

## IDS 561 – Assignment 2

---

### Please read before running the code:

- There are two inputs to the program. The input file and the centroid file. Input file name is `input.txt` and centroid file name is `centroid.txt`.
- You have to specify three command line arguments for the program
  - Arg 0 – name of input directory
  - Arg 1 – name of output directory
  - Arg 2 – name of centroid directory.
- The number of lines present in the centroid file determines value of K.
- Before running the program, make sure there are no output directories present in the file system.
- Number of mappers is 3
- Number of reducers = K

### Result after running:

- All intermediate results will be stored in the file system in folders output1, output 2, output3 ... output(x) etc.

### Metadata about input:

- Contains 4020 data points.
- Source: [http://cs.joensuu.fi/sipu/datasets/data\\_dim\\_txt.zip](http://cs.joensuu.fi/sipu/datasets/data_dim_txt.zip)