Electricity Consumption Datasets

A Position Paper

Christoph Klemenjak, Andreas Reinhardt, Lucas Pereira, Mario Bergés, Stephen Makonin, and Wilfried Elmenreich

A Story of Successful Collaboration

UNIVERSITÄT KLAGENFURT

- Christoph Klemenjak → SynD
- Andreas Reinhardt → Tracebase, AMBAL
- Lucas Pereira → SustData, SustDataED, NILMPEds
- Stephen Makonin → AMPds, RAE, HUE, ODDs
- Mario Bergés → BLUED, PLAID
- Wilfried Elmenreich → GREEND

Pereira, Lucas, and Nuno Nunes. "Performance evaluation in non-intrusive load monitoring: Datasets, metrics, and tools—A review." Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 8.6 (2018): e1265.

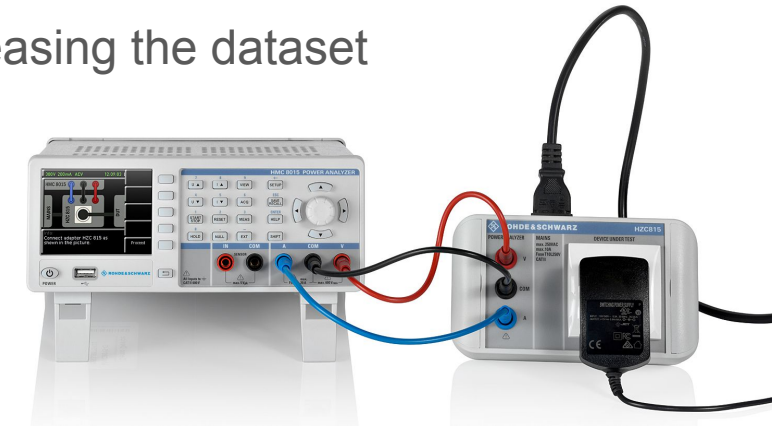
Motivation: Spark a Discussion!

UNIVERSITÄT KLAGENFURT

- Real-world datasets are crucial for R&D
- Scope is limited to experimenter's objectives and time constraints
  - This heterogeneity poses a comparability problem
  - We identify a lack of widely-agreed best practices
- Motivation of this position paper
  - Foster the creation of datasets with increased usability
- We present recommendations for:
  - Data collection
  - Data storage
  - Data provision

Instrumentation: on Rates & Verification

- Collection of microscopic data should be favoured over macroscopic data
  - Reporting rates beyond 1 kHz
- Provide a down-sampled version of the dataset (e.g., one second)
  - Convert from microscopic to macroscopic
- Conduct manual cross-checks before releasing the dataset
  - Deal with gaps and irregularities in the data
- Think beyond your application



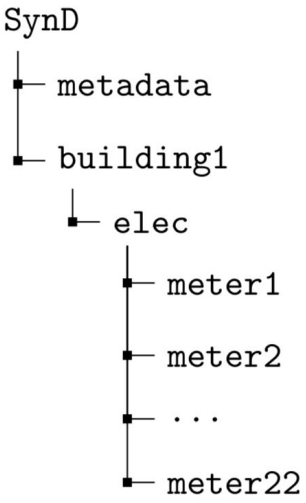
Metadata: Collect, Hoard, and Annotate!

- Photograph type tags or stickers
  - Collect relevant documentation
- Log activities and events during the campaign
  - Serves to annotate power traces with event information
  - Make remarks in case appliances are removed or replaced
  - Consider logging geographic and ambient features
- Consider machine-readable metadata formats
  - Facilitates simple processing
  - NILM metadata schema to the rescue

Kelly, Jack, and William Knottenbelt. "Metadata for energy disaggregation." 2014 IEEE 38th International Computer Software and Applications Conference Workshops. IEEE, 2014.

File Formats: a Plethora of Options….

- We observe a variety of file formats:
  - Comma-separated values (CSV),
  - Hierarchical data format (HDF5),
  - Matroska media containers,
  - FLAC, WAVE, etc.
- HDF has emerged as a viable candidate
  - Supports metadata annotations
  - Efficient data storage
  - Is supported by most scientific computing frameworks



Klemenjak, Christoph. 2019

Dataset Provision: Trust the Cloud (once)

UNIVERSITÄT KLAGENFURT

- Consider public hosting services to ensure long-term availability
  - Harvard Dataverse, IEEE DataPort, …
  - Self-hosting solutions pose a threat to availability
    - Contributions evaluated on such datasets can no longer be validated
- Data should be provided with a permissive license
  - ...but take necessary precautions to protect the privacy of users
  - Clearly state the conditions under which the data may be used and distributed

Conclusions

- We elaborate important issues of energy datasets (for NILM)
  - We bring forward 17 suggestions for future investigators
  - Basis for planning new datasets
  - Existing datasets can benefit from a conversion
    - Achieve a wider use in the development of algorithms
- What could be done next?
  - Consider retrofitting of existing datasets (e.g. GREEND)
  - Conduct a SWOT study on file formats to identify the best option

klemenjak@ieee.org - klemenjak.github.io - @CKlemenjak