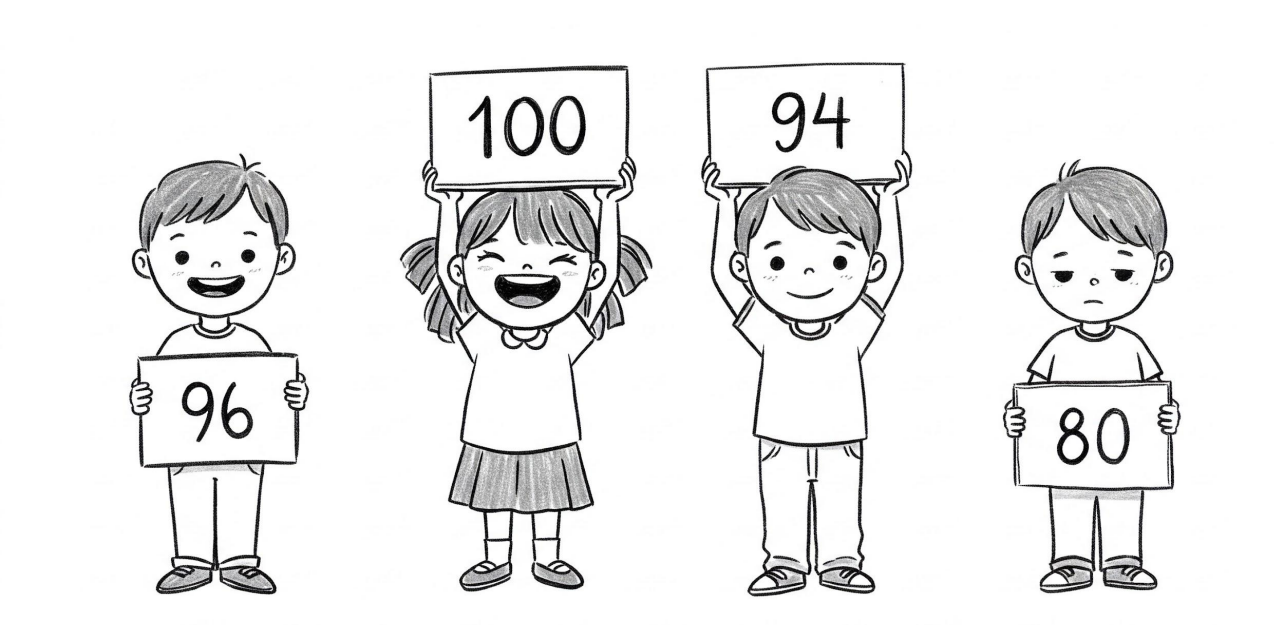
Simplified Streaming Anomaly Detection with Apache Beam's Latest Transform

Shunping Huang

BEAM SUMMIT NYC 2025

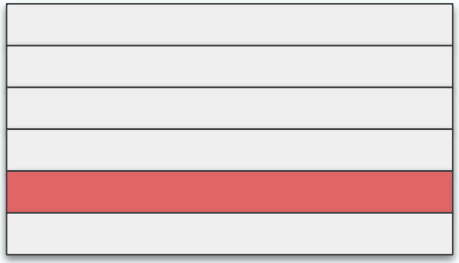
Anomaly Detection

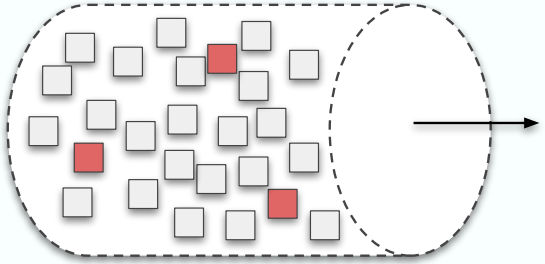


- **Anomaly**: the data instance that is **different** from the **normal** ones.
- **Anomaly Detection**: the task to identify the anomalies.

