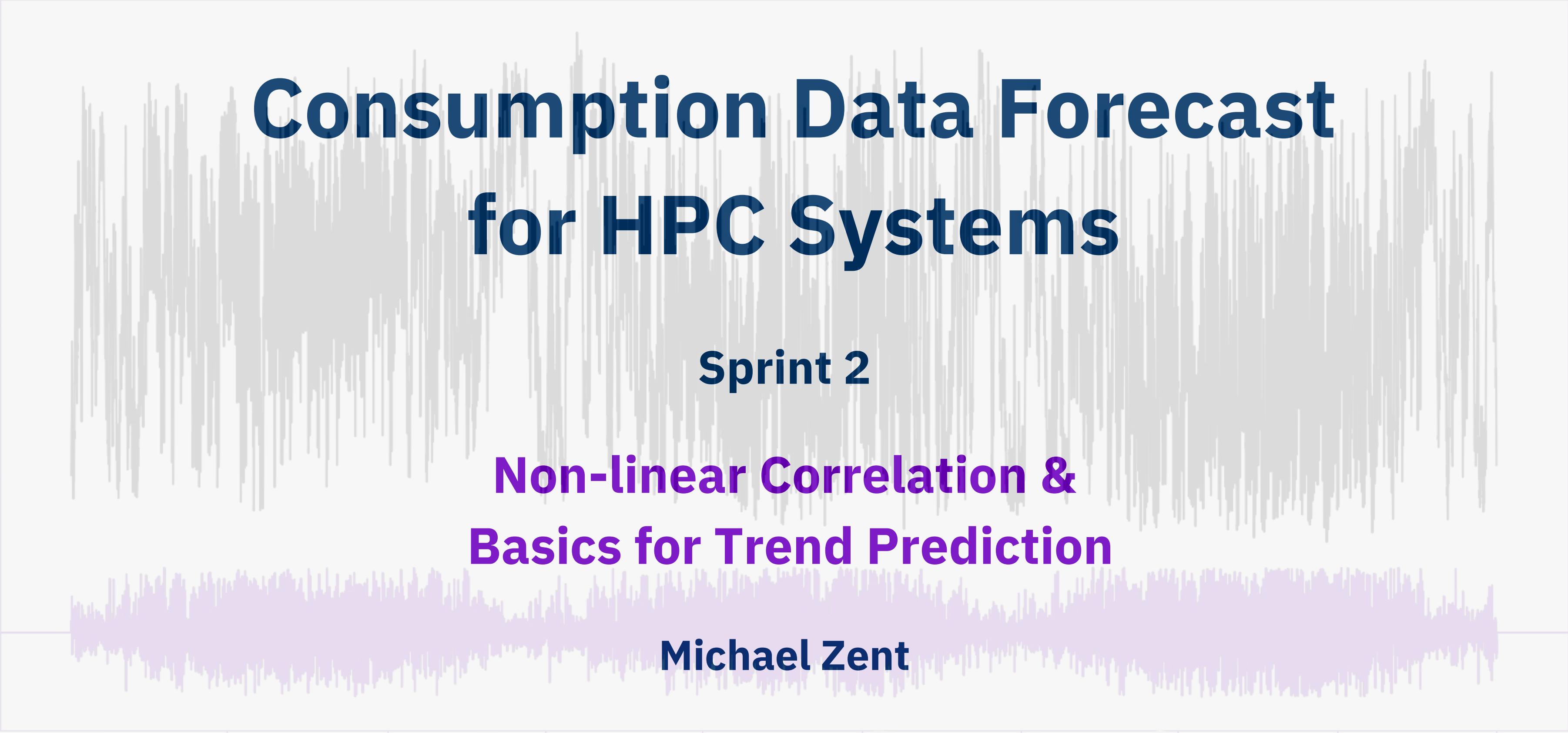
Justus Purat & Alexander Kammeyer
Software Project Distributed Systems



Consumption Data Forecast for HPC Systems

Sprint 2

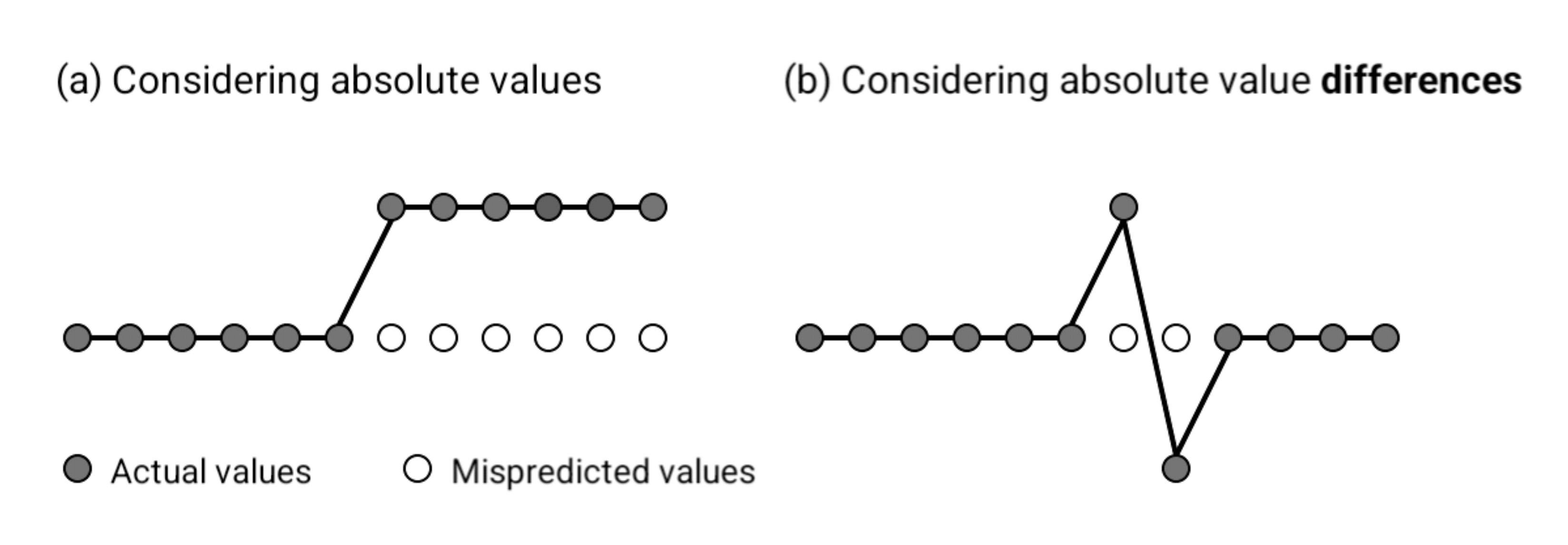
Non-linear Correlation & Basics for Trend Prediction

Michael Zent

Institute for Computer Science
SS 2025, FU Berlin

Lessons from SWP 2023/24

- Find and exploit meaningful **Correlations**, i.e. with causality
- **Predict Trends**
  - Abolute values can become easily obsolete, e.g. by political choices, but relative changes are usually more stable
  - Ruptures in absolute data with long-term impact are only outliers, when looking at differences


(a) Considering absolute values
(b) Considering absolute value **differences**
● Actual values
○ Mispredicted values

Non-linear Correlations

- Pearson Correlation (Cor) catches only linear correlations
- **Distance Correlation (dCor)** also measures non-linear correlations [1]

