Technical Challenge (Capitole RD Sr. Data Scientist)

This challenge involves working with a time-series dataset containing **weather parameters*** such as precipitation, temperature, wind features, etc. The participant must tackle one (encouraged to do both) of the two key tasks:

1. **Data Imputation:** Simulating missing data and developing methods for reconstruction while **comparing results with established academic findings**. Why is this relevant? Data Quality precautions will always be required in any project. Building and integrating these methods will ensure our continuous integration of these into several different projects.
2. **Dimensionality Reduction and component participation over time:** Understanding how components behave over time. The motivation of this comes from the fact that the system's time-wise behavior might change over time and some components might be more relevant at a given point, while others remain silent. Say you have successfully reduced (with novel or traditional methods) the space to three components. What is their contribution to the overall system's time-wise behavior? Do they switch states often? Tip: sliding-window techniques might be your friend. **Compare and contrast results with what the state of the art says**

This Kaggle data set is a starting point for motivation. If you find other(s) that are better for the tasks described, feel free to use them.

Task 1: Data Imputation

Objective

- Randomly remove parts of the dataset and use imputation techniques to reconstruct missing values. Remember, if variables are correlated, that might give you a hint of where to start. Avoid simple imputation techniques such as mean or mode.
- Compare results and analyze them in relation to existing research.

Implementation Requirements

1. **Amputation:** Randomly remove a percentage (say 10%) of all time-series data points. Doing it randomly avoids bias and will ensure a sound statistical performance analysis.
2. **Imputation Methods:** Implement at least two imputation techniques. One is your baseline (linear, cubic, etc.) and the other should be using an ML model / AI method (Regression, LSTM, etc.).

    TIP: If variables present statistical dependencies, you might want to use this information to impute *y* based on collection of *X*. Here *X* is the entire data set minus one variable to be imputed, namely, *y*. You might've already guessed that this should be an iterative process so all variables can be imputed. The challenge here is that given that *X* will be pruned (by

randomly removing fields), you need to come up with a way to initialize this algorithm. This is not a new field, so do doing research will help.

3. **Performance Evaluation:** Compare imputed vs. actual values using and benchmarking the applied methods. To do this, it is important to store your actual deleted values, as they are now your ground truth.
4. **Visualization:** Show actual vs. imputed values and compare methods
5. **Theoretical Reflection:** Candidates must link results to literature.

Task 2: Dimensionality Reduction for Understanding time-wise state transition

Objective

- Use dimensionality reduction techniques to identify core states / components in a time-wise manner. TIP: Once you've featured-engineered what each point in time means (a vector, a matrix etc). Apply some unsupervised learning techniques to such data to explore how uncovered clusters behave over time. Each cluster then is a proxy for your state.
- Evaluate findings in the context of these methods.

Implementation Requirements

1. **Apply dimensionality reduction / clustering techniques**: Show results of your approach in a clear and straight-to-the-point manner. TPM (transition probability matrices) applied to time series might give you a hint (a hint really, not a requirement) of where to start. Plotting states over time and explaining how they behave might also be a good idea.
2. **Theoretical Reflection:** Again, candidates must compare findings with academic literature.

Deliverables

Participants should submit results in **one or more** of the following formats:

1. **PowerPoint Presentation:**
    - Clearly explains methodology, results, and their links to what academia says about the matter.
2. **Python Notebook (Jupyter):**
    - Implements model and processing pipelines, visualizes results, and includes academic discussion.
3. **EXTRA: Live Demo (Streamlit/Gradio) (Optional but Encouraged):**
    - Streamlit or gradio might be your friends here

Final Note

Remember, you must **explicitly connect findings to academic research**. You should **not just present results** but also **explain how these compare with past studies**, identifying consistencies, discrepancies, and possible reasons for deviations as well as future areas of interest.