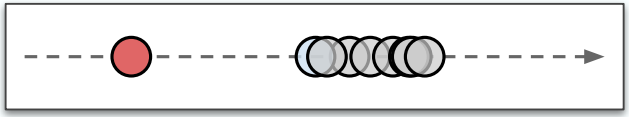
Anomalies in Data

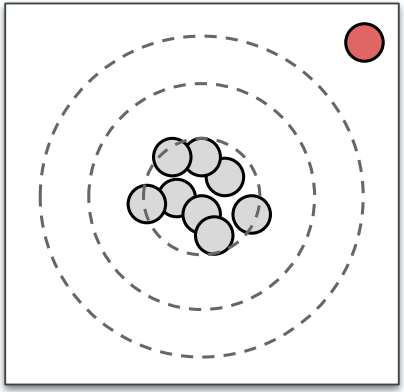
- Bounded vs. Unbounded





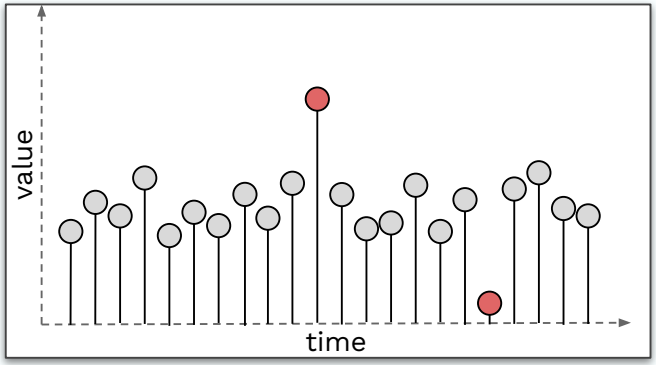
- Univariate vs. Multivariate

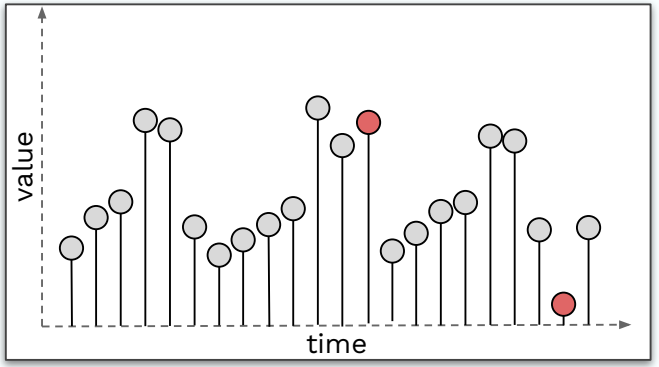




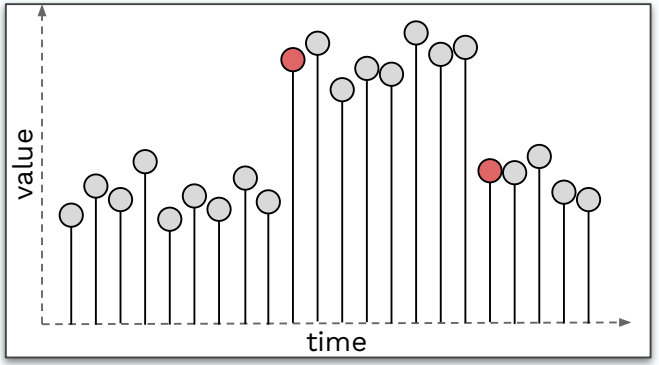
Anomalies in Data (cont.)

