 **NAMAL UNIVERSITY**

**Department of Computer Science**

**BS Computer Science Program**

**6th Semester, Session 2024 Semester Project**

**CS-352 Data Analysis and Visualization**

**Data Analysis and Visualization Project Proposal**

# Dataset Source and Description

We are using the **PIRvision Office Dataset**, a publicly available dataset containing timestamped PIR (Passive Infrared) motion sensor readings along with occupancy labels. The dataset captures occupancy patterns in a real-world office setting and was used in the research paper:

**"Promoting Occupancy Detection Accuracy Using On-Device Lifelong Learning"**

In this paper, the authors developed a **lifelong learning model** for occupancy detection called **LL-E (Lifelong Learning with Experience replay)**, designed to improve model accuracy over time by continually adapting to new data in a lightweight manner suitable for edge devices like Raspberry Pi. In our case, we will implement the same model on a standard computing environment (PC) to replicate and evaluate the results.

# Research Questions

This project aims to answer the following research questions using data analysis and machine learning techniques:

**1. Can we successfully reproduce the Lifelong Learning model (LL-E) proposed in the research paper and evaluate its performance on the PIRvision dataset?**

We will implement the LL-E model using Python and evaluate its performance using accuracy, precision, recall, and F1-score.

**2.What are the patterns of occupancy throughout the day and across different weekdays?**

Using Exploratory Data Analysis (EDA), we will visualize trends using line plots, heatmaps, and box plots to identify peak occupancy times and usage behavior over the week.

**3.How do baseline models like KNN, Decision Tree, Naive Bayes, and SVM perform on the PIRvision dataset compared to the LL-E model? Can additional models such as logistic regression or random forest provide further insights?**

We will compare multiple models' classification performance to evaluate the effectiveness of LL-E and identify any models that perform competitively or better under certain conditions.

**(Optional – Time permitting)**

**4.Which PIR sensors or features contribute most significantly to predicting occupancy?**

Feature importance techniques such as those provided by tree-based models (e.g., XGBoost) or SHAPvalues will be used to interpret the model's decision-making.

# Preliminary Thoughts on Challenges and Solutions

 **Model Details Missing**: If full algorithmic details are missing in the paper, we will reconstruct the LL-E model based on described methodology and adapt it as needed.

 **Time-based Data Analysis**: Occupancy data is time-sensitive; careful time-indexed pre-processing and resampling will be applied.

 **Validation**: We will split the dataset into training and testing sets to validate generalization.