LÄNDLE MEETS BUDAPEST

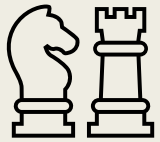
Aisha, Alex, Alik, Braian

TU Vienna – Hackathon „Watt‘s Up“ 2025

OUR TEAM:

AISHA AND BRAIAN
FROM “LÄNDLE”
&
ALEX AND ALIK
FROM BUDAPEST

Objective



- Generate a synthetic dataset that mirrors the statistical patterns of real energy usage data of 8760 hours (a full consecutive year)
- Ensure individual user profiles remain unidentifiable in the synthetic dataset
- Validate the trustworthiness of the synthetic data through rigorous testing to provide evidence that the info on single profile levels cannot be inferred from the real data

Approach #1

COMPRESS DATA

DESENSITIZE PERSONAL INFO

ANALYZE PATTERNS

DECOMPRESS DATA

Approach #2

PREDICT FUTURE CONSUMPTION

GENERATE SYNTHETIC DATA

VALIDATE ACCURACY

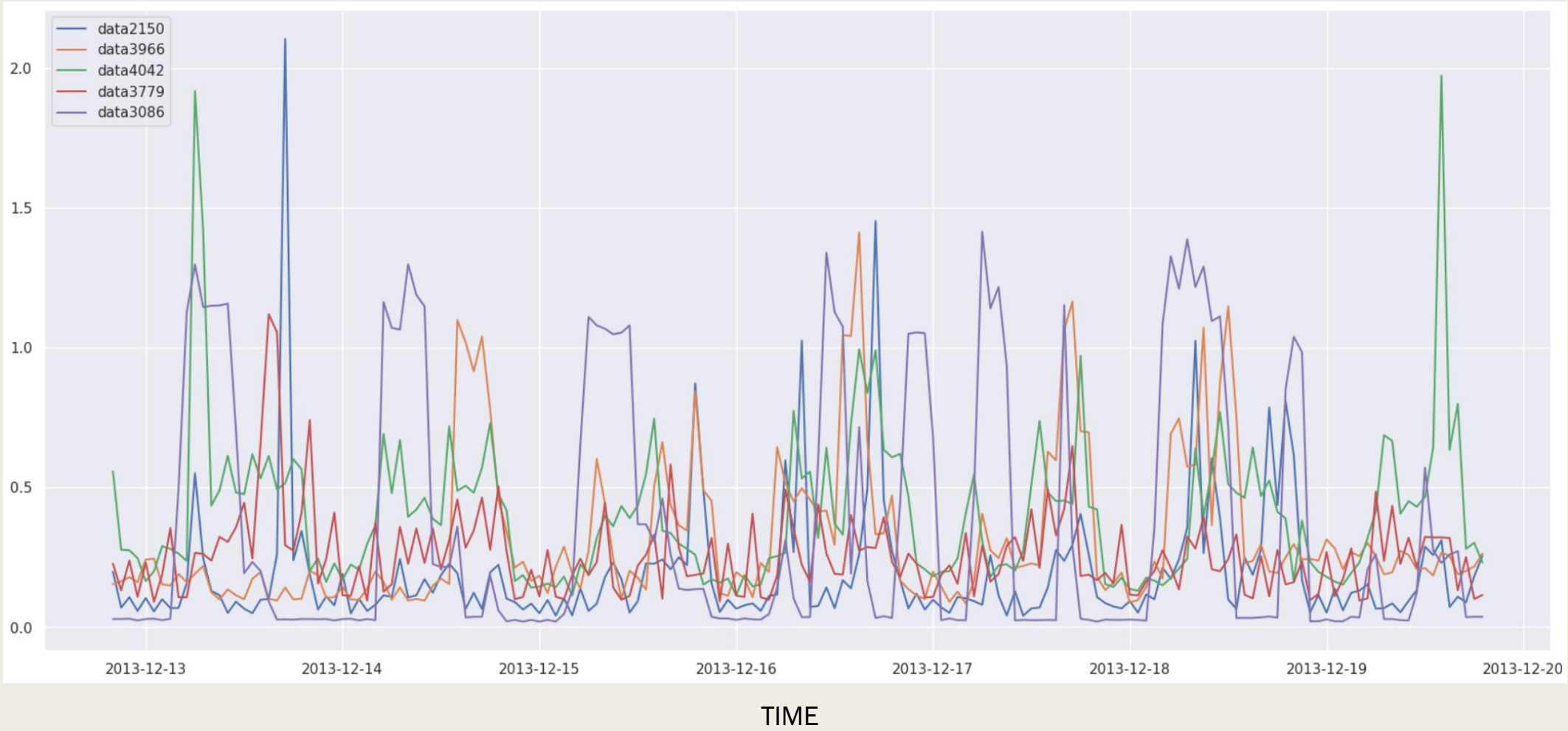
IMPLEMENTATION

Resources

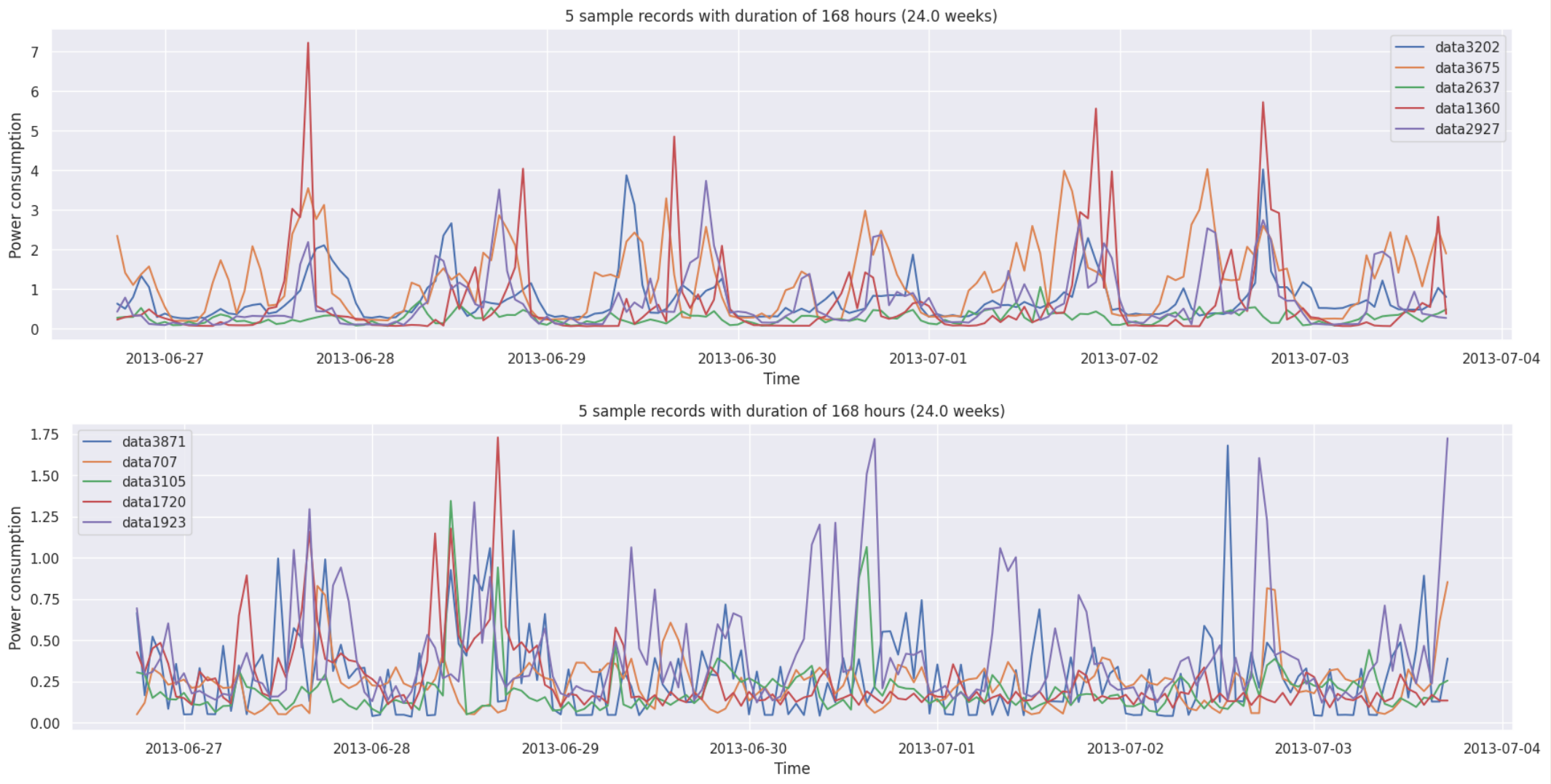
- For this project, following applications were used:
  - *Google Colab*
  - *Jupyterhub*
  - *Github*
  - *Perplexity*
  - *Dependencies*
    - PyTorch
    - NumPy
    - pandas
    - Tensorflow
    - scikit-learn

