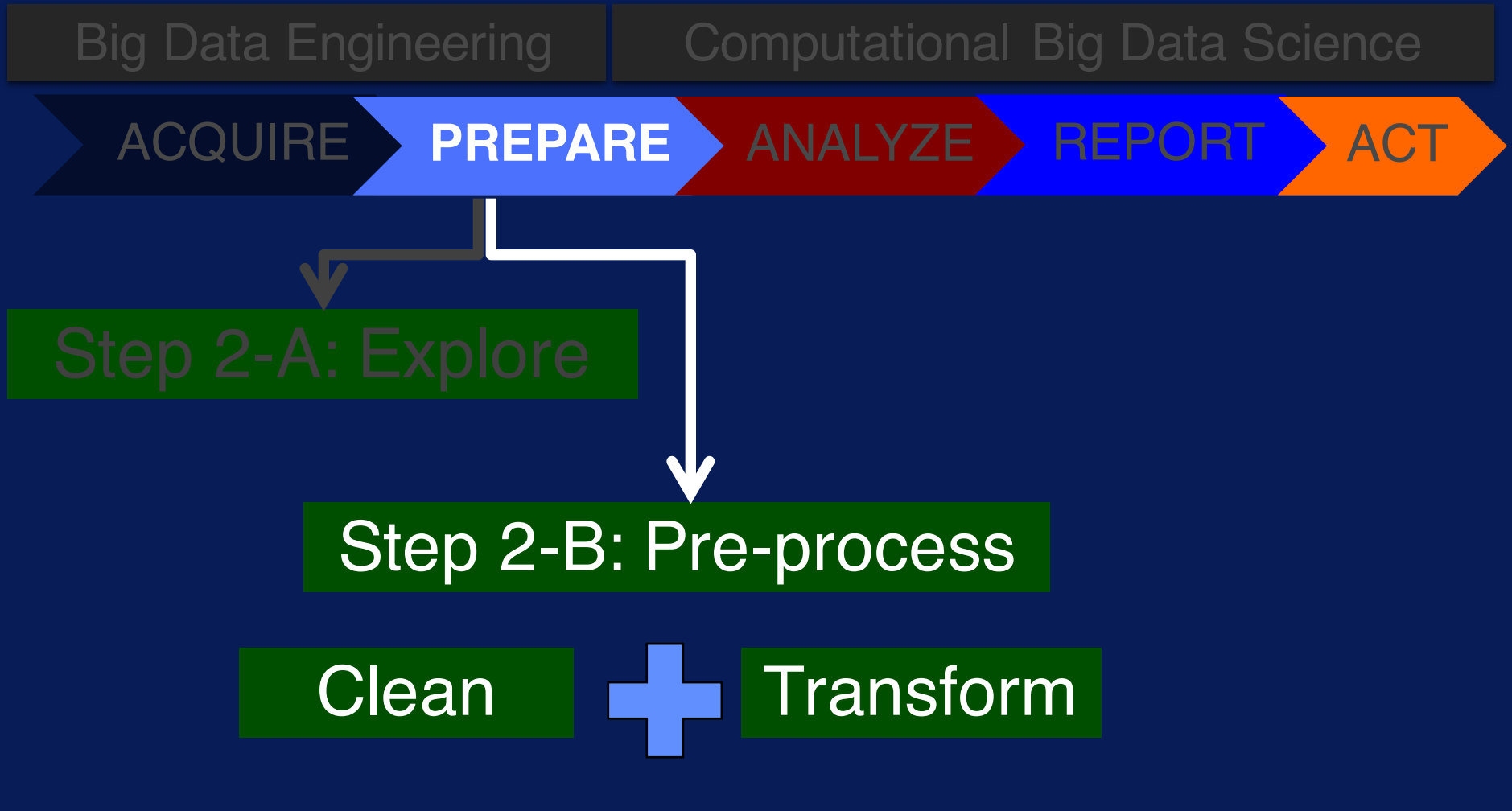
Step 2-B:

