**EPI-USE - Intern Project by Alec Han**

**Purpose:** Outline the required structure and constraints for using the Time Compare properly. It ensures proper processing, comparison logic, and accurate outputs.

1. **Required Inputs**  
   The tool requires two input files:
   * Legacy System File
     + File format must be .csv or .xlsx
     + Must contain the following columns with these exact names:
       - EMPLOYEE\_ID
       - WORK\_DATE (date only)
       - START\_DATE (date only)
       - START\_TIME (time only)
       - END\_DATE (date only)
       - END\_TIME (time only)
       - ACTIVITY\_CODE (categorical identifier for type of work)
       - DURATION\_MINUTES (numeric, float or int)
   * WorkForce Software (WFS) File
     + File format must be .csv or .xlsx
     + Must contain the following columns with these exact names:
       - EMPLOYEE\_ID
       - WORK\_DATE
       - START\_DTTM (combined datetime of start time)
       - END\_DTTM (combined datetime of end time)
       - PAY\_CODE (WFS equivalent of activity code)
       - HOURS (float, can be used to backfill missing duration)

Note: The tool automatically splits START\_DTTM and END\_DTTM into START\_DATE, START\_TIME, END\_DATE, and END\_TIME internally (this is how the tool compares the two data sets).

1. **Mapping File (Optional)**  
   If using the “Upload Mapping File” option, your mapping file must:
   * Be .csv or Excel format
   * Contain exactly two columns:
     + ACTIVITY\_CODE\_legacy
     + PAY\_CODE\_wfs

Rules:

* Every legacy activity code must appear once
* If a legacy activity code maps to multiple WFS pay codes, the code must be labeled as “OTHER” in the ACTIVITY\_CODE\_legacy column, and the associated WFS codes should appear in separate rows.

1. **Threshold for Comparison**
   * The user must specify a threshold between 0.00 and 1.00
   * This represents the minimum similarity required for a row to not be flagged as a mismatch (e.g., 0.98 = 09% similarity)
   * Applies to both:
     + Numeric comparisons (e.g., duration)
     + Floating-point time comparisons (for START\_TIME and END\_TIME after they are converted from timedelta objects)
2. **Matching Rules**
   * DURATION\_MINUTES values are compared relatively using the formula: abs(legacy – wfs) / legacy > (1 – threshold)
   * NaN values are handled explicitly. A mismatch is flagged if one system has a value and the other has NaN.
   * ACTIVITY\_CODE and PAY\_CODE are matched via a user-defined mapping, not simple equality.

**Limitations/Warning**

* Both files must have the same number of rows, and records must be in the same order.
* The tool does not perform partial matching or sorting. It assumes the datasets are aligned row-wise (i.e., record 1 in the legacy system is record 1 in WFS).
  + Suggestion for the future: Order the data sets by EMPLOYEE\_ID in ascending order then for any ties sort by the WORK\_DATE in ascending order.
* Field names are case-sensitive. Make sure to follow the exact casing and naming conventions described above.