Link to our GitHub:
https://github.com/am11001/watt_s_up

Sample Data – 5 random profiles in a week



Sample Clastered Records



Different approaches, same goal

Privacy-Preserving Compression

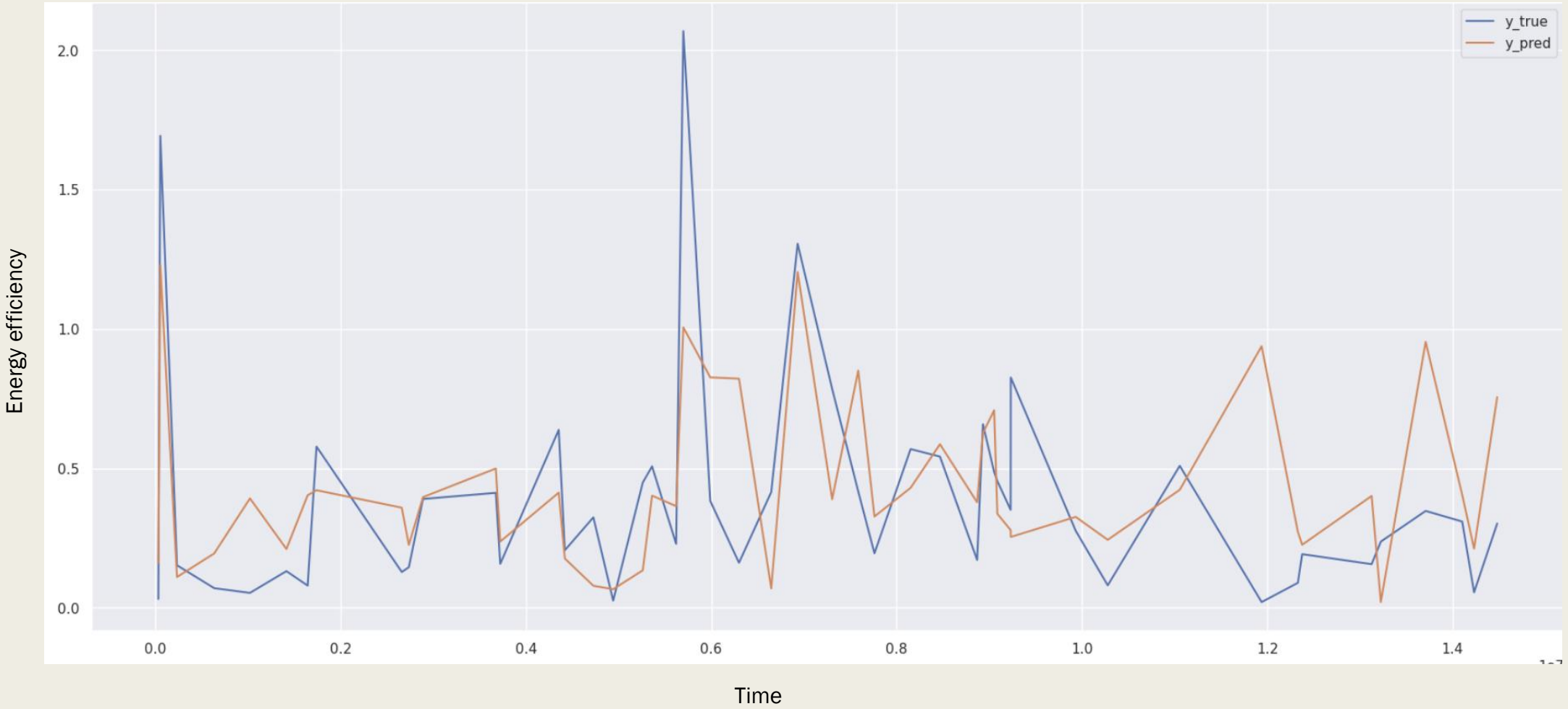
- Focuses on reducing data size while trying to ensure personal information is desensitized.
- Through the regeneration of the decompressed data after desensitization, we create synthetic data

