Tools for Evaluating, Developing and Testing Water Demand Disaggregation Algorithms

Nour A. Attallah[1]; Jeffery S. Horsburgh[2]; Camilo J. Bastidas Pacheco[3]
[1] nour.atallah@usu.edu; [2] jeff.horsburgh@usu.edu; [3] camilo.bastidas@usu.edu

1552444

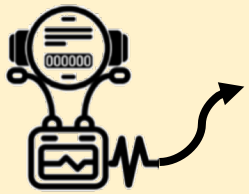
1. Motivation – Existing studies are not reproducible



- Resource and time intensive
- Code and data are not available
- Difficult or impossible to reproduce
- Non label-intensive technique

2. Data Collection – Generating a detailed HR dataset





CIWS Datalogger



High-performance Time Series Database

influxdb

- Data collected on existing residential water meters
- Local data storage
- QC datasets are sent to influxDB

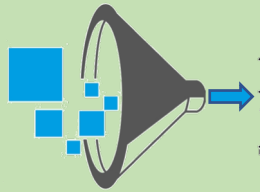
3. Methods – Developing a new open source algorithm



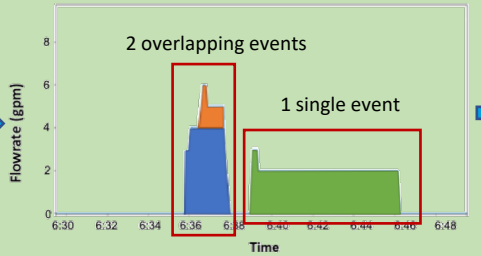
- Implemented in Python
- Open-Source
- Reproducible
- Verified using actual labeled events

Data Processing Script

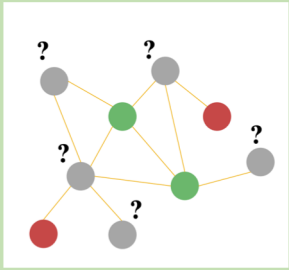
scikit learn



1. Data Filtering



2. Disaggregation

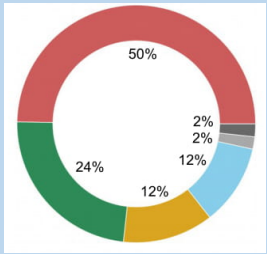


3. Semi-Supervised Learning

4. Results



- Tested on 5 residential houses with different meters
- Overall accuracy of the tool was 98.3% based on resident's labeled events
- Highest accuracy observed (100%) for clothes washer and faucet/dishwasher





5. Contribution



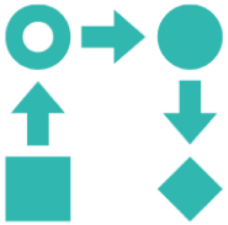
- **Open source, and reproducible tool***
- Overcomes the shortcomings of supervised techniques (expensive) and unsupervised techniques (unknown labels)
- Useful info to researchers, water providers, and customers

*Attallah, N., J. S. Horsburgh, C. J. Bastidas Pacheco (2021). Tools for Evaluating, Developing, and Testing Water End Use Disaggregation Algorithms, HydroShare, http://www.hydroshare.org/resource/1521eba67f1d4571ac5fe2b8c5e01035

Challenges









- **Algorithms for water end use**
  - Identifying actionable information extracted from data
  - Algorithms for identifying and disaggregating end uses from data
  - Where to most efficiently execute data analysis algorithms
- **Existing Tools**
  - Resource and time intensive
  - Require a large dataset of labelled events
  - Code and/or data are not available for testing
  - Difficult or impossible to reproduce the work

Introduction

- Collaborative work to enhance the current **Cyberinfrastructure**

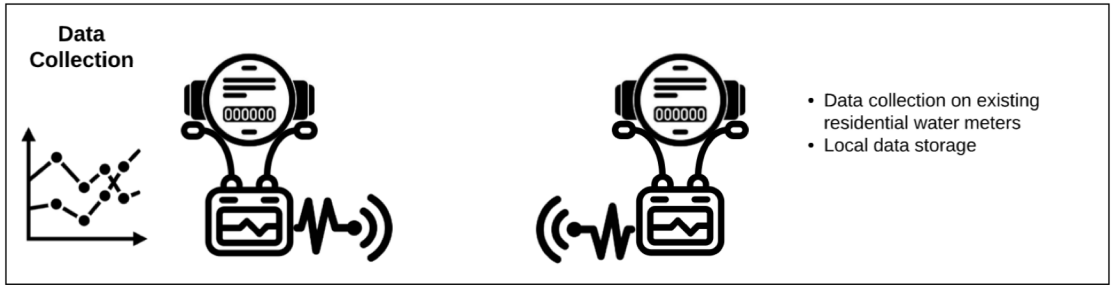
Cyberinfrastructure = ***infrastructure*** + ***Information Technology***

Smart water meter ~ High temporal data → ***Tools for Evaluating, Developing and Testing Water Demand Disaggregation Algorithms***

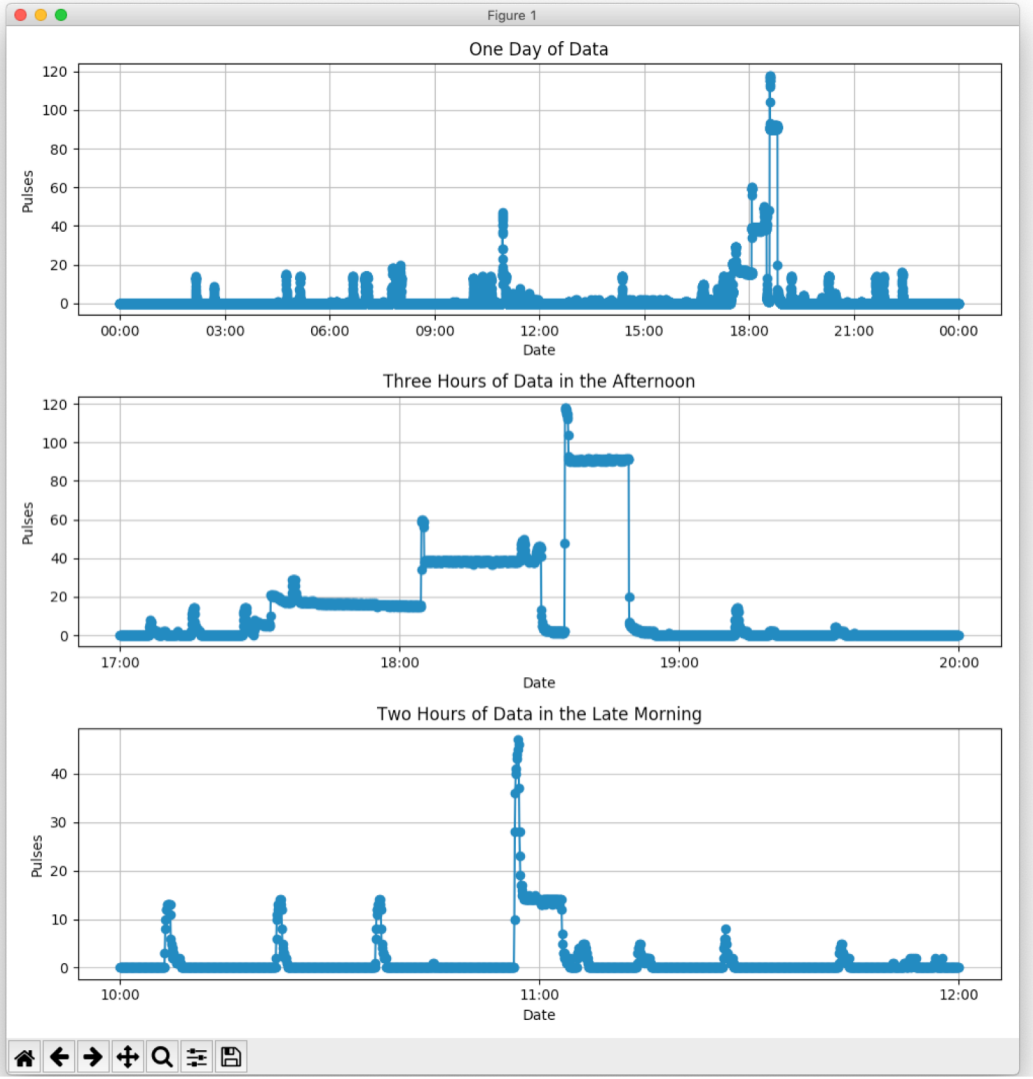
CIS We make IT possible | Cyberinfrastructure



