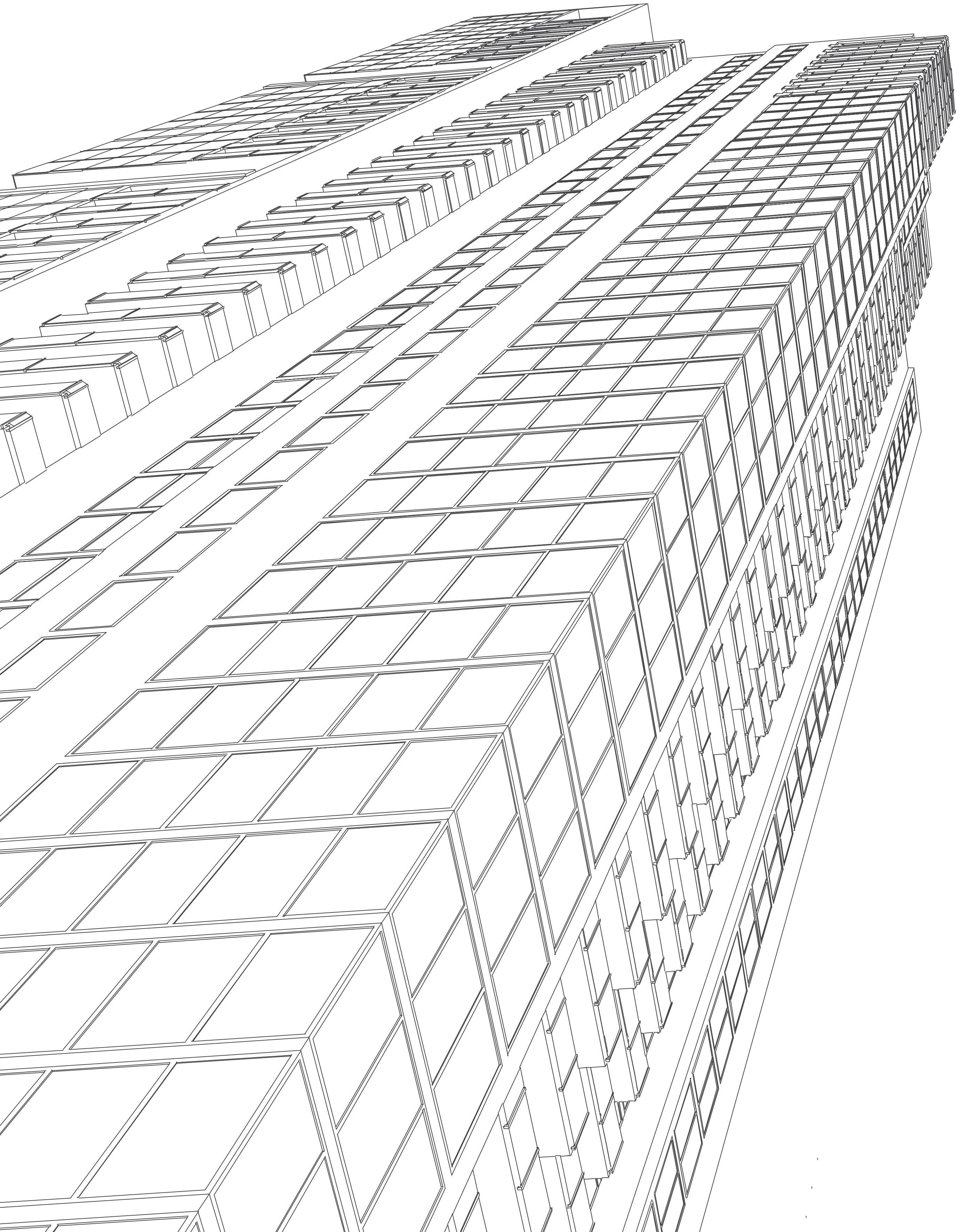
BTS: Building Timeseries Dataset: Empowering Large-Scale Building Analytics

Arian Prabowo[1], Xiachong Lin[1], Imran Razzak[1], Hao Xue[1], Emily W. Yap[2], Matthew Amos[3], Flora D. Salim[1]

[1]University of New South Wales (UNSW)
[2]University of Wollongong (UoW)
[3]Commonwealth Scientific and Industrial Research Organisation (CSIRO)

Motivation



Building play a crucial role in human well-being, influencing occupant comfort, health, and safety. Additionally, they contribute significantly to global energy consumption, accounting for one-third of total energy usage, and carbon emissions.

Optimizing building performance presents a vital opportunity to combat **climate change** and **promote human flourishing.**

However, research in building analytics has been hampered by the lack of accessible, available, and comprehensive real-world datasets on multiple building operations.

BTS is a multi-year, multi-building, multi-modal operational focus dataset for bridging this research gap.

UNSW SYDNEY

Building Timeseries Dataset

3 Buildings

3 Years

>15,000 Time Series

>300 Classes

Knowledge Graph for Building Metadata

Literature Gap

Despite the importance and urgency of advancing building analytics, there is a lack of required datasets on buildings with properties. These properties include:

Publicly Available
Does not require permission from the data provider.

Real-world
Enable real-world insights, not simulated

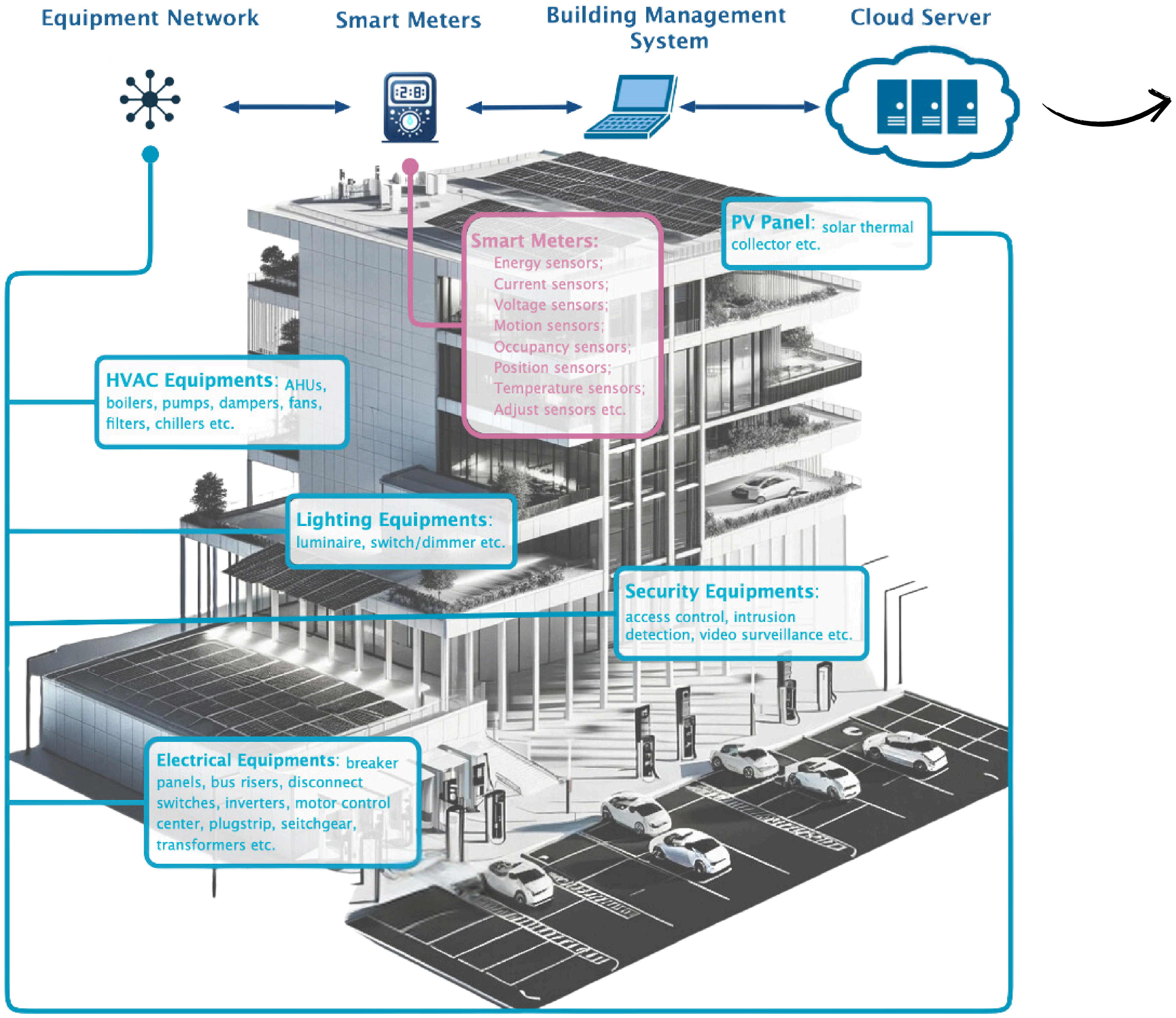
Freely Accessible
No financial cost for acessing the dataset.

Comprehensive
Provide comprehensive information for building activities, such as comfort factors, HVAC, electricity usage, occupant data etc.

Operational Focus
On building operations, e.g. not blueprints

	Datasets
Private	HVAC [70, 35, 79, 72, 32, 31, 21], energy use [62, 63], timeseries ontology classification [36, 37, 25, 44, 45, 68], and simulation [78].
Paid	Pecan Street [15].
Upon discretion of the data provider	ecobee [22]. Mortar [24] (Not freely available from the website (https://mortardata.org/) as per 13 August 2024, awaiting imrpovement in infrastructure).
Static	EUBUCCO [55], PLUTO [18], GBMI [10], Roofpedia [90], HBD3D [9], and Google Research's Open Buildings [76].
Corase temporal granularity (more than daily)	CBECS [17], BERTOOL [77], CENED+2 [69],
Simulation-based	BEM4CBECS [2, 94, 95, 93], ResStock [87], ComStock [60], CityLearn Challenge Series [84, 56, 59, 58], BuildingBench [23], and hardware-in-the-loop laboratory [67, 66].
Limited scope	SLRHOME [5], LCLD [81], and UCI [80]
NILM	Non-intrusive load monitoring (NILM) is task and many dataset have been made for this task check this recent survey [61] that list publicly available dataset. However, since the datasets are only made for this specific task in mind, the scope is limited to only electricity submetering. Other datasets with focus on submetering: BDG [54] and BDG2 [53].
Occupant behaviour	From AshraeOB [19, 49] website: "The ASHRAE Global Occupant Behavior Database aims to advance the knowledge and understanding of realistic occupancy patterns and human-building interactions with building systems. This database includes 34 field-measured occupant behavior datasets for both commercial and residential buildings, contributed by researchers from 15 countries and 39 institutions covering 10 different climate zones. It includes occupancy patterns, occupant behaviors, indoor and outdoor environment measurements."
Comprehensive	Lawrence Berkeley National Laboratory building 59 (LBNL59) [38, 51] and **BTS (ours)** https://github.com/cruiseresearchgroup/DIEF_BTS.

Data Collection



Data scientist can access the building data through the cloud server and further analyze.

However, data collected from different IoT devices may follow various protocols for data storage, leading massive information which is hard to organize.

To address this problem, building engineers involve the **Centralized Management Tool** (i.e., Brick Schema), with standardized ontologies to format the building metadata using a Knowledge Graph, ensuring consistency and compatibility across analyses.

BTS is comprised of data collected onto CSIRO's Data Clearing House (DCH) digital platform. Connecting to the Building Management Systems (BMS), times series data is collected from IoT devices within the buildings and uploaded using Message Queuing Telemetry Transport Secured (MQTTS).

Building Timeseries Dataset

The Building Time Series (BTS) dataset is a **multi-year time series** dataset collected from **three anonymized buildings** in Australia. The dataset includes the following components:

Multivariate Time Series

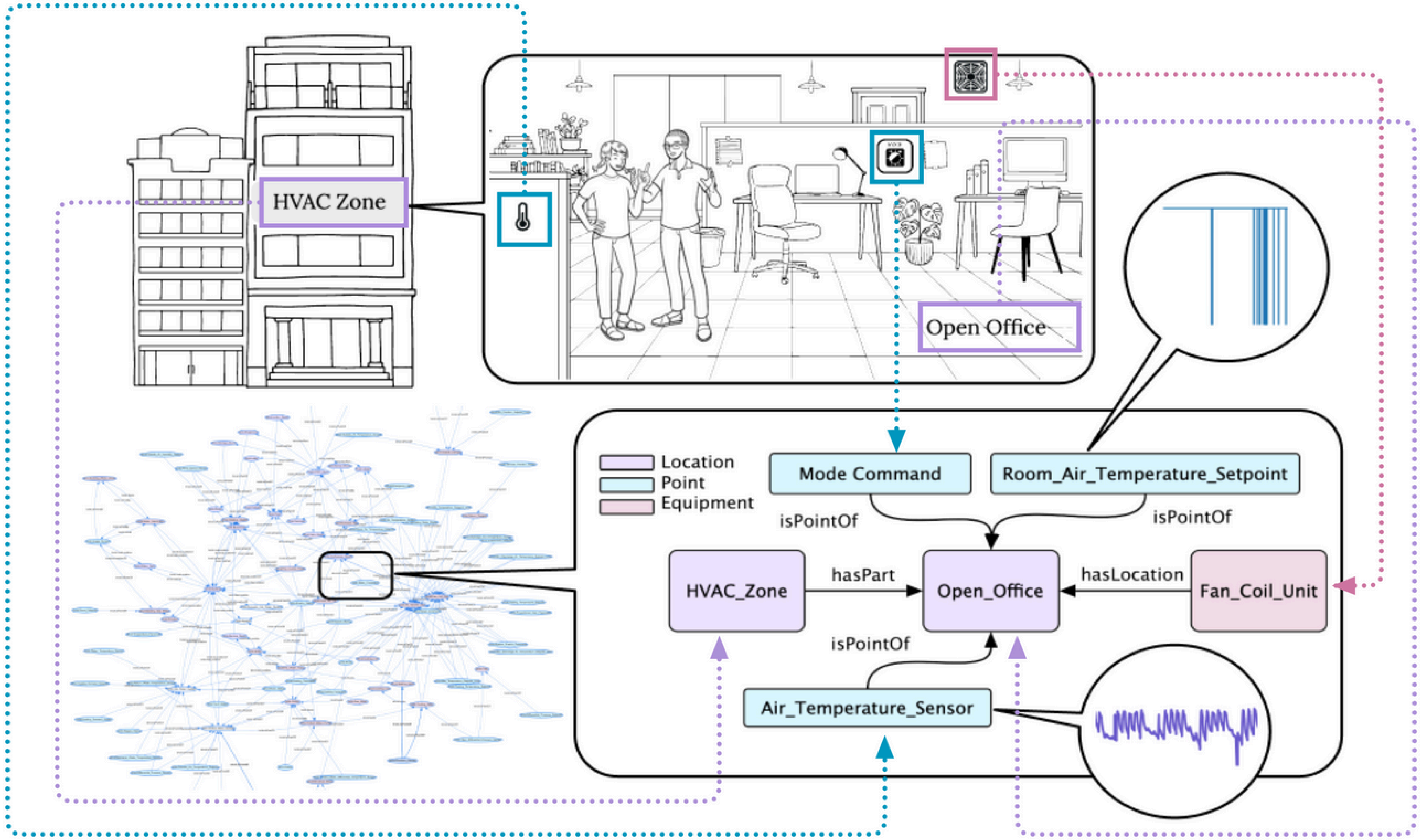
Spanning 2021 to 2024, contains over **15,000 time series** across **300 unique classes**.

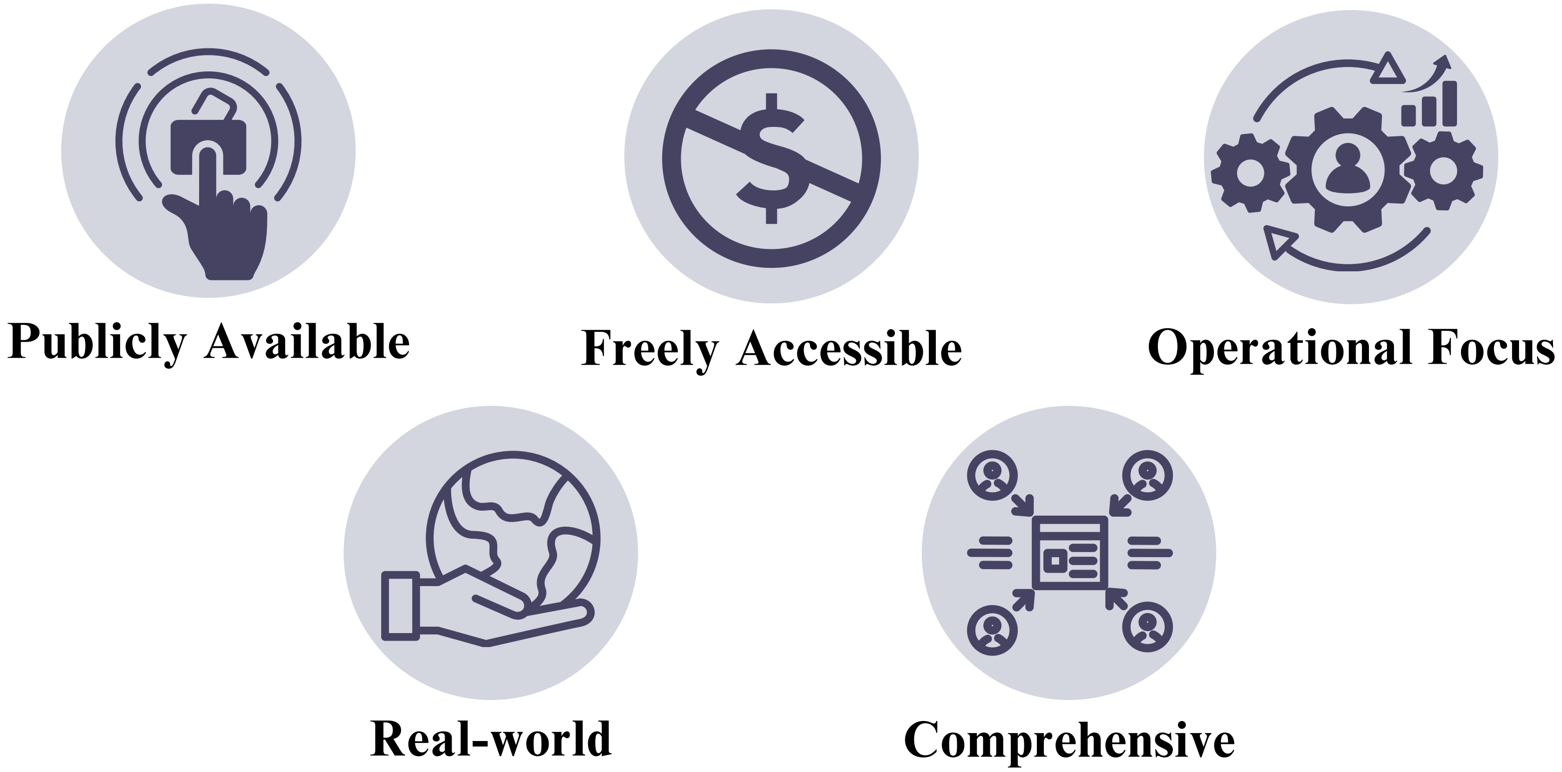
Knowledge Graph

In addition to the time series data, BTS includes a **metadata** schema in the form of a knowledge graph that captures the **relationships** between time series and their associated physical, logical, and virtual entities.

We use the Brick schema:

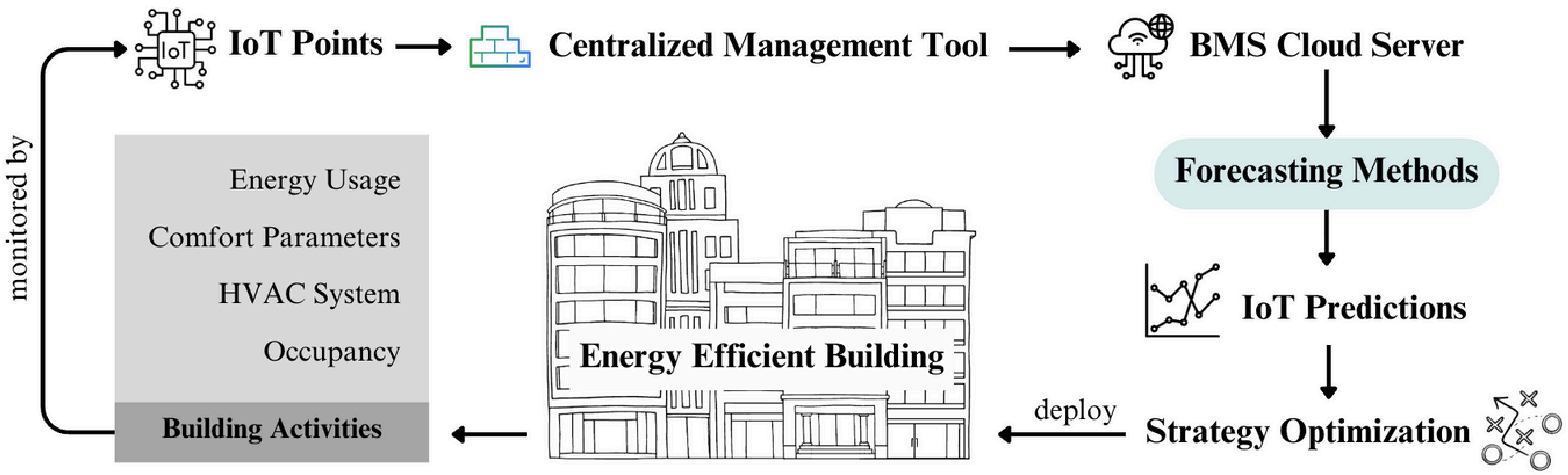
Brick: A Uniform Metadata Schema for Buildings





Potential applications include but not limited to:

- Optimizing energy, emissions, and occupant comfort.
- Developing AI-powered chat systems and copilots for smart buildings.



THREE Reasons to Use BTS

1

BTS contributes to building analytics and sustainable future studies.

2

BTS captures the complexities of real-world building operations.

3

BTS enables fundamental machine learning research.

UNSW SYDNEY

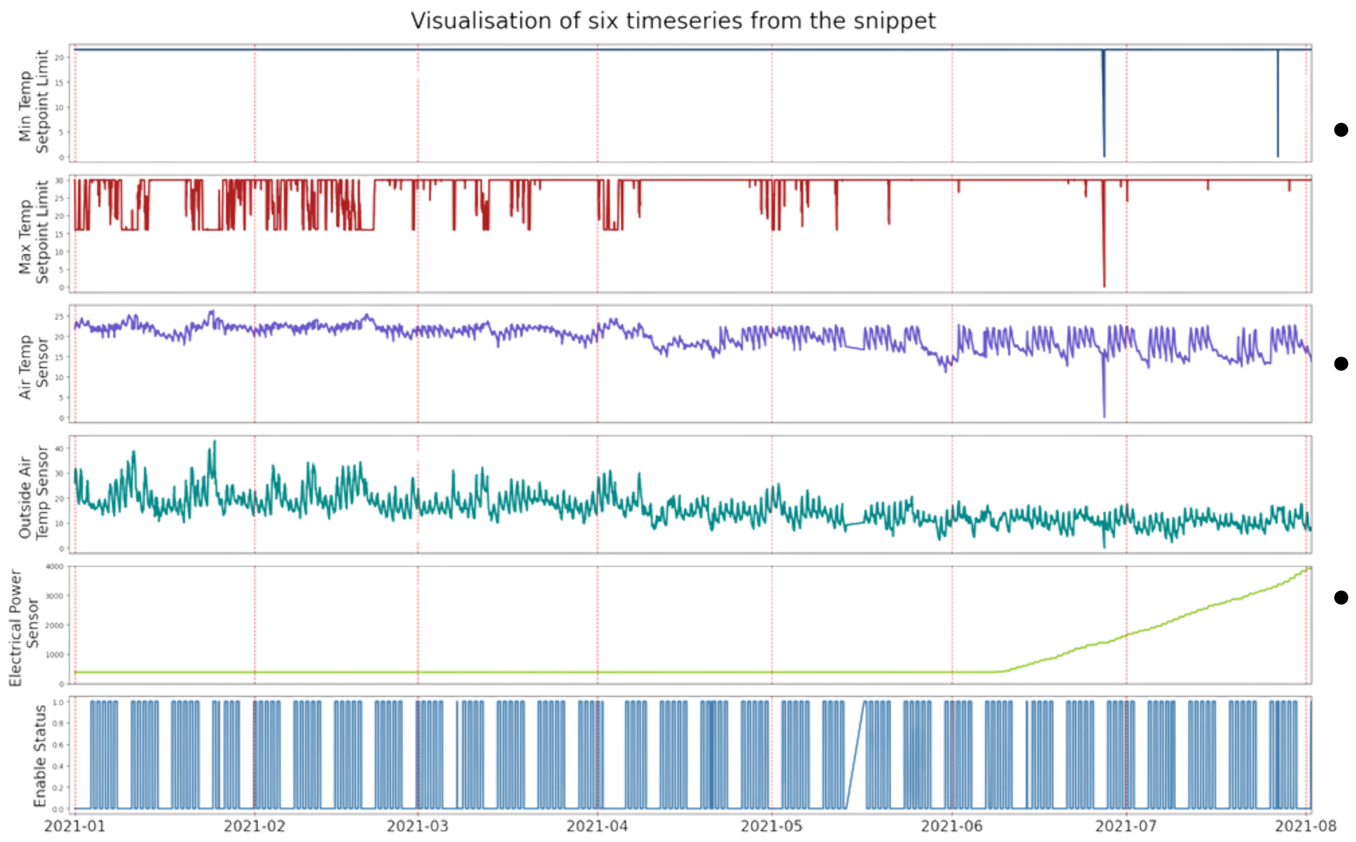


Figure1: Visualization of Sample Time Series

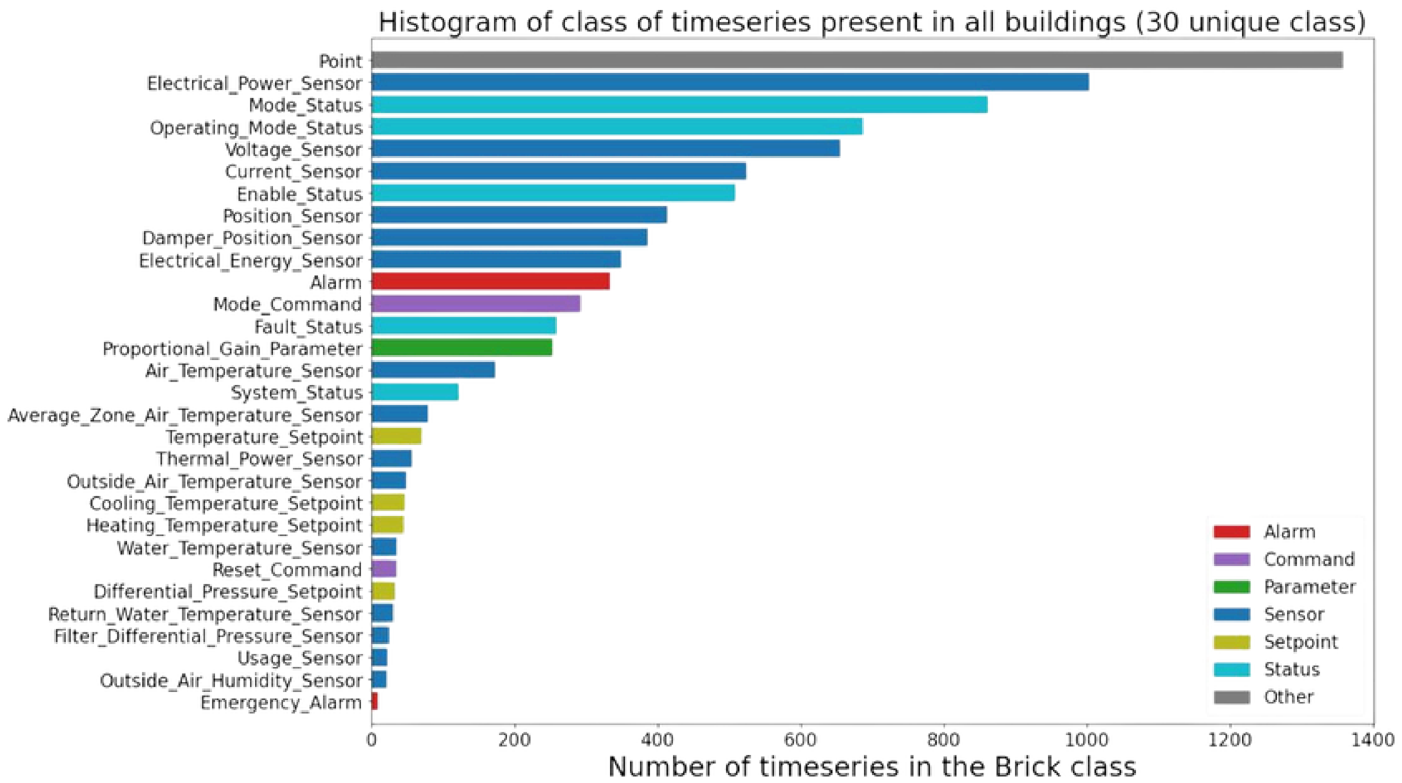


Figure2: Histogram of Class of Time Series by Building

Real-world Challenges

- Temporal Irregularity
  - Irregular sampling frequency
- Spatial Heterogeneity
  - Domain shift in inter-building scenarios
- Long-tail Distribution
  - Data: As shown in *Figure1*, data streams demonstrate various fluctuating shapes. e.g., binomial, normal, lognormal, or display as a flatten line.
  - Class: Sensors utilization rates are varied, as *Figure 2*, some sensors (e.g., temperature sensor) are very common, while some (i.e., dewpoint sensor) are rare.

THREE Reasons to Use BTS

UNSW SYDNEY

1

BTS contributes to building analytics and sustainable future studies.

2

BTS captures the complexities of real-world building operations.

3

BTS enables fundamental machine learning research.

	Count (Unique)	LBNL59		BTS_A		BTS_B		BTS_C	
Top Level	Collection	0	(0)	4	(2)	2	(2)	8	(1)
	Equipment	59	(3)	547	(24)	159	(25)	963	(41)
	Location	73	(3)	481	(9)	68	(17)	381	(26)
	Point	230	(11)	8374	(126)	851	(57)	10440	(159)
	Timeseries	337		8349		851		5347	
Point Subclass	Alarm	0	(0)	798	(16)	5	(2)	109	(8)
	Command	0	(0)	363	(6)	97	(5)	785	(13)
	Parameter	0	(0)	79	(6)	36	(2)	935	(17)
	Sensor	144	(8)	4396	(56)	266	(25)	4062	(68)
	Setpoint	86	(3)	772	(26)	232	(16)	1629	(41)
	Status	0	(0)	1628	(17)	110	(6)	2187	(19)
	Location	Berkeley, USA		Undisclosed locations in Australia					
	Start Date	01-01-2018		01-01-2021		01-01-2021		23-06-2021	
	End Date	31-12-2020		31-12-2023		31-12-2023		18-01-2024	
	Duration (Days)	1094		1094		1094		939	
	Size Zipped (GB)	0.26		8.48		1.31		8.98	

Compared to the only comparable dataset in existing literature, LBNL59, BTS demonstrates a **larger scale**, a **higher number of buildings**, and **greater diversity in building classes** (unique counts provided in brackets).

THREE Reasons to Use BTS

1

BTS contributes to building analytics and sustainable future studies.

2

BTS captures the complexities of real-world building operations.

3

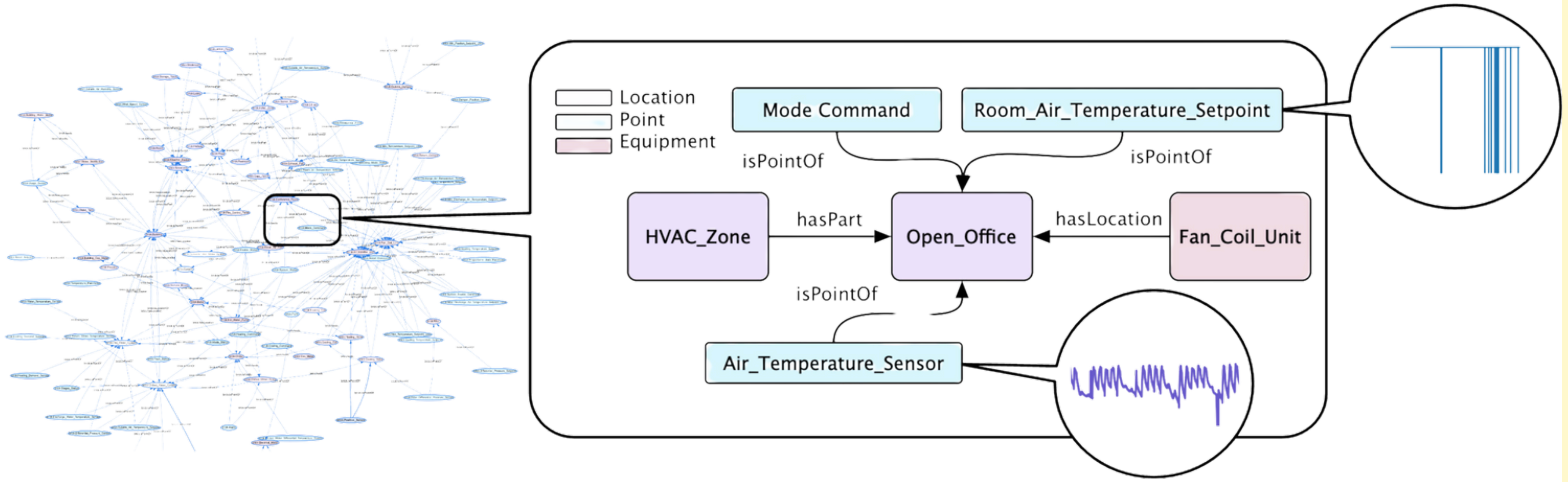
BTS enables fundamental machine learning research.

UNSW SYDNEY

Multi-modal

Knowledge Graph + Time Series

- Exploring multimodal learning with knowledge graphs.
- Tackling unbalanced multivariate time series with long-tail distributions.



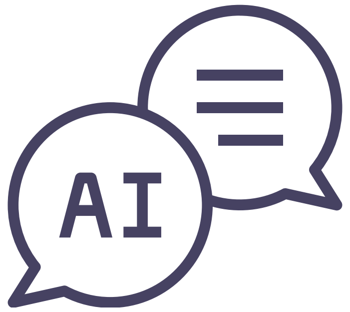
Multi-building

Three Buildings with multi-modal details

- Addressing domain shift and domain adaptation.

Potential Task

- Forecasting
- Anomaly Detection
- AI Chat for Building or Building Q&A



THREE Reasons to Use BTS

1

BTS contributes to building analytics and sustainable future studies.

2

BTS captures the complexities of real-world building operations.

3

BTS enables fundamental machine learning research.

UNSW SYDNEY

Benchmark1: Multi-label Classification

Brick schema was developed to aid in data interoperability across buildings. However, constructing the Brick schema for each building requires expensive and error prone manual expert labor to classify time series data into the correct Brick ontology. Past studies have attempted to automate this process with ML relied on private data and did not release their code. This benchmark is the first to address the task using publicly available data. We formulated this task as a multi-label classification task, where a label will also return true for all super-classes and return as zero for all subclass (see the figure in the left bottom).

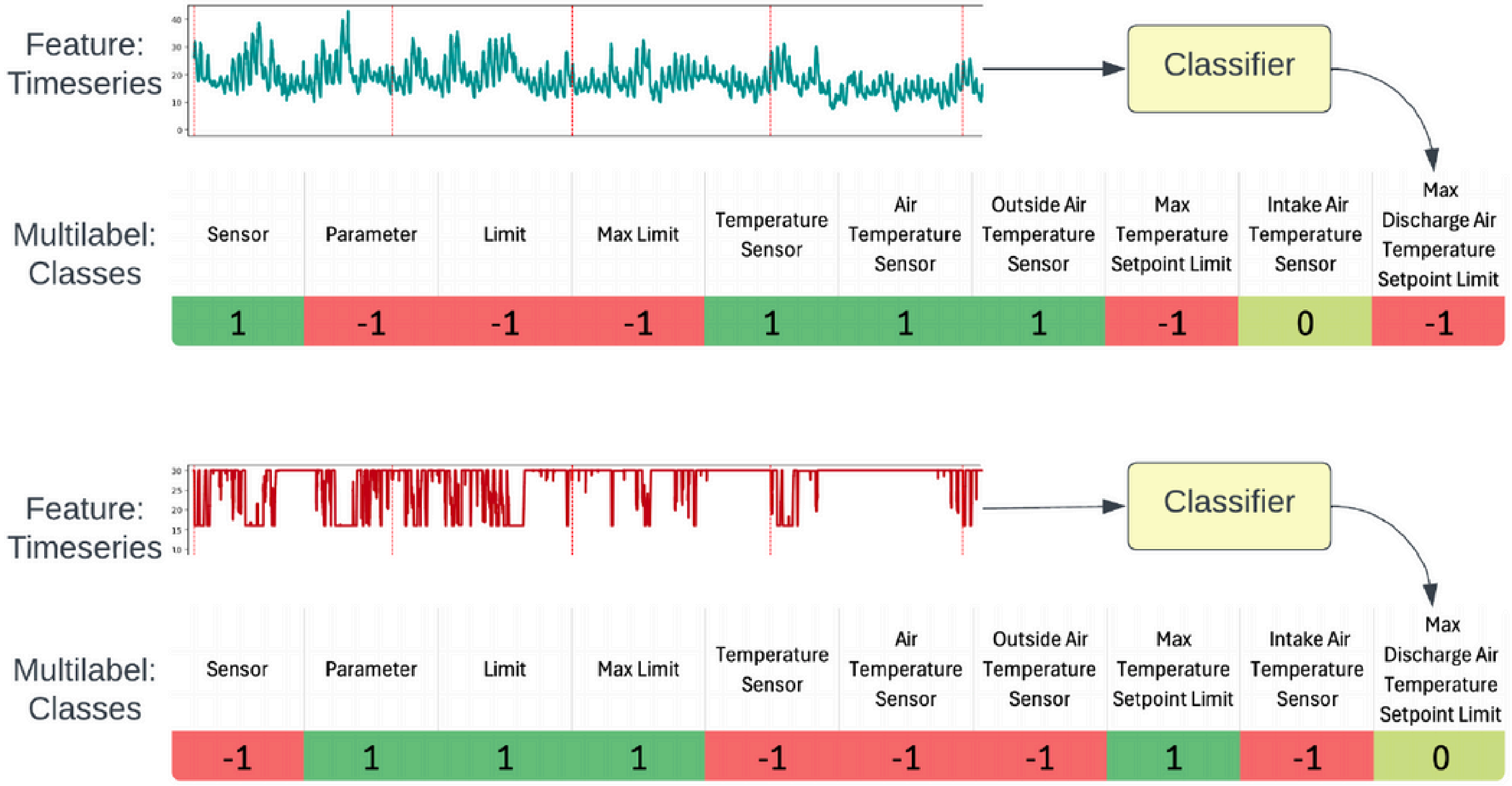


Figure: Visualization of the multi-label time series classification task.

Method	Accuracy		F1		mAP	
Zero	0.8484	±N/A	0.0000	±N/A	0.0000	±N/A
Mode	0.8592	±N/A	0.1296	±N/A	0.0990	±N/A
Random Proportional	0.8147	±0.0001	0.1487	±0.0002	0.1520	±0.0001
Random Uniform	0.4999	±0.0002	0.1813	±0.0002	0.1520	±0.0001
One	0.1516	±N/A	0.2234	±N/A	0.1516	±N/A
LR	0.2366	±N/A	0.0882	±N/A	0.0497	±N/A
XGBoost	0.8593	±N/A	0.2697	±N/A	0.2627	±N/A
Transformer (default)	0.7807	±0.0139	0.3360	±0.0116	0.3171	±0.0078
Transformer (HP tuned)	0.8052	±0.0074	0.3615	±0.0079	0.3489	±0.0057
Informer	0.7627	±0.0010	0.3162	±0.0019	0.2849	±0.0030
DLinear	0.7030	±0.0042	0.2499	±0.0020	0.2494	±0.0010
PatchTST	0.7534	±0.0017	0.2981	±0.0014	0.2721	±0.0013

Table: Benchmark results on the time series ontology multi-label classification task. Deterministic methods do not have standard deviation.

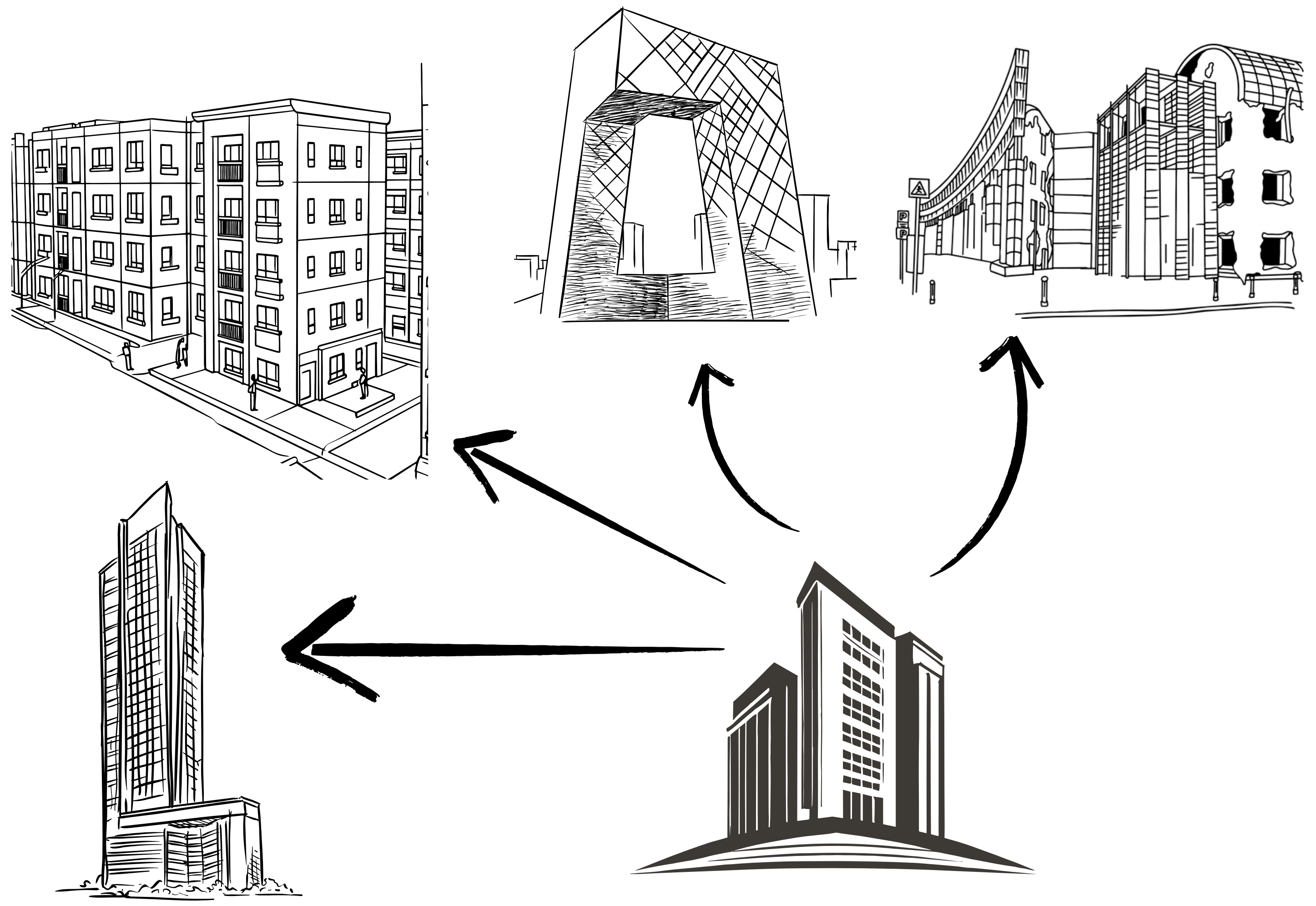
UNSW SYDNEY

Benchmark2: Zero-shot Forecasting

The advent of building digitalization presents significant opportunities for leveraging deep learning methods in building management systems for accurate forecasting. In practical applications, it is crucial for well-trained models to be applicable across diverse building scenarios without retraining costs. However, specific building constraints, operational variances, functionality differences, and data heterogeneity pose significant challenges in real-world settings. Models must adapt to dynamic ontology changes when applied to different buildings. Previous studies often rely on identical features and well-processed data, not reflecting the complexity of real-world scenarios.