- Data Streams vs. Time Series

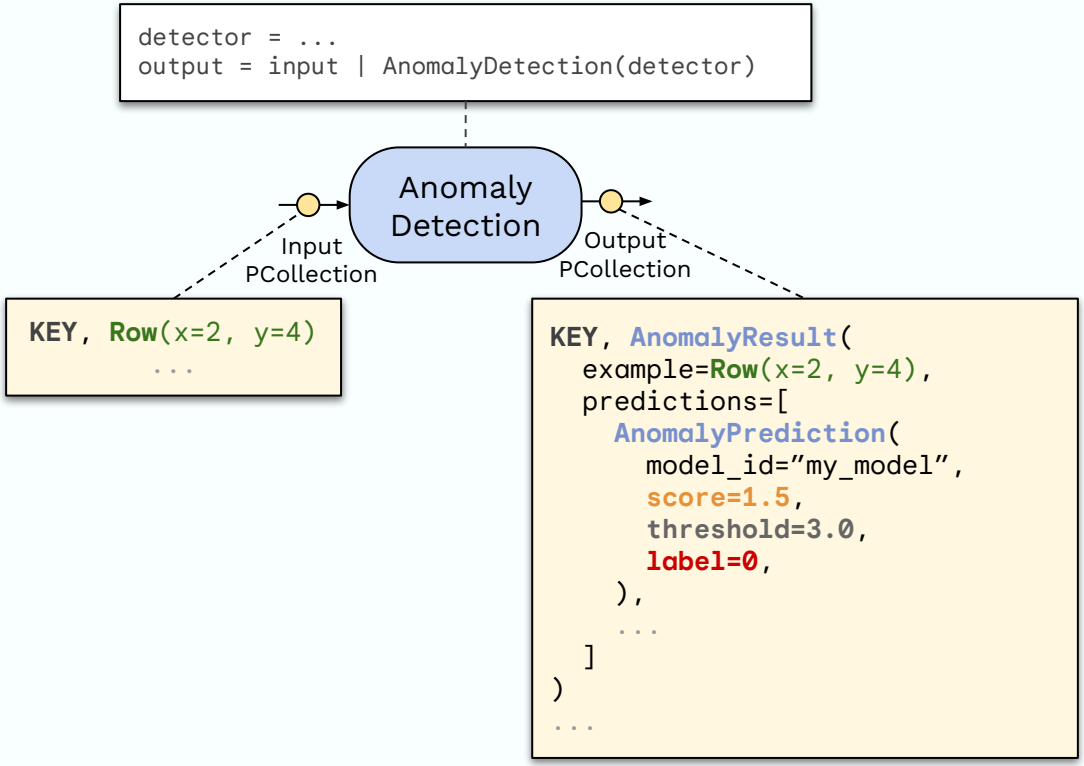




- Data Streams with **Concept Drift**

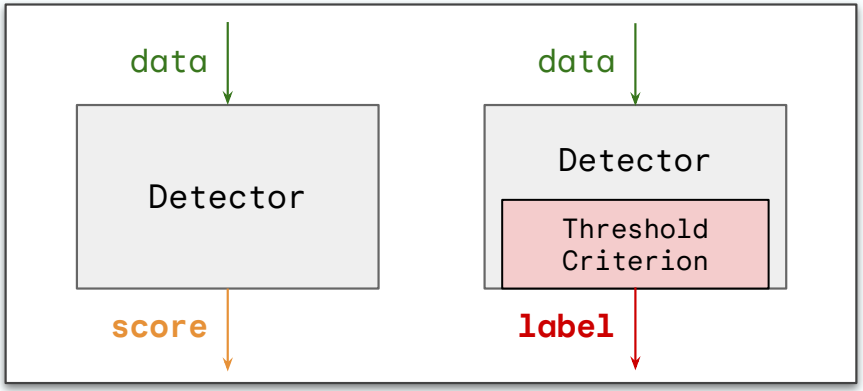


PTransform for Anomaly Detection



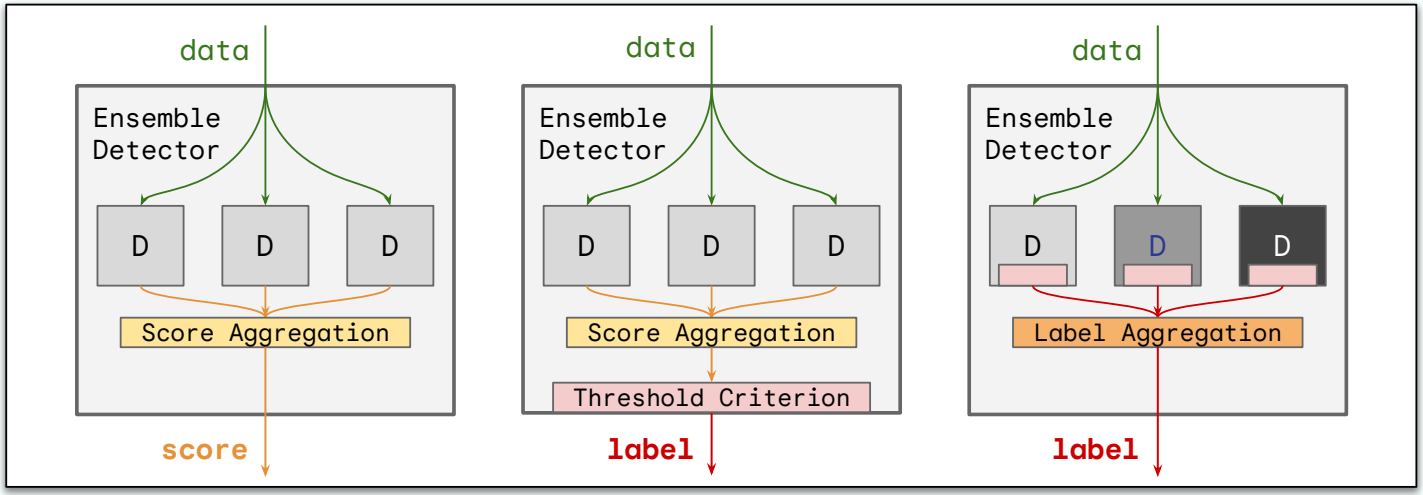
Internals

- Individual Detectors
  - Run a specific anomaly detection method and generate scores or labels
  - e.g. Incremental Z-Score, IQR, and offline models (such as isolation forests, LOF, one-class SVM, etc) supported by PyOD.
- Threshold Criteria
  - Fixed Thresholding, Quantile Thresholding



Internals (cont.)

