MIS0855: Data Science
In-Class Exercise: How Data Gets Dirty

Objective: Analyze and understand the process of evaluating data quality.

Learning Outcomes:
- Identify threats to data quality
- Design mechanisms to identify quality problems in collected data
- Develop remedies to prevent future data quality problems

Step 1: Identify potential sources of data quality problems (15 minutes)
In groups of three, identify sources of data quality problems in these two medical forms:
- Adult Complete Physical Examination (completed by the physician)
- Adult Health History Form (completed by the patient)

In each case, many different doctors and patients will complete these forms. Assume that this data is promptly entered into a database. With that in mind, identify opportunities for:
- Measurement issues (the measurement can be wrong or inaccurate)
- Instrument issues (the right question may not have been asked)
- Consistency issues (the question can be answered in multiple ways)

For example, the first question on the history form has an instrument issue:

PERSONAL MEDICAL HISTORY: Do you currently have or have had in the past (mark all that apply)...

This question does not allow for the differentiation between a current or past condition.

Step 2: Remedy the data quality issues (15 minutes)
In your groups, also outline how each of the data quality issues you have identified can be dealt with. Issues might be dealt with either before or after data collection. For example:
- How would you verify accuracy of these questionnaires (measurement issues)?
- How would you ensure that you were not missing critical data (measurement issues)?
- How would you verify that the questions were answered the similar ways across doctors and patients (instrument and consistency issues)?

Step 3: Class Discussion (15 minutes)
- What kind of data quality problems require remedies before data collection?
- What kind of data quality problems require remedies after data collection?