Data Collection







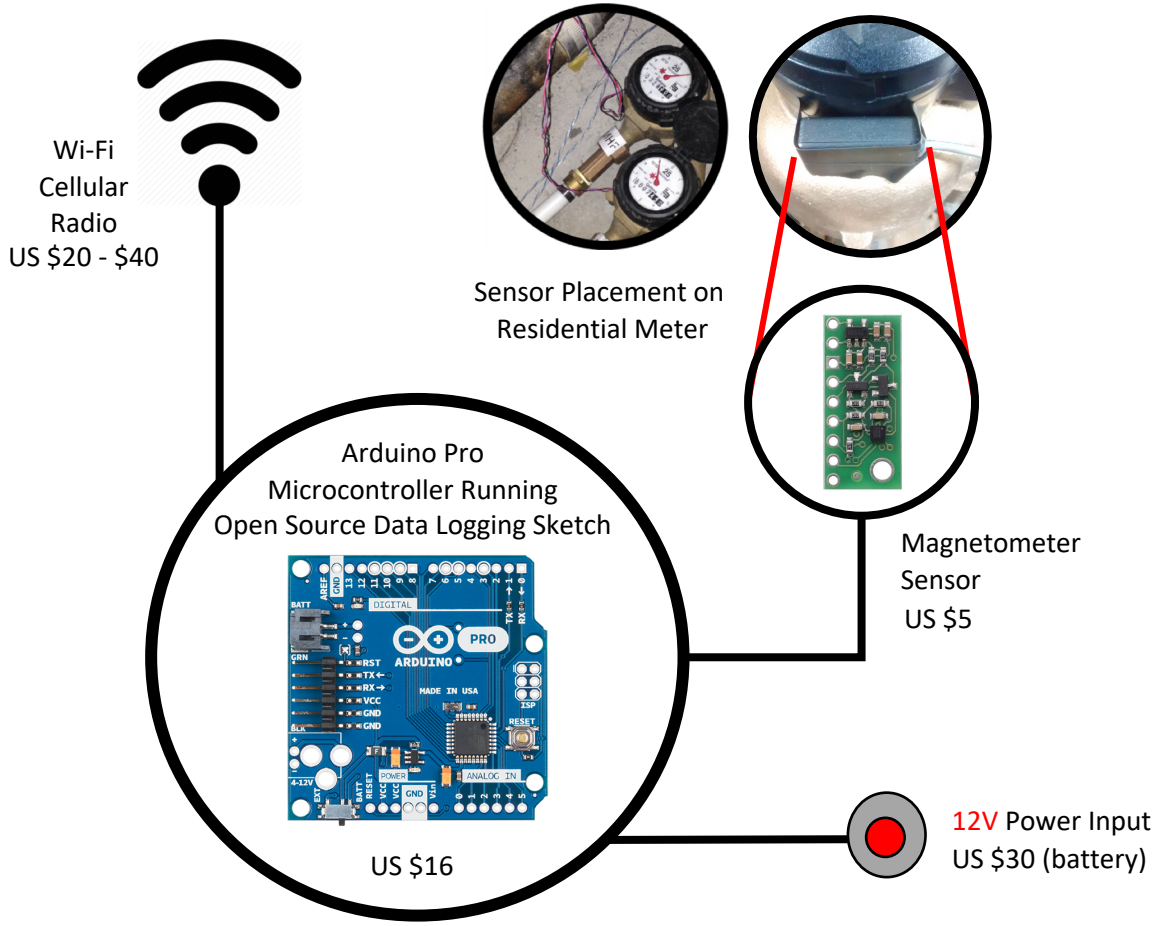
High frequency data enables:

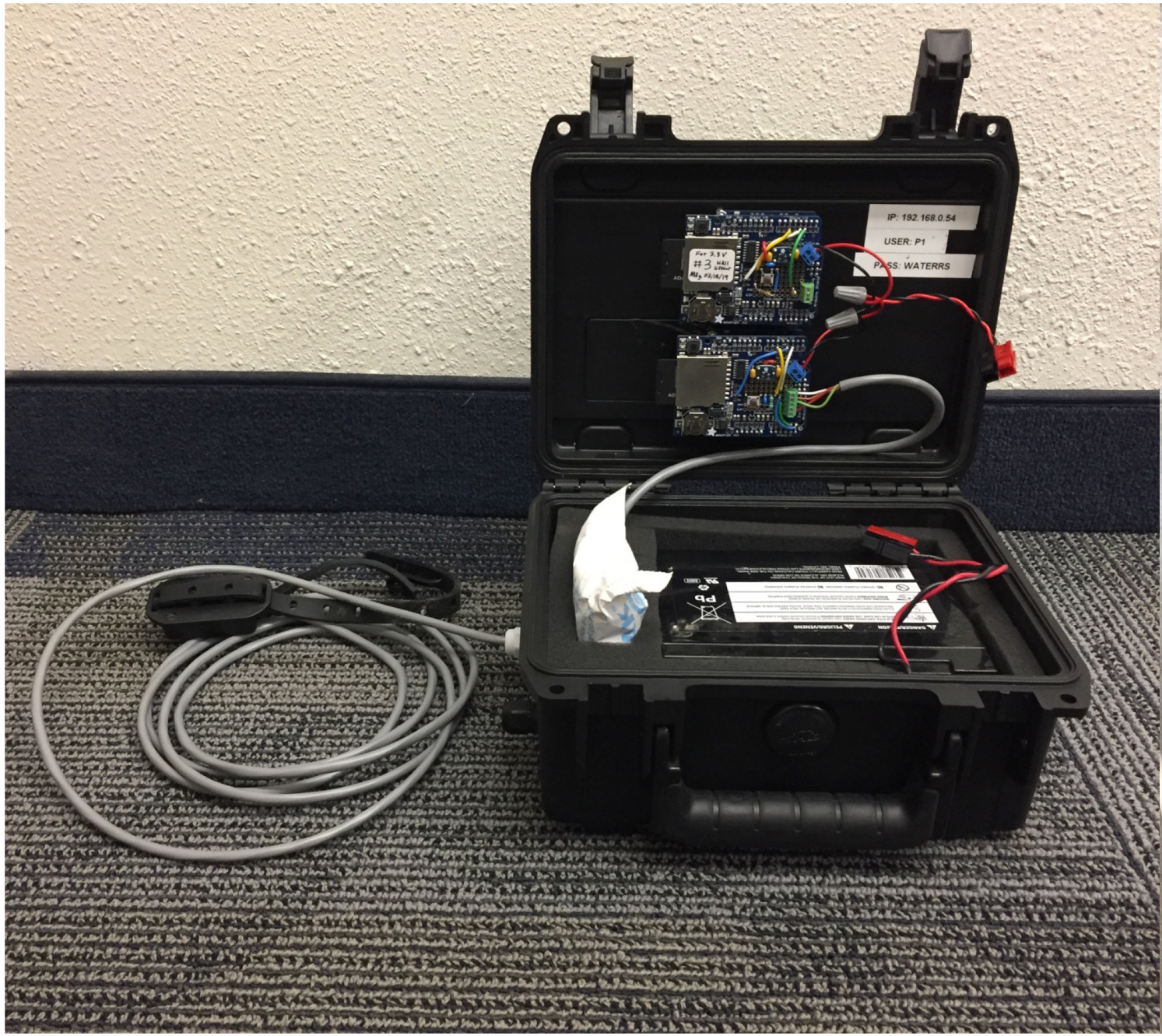
- Identifying and characterizing individual water end uses
- Providing information to utilities to better manage water demand
- Providing feedback to water users
- Estimating water-related energy use