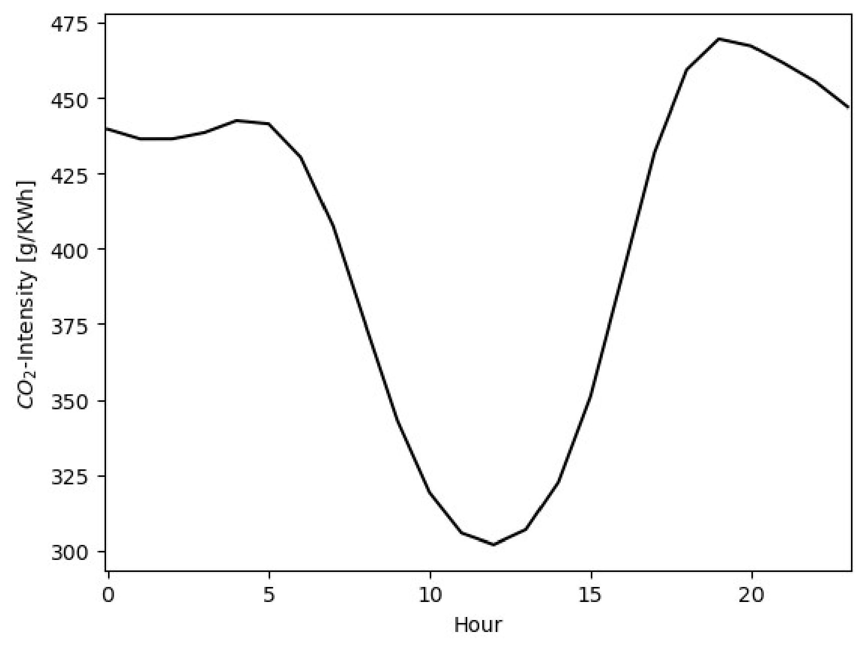
  = 0 : data vectors are independent

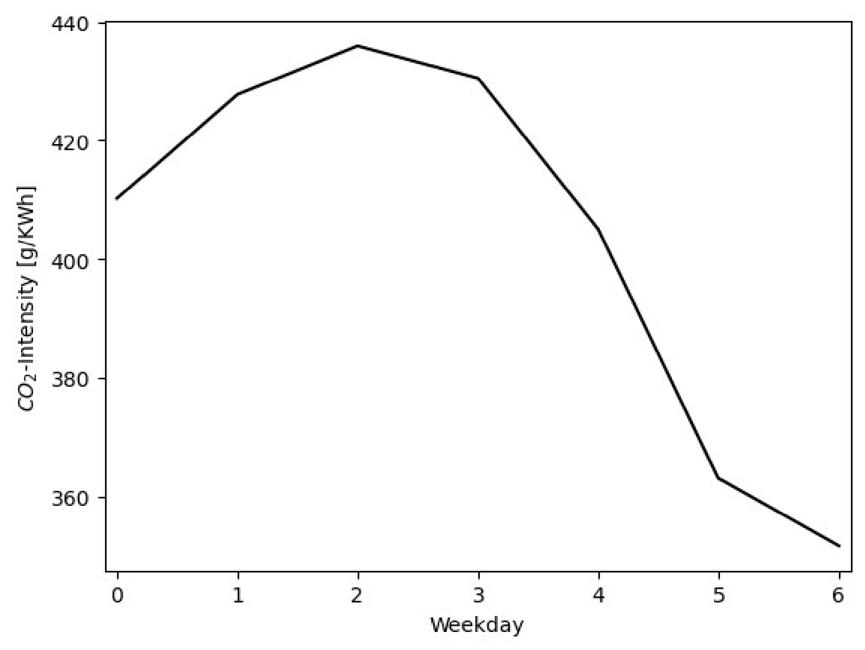
  = 1 : linear correlation

Corr. Type	Cor	dCor
linear	1	1
quadratic	0	0.5
cubic	0.9	0.9
sinusoid	0	0.5
circular	0	0.2

Non-linear Correlation

- Two-year mean of CO_2-Intensity per Hour resp. Weekday





Hours:

Corr = 0.09

dCorr = 0.53

Weekdays:

Corr = -0.77

dCorr = 0.85

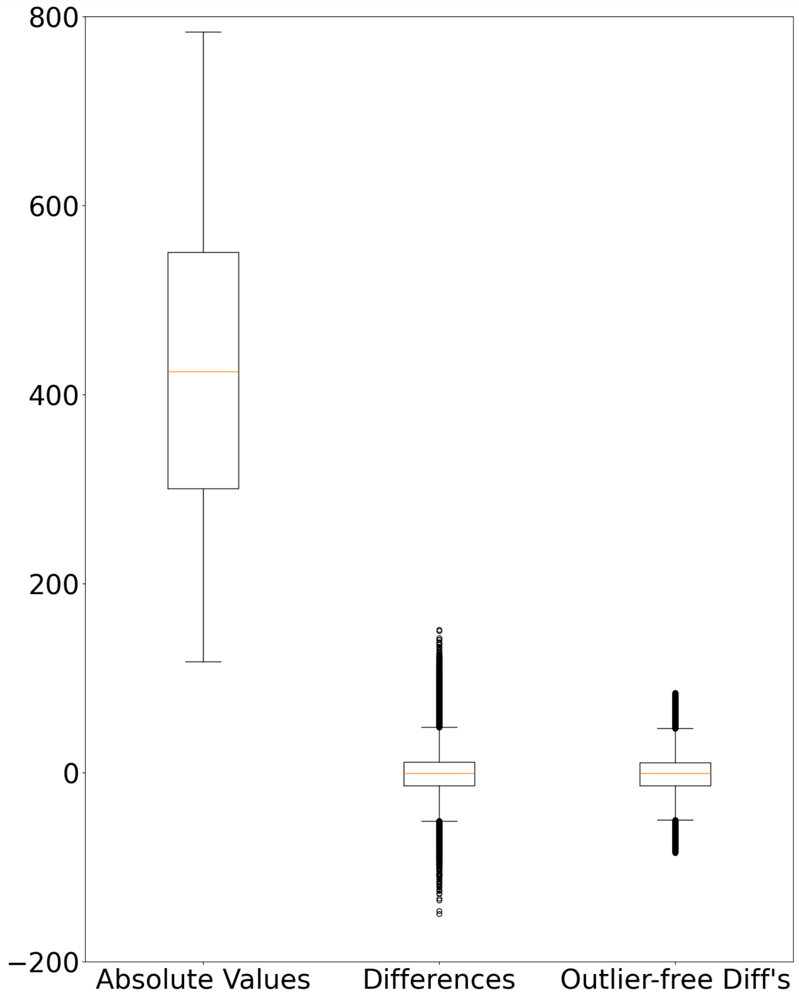
Problem Reduction

- Reduce the problem from predicting absolute values to predicting trends
  - Calculate differences in the data
  - Remove outliers, i.e. abrupt trend changes by Z-Score

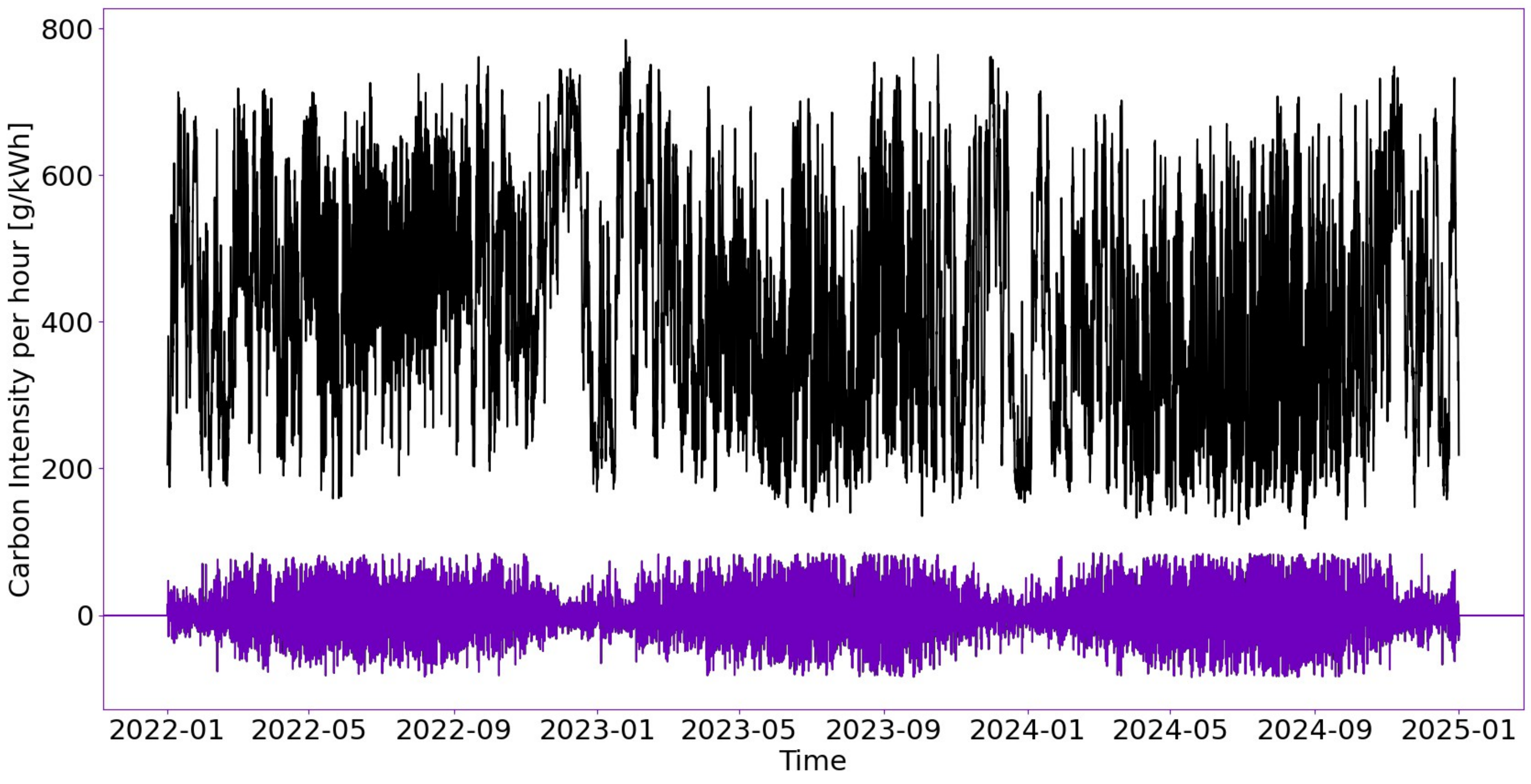
- Z-Score [2]
  - Number z of standard deviations σ by which a measurement value x is away from the mean μ

$$z = \frac{x - \mu}{\sigma}$$

  - Values with Z-Score > 3 are outliers



Result of Problem Reduction



Problem-reduced **Trend Data** (bottom) compared to the **Raw Data** (top)

Conclusion

- Analyis on **Non-linear Correlations** enables to detect an exploit a wider range of intra-data correlations without loss in detection of linear correlations
  - Especially of interest – Periodicity

- **Trend Analysis** reduces the problem to centered *changes* in the data, such easying outlier removal and compressing the data span which is to learn on.

- **Next**
  - Feeding the trend data into forecast models
  - Search for, and refine with further non-linear correlations

Literature

[1] G.Székely et al. (2007), *Measuring and testing dependence by correlation of distances*, The Annals of Statistics, 35(6):2769-2794

[2] C.A.Mertler & R.V.Reinhart (2017), *Advanced and Multivariate Statistical Methods*, 6th ed., Routledge, pp.29-32

[3] A.C.Elliott et al. (2017), *Applied Time Series Analysis*, 2nd ed., CRC

[4] R.E.Chandler & E.M.Scott (2011), *Statistical Methods for Trend Detection and Analysis in the Environmental Sciences*, Wiley

[5] W. Palma (2016), *Time Series Analysis*, Wiley

This slides and the corresponding Python code at

git.imp.fu-berlin.de/timeout/swp-distributed-systems-t5-ml

or

timeout.userpage.fu-berlin.de/hpc/consumption-data-forecast

Thank You