Spring 2020 - **Knowledge Discovery in Data at Scale Technologies** - Homework #1

Let $(x_1, x_2, \ldots, x_n)$ be a sequence of column vectors:

$$x_i = \begin{bmatrix} x_i^1 \\ \vdots \\ x_i^m \end{bmatrix}, \qquad i = 1, \ldots, n.$$

In statistics one often has to compute the **sample mean** vector

$$X = \frac{1}{n} \sum_{i=1}^{n} x_i$$

and the **sample covariance** (or **variance-covariance**) matrix

$$V = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - X)(x_i - X)^T,$$

where x^T is the transpose of x.

What **canonical** form of information would you suggest to represent the sequence $(x_1, x_2, \ldots, x_n)$ in order to compute the sample mean vector and the sample covariance matrix?

Verify that all the "desirable" properties of canonical information are satisfied:

1. **Existence and Uniqueness**. Can any sequence of raw data $(x_1, x_2, \ldots, x_n)$ be represented by canonical information in a unique way? Does this representation depend on the order of vectors x_i in the data sequence?

2. **Completeness**. Canonical information should retain ALL the information which was present in the original raw data. Specifically, an algorithm applied to canonical information (deployment phase) should produce the same results as the original algorithm applied to the original raw data.

   Does your canonical representation of data conform this requirement? How would you compute X and V using only collected canonical information?

3. **Elementary** canonical information. Does canonical information exist for a single observation?

4. **Empty** canonical information. Does canonical information exist for an empty sequence of observations?

5. **Combination** (or composition) operation. How would you define composition of pieces of canonical information? Does it satisfy axioms for a commutative monoid? (commutativity, associativity, neutral element)

6. **Update** operation. How is canonical information updated when a new observation vector x arrives?

7. **Compactness and Efficiency.** What can you say about compactness (or minimality) and efficiency of your canonical form of information in terms of storage requirements and complexity of combination, update, and deployment operations?

8. What is the **minimum number** of observations n for which X and V are defined?