- Ensemble Detectors
  - Run a set of sub-detectors in parallel and aggregate scores or labels
- Aggregation Strategies
  - For Scores: Average Score, Max Score
  - For Labels: Majority Vote, All Vote, Any Vote



Demo

DEMO


IBM Cloud
Pak for Data
An enterprise platform to operationalize data, analytics, and AI
Mara
BEAM SUMMIT
Machine Learning Engineering With MLflow
Google Open Source

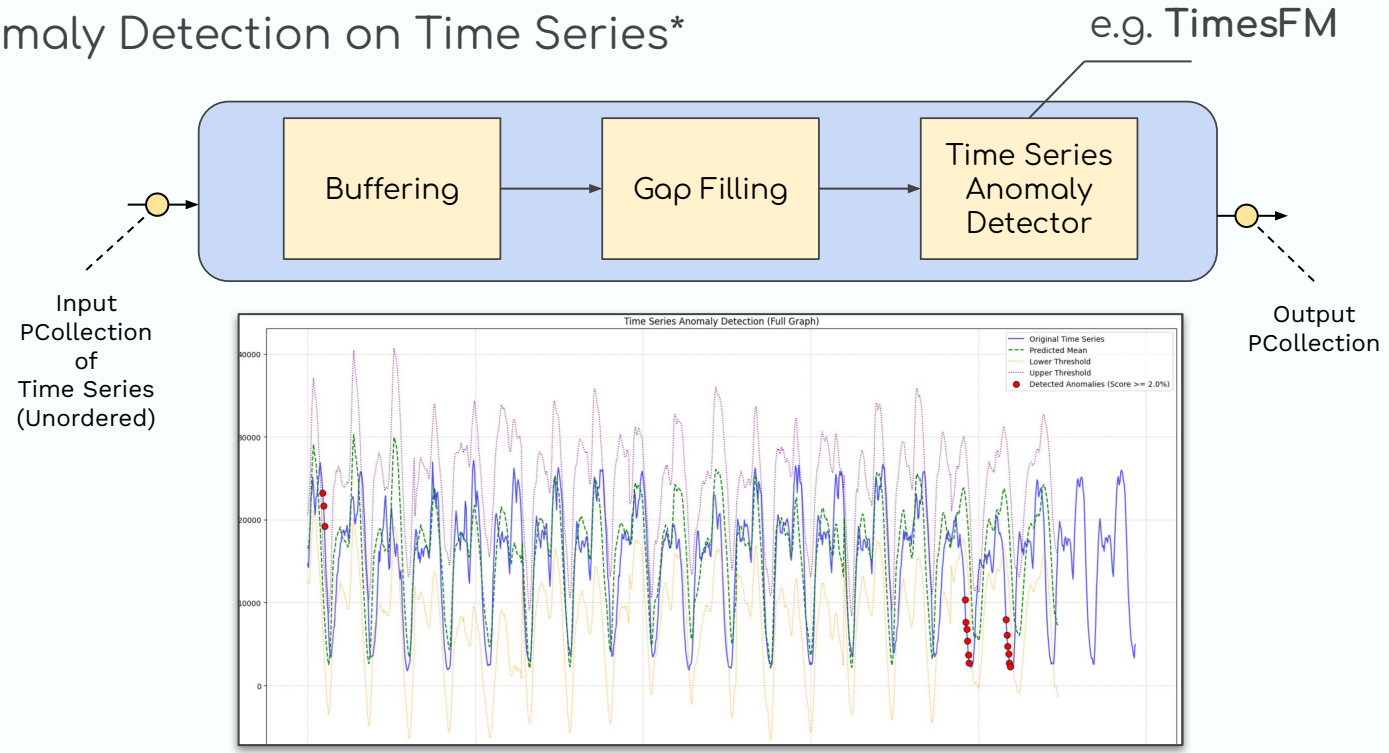
Upcoming Features

- Calling Anomaly Detection via YAML

```yaml
type: chain
  transforms:
    - type: AnomalