data preprocessing

aggregation = combining multiple attributes/objects into a single attribute/object

sampling = selection of a subset from data, reduction

↳ used because processing the entire set is too expensive or time consuming

↳ representative = if the sample has almost same properties as the original set of data

- simple random sampling = equal probability of selecting any particular item
  - ↳ sampling without replacement = as item is selected, removed from the population, no duplicate
  - ↳ sampling with replacement = selected items are not removed same object can be selected multiple times
- stratified sampling = split the data into several portions, select random object from each partition

discretization = process of converting a continuous attribute into an ordinal attribute

↳ used in both supervised and unsupervised settings

binarization = maps a continuous or categorical attribute into one or more binary variables

Table 2.6. Conversion of a categorical attribute to five asymmetric binary attributes.

Categorical Value	Integer Value	x_1	x_2	x_3	x_4	x_5
awful	0	1	0	0	0	0
poor	1	0	1	0	0	0
OK	2	0	0	1	0	0
good	3	0	0	0	1	0
great	4	0	0	0	0	1

attribute transformation = function that maps the entire set of values of a given attribute to a new set of replacement values → x^k, $\log(x)$, $|x|$

↳ normalization = various techniques to adjust to differences among attributes in terms of freq of occurance, mean, variance, range

↳ standardization = subtracting off the means and dividing by std dev

dimensionality reduction = to avoid curse of dimensionality, to reduce amount of time and mem required, visualize more easily, eliminate irrelevant features

↳ principal component analysis, singular value decomposition, supervised...

feature subset selection = redundant features = when two attributes have almost the same feature

↳ irrelevant features = contains no useful information for the task

feature creation = feature extraction, feature construction, mapping data to new space