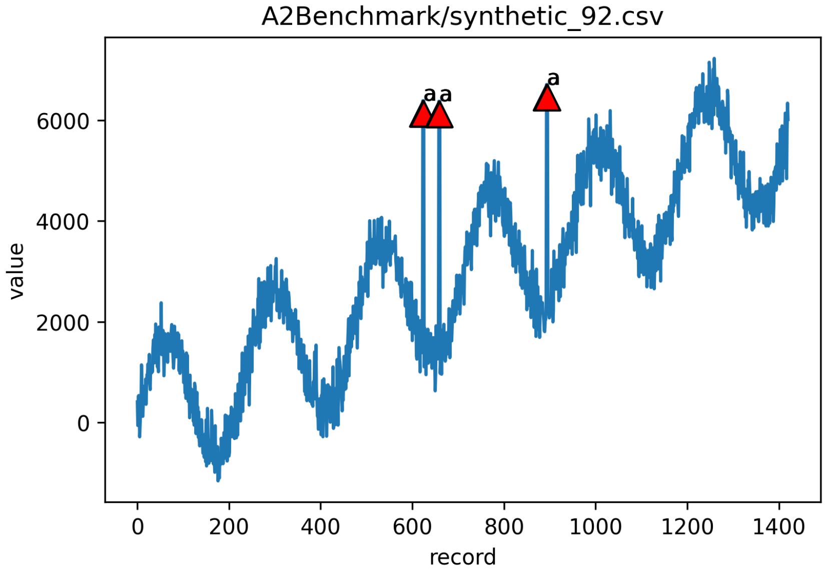
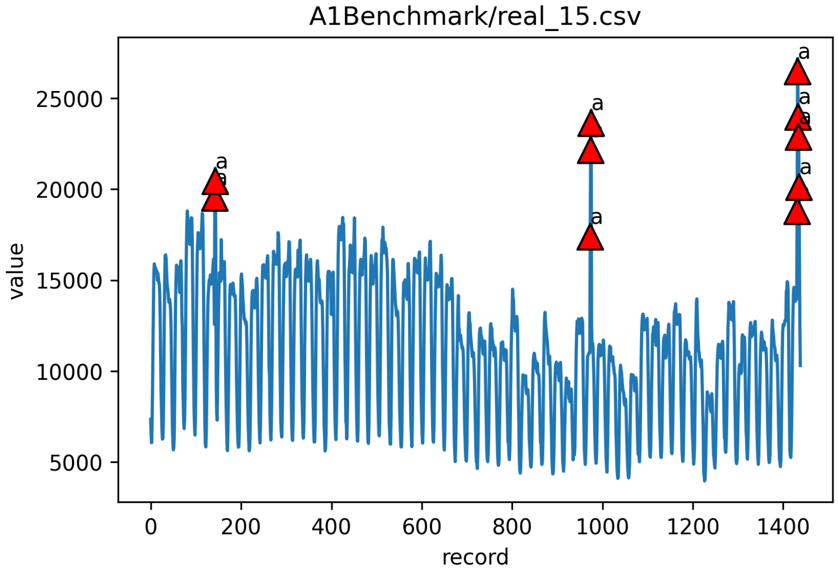
Dataset Summary

