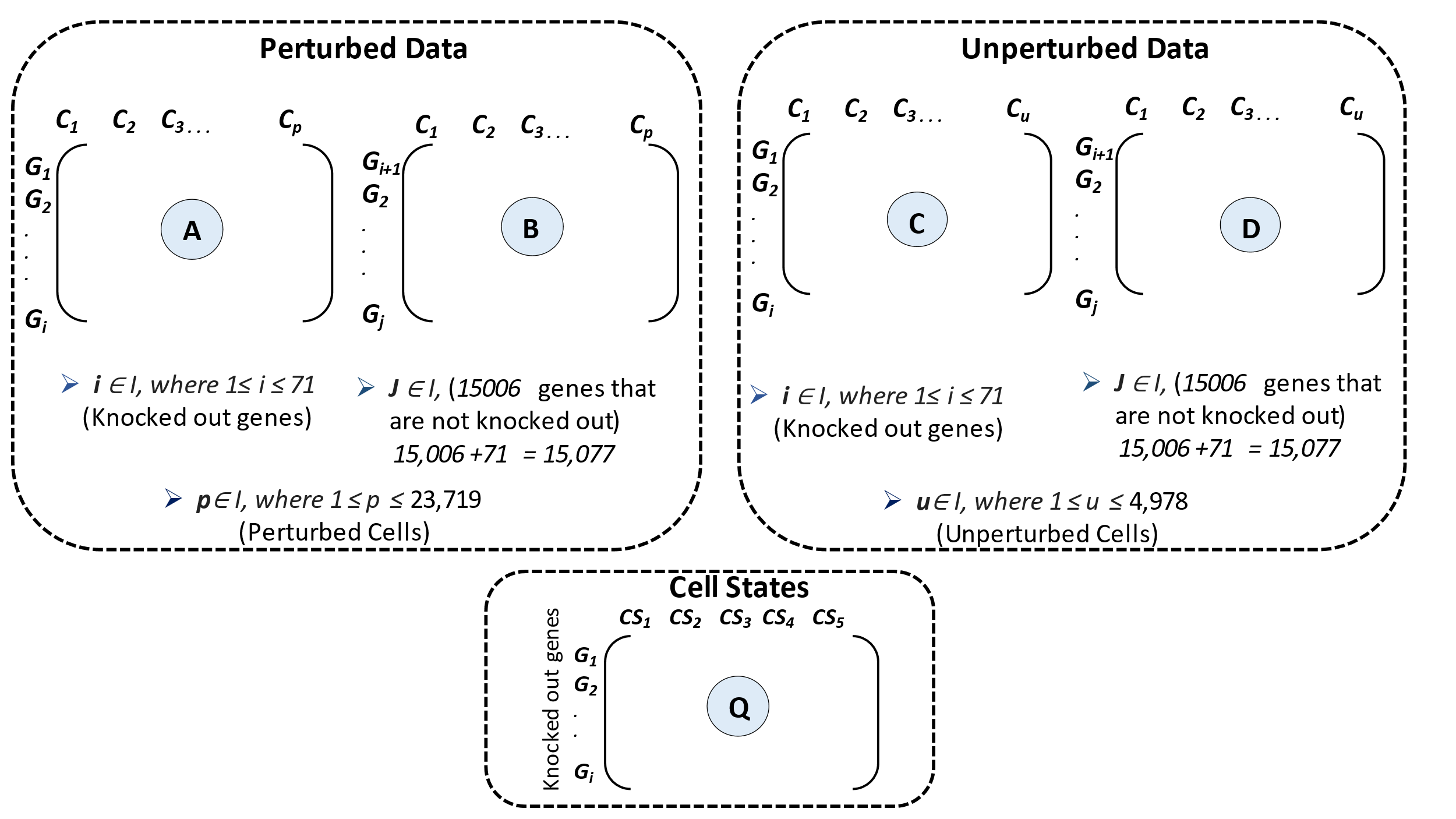
**Cancer Immunotherapy Grand Data Science Challenge - Challenge 3**

**3rd Feb,2023**



**Procedure:**

1. From the matrix A: Collect the gene expression vector of *Gene*i, i.e., gene expression of the cells that are perturbed with *Gene*i

-- This gives a vector (V) of values based on the number of cells the gene is perturbed with

1. From the matrix B: Subset the columns with the cells identified in *step-a* (for *Gene*i)

-- This step outputs a matrix (M) of [J \* n] dimension,

where J is non-knocked-out genes (15,077 – 71)

n is cells that are perturbed with *Gene*i

1. Compute the correlation between the V and each gene in M and find the most dissimilar genes (Threshold (T) to be computed empirically by verifying the distribution)

-- This step outputs a set of genes that are highly dissimilar from *Gene*i

1. From the matrix C & D: Compute the correlation between *Gene*i from matrix C to all other genes in D, to find the most similar genes (Threshold (T) to be computed empirically by verifying the distribution)

-- The intersection of most dissimilar genes in *step-c* and similar genes in *step-d* can be considered as genes of similar type.

1. Using cell state matrix Q: Assign a cell state for *Gene*i by considering the maximum value of all cell states. The genes identified in *step-d* can be associated with the same cell state of *Gene*i.

-- Limitations**: -** Possibility of multiple cell states for the same gene. Ranks/probability-score for each cell state is needed

1. Repeat *steps, a - e* with all genes in matrix A
2. Check if all genes are assigned to a cell-state (all 15,77 genes). If not, Assign ‘other’ for the missing genes.