L3S Research Center
Large Scale Data Mining, SS 2016
Dr. Avishek Anand
Assignment 3, due: **06 May 2016**

You must return your assignment sheet and have a correct solution in order to present in the exercise groups. Please write legibly! Do no forget to put your name and matriculation number on your solution!

Problem 1. Suppose we have a stream of tuples with the schema:

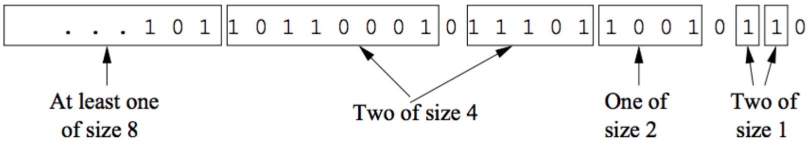
Grades(university, courseID, studentID, grade)

Assume universities are unique, but a courseID is unique only within a university (i.e., different universities may have different courses with the same ID, e.g., CS101) and likewise, studentID's are unique only within a university (different universities may assign the same ID to different students). Suppose we want to answer certain queries approximately from a 1/20th sample of the data. For each of the queries below, indicate how you would construct the sample. That is, tell what the key attributes should be.

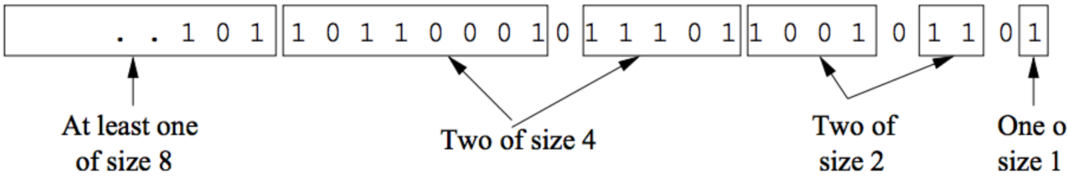
1. For each university, estimate the average number of students in a course.
2. Estimate the fraction of students who have a GPA of 3.5 or more.
3. Estimate the fraction of courses where at least half the students got 'A'.

Problem 2. Consider the DGIM algorithm for counting 1's in a stream. Suppose the window is as shown in the figure. Estimate the number of 1's the the last k positions, for k = (a) 5 (b) 15. In each case, how far off the correct value is your estimate?

```
. . 1 0 1 1 0 1 1 0 0 0 1 0 1 1 1 0 1 1 0 0 1 0 1 1 0
```



Problem 3. Describe what happens to the buckets if three more 1's enter the window represented by the figure below. You may assume none of the 1's shown leave the window.



Problem 4. Now let us consider an algorithm which uses fixed window sizes to count 1's (say

window width w), then what is the:

- estimated space requirement for storing windows and their representations?
- maximum error you can expect when querying for the last k bits where $k \leq N$?