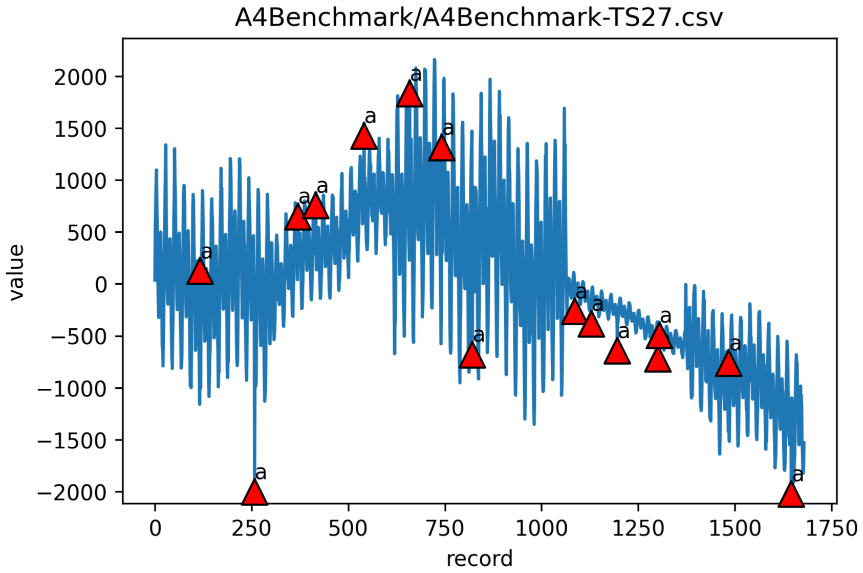
Anomaly Types

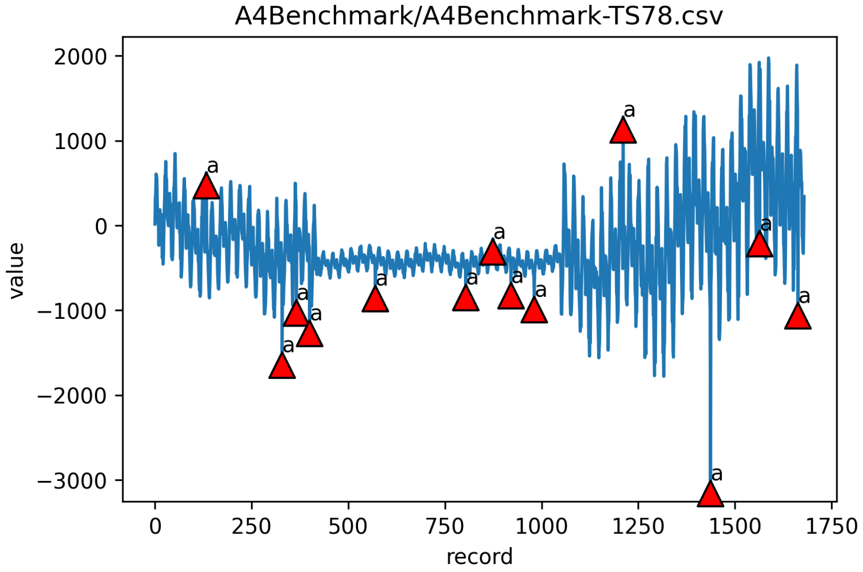
- **Point Anomalies**: These anomalies can only be identified given a specific context, but not otherwise.





- **Changpoint Anomalies**: This type of anomaly indicates an anomalous behaviour on a more global scale, for example in terms of trend and seasonality.

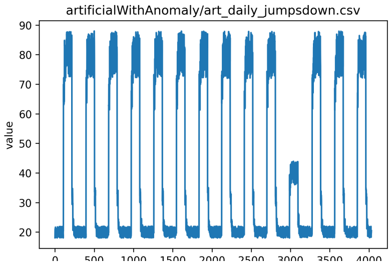


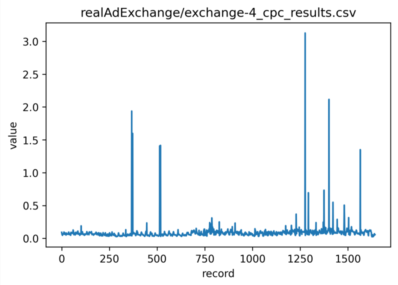
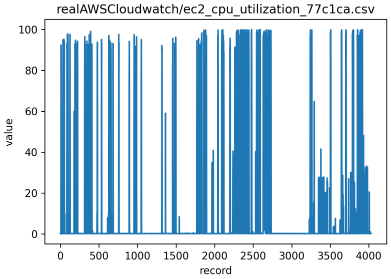
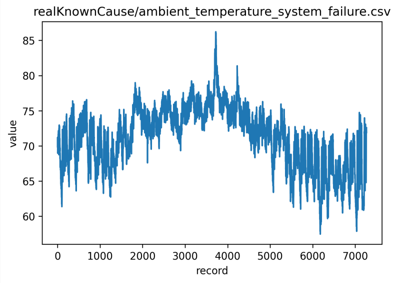
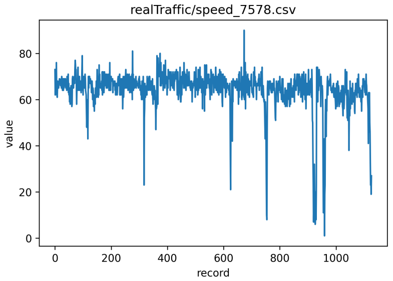
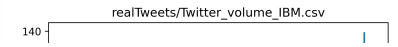


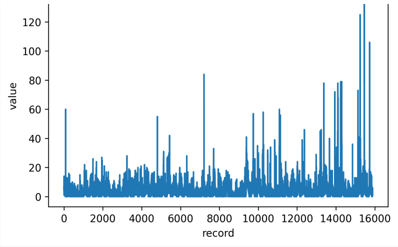
- **Sequential Anomalies**:

Numenta Anomaly Benchmark(NAB)

Each CSV data file consists of **two time series**, one of them being a series of timestamp values and the second one being series of a input values. Overall, there are **58** data files in NAB. Each time step in real datasets represent **5 minutes** of aggregated traffic.