Predictive Accuracy with Synthetic Data

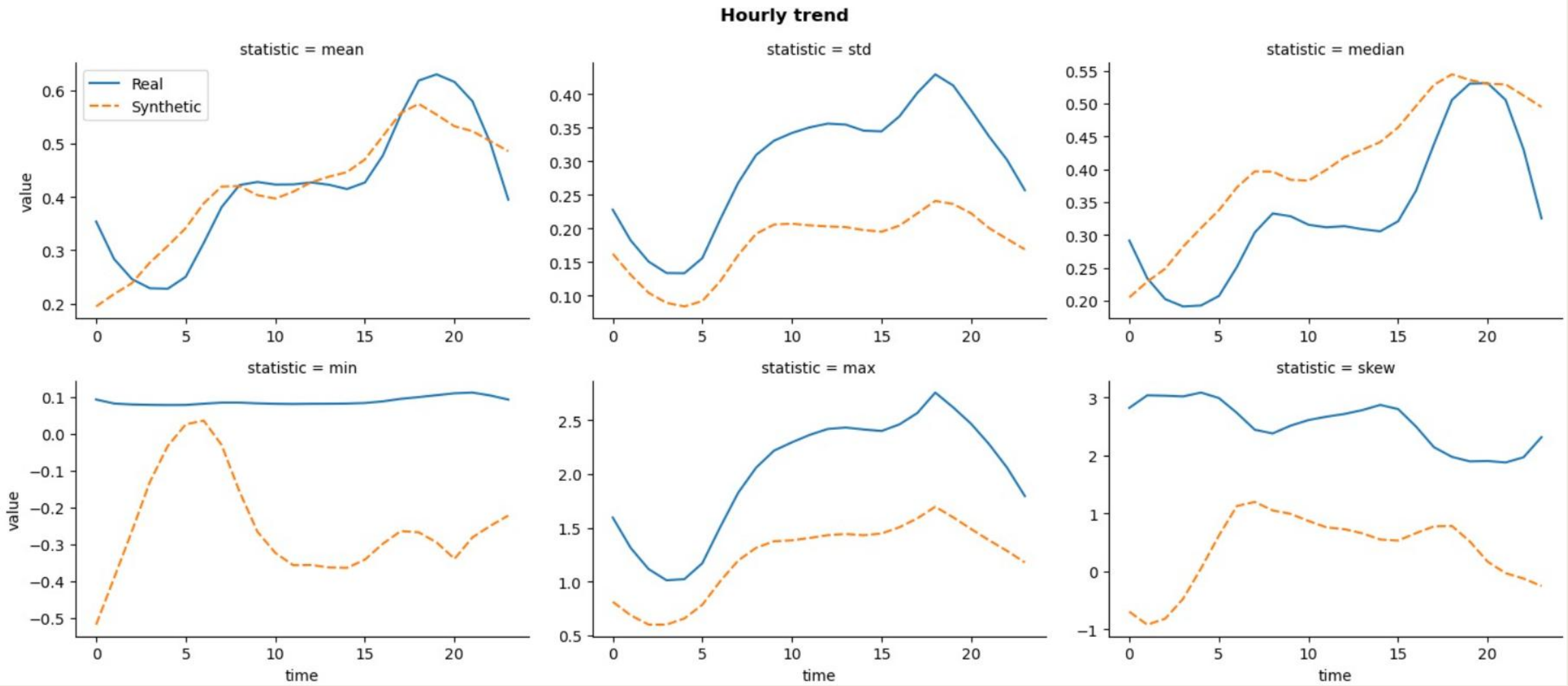
- Emphasizes testing the accuracy of synthetic data by comparing predictions to real data
- Synthetic data is generated from the use of future predictions through ML and randomization

