# Data Preparation

## Data Cleaning

* *Inconsistent Data*
* *Missing Data (Imputation)*
* *Noisy Data (Outliers)*
* *Systematic vs Human Error*

## Data Integration

* *Conflicting Data*
* *Redundant Data*

## Data Transformation

* *Noise Reduction*
* *Normalization/Standardization*
* *Aggregation*
* *Generalization*

## Data Reduction

* *Dimensionality Reduction*
* *Data Cube Aggregation* 
  + *Can the data be aggregated by any level*
* *Numerosity* 
  + *Can we represent the population with a smaller sample*
* *Sampling (Systematic, Stratified, Random, Cluster)*
  + *Systematic – Start at a random location, and then sample every Kth observation*
  + *Stratified – Cluster into homogeneous groups (hair color, eye color, etc.) and then sample from each cluster (Each strata is sampled)*

*Forcing it to closely represent the population (Imposed by researcher)*

* + *Random – Shuffle and choose N samples*
  + *Cluster – Create clusters within the data, then randomly sample from the cluster (Clusters may or may not be sampled from)*

*May or may not represent the population (Naturally occurring)*