Making Existing Meters Smarter



+




IP: 192.168.0.54
USER: P1
PASS: WATERRS


NEPTUNE
GALLONS

Data Collection

- Identify participants using Monthly billing data
- Divide users into categories, randomly sample from each category
- Conduct a walk-through inside each house to identify their water end uses



Soliciting Participants



Device Installation

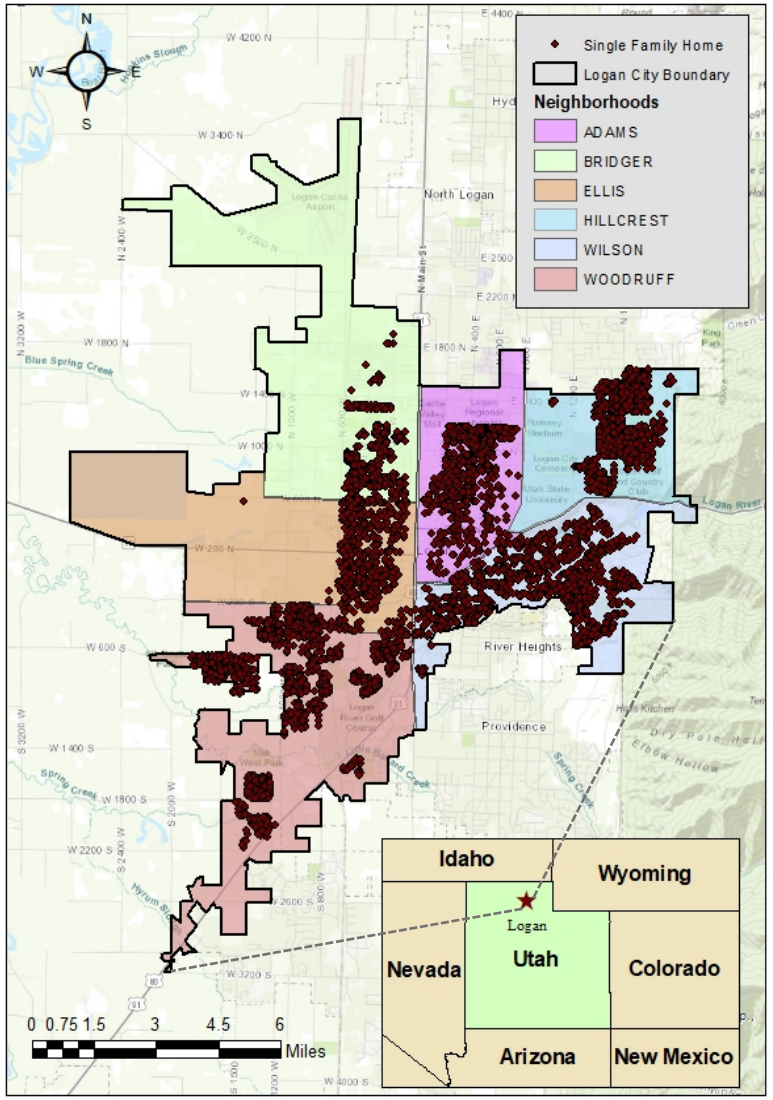


Walk-Through

Data Collection

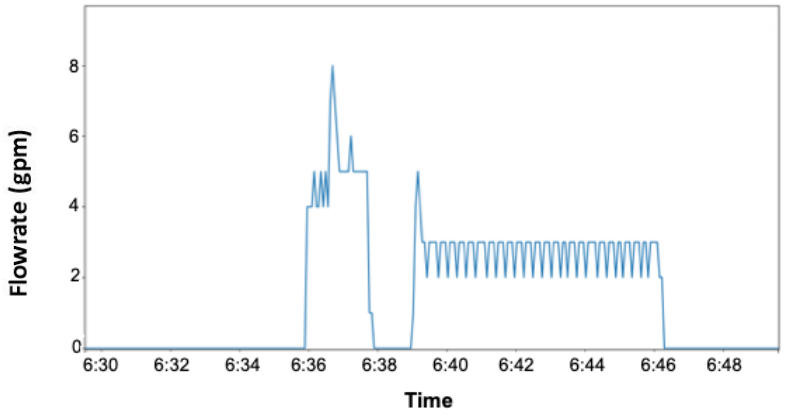
- Distribution of Houses
- Data collection:
  - 2 weeks in summer, 2 weeks in winter
  - A sample of 31 participants
  - A dataset consisting of high-resolution water use data along with household characteristics identified during the walk-through