Pre-processing Data


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

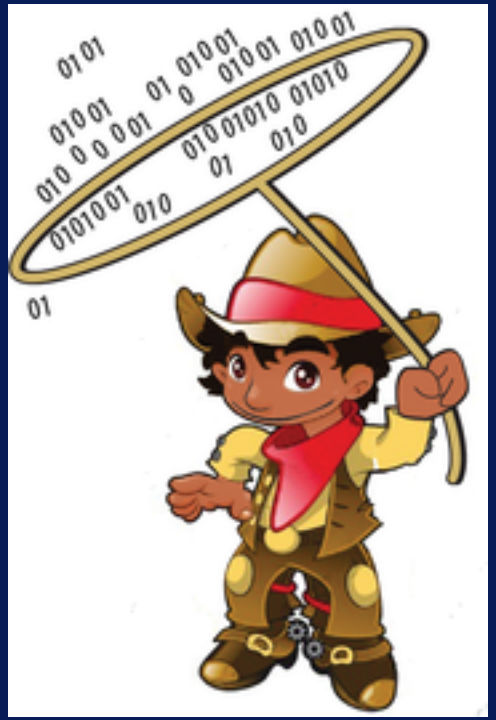
Remove outliers

Domain Knowledge

Getting Data in Shape

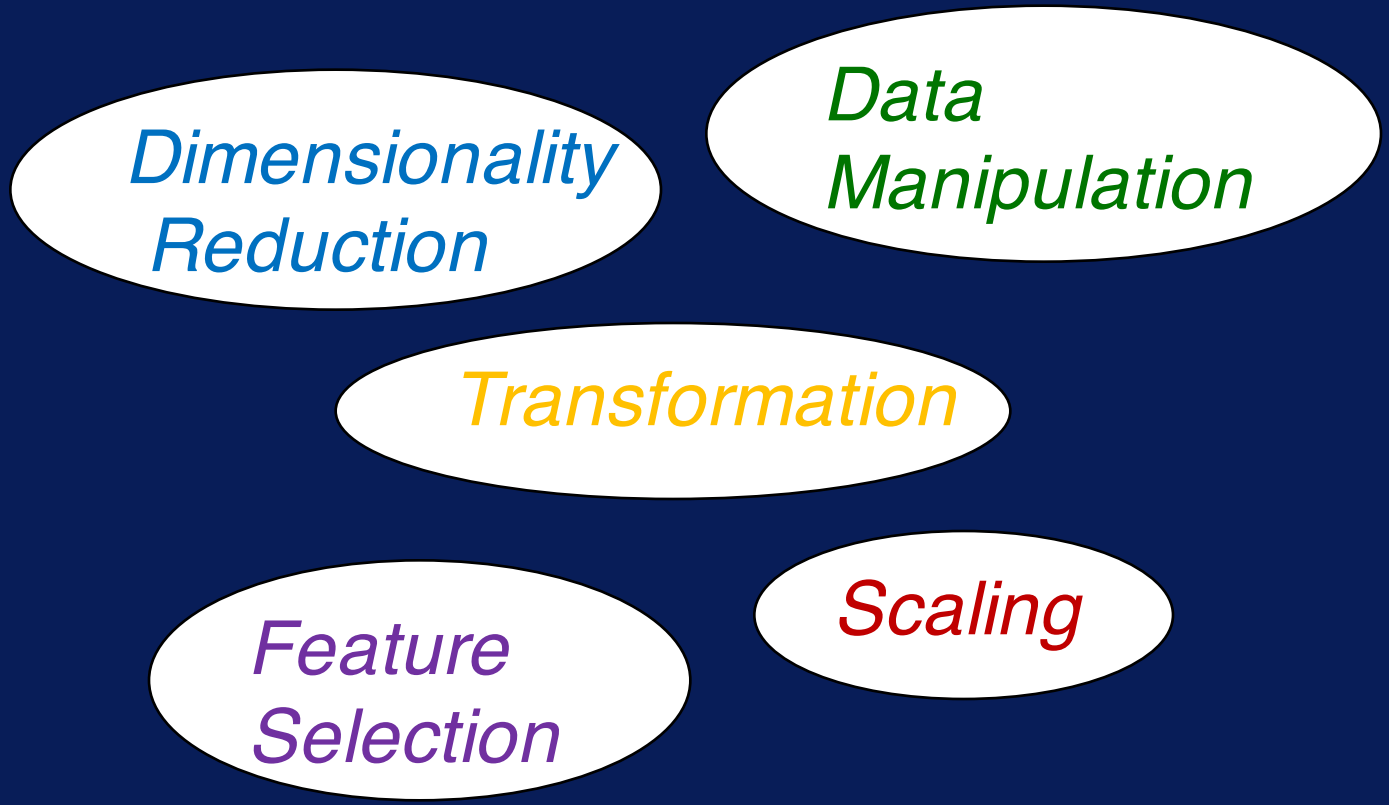
Data Munging

