

**1st Annual Big Data Health Science Case Competition Case February 7 – February 9, 2020**

# CASE SYNOPSIS

The primary objective of this year’s case competition is to develop analytic tools that mitigate the human and financial cost of similar incidents in the future. The challenges with the data are carefully laid out below. During this event, health systems were overwhelmed with the number of patients and symptoms affecting the patients. The greatest challenge was identifying which symptoms are associated to which chemicals for each patient. This problem will be reflected in the final deliverables for the case, with patient row data.

Team members will work together to present their methods and analysis results at the Big Data Health Science Conference. A panel of industry and academic experts will judge the presentations based on teams’ use of the full analytics process, from framing the problem to methodology selection, data use, model building and innovation. Only teams of students can work together on the case when it is released on Friday 7th, 2020.

# Data File

Each team will receive a jump drive with the following data structure

* 311 unique chemicals
* Each chemical can elicit up to 79 different clinical signs and symptoms
* Each patient information is on one row.
* Each row consists of signs and symptoms for one patient.
* Be aware that some patients may not experience any chemical exposure related to the incident.

# The CaseCompetitionPatientData.csv consist of patient records on each line. Each line (patient record), will consist of the presence or absence of 79 signs and symptoms. The presence is indicated by a 1, and the absence is indicated by a 0. The Chemicals.csv file consists of 311 lines, each capturing the 79 signs and symptoms that will be observed for each of the chemicals. Again, the presence of a symptom is indicated by a 1 and the absence is indicated by a 0.

# PHASE 1: DELIVERABLES

Using the data provided, for the first round, all contestants at a minimum are required to do the following:

1. Return a single csv file containing 1,010,000 integer numbers (between 1 and 311) that indicate the assignment of each patient to a chemical listed in the Chemicals data file.
2. Use analytic techniques to identify and describe ***the number*** and ***types of chemicals*** affecting the patient population using the datasets provided.
3. Present the essential facts of the data and your findings visually to the audience and judges
4. Identify the solutions and methods you used to solve this problem
5. Account for limitations within the data.
6. Deliver it as a workable solution to the health system swamped by the critical incidents.

**Logistics for Submission**

The final slides and single CSV file for each team must be submitted back to the faculty monitors using the assigned jump drive. The final slides and CSV slide should be labeled as follows

**TeamAA##\_CSV file**

**TeamAA##.ppt(x)**