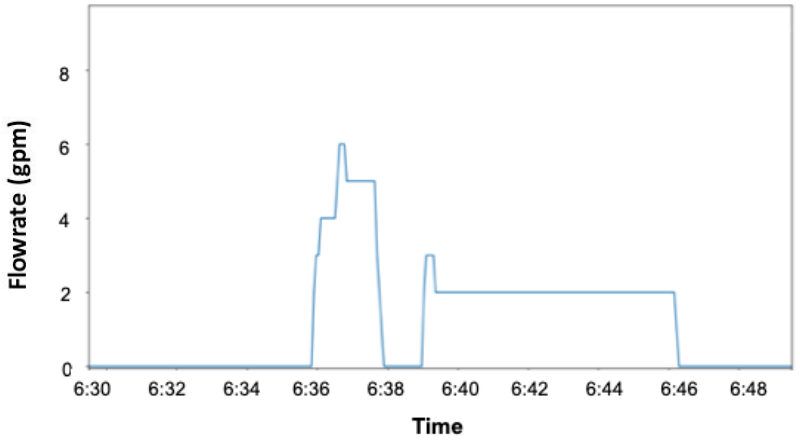
Methods

Raw Data



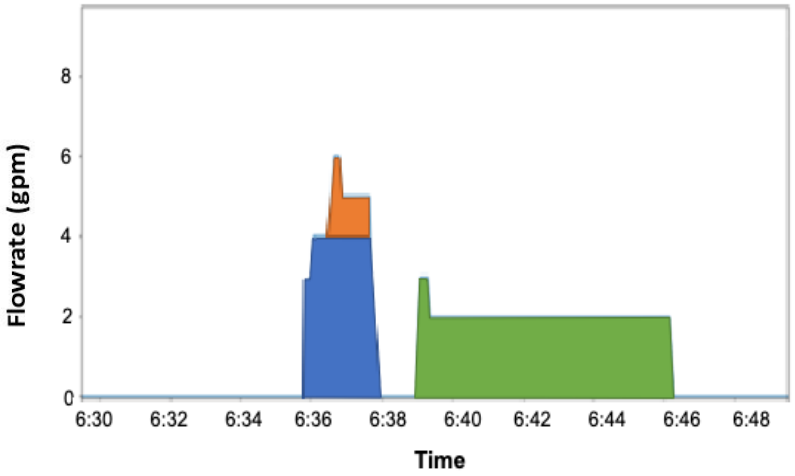
Filtering

Filtered Data



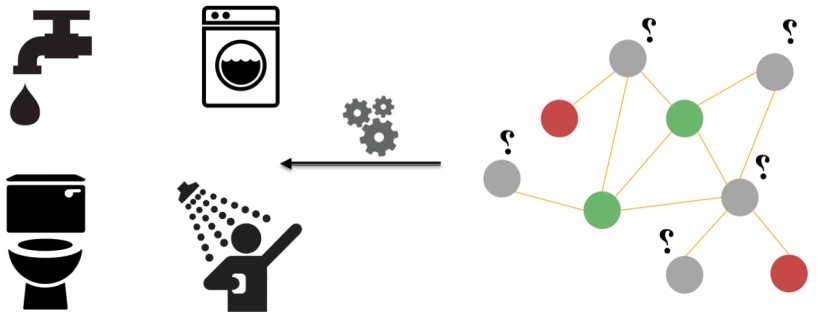
Disaggregation

Separated Events



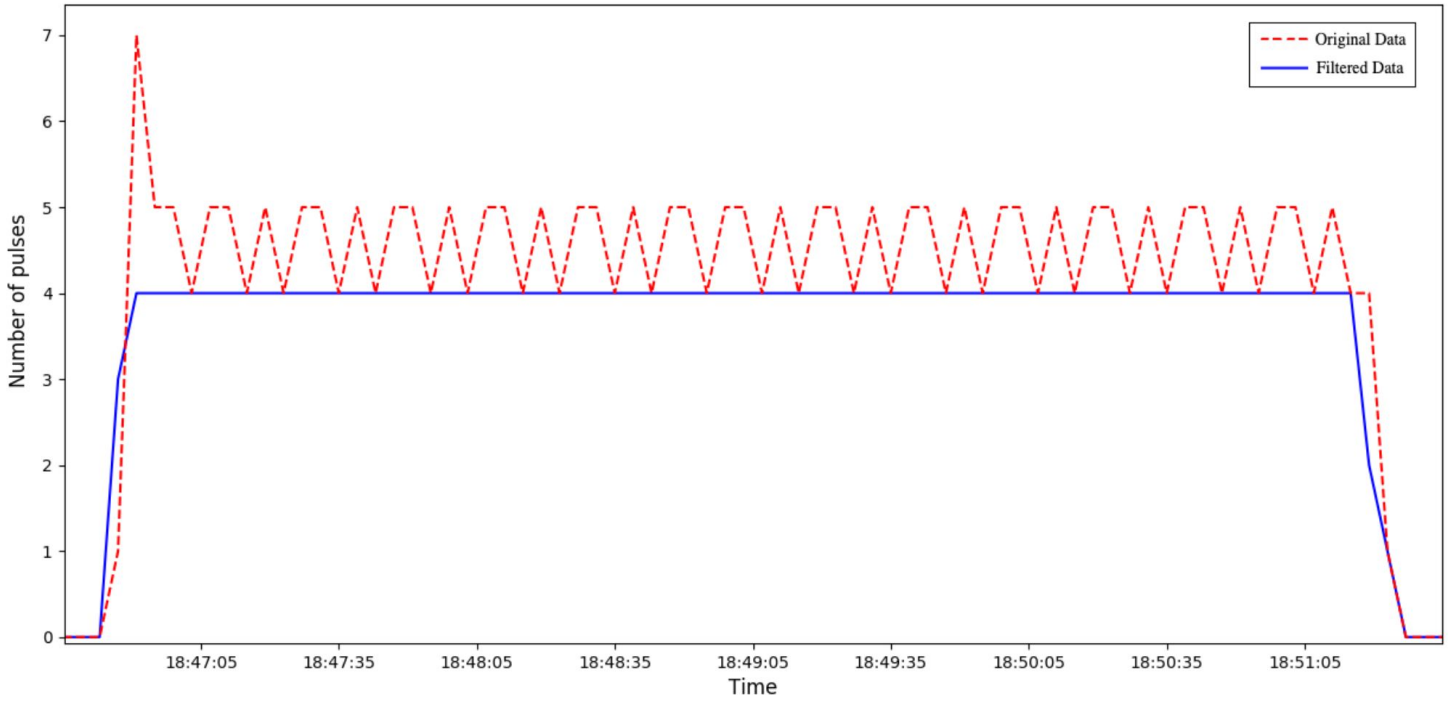
Classification

Classified Events



Methods: Data Filtering

- Low Pass filtering technique
- Enhance the process of extracting water use events and their associated physical features
- The filter modifies the raw trace data to accentuate event start and end times and the number of vertices of each event in the trace



$$P`_j = \begin{cases} \dfrac{\sum_{i=1}^{n} P`_{j-i} + \sum_{i=0}^{n} P_{j+i}}{2n+1} & for\ center\ data\ points \\ \dfrac{\sum_{i=0}^{n} P_{j\pm i}}{n+1} + 1 & for\ edge\ data\ points \end{cases}$$