Data Preprocessing



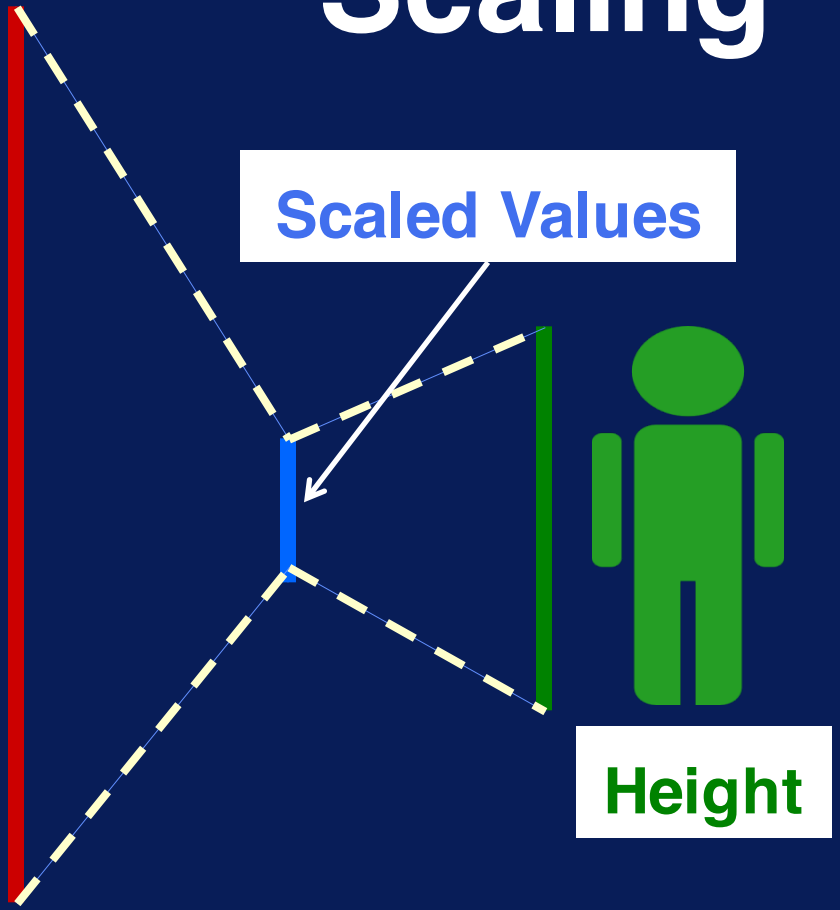
Data Wrangling

Data Munging



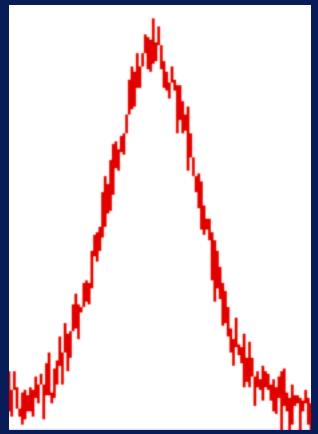
Scaling

Weight

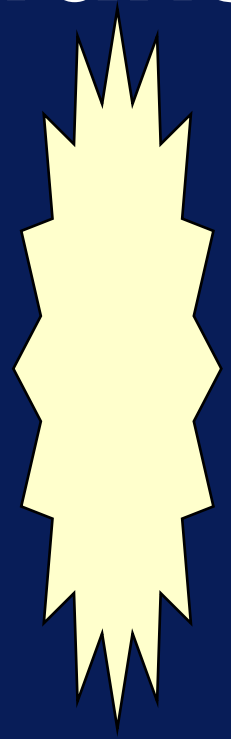


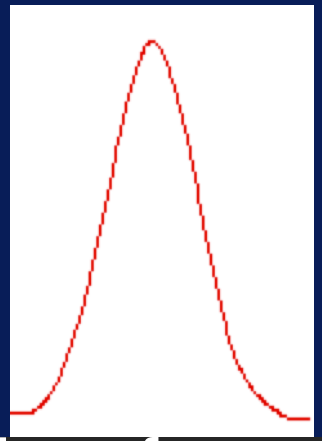
