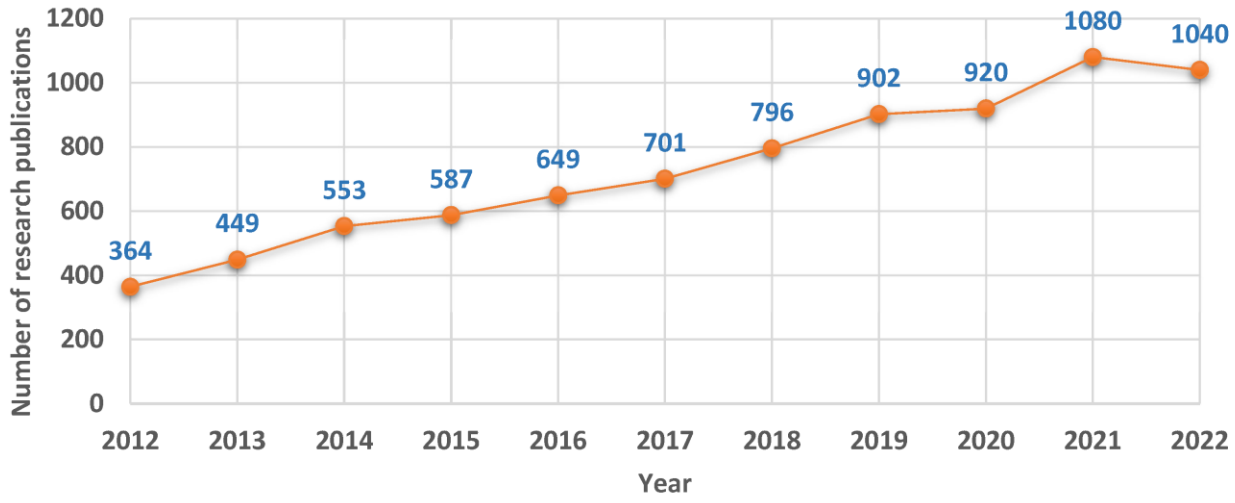
DATA SCIENCE CHALLENGE | ACTIVITY RECOGNITION

Pedro Matias

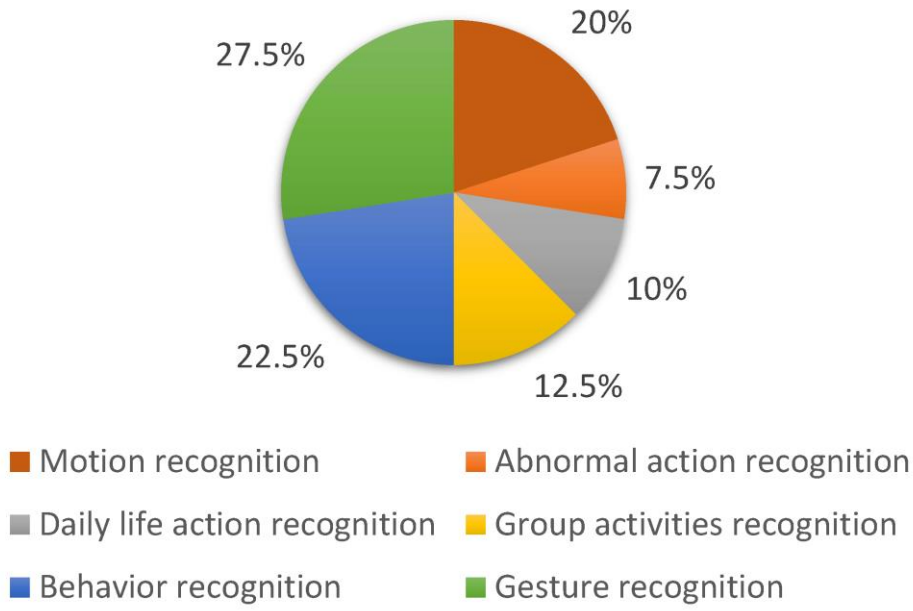
DATA SCIENCE ASSIGNMENT

MOTIVATION

Number of publications over time



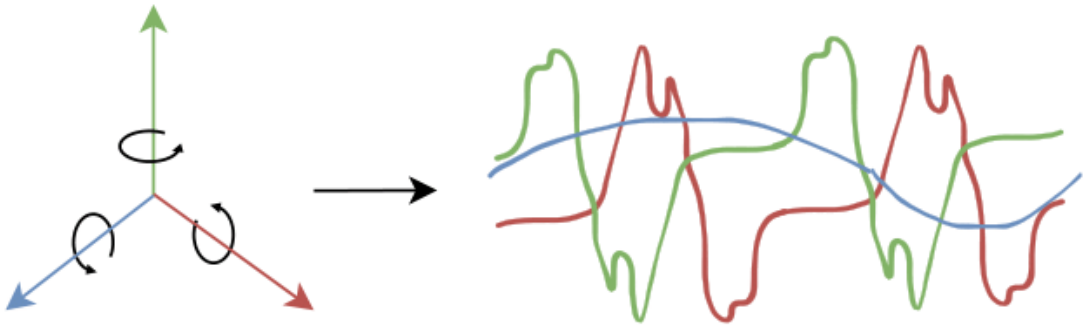
Downstream Task



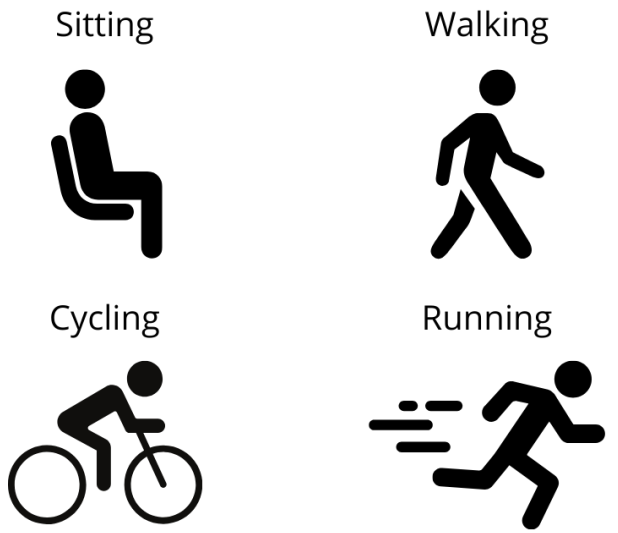
DATA SCIENCE ASSIGNMENT

PROBLEM

Accelerometer Data



Target Daily Activities



DATA SCIENCE ASSIGNMENT

DATASET

Characteristics

A. 13 users

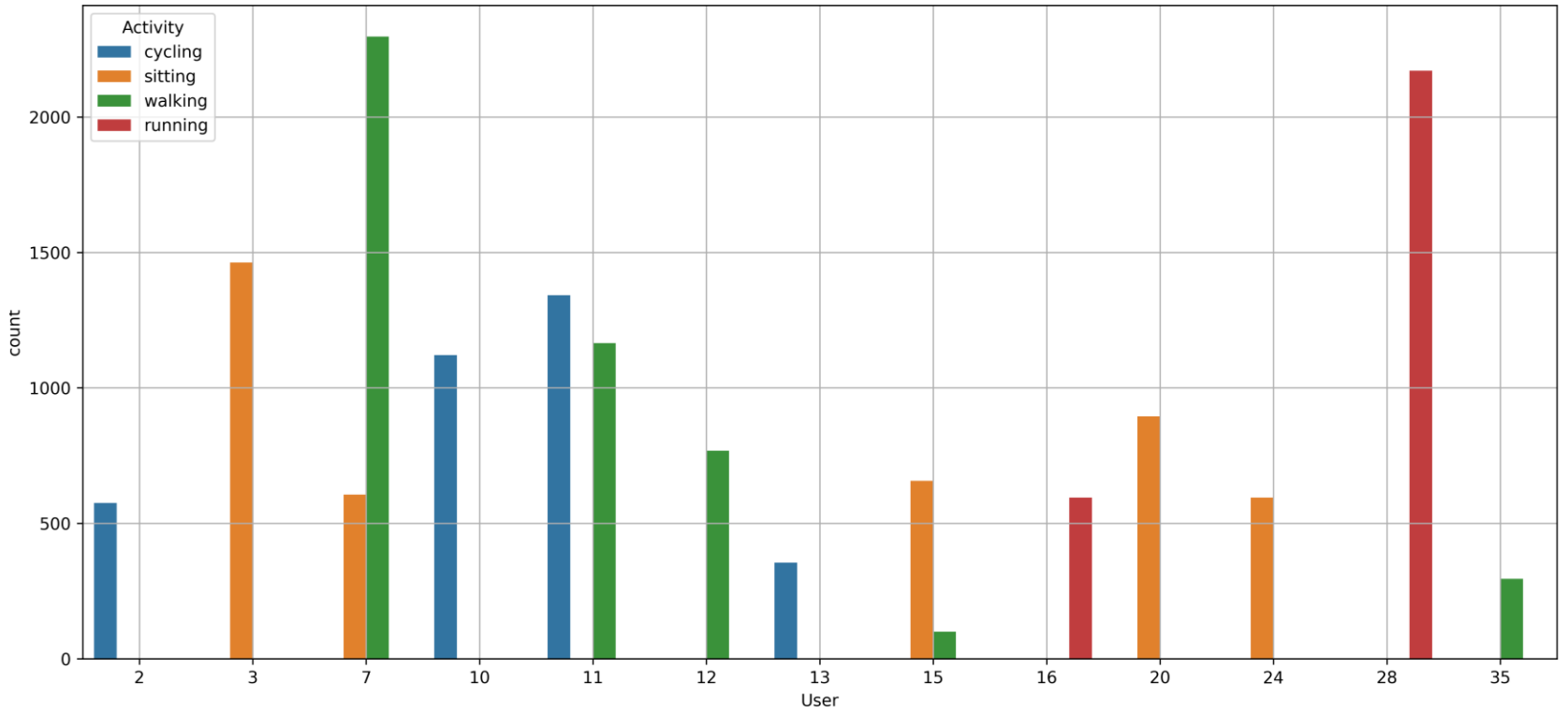
B. 24 sessions

C. 4 human activities

D. 10.7 ± 4.4 min (duration)

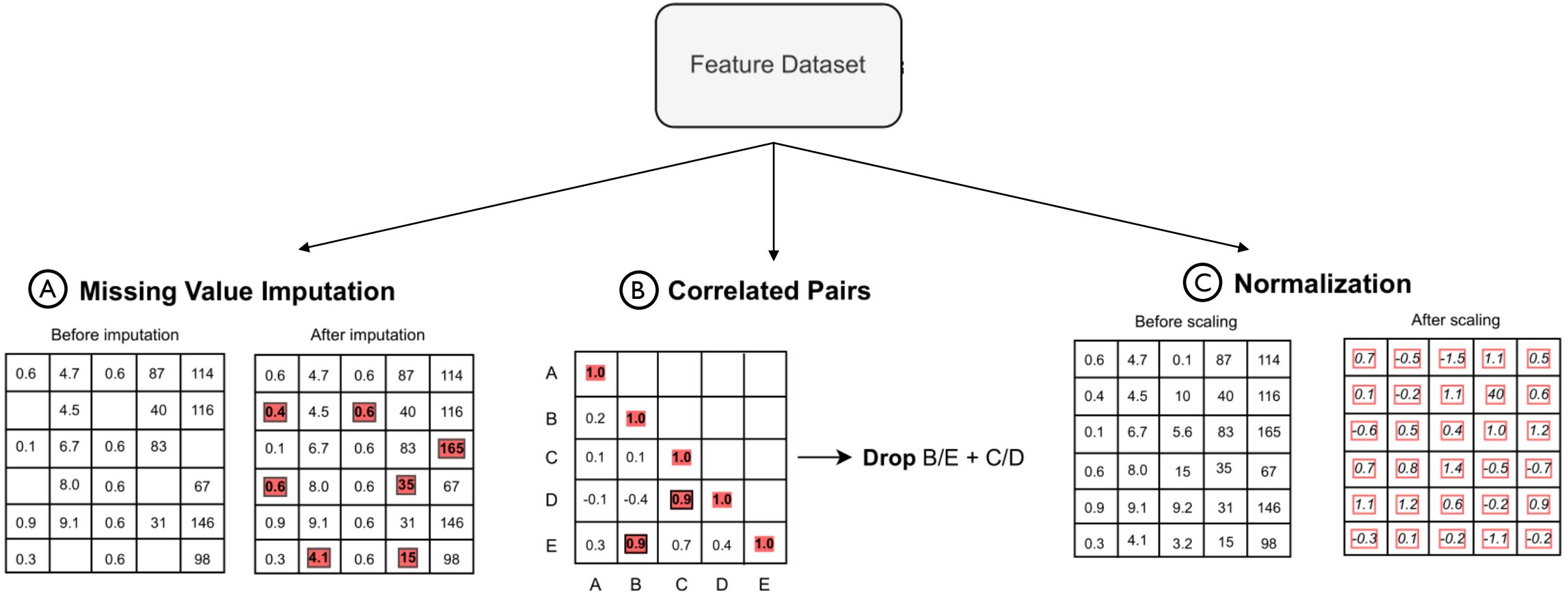
E. 15 000 samples

F. 40 features



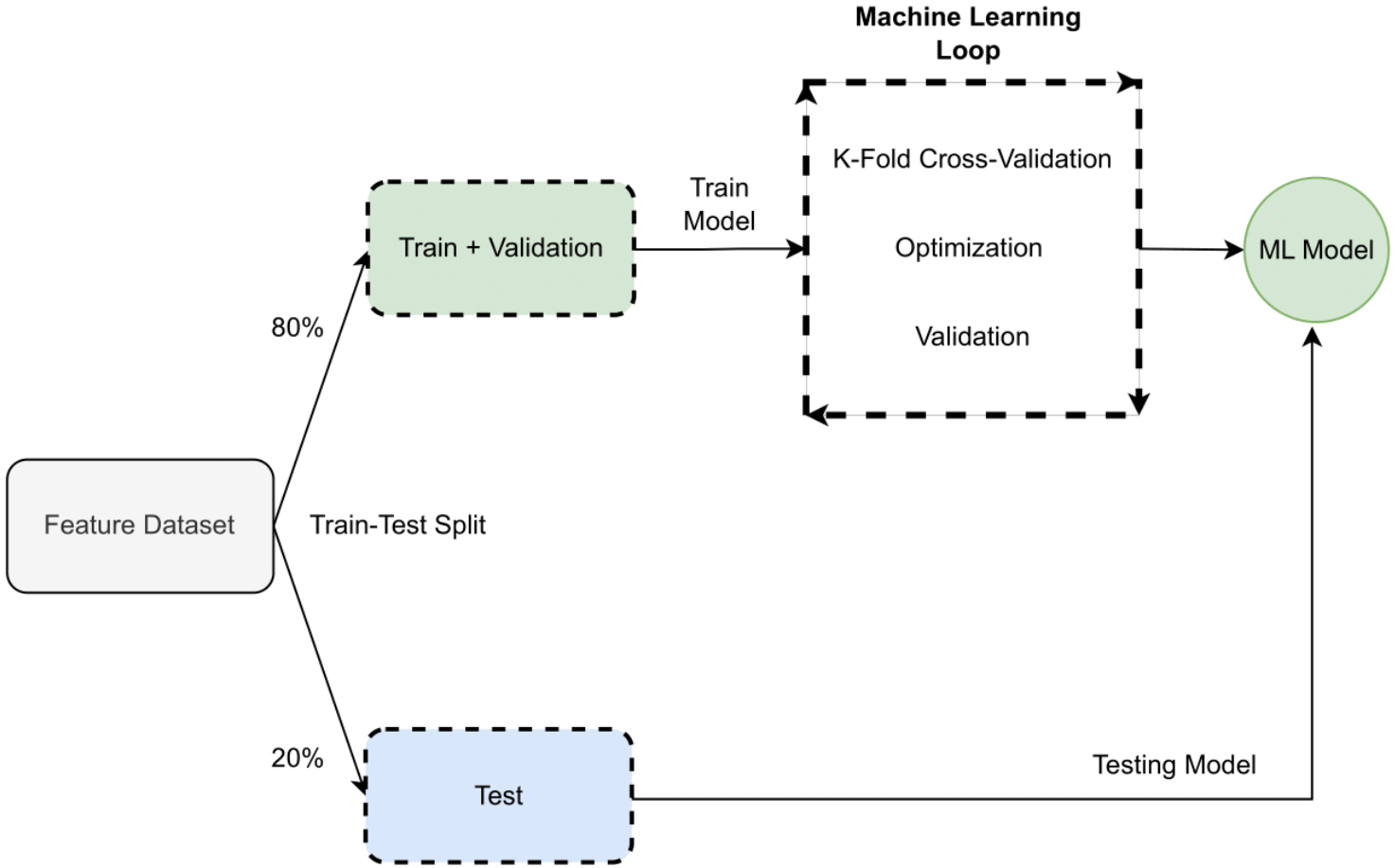
DATA SCIENCE ASSIGNMENT

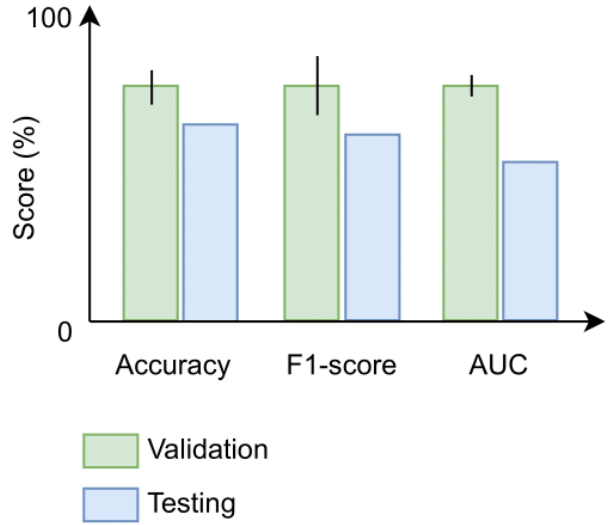
PREPROCESSING PIPELINE



DATA SCIENCE ASSIGNMENT

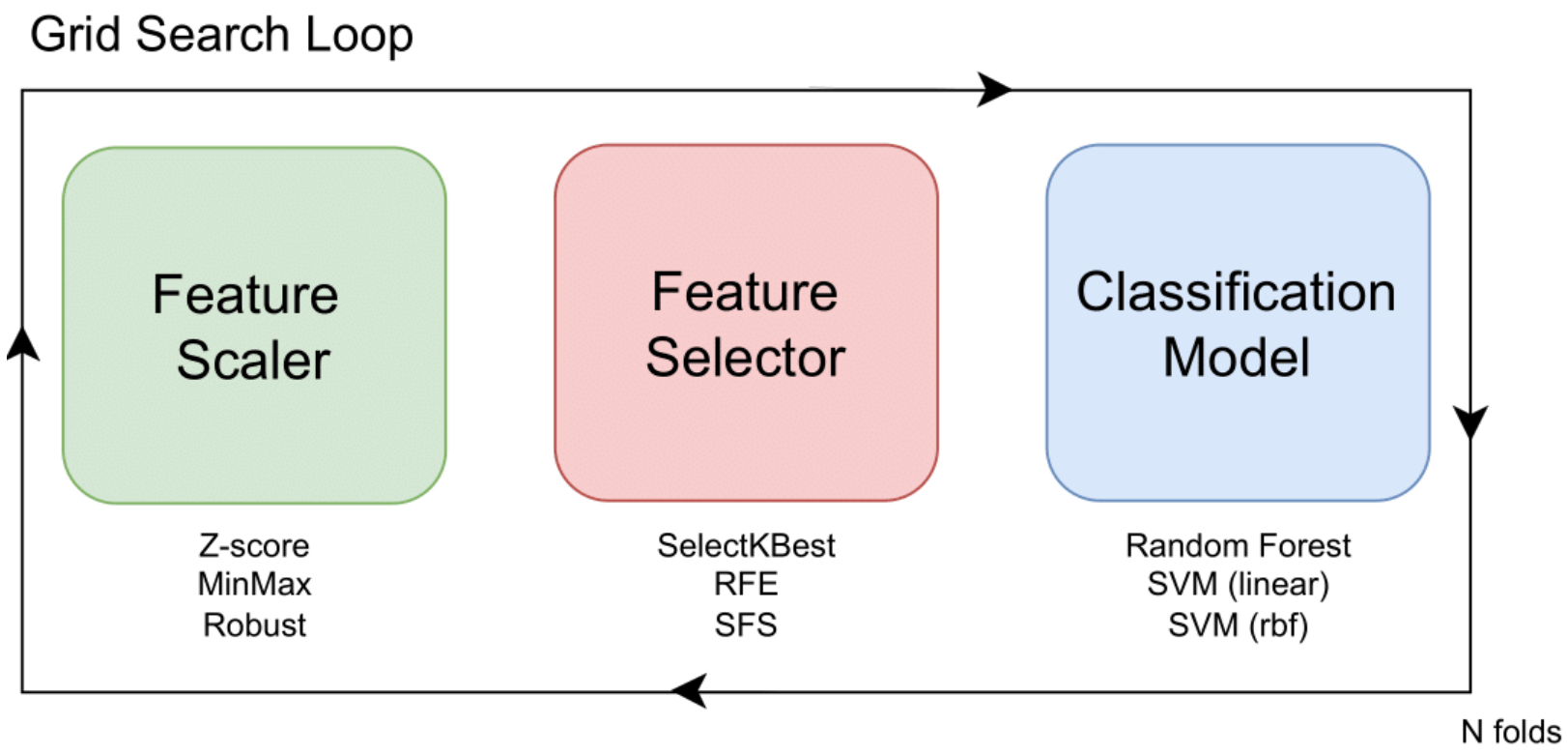
ML PIPELINE





DATA SCIENCE ASSIGNMENT

GRID-SEARCH PIPELINE



Notes

5 folds (80/20%)