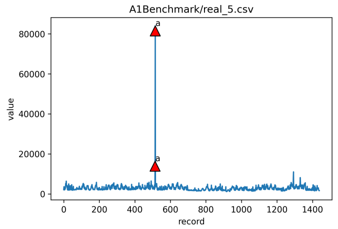
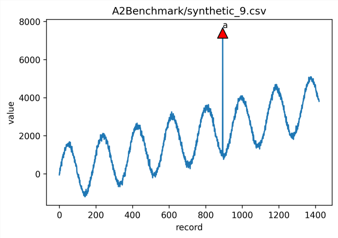
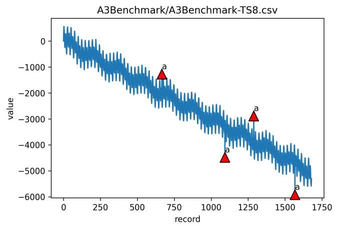
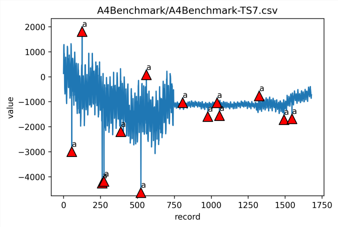
Dataset	Data Files	# of Records	Description	Example
artificialWithAnomaly(synthetic)	artificial	4032	Artificially-generated data with varying types of anomalies.	

				0 500 1000 1500 2000 2500 3000 3500 4000 record
realAdExchange(real)	1. cost-per-click (CPC) 2. cost per thousand impressions(CPM)	1538 ~ 1643	Online advertisement clicking rates	
realAWSCloudwatch(real)	1. EC2/RDS CPU Utilization 2. EC2 Network Bytes In 3. EC2 Disk Read Bytes 4. ELB Requests	4032~4730	AWS server metrics as collected by the AmazonCloudwatch service.	
realKnownCauses(real)	1. ambient temperature in an office setting 2. average CPU usage across a given cluster 3. request latency from a server in Amazon's East Coast datacenter 4. temperature sensor data of an internal component of a large, industrial mahcine 5. the total number of NYC taxi passengers into 30 minute buckets 6. timing the key holds for several users of a computer 7. timing the key strokes for several users of a computer	7267, 18050, 4032, 22695, 10320, 1882, 5315	This is data for which we know the anomaly causes; no hand labeling.	
realTraffic(real)	1. occupancy(persons per vehicle) 2. speed 3. travel time	2380 ~ 2500,1127 ~ 2500, 2162 ~ 2500	Real time traffic data from the Twin Cities Metro area in Minnesota.	
				

realTweets(real)	the number of mentions for a given ticker symbol every 5 minutes.	15831 ~ 15902	A collection of Twitter mentions of large publicly-traded companies such as Google and IBM.	

Yahoo! S5 Dataset

This is a **labeled** anomaly detection dataset. The dataset consists of real and synthetic time-series with tagged anomaly points. The dataset tests the detection accuracy of various **anomaly-types including outliers and change-points**. The synthetic dataset consists of time-series with varying trend, noise and seasonality. The real dataset consists of time-series representing the metrics of various Yahoo services. Each time step represent **a hour** of aggregated traffic.

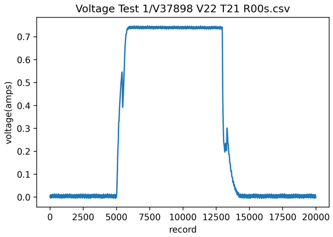
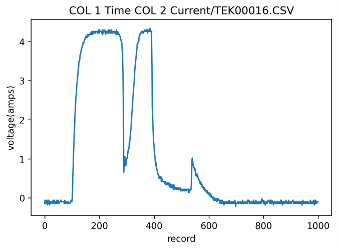
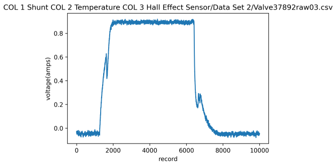
Data class	# of data files	Anomaly types (total frequency)	Description	# of Records	Contamination	Example
A1 (real)	67	contextual and/or collective (1669)	Both **point and window anomalies** occur in these data files.	741~1461	0.0176	
A2 (synthetic)	100	contextual (466)	All metrics from this data class have a constant trend as well as a constant seasonality, only noise is added.	1421	0.0033	
A3 (synthetic)	100	contextual (943)	Input values time series show a varying trend as well as three different seasonalities.	1680	0.0056	
A4 (synthetic)	100	contextual (1045, containing 208 changepoints)	Anomalies are mainly **sudden step changes**.	1680	0.0062	