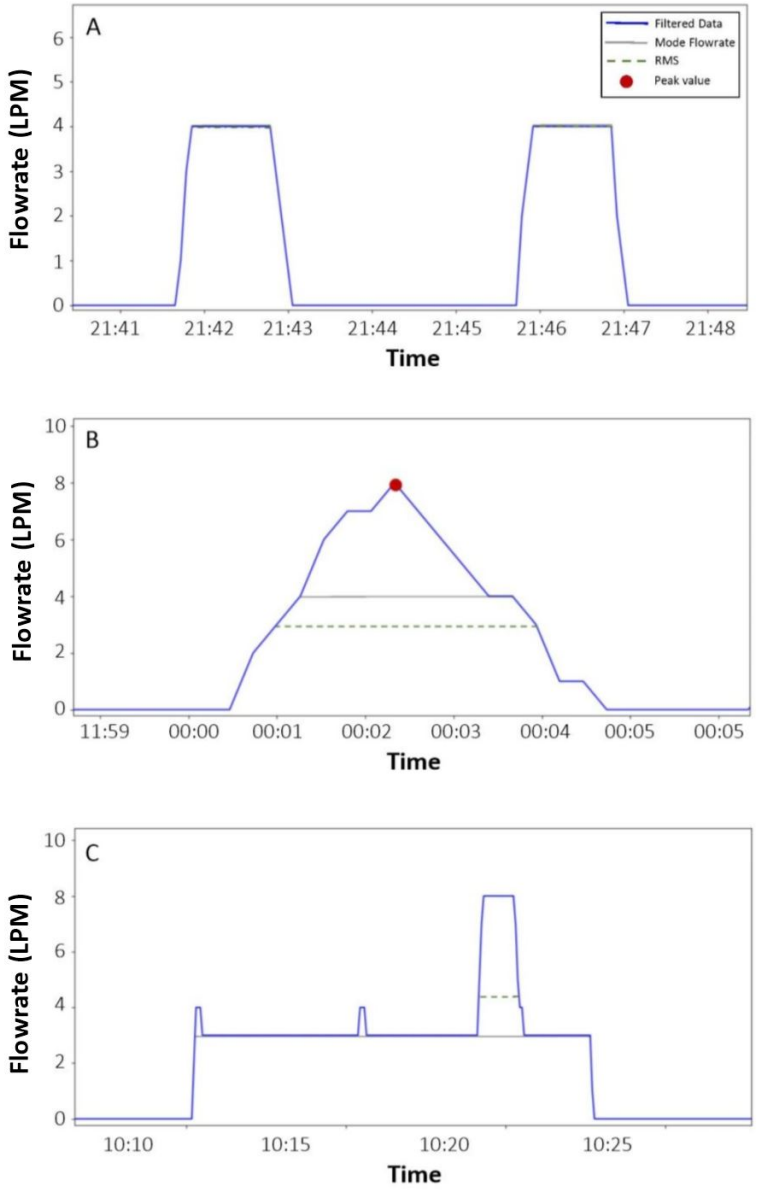
Methods: Disaggregation

Feature	Definition	Data used
Volume (L)	The summation of water use volume between the beginning and end of a set of consecutive non-zero values	Raw
Start time	The time at which the water use volume changes from zero to any positive value	Raw
End time	The time at which the water use volume transitions back to zero	Raw
Duration (min)	The elapsed time of the event calculated as the difference between start and end times	Raw
Flow rate (LPM)	The volume of an event divided by its duration	Raw
Peak flow rate (LPM) – original	The maximum rate of water flow for any time step within the event	Raw
Peak flow rate (LPM) – filtered	The maximum rate of water flow for any time step within the event	Filtered
Peak flow rate frequency – original	The number of occurrences of the maximum flow rate within an event	Raw
Peak flow rate frequency – filtered	The number of occurrences of the maximum flow rate within an event	Filtered
Mode flow rate (LPM)	The flow rate value that appears most often within an event	Filtered
Mode flow rate count	The number of occurrences of the mode flow rate within an event	Filtered
Root mean square (RMS)	The square root of the mean square of flow rate (the arithmetic mean of the squares of the flow rates within the event)	Filtered
Number of vertices (V)	The number of points where the flow rate changes from one non-zero value to another non-zero value within the same event	Filtered
Irregularity measure I (IR1) (LPM)	The difference between the mode flow rate and the RMS value	Filtered
Irregularity measure II (IR2) (LPM)	The difference between peak flow rate and mode flow rate value	Filtered
Complexity measure (CX)	The count of the mode flow rate divided by the number of vertices	Filtered

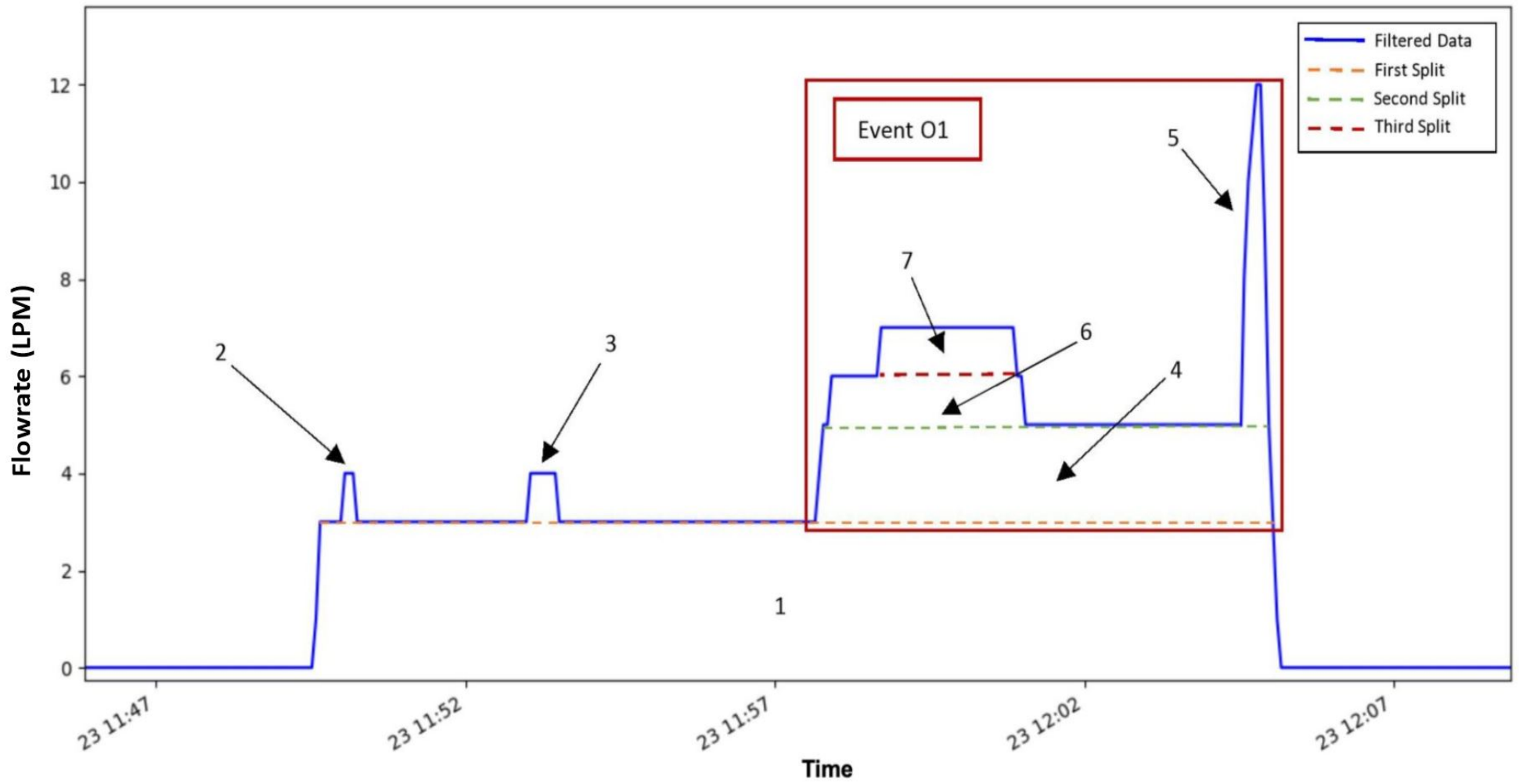
Methods: Disaggregation

- Classify events as single or overlapping events using a combination of features
- These features reflect the shape, irregularity, and complexity of each event
- Disaggregate overlapping events into single events

$$Event = \begin{cases} Overlapping & if\ IR1\ \neq 0\ and\ IR2\ \neq 0\ and\ CX > 1\ \ and\ V > 4 \\ Single & all\ other\ events \end{cases}$$

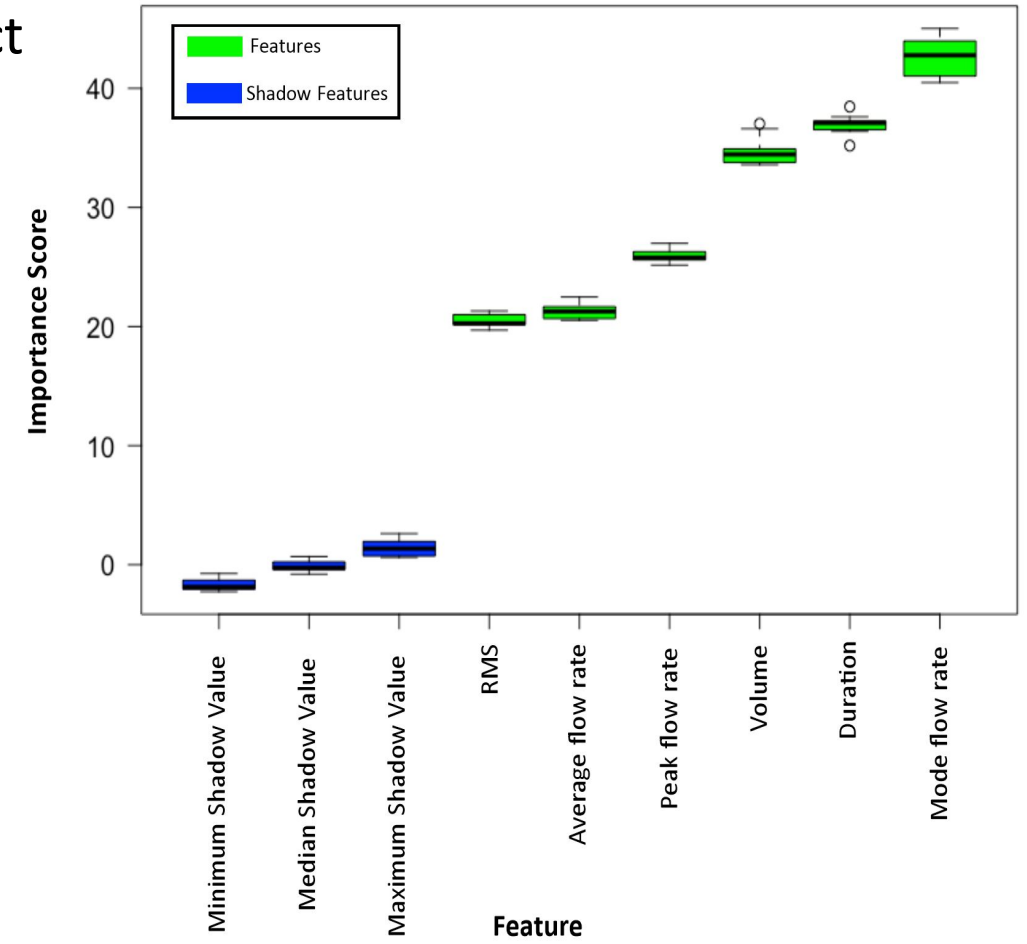


Methods: Overlapping Events Decomposition



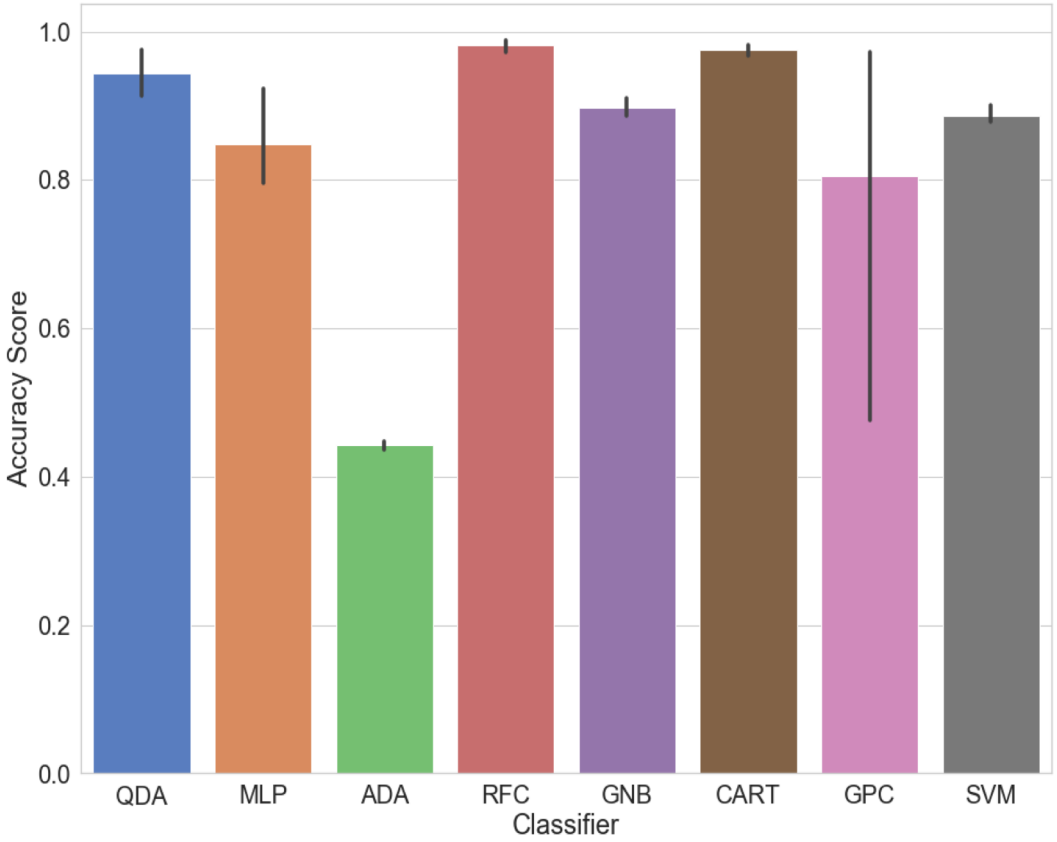
Methods: Feature Selection

- Subset a feature combination that are relevant to predict the label of each event
- The Boruta algorithm was used
- Mode flow rate, duration, volume, peak flow rate, and average flow rate are the most relevant features in predicting the end use labels for each event



Methods: Semi-Supervised Classification

- Utilizes unsupervised clustering to detect and eliminate outlier events that deviate from other events in the data space
- Predicts the category label of each remaining unlabeled event in the developed clusters using semi-supervised classification
- Random forest produced the highest accuracy score

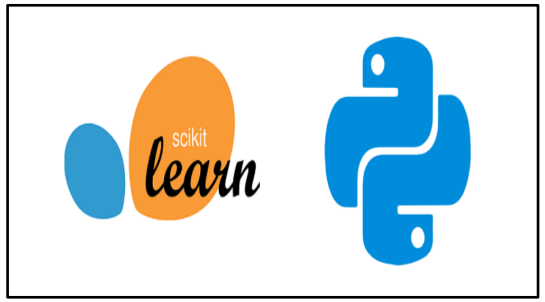


Methods: Semi-Supervised Classification

Let $\mathbf{L} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^{l}$ be the input manually labeled training dataset where *x* is the set of labeled events, *y* is the label set, and *l* is the number of labeled events. Let $\mathbf{U} = \{\boldsymbol{x}_j\}_{j=l+1}^{l+u}$ be the set of *u* unlabeled events for a home.