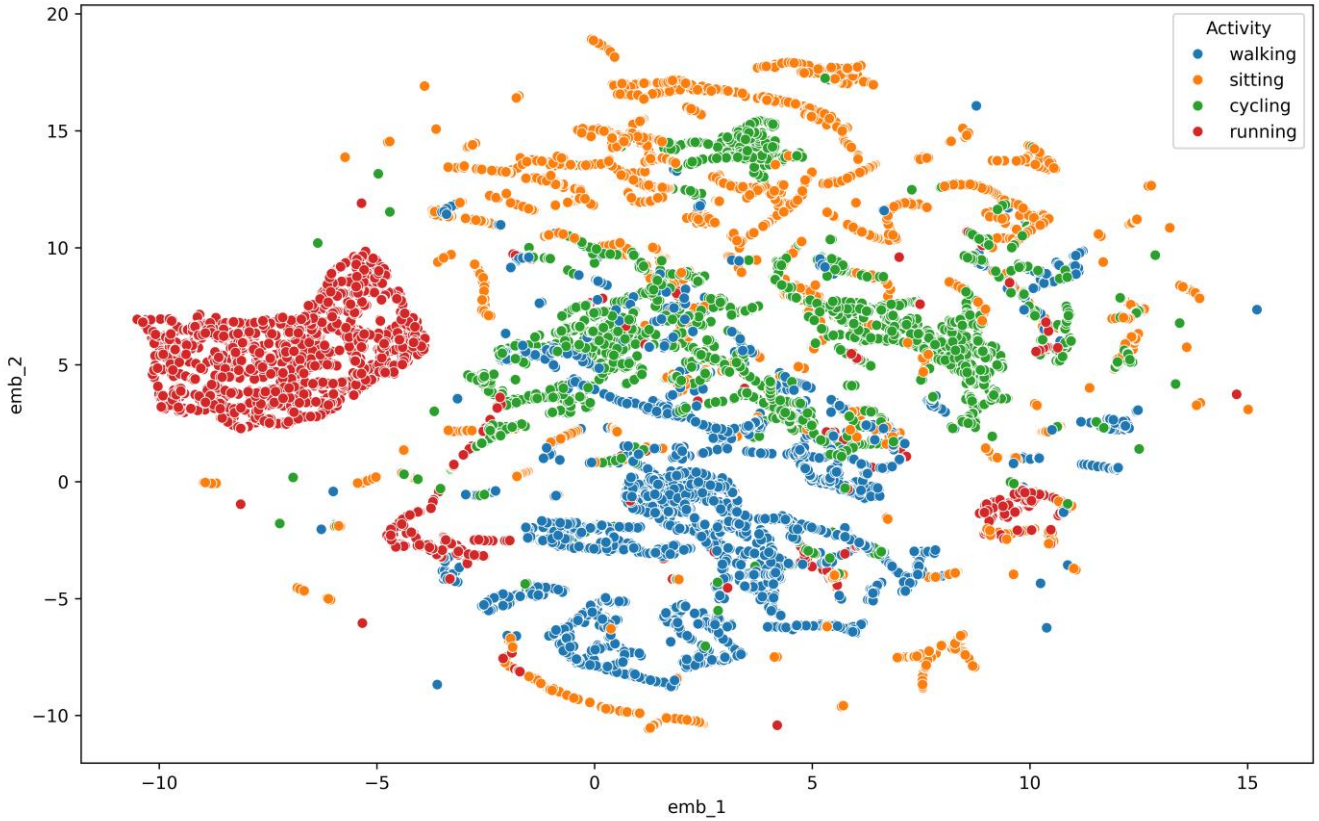
Split by Session

Optimization by F1-score (macro)