A4 Data Class

- Timestamps: the UNIX timestamp marks every hour (hourly sampled data)
- Value: time series value at relevant timestamp
- Anomaly: for an outlier value will be 1
- Changepoint: if the change point was there, the value will be 1
- Trend: the additive trend value for this timestamp
- Noise: the additive noise value for this timestamp
- Seasonality1: seasonality value for a period of twelve hours
- Seasonality2: calculated seasonality value for the daily period
- Seasonality3: calculated seasonality value for the weekly period

NASA Shuttle Valve Data

The time series data are solenoid current measurements on a Marotta MPV-41 series valve as the valve is cycled on and off under various test conditions in a laboratory. The valves are used to control fuel flow on the Space Shuttle.

Data Class	# of Data Files	# of Records(Sampling Rate)	Description	Example
Voltage Test 1(COL 1 Shunt | **COL2** Hall Effect sensor)	27	20K samples(1 sample/ 0.1 ms)	The data is ASCII text floating point numbers in two columns; Column 1 is current from one shunt resistor and column 2 is the current as detect by a Hall Effect sensor. The current is in amps.	
COL1 Time **COL2** Current	TEK00000 ~ TEK00003: normal TEK00010 ~ TEK00017: **abnormal**	1K samples(1K samples / s)	This is just waveform data recorded for various forced failures. There are *CVS files which are just raw data. Column 1 is the time of the sample in seconds and column 1 is the current in amps.	
COL1 Shunt COL2 Temperature COL3 Hall Effect Sensor	268	20K samples (10K samples / s)	All the file have the same format: ASCII text floating point numbers. The first column is current data detected by the shunt resistor. The second column is temperature data in Kelvin/100. This temperature is the temperature of the Hall Effect Sensor not the valve solenoid. The third column is the current data as detected by the Hall Effect sensor. The current data is in amps.	

OmniAnomaly Server Machine Dataset(SMD)

SMD (Server Machine Dataset) is a new 5-week-long dataset which was collected by OmniAnomaly authors from a large Internet company, and it was publicly published on Github. The SMD dataset is divided into two subsets of equal size: the first half is the training set and the second half is the testing set. Anomalies and their anomalous dimensions in SMD testing set have been **labeled by domain experts** based on incident reports. Paper: https://netman.aiops.org/wp-content/uploads/2019/08/OmniAnomaly_camera-ready.pdf

Dataset	# of Data Files	# of Dimensions	Training Set Size	Testing Set Size	Anomaly ratio(%)	Metrics
SMD	28	38	708405	708420	4.16	CPU load, network usage, memory usage, etc.

CTF DataSet

CTF_dataset is collected from a top global Internet company, where geo-distributed data centers serve global users. The businesses running on the infrastructure are typical Internet services (e.g., news, advertisement, videos). It contains 533 machine entities, and each is monitored with 49 KPIs. KPIs are collected every 30s spanning 13 days (from April 18th to April 30th). Github: https://github.com/NetManAIOps/CTF_data, Paper: https://netman.aiops.org/wp-content/uploads/2021/02/paper-INFOCOM21-cfp.pdf

Category	Metrics Count	Metrics
CPU	15	CPU idle rate, CPU busy rate, CPU utilization at user or system level, CPU load, etc.
Memory	10	Memory usage or free or available rate, etc.
Sockets	6	Sockets established or closed or orphaned, etc
UDP	7	count of UDP packets sent or received, count of UDP buffer errors sent or received, etc.
TCP	11	TCP retransmisstion rate, TCP listen drops, TCP listen overflows, TCP delayed ACK locked, etc.

NASA SMAP and MSL Datasets

Dataset	# of Data Files	# of Dimensions	Training Set Size	Testing Set Size	Anomaly ratio(%)	Metrics
Soil Moisture Active Passive satellite(SMAP)	55	25	135183	427617	13.13	Telemetry data: radiation, temperature, power, computational activities, etc.
Mars Science Laboratory rover	27	55	58317	73729	10.72	Telemetry data: radiation, temperature, power, computational activities, etc.