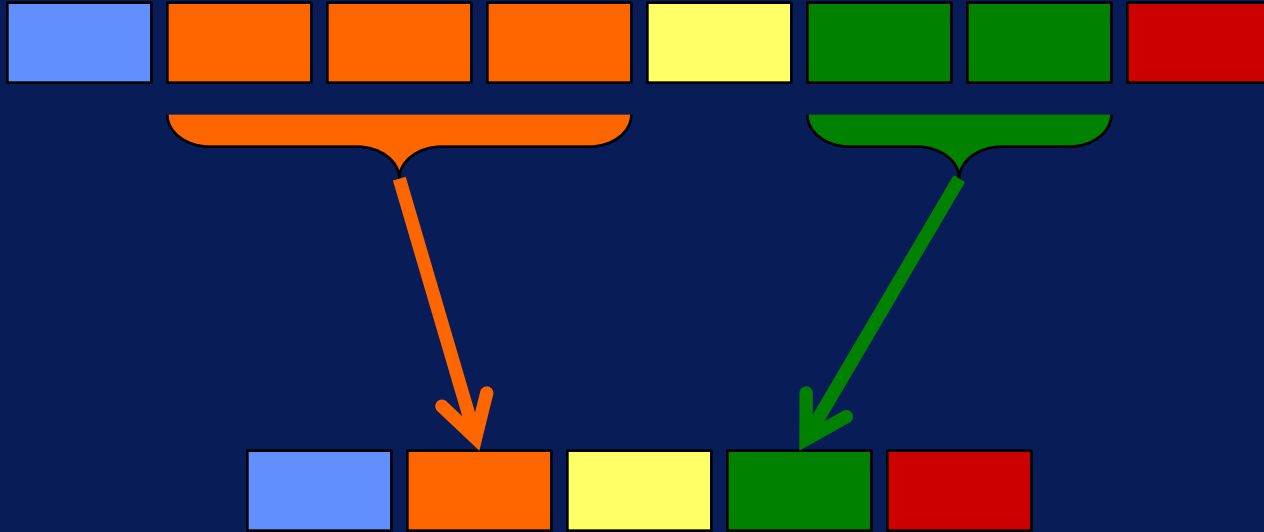


3D



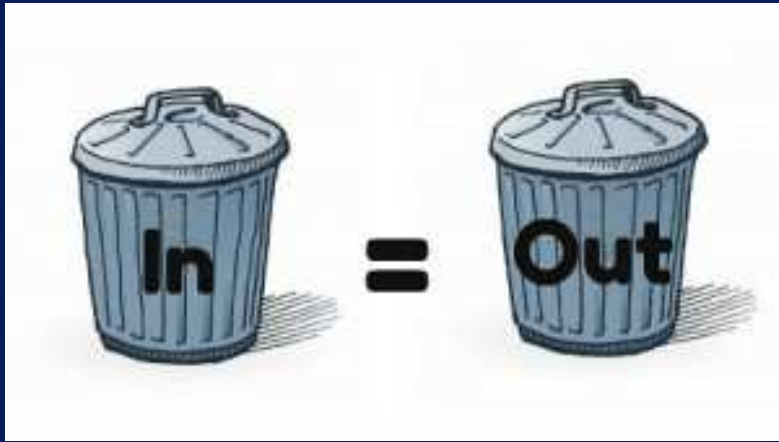
2D

Data Manipulation



Always Remember!

Garbage in = Garbage out



Data preparation is
very important for
meaningful analysis!