		BTS-A		BTS-B		BTS-C	
		MAE	SMAPE	MAE	SMAPE	MAE	SMAPE
Previous Day Persistence		0.5377	48.1539	0.4976	43.2985	0.5458	45.7014
Previous Week Persistence		0.6190	57.2713	0.5918	51.3867	0.6499	58.1922
BTS-A	DLinear	N/A		0.4324	35.9846	0.4262	36.2734
	PatchTST	N/A		0.3748	29.2570	0.3712	29.5552
	Informer	N/A		0.5968	49.2217	0.5920	51.9745
	iTransformer	N/A		0.4026	31.1924	0.3842	30.1102
BTS-B	DLinear	0.4940	41.2264	N/A		0.4206	35.3121
	PatchTST	0.4575	36.7689	N/A		0.3711	29.2135
	Informer	0.5233	45.9279	N/A		0.4592	39.7068
	iTransformer	0.4783	37.5907	N/A		0.3901	29.9940
BTS-C	DLinear	0.4858	40.7421	0.4158	34.1473	N/A	
	PatchTST	0.4542	36.9451	0.3723	28.9325	N/A	
	Informer	0.5213	46.6112	0.4602	39.7162	N/A	
	iTransformer	0.4859	39.5158	0.4262	32.6550	N/A	

Table: Benchmark results on the zero-shot forecasting task. The columns refer to the training set, whereas the row represents the testing set. Please find the paper appendix for the standard deviations.

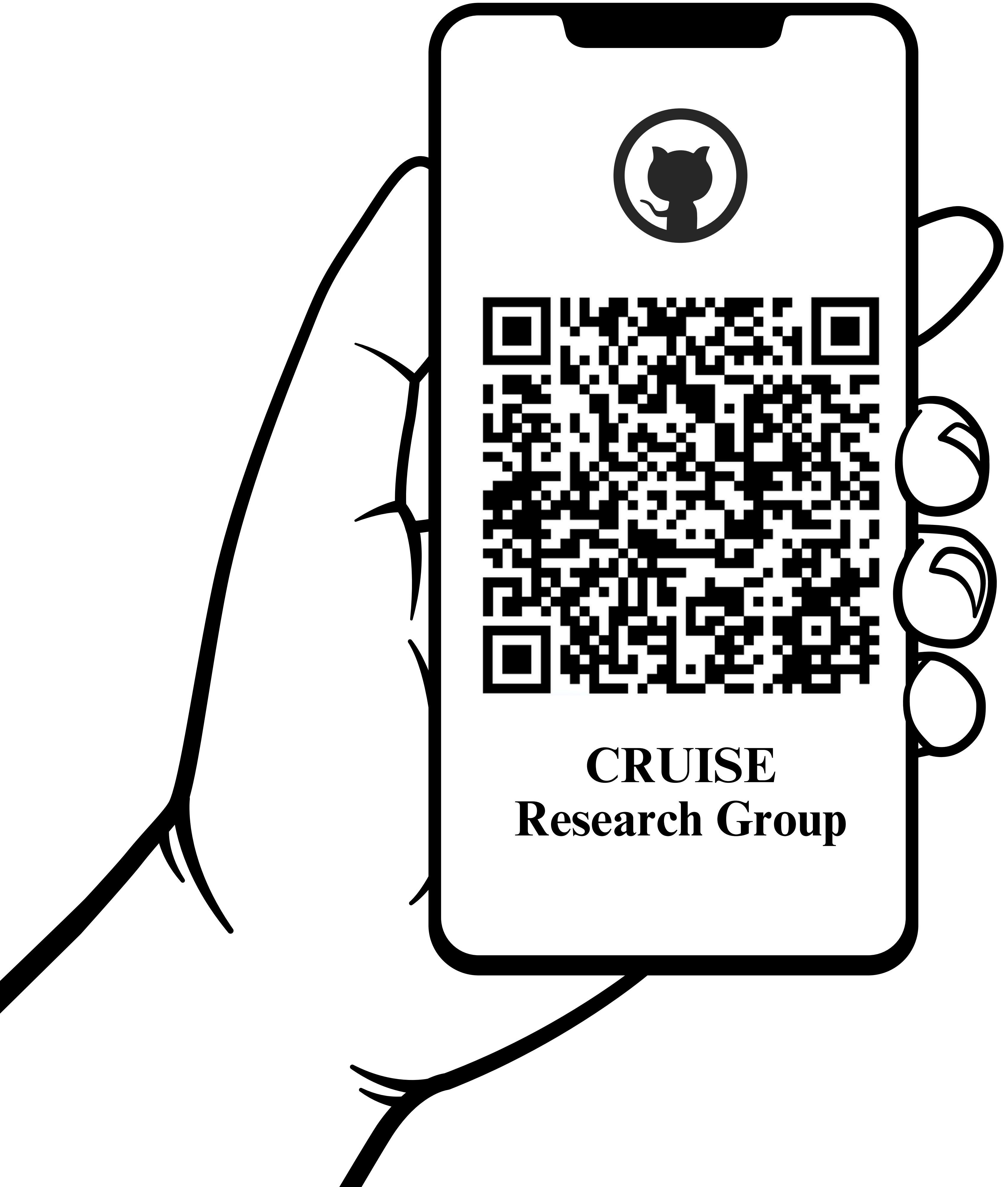


Call For Challengers

Brick-by-Brick: Automating Building Data Classification Challenge
is a time series classification challenge based on this dataset.

https://www.aicrowd.com/challenges/brick-by-brick-2024

We welcome researchers from all over the world
to participate in this challenge and have fun.



Scan to follow our GitHub repository and access the dataset, benchmark experiment code, visualization tools, and competition details.

flora.salim@unsw.edu.au

github.com/cruiseresearchgroup/DIEF_BTS

UNSW SYDNEY

Acknowledgement

This research is supported by the NSW Government through the CSIRO's NSW Digital Infrastructure Energy Flexibility (DIEF) project, funded under the Net Zero Plan Stage 1: 2020-2030.

This project is also funded by the Reliable Affordable Clean Energy for 2030 (RACE for 2030) Cooperative Research Centre.

This research was undertaken with the assistance of resources from the National Computational Infrastructure (NCI Australia), an NCRIS enabled capability supported by the Australian Government.

Click for NSW DIEF webpage

Click for Race2030 webpage

Click for NCI webpage

UNSW SYDNEY

Thank You