Height

Transformation



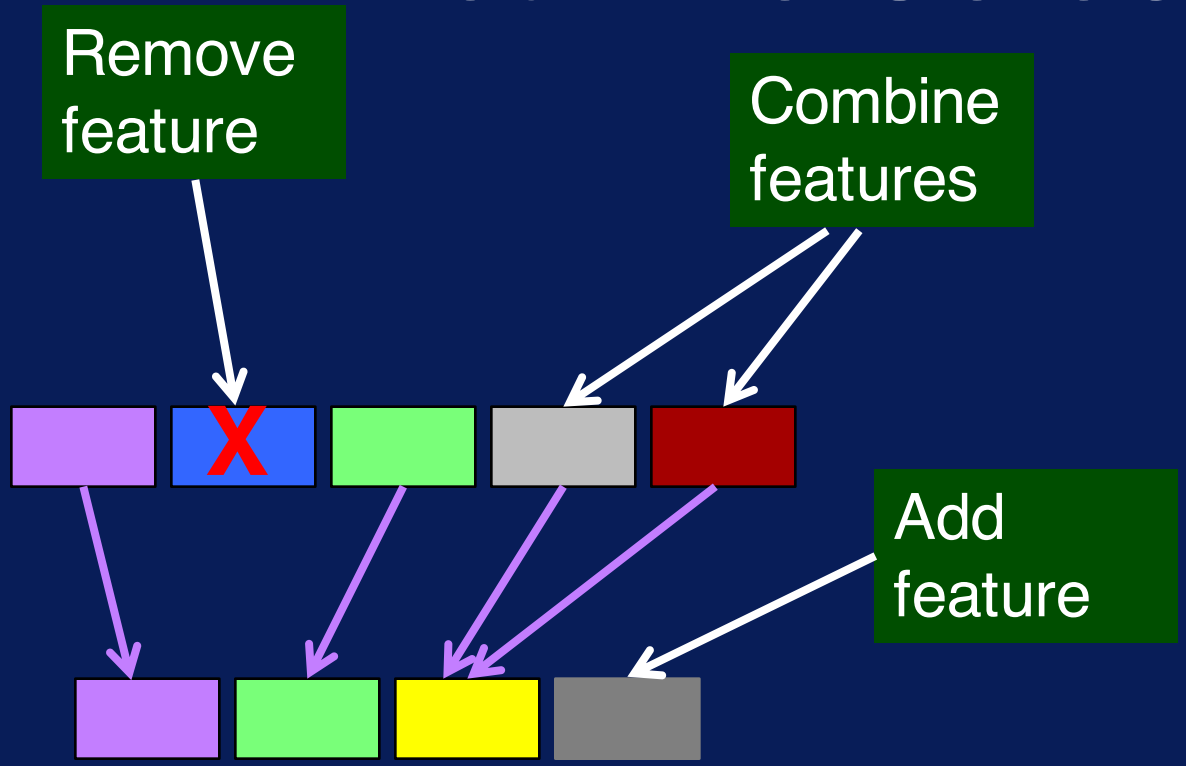
Original Data



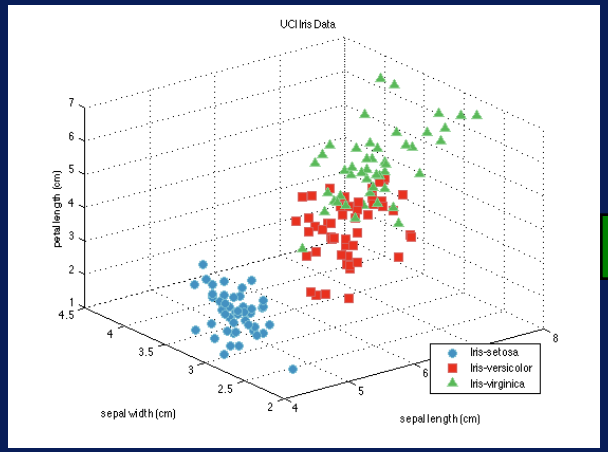


Transformed Data

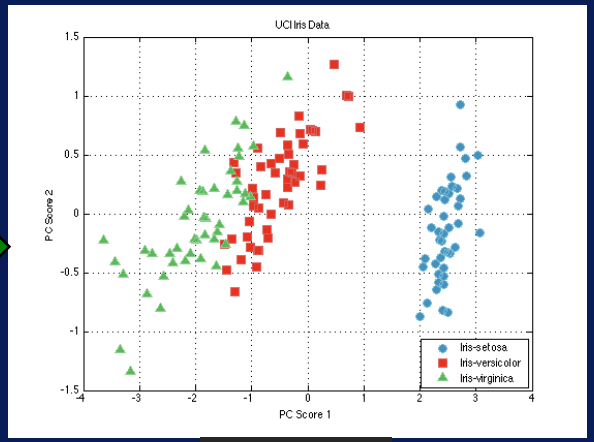
Feature Selection



Dimensionality Reduction