DATA SCIENCE ASSIGNMENT

EXPLORATORY DATA ANALYSIS

Normalization + UMAP 2D



Notes

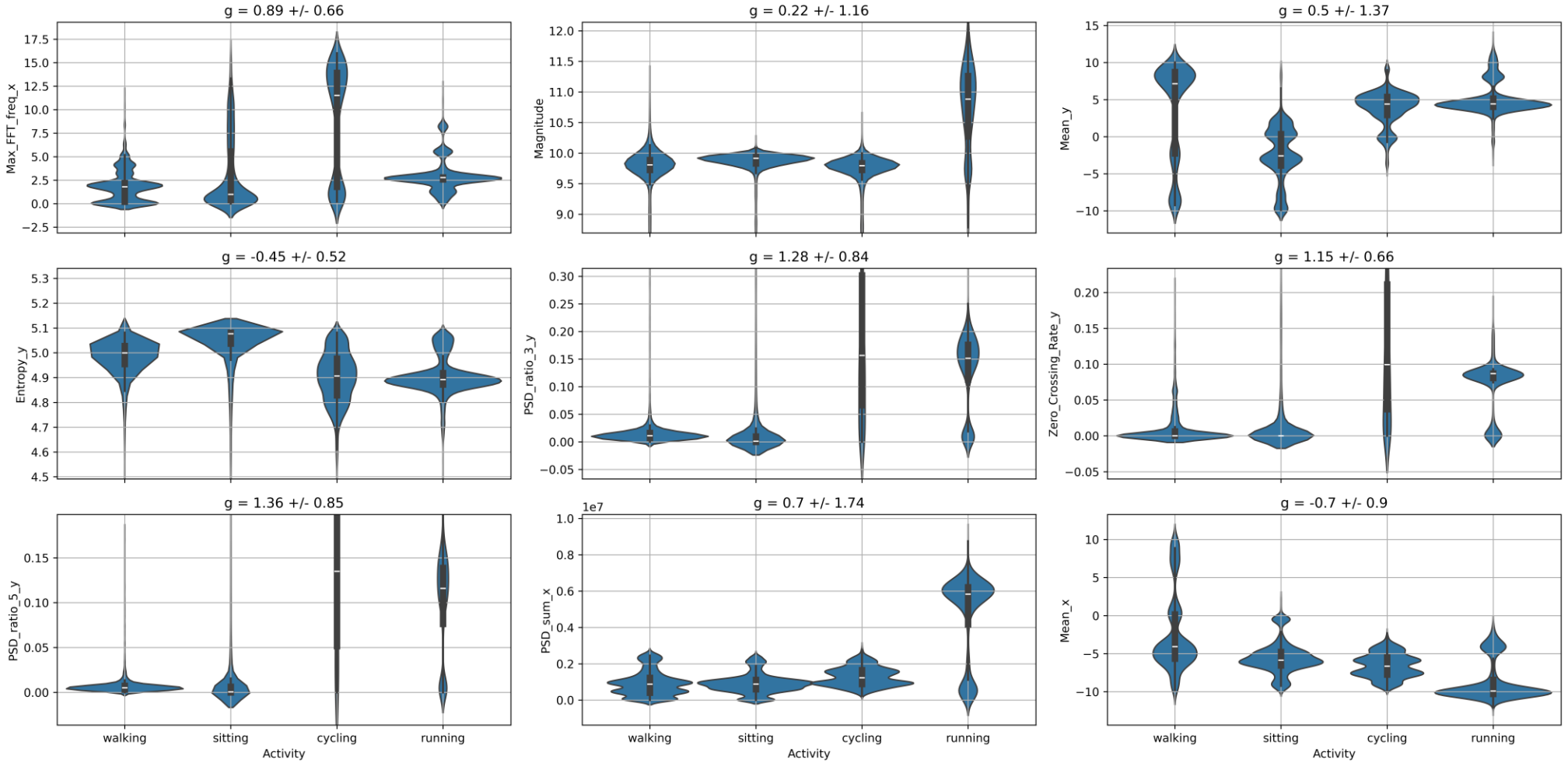
Running potentially easier to identify;

Walking and **Sitting** seem to be distinguishable as well;

Cycling and **Sitting** seem to be more overlapped. Physiological component can play and important role here.