1. Apply DBSCAN on $\mathbf{U}$ to generate an initial set of clusters
2. Using the initial DBSCAN clusters, identify outlier data points and remove them from $\mathbf{U}$
3. Train a predictive random forest classifier $\boldsymbol{f}$ using $\mathbf{L}$ as a training dataset
4. Predict the label of all unlabeled events in $\mathbf{U}$
5. Calculate the similarity score for the newly labeled events in $\mathbf{U}$
6. Add $\{(\boldsymbol{x}, \boldsymbol{f}(\boldsymbol{x}))\mid \boldsymbol{x} \in \mathbf{S}\}$ to $\mathbf{L}$, where $\mathbf{S}$ is the subset of newly labeled events in $\mathbf{U}$ with similarity score > 90%
7. Remove $\mathbf{S}$ from $\mathbf{U}$
8. Retrain the classifier $\boldsymbol{f}$ using the enhanced training dataset $\mathbf{L}$
9. Repeat steps 4-8 until there is no $(\boldsymbol{x}, \boldsymbol{f}(\boldsymbol{x}))$ with similarity score > 90% in $\mathbf{U}$, or until $\mathbf{U} = \emptyset$

Data Processing Script

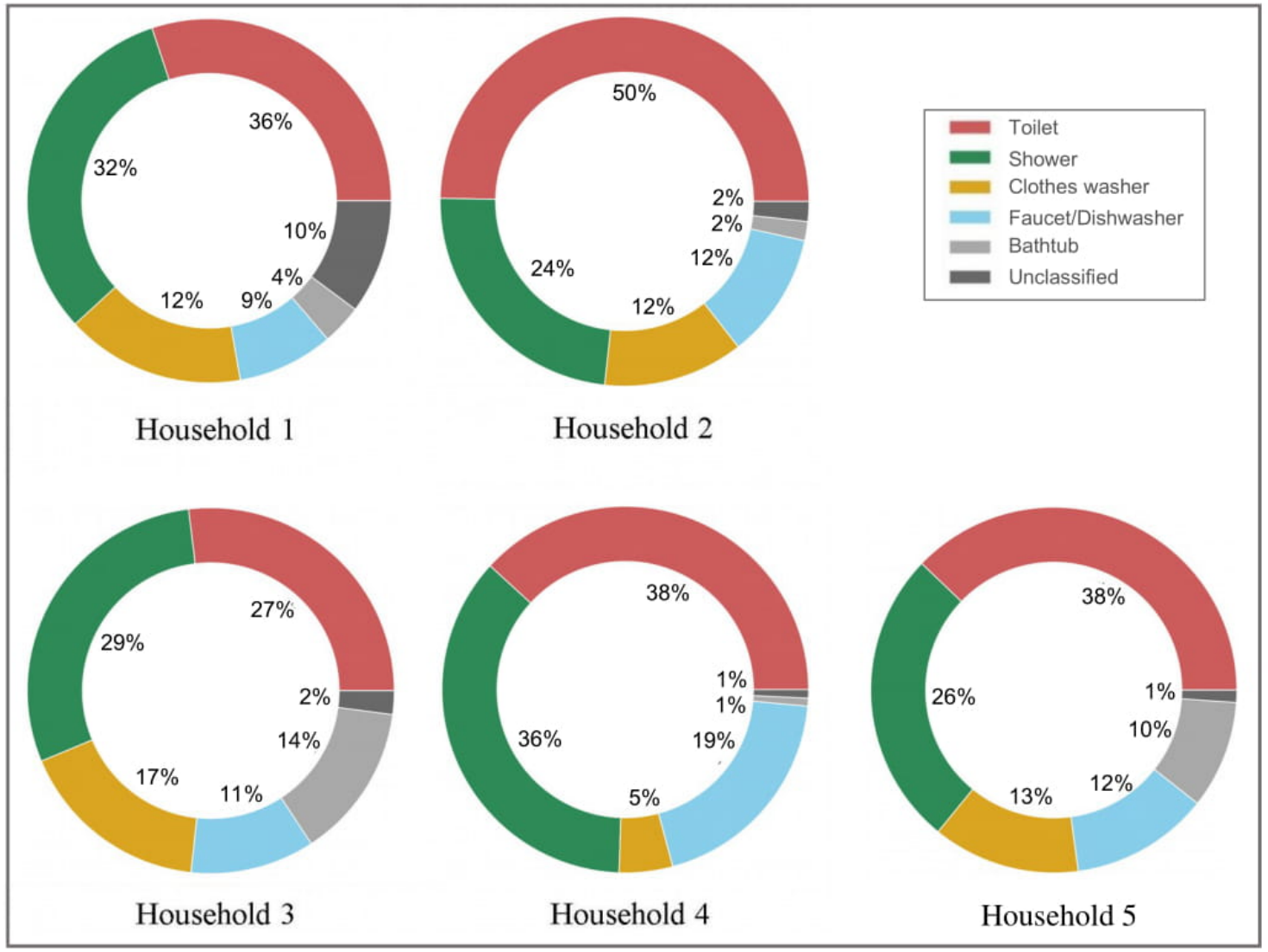


Results: Classification Accuracy

End use type	Number of events	Classification accuracy (%)
Bathtub	15	66.7
Clothes washer	24	100
Faucet and Dishwasher	191	100
Shower	59	98.3
Toilet	171	99.4
All categories	460	98.3

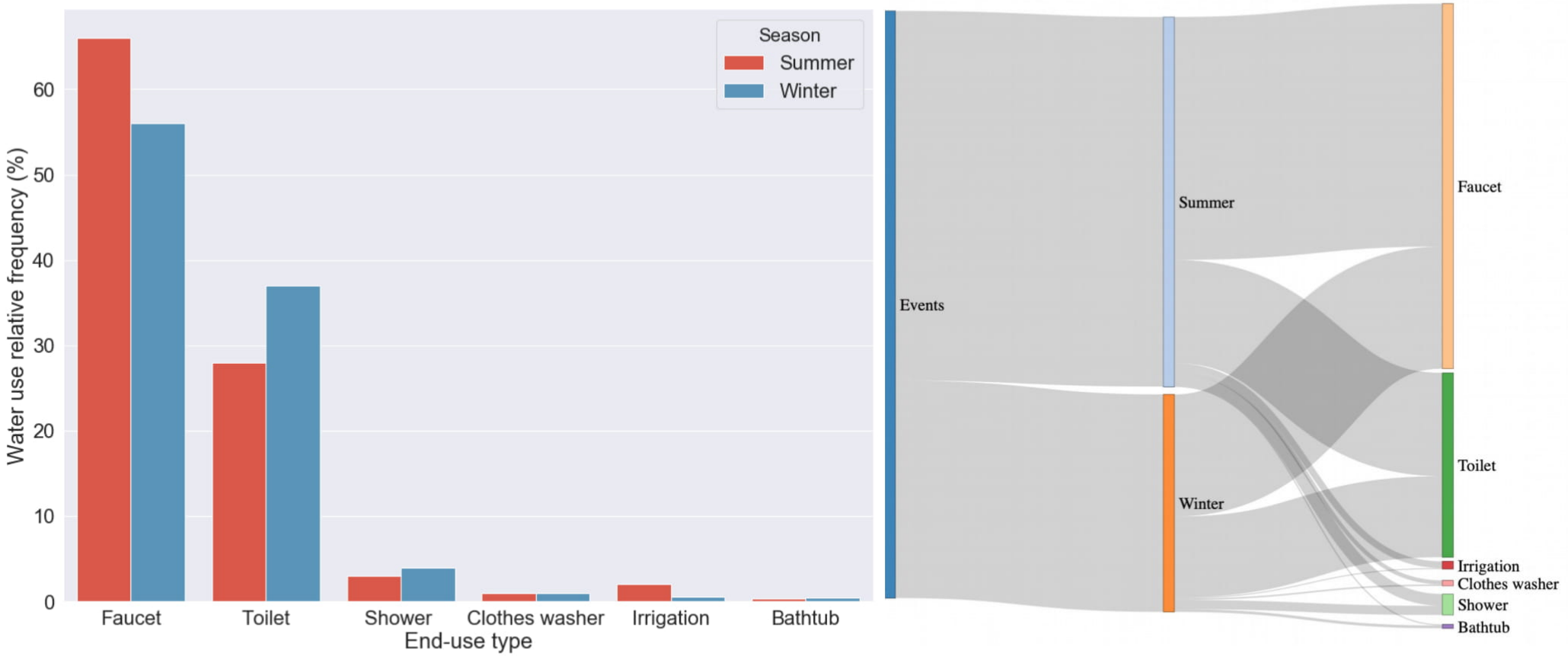


Results: Overall Water Use





Results: Characterizing Individual End Uses



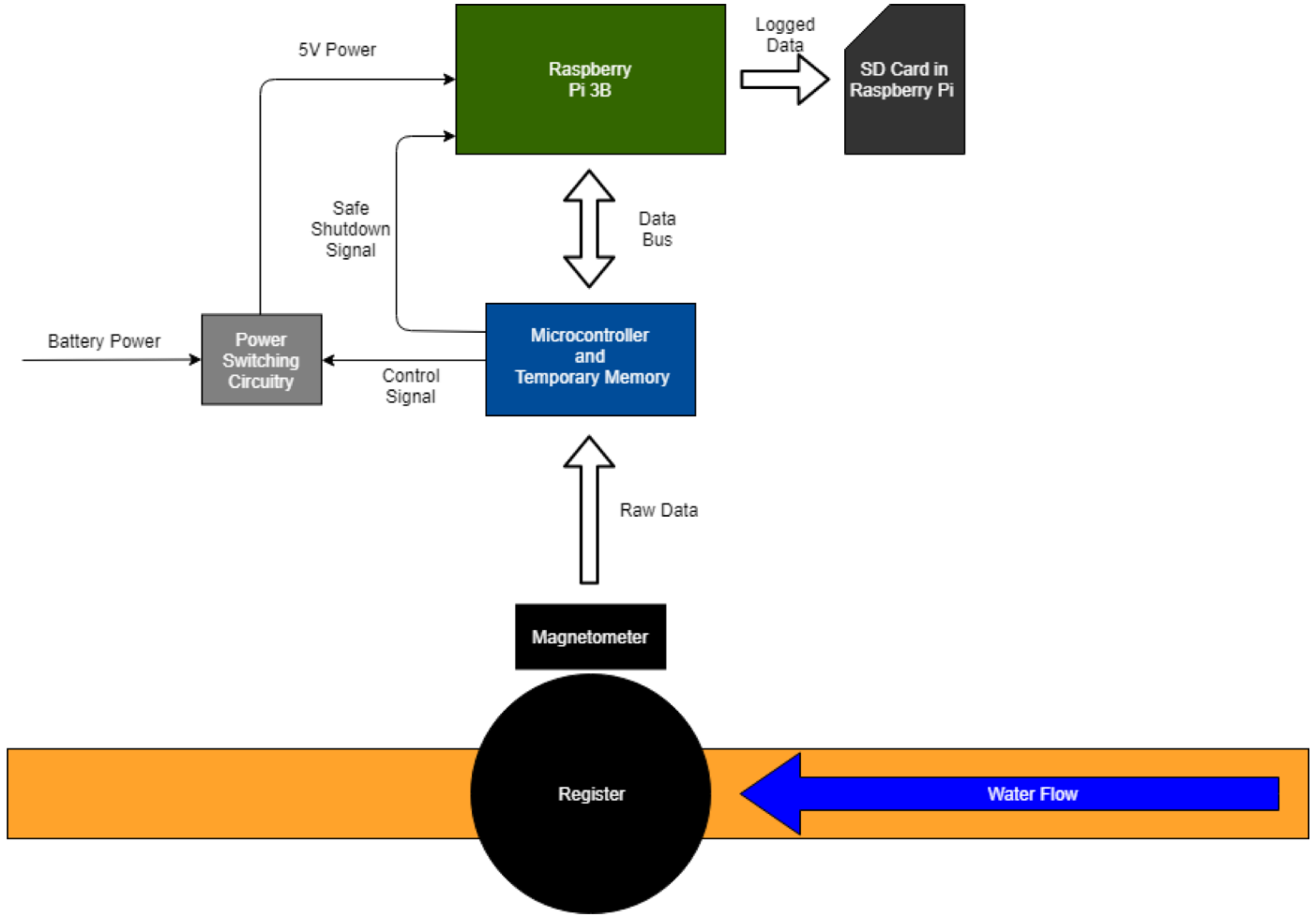
Conclusions

- Open source and semi-supervised water end use disaggregation and classification tool
- The tool employs non-intrusive monitoring data collected at high temporal resolution from a residential home's water meter along with machine learning techniques
- Driven by the fact that, for most other studies that have worked on end use disaggregation algorithms, neither the source code nor the data are available for testing or further advancement



What's coming?

- **<u>Edge Computing</u>**:
  - Computation and disaggregation of individual water use events at the data collection site
  - "Shrink" the Big Data produced by smart meters
  - Overcome storage and communication bandwidth limitations and
  - Provide more timely, decision-relevant information



Our Team

- **Faculty**:
  - **Jeff Horsburgh**
- **Graduate students**:
  - **Camilo Bastidas**: PhD working on smart meter development and residential water use studies
  - **Nour Attaallah**: Phd working on end use disaggregation algorithms and modeling
- **Undergrads**:
  - **Josh Tracy**: Electronics prototyping for logging water use data
  - **AJ Backwith**: Electronics prototyping for processing water use data remotly



Questions?

Jeffery S. Horsburgh, Nour A. Atallah, and Camilo Bastidas

jeff.horsburgh@usu.edu, nour.atallah@usu.edu, camilo.bastidas@usu.edu

Utah Water Research Laboratory

Department of Civil and Environmental Engineering

Utah State University

COLLEGE of ENGINEERING
UtahStateUniversity

UtahStateUniversity
UTAH WATER RESEARCH LABORATORY

Civil & Environmental ENGINEERING