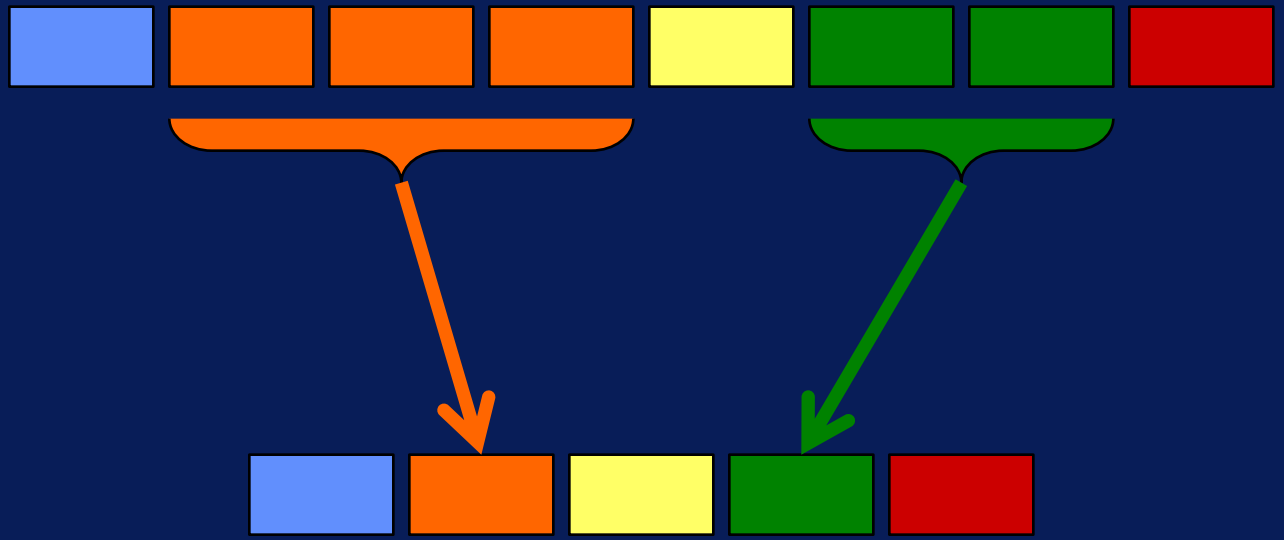


3D



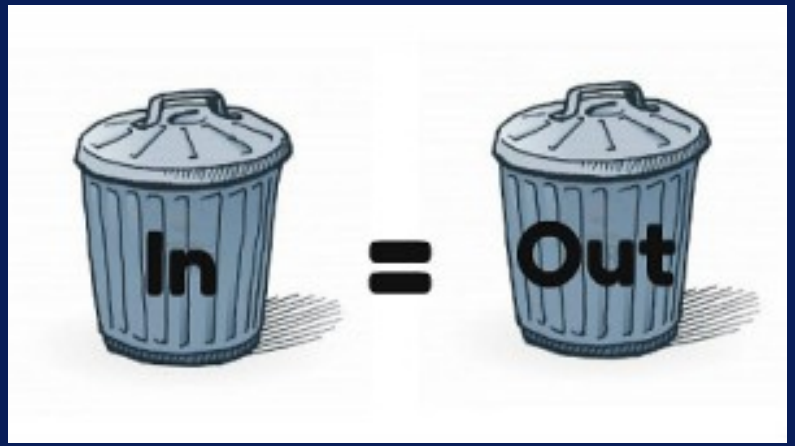
2D

Data Manipulation



Always Remember!

Garbage in = Garbage out



Data preparation is very important for meaningful analysis!