**Data Mining : Intuitive Partitioning of Data or 3-4-5 Rule**

* Intuitive partitioning or natural partitioning is used in data discretization.
* Data discretization is the process of converting continuous values of an attribute into categorical data or partitions or intervals.
* Discretization helps reducing data size by reducing number of possible values. Instead of storing every observation we can only store partition range in which each observation falls. One of the easiest ways to partition numeric values is using intuitive (natural) partitioning. It works using following rules:

**Intuitive partitioning or The 3-4-5 Rule for Data Discretization:**

1. If an interval covers 3, 6, 7 or 9 distinct values at most significant digit, then create 3 intervals. (There can be 3 equal width intervals for 3, 6, 9; and 3 intervals in the grouping of 2-3-2 each for 7).
2. If it covers 2, 4 or 8 distinct values at most significant digit, then create 4 sub-intervals of equal-width.
3. If it covers 1, 5, or 10 distinct values at the most significant digit, then partition range into 5 equal-width intervals.

**Example:**

* Suppose that profit data values for year 2017 for a company range from -3,51,976 to +4,70,00,896.
* For practical purpose of avoiding noise, extremely high or extremely low values are not considered. So first we need to smooth out our data. Let’s discard bottom 5% and top 5% values.
* Suppose after discarding above data new values for LOW = -159876 and HIGH = 1838761.
* Most Significant Digit or MSD is at million position, see highlighted digit : –**1**59876 and **1**838761.
* Next step is to *round down* LOW and *round up* HIGH to MSD that million position. So LOW = -1000000 and HIGH = 2000000. -1000000 is nearest down million to -159876 and 2000000 is nearest up million to 1838761.
* Now let’s identify range of this interval. Range = HIGH – LOW that is 2000000 – (-1000000) = 3000000. We consider only MSD here which is 3.
* Now that we know range MSD = 3, we can apply rule #1.
* Rule #1 says that we can divide this interval into three equal size intervals:
  + Interval 1 : -1000000 to 0
  + Interval 2 : 0 to 1000000
  + Interval 3 : 1000000 to 2000000
* Question : How 0 can be part of multiple intervals?
* Answer : We should represent it as follows:
  + Interval 1 : (-1000000 … 0]
  + Interval 2 : (0 … 1000000]
  + Interval 3 : (1000000 … 2000000]
  + Here (*a … b*] denotes range that excludes *a* but includes *b*. ( , ]  is notation for half-open interval.