RESULTS

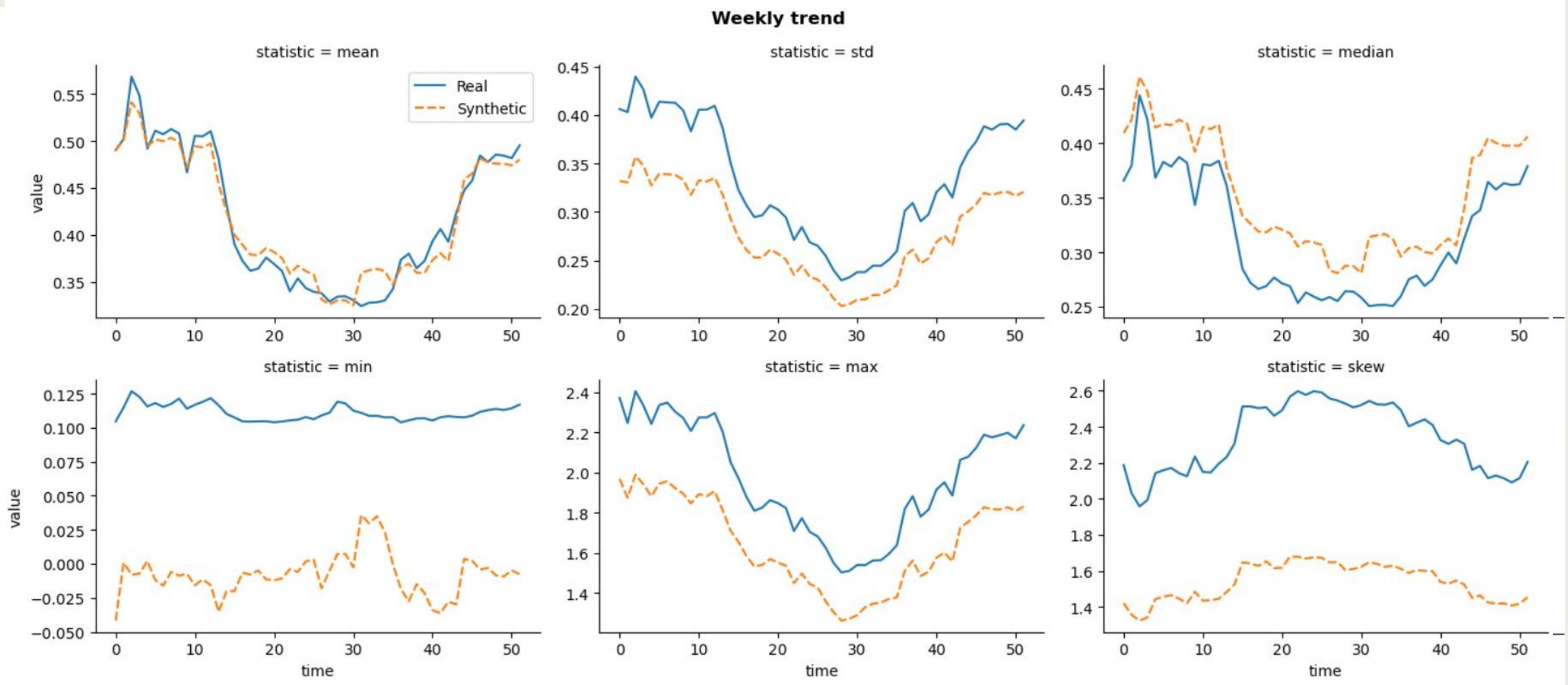
Real vs Synthetic