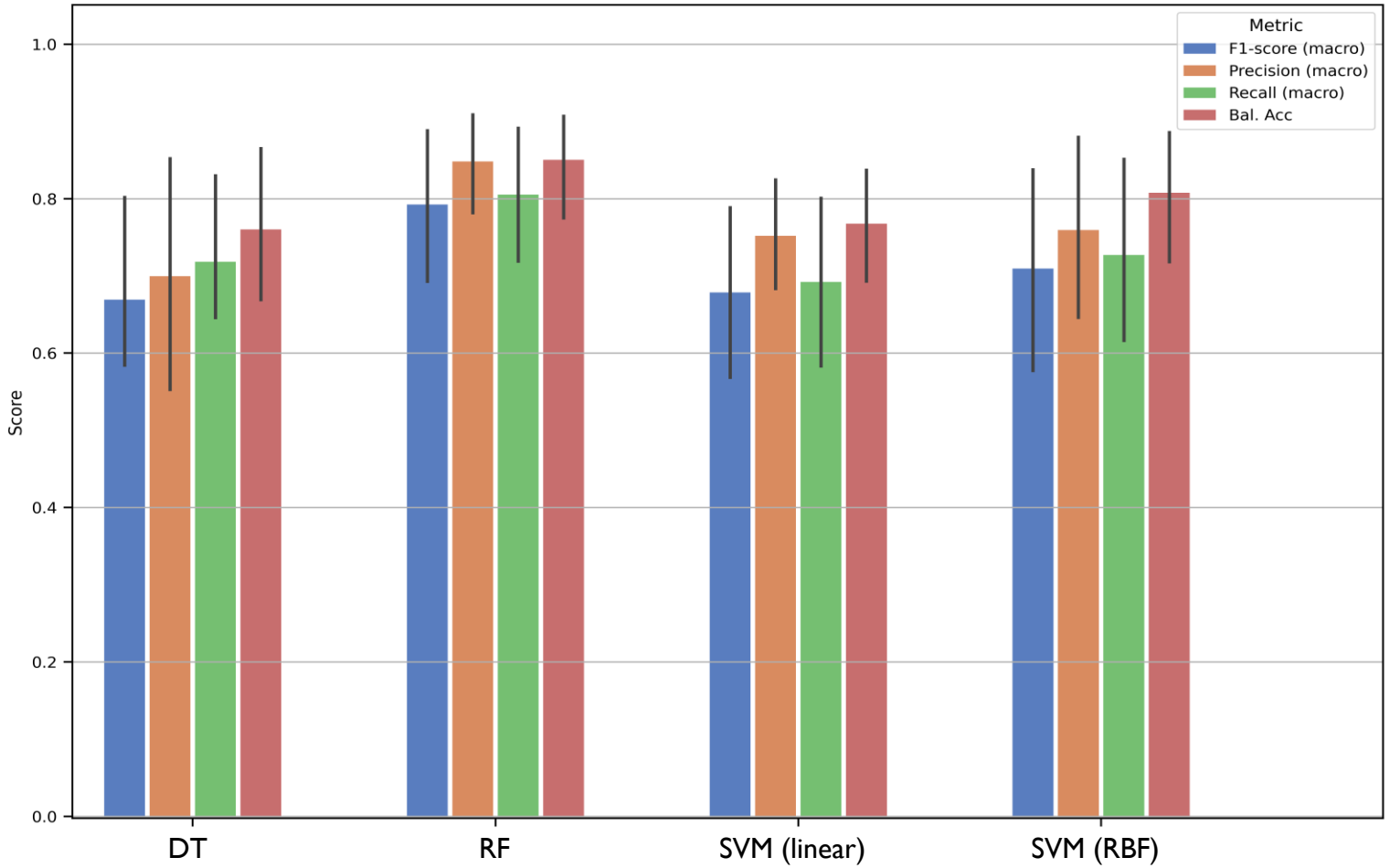
DATA SCIENCE ASSIGNMENT

EXPLORATORY DATA ANALYSIS



DATA SCIENCE ASSIGNMENT

ML CLASSIFICATION RESULTS



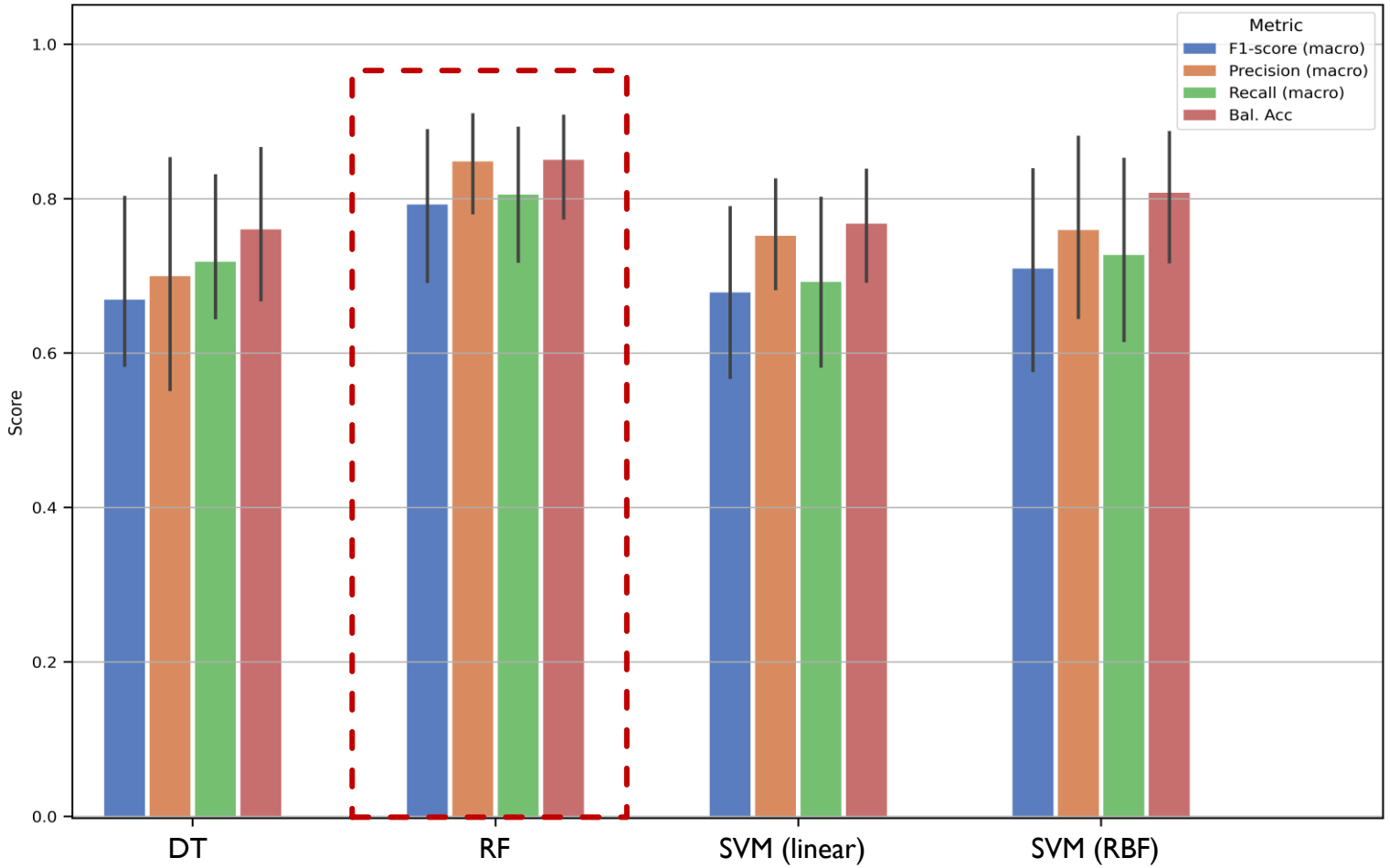
Remarks

Random Forest performs better (79.2% ± 12.5% F1-score)

To deploy: model should be trained with all data, and a single external testing set should be used to report.

DATA SCIENCE ASSIGNMENT

ML CLASSIFICATION RESULTS



Remarks

Random Forest performs better (79.2% ± 12.5% F1-score)

To deploy: model should be trained with all data, and a single external testing set should be used to report.

DATA SCIENCE ASSIGNMENT

PERFORMANCE IN REAL-WORLD SCENARIOS

- **Performance** of ML models in **free-living** scenarios often **decreases** when compared with validation scores;
- Some **features** are not **position-invariant.** This can affect real-world performances, if new incoming samples are collected with different device orientations;
- **Reporting** model performances **transparently** is crucial when deploying model into the real-world. **Some examples**: female vs male, young vs adult vs elderly, diseased vs healthy, device position, walking style;
- **To improve performance:** opening the context window (post-processing) may help reduce model failures (e.g., classification consistency over N consecutive windows; steps detected in walking windows; intra-user variability may help figure out most likely activities in specific times of day);
- **Other approaches**: **position-invariant** features (e.g., same metrics over signal magnitude), **more activities** (e.g., stand-to-sit/sit-to-stand; jumping; laying; indoors vs outdoors), **ensemble methods** or layered learning (model A to detect activity levels, model B to identify model ensembles), **DL models** (not in real-time but for offline processing).

THANK YOU

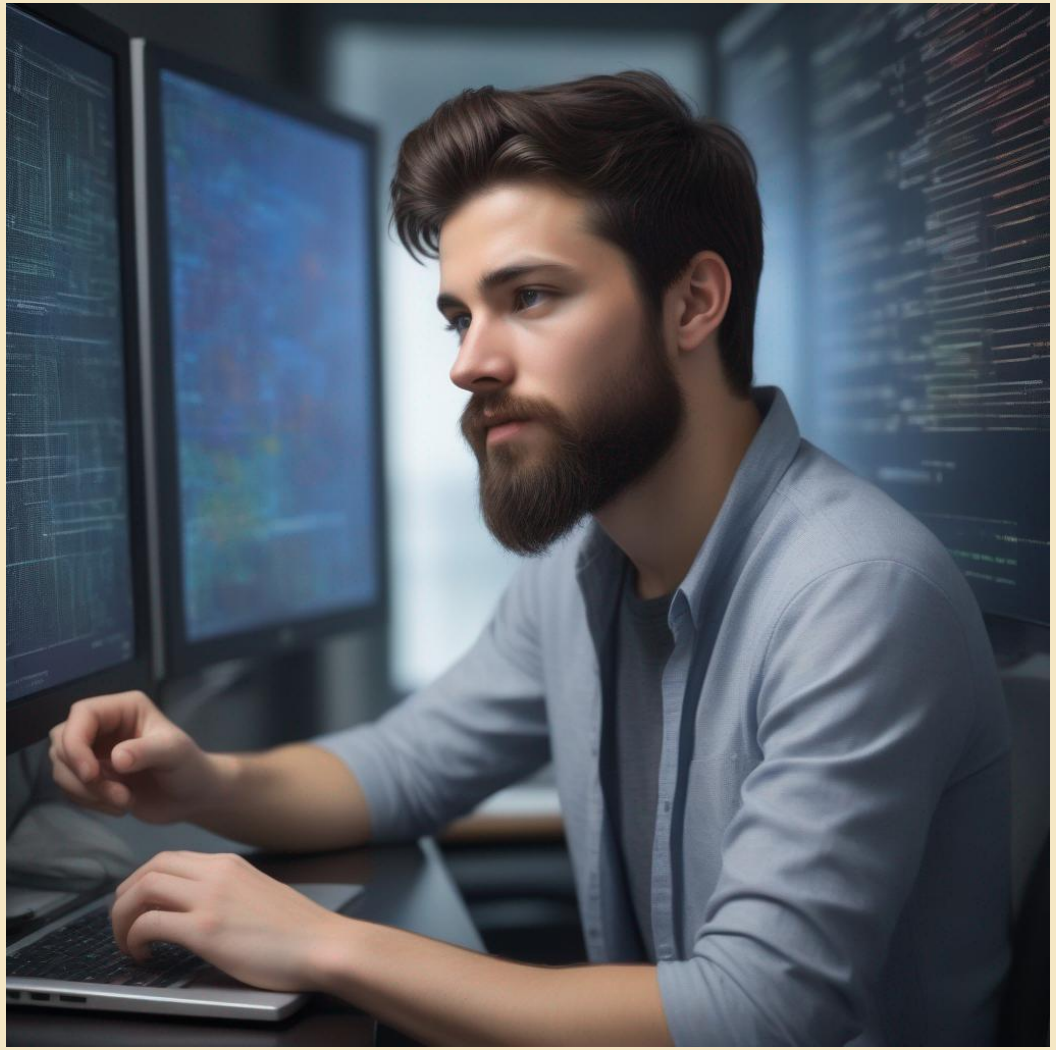
Contacts

matiaspedro97@gmail.com

Pedro Matias

github/matiaspedro97

pedromatias



dataautogpt/OpenDalleV1.1