Data Matrix/Regular Table (dataset)

* Data item (row)
  + The samples (observations) we obtained from the population of all instances
* The variables
  + The attributes or properties of dataset

Data Cleaning (wrangling)

* Fill in missing value
  + Data is not always available
  + Reason
    - Equipment malfunction
    - Inconsistent with other recorded data and deleted
    - Data not entered
    - …
  + Method to estimate miss value
    - Average over all data, average based on given condition
    - Probabilistic method
    - Ignore/default value
    - ML
* Smooth noisy data
  + Random error in measured variable
    - Faulty data collection instruments
    - Data entry problems
    - Data transmission problems
    - Technology limitation
    - Inconsistency in naming convention
  + Method
    - Binning method
    - Clustering
      * Detect and remove outlier
    - Semi-automated method
      * Combining computer and human inspection
      * Detect suspicious value and check manually
    - Regression
      * Smooth by fitting the data to a regression function
  + An outlier may not be noise
    - It may be an anomaly that is very valuable
* Identify or remove outliers
* Resolve inconsistencies
  + Inconsistencies in naming conventions or data codes
  + Redundant data
    - Duplicate tuples, which were received twice should be removed
* Standardize/normalize data
  + Min/max normalization
  + Standardization/z-score
  + Clipping tails and outliers, reasonable cut-off point
  + Not sensitive to outliers
* Fuse/merge data
  + Can gain more comprehensive insight (value > sum of parts)
  + Attribute with different labels may have same meaning
  + Attribute with same label may have different meaning
  + Goal
    - Enable deeper and more far fetching insights
    - Can open valuable opportunities for research and $$$
  + Procedure
    - Start with a first dataset
    - Pick an attribute
    - Join with another dataset that has same attribute
  + Problem
    - Privacy
  + Solution: to obfuscate for protection
    - K-anonymity (generalize)
    - Make data less specific – binning
    - Make blobs instead of points

Data Reduction

* Purpose
  + Reduce data to a storable size
  + Reduce data so a mining algorithm can be feasibly run
* Sampling
  + Goal
    - Pick a representative subset of the data
  + Random
    - Pick sample points at random
    - Work if the points are distributed uniformly
    - Outliers will be missed
    - Is not representative
  + Adaptive Sampling (Stratified sampling)
    - Pick the samples according to some knowledge of the data distribution
      * Create a binning of some sort
      * The size of each bin represents its percentage in the population
      * It guides the number of samples – bigger bins ger more samples

Data augmentation (When dataset is too small)

* Strategy to artificially synthesize new data from existing data

Jittering

* Definition from dictionary
* Act nervously
* Small random noise about a steady signal

Data Summarization

* Binning
* Clustering
* Dimension reduction