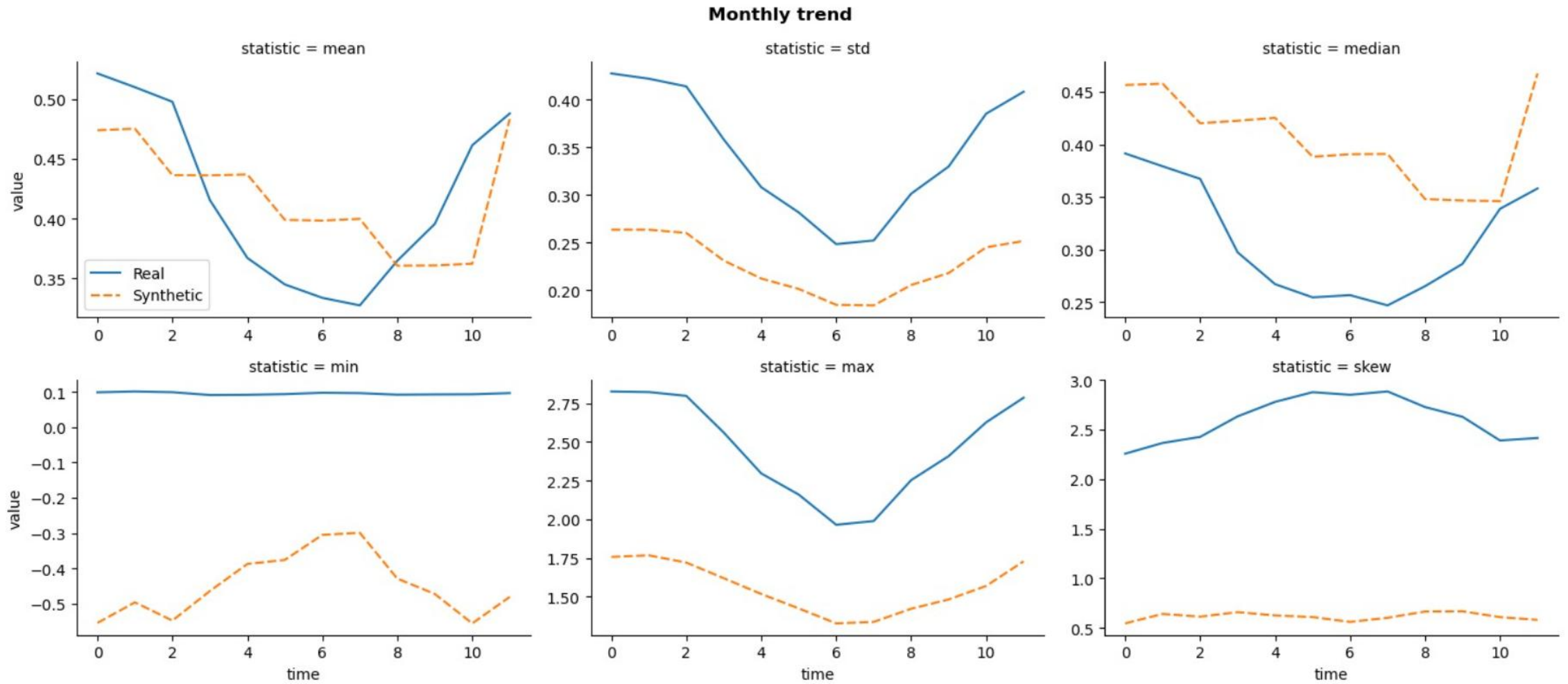
Hourly Trend Comparison – Real vs Synthetic



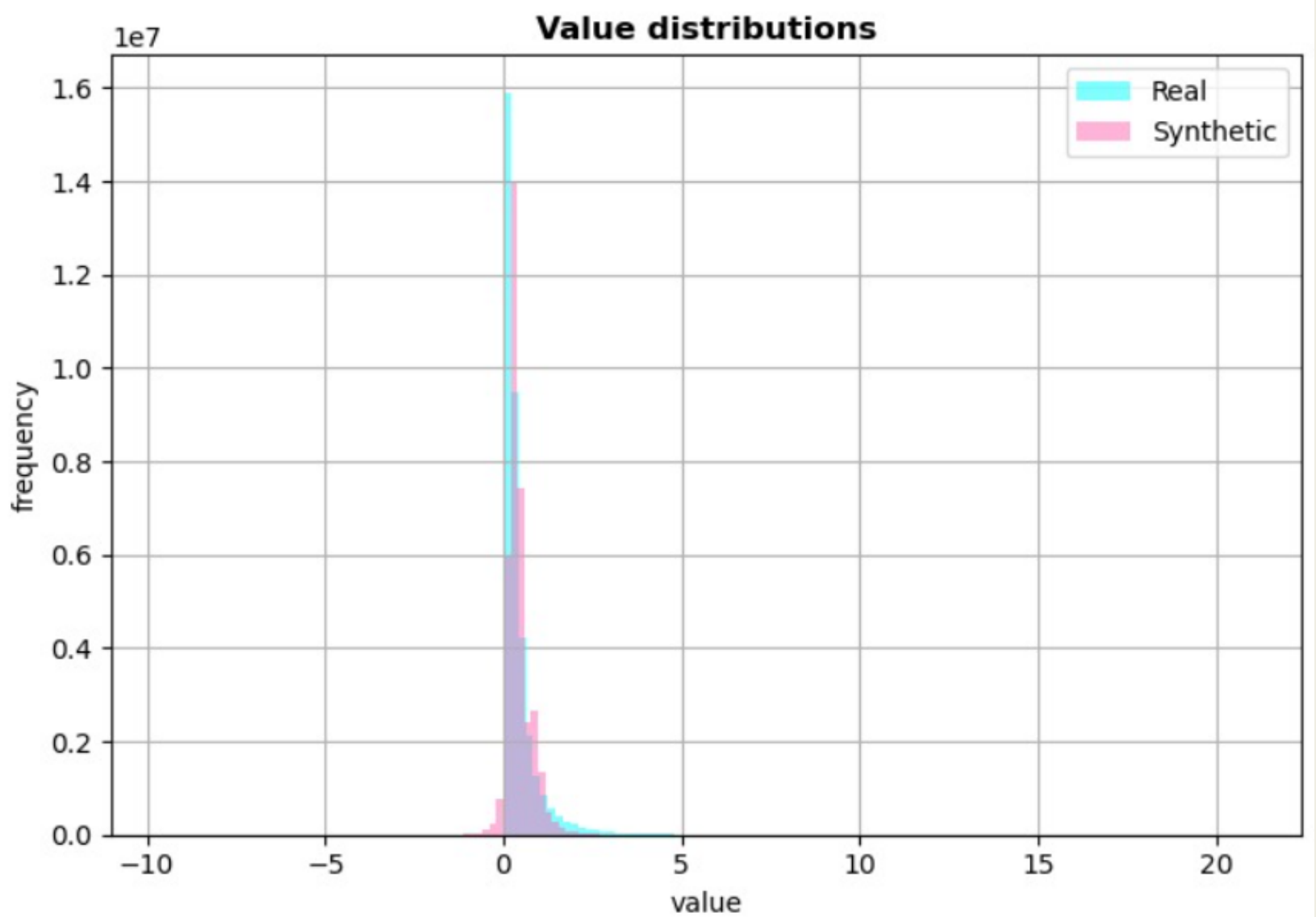
Weekly Trend Comparison – Real vs Synthetic



Monthly Trend Comparison – Real vs Synthetic



Value distributions of Real vs Synthetic



ANSWERING THE COACHES

Can we, in fact, prevent inferring real single profile data from synthetic data?

Completely preventing inference is very difficult. However, we can significantly reduce the risk.

- Techniques:
  - **Differential privacy**: Add controlled noise to the data during the generation process. This provides a mathematical guarantee of privacy.
  - **Data aggregation**: reduce risk of inferring individual profiles by grouping, e.g. instead of providing individual household energy usage, provide averages for neighborhoods or regions.
  - **Mark direct identifiers & generalize** e.g. replace exact ages with age ranges

How could one determine if synthetic data is “trustworthy”?

In our use case we understand this as the assessment of the quality, fidelity, and utility of an open-source platform, while ensuring it doesn’t expose sensitive information.

- Key approaches:
  - **Statistical similarity** by doing tests e.g. Kolmogorov-Smirnov test, Chi-squared test
  - **Machine Learning** to train model with both real & synthetic data to compare their performance on a validation set of real data. If the performance is similar, the synthetic data is likely trustworthy.
  - **k-anonymity and l-diversity** to evaluate if the synthetic data provides sufficient anonymity
  - **Watermarking:** One could consider adding a digital watermark to the synthetic data to indicate its origin and prevent unauthorized modifications

THANK YOU

Any Questions?