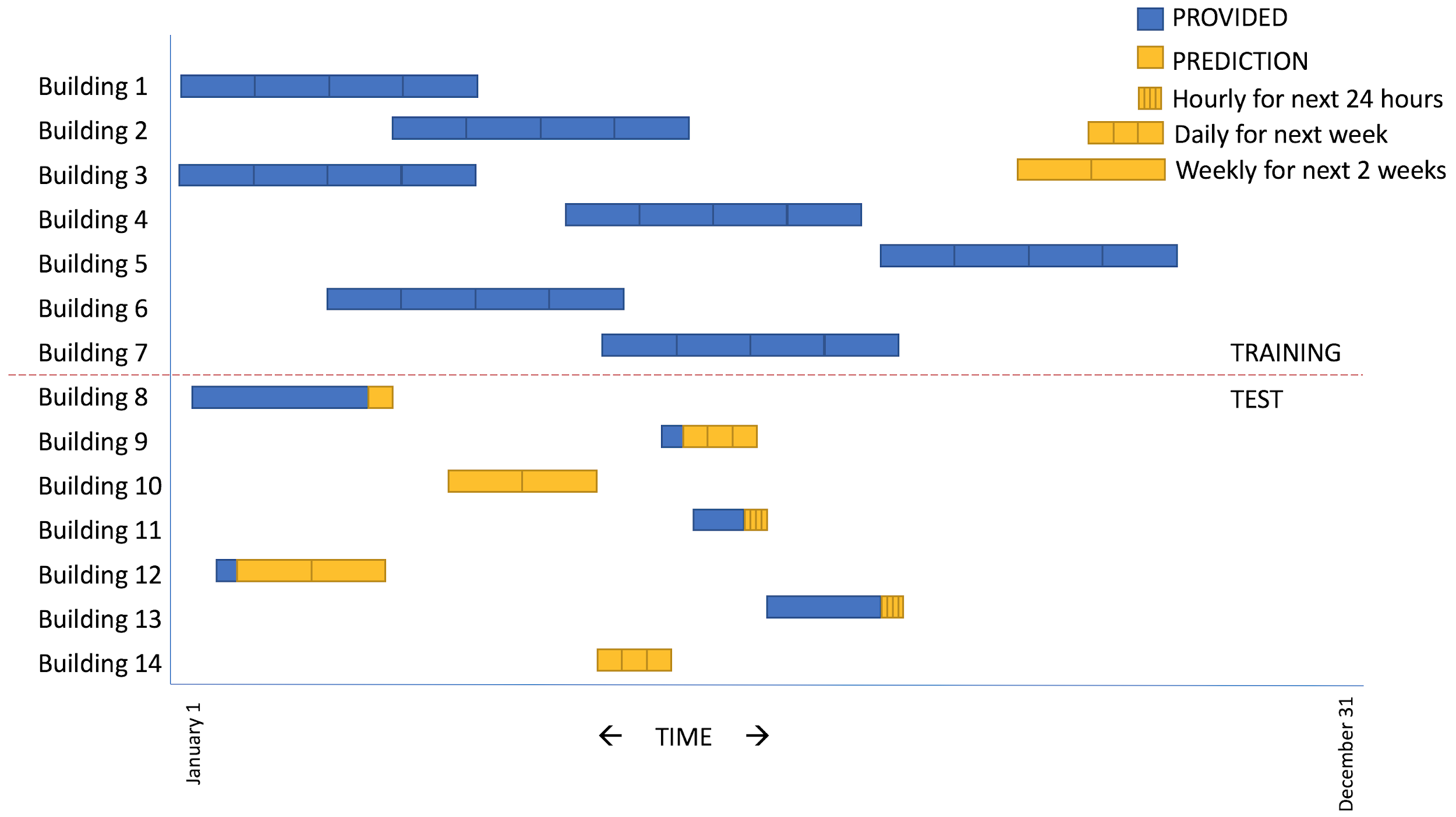
**Data Science Task for METCO**

The objective of this task is to forecast energy consumption from varying amounts of "cold start" data, and little other building information. That means that for each building in the test set you are given a small amount of data and then asked to predict into the future. Since this format of test data is a bit unusual, we'll give an overview here, with more detail below. Here is a visual representation of the data setup for this problem:



**Overview:**

Three time horizons for predictions are distinguished. The goal may be either:

* To forecast the consumption for each hour for a day (24 predictions).
* To forecast the consumption for each day for a week (7 predictions).
* To forecast the consumption for each week for two weeks (2 predictions).

In the test set, varying amounts of historical consumption and temperature data are given for each series, ranging from 1 day to 2 weeks. The temperature data contains a portion of wrong / missing values.

In the training set, 4 week series of hourly consumption and temperature data are provided. These series can be used to create different cold start regimes (varying amounts of provided data and prediction resolutions) for local training and testing.

Basic building information such as surface area, base temperature, and on/off days are given for each series in the training and test sets.

**Datasets:**

**Historical Consumption**

Time series data of consumption and temperature data identified by their series\_id.

Training Data.csv

* series\_id - An ID number for the time series, matches across datasets
* timestamp - The time of the measurement
* consumption - Consumption (watt-hours) since the last measurement
* temperature - Outdoor temperature (Celsius) during measurement from nearby weather stations, some values missin

**Buildings Metadata**

Additional information about the included buildings in the train and test set.

Building Metadata.csv

* series\_id - An ID number for the time series, matches across datasets
* surface - The surface area of the building (ordinal)
* base\_temperature - The base temperature that the inside of the building is set to (ordinal)
* monday\_is\_day\_off - Whether or not the building is operational this day
* tuesday\_is\_day\_off - Whether or not the building is operational this day
* wednesday\_is\_day\_off - Whether or not the building is operational this day
* thursday\_is\_day\_off - Whether or not the building is operational this day
* friday\_is\_day\_off - Whether or not the building is operational this day
* saturday\_is\_day\_off - Whether or not the building is operational this day
* sunday\_is\_day\_off - Whether or not the building is operational this day

**Cold Start Test Set:**

Test data used to start a forecast. Includes metadata about prediction window as well as time series data on consumption. Being time series data, the test data is used to initiate the predictions required for the submission file.

Test Data.csv

* series\_id - An ID number for the time series, matches across datasets
* timestamp - The time of the measurement
* consumption - Consumption (watt-hours) since the last measurement
* temperature - Outdoor temperature (Celsius) during measurement from nearby weather stations, some values missing

**Desired Predictions:**

Predictions.csv defines the required predictions. series\_id together with timestamp is unique (no timestamp appears twice in the same series). We have provided the index pred\_id for simplicity. Additionally, there is the consumption column **where your hourly, daily, or weekly predictions will go**. The temperature column reports the average temperature for that time period. The prediction\_window column includes how often predictions need to be made (can be inferred from the timestamps).

**Guidelines:**

* The goal of this model is to predict energy consumption, based on some historical information (i.e., ae “cold start”) In the training process you should provide the cold start yourself, in the validation process it is given in the test dataset.
* The goal of the validation process is to validate the model’s feasibility - i.e., its ability to provide predictions given a test dataset.
* The problem description and the data at hand are not our operational problem/data, but they do describe similar characteristics to the problems solved in METCO.
* You may either use R or Python 3.x

**Solution:**

Your solution will be presented during the interview. Please prepare some graphic aids for that matter.

You are also required to submit us the solution. Please follow the submission guidelines:

* Submit a single PDF file.
* The PDF file must include all source code you’ve used. Please organize the code chronologically.
* Add auxiliary results and graphics in the code, where they are relevant, at the places they were generated.
* Add all graphics presented in the interview.
* Add answers to the following questions:

1. What are the results of your descriptive statistics?
2. How did you manipulate the data, and why? Illustrate your answer with plots.
3. How did you model the problem, and why?
4. How did you address the timeliness property of the problem?
5. How did you evaluate your model? What were the results of the evaluation?
6. Provide a brief plan description of solving the problem with a different method than you did (e.g.,you solved it with LSTM, and provide a plan for solving it with GBM, etc.)
7. If you had extra time, what would you do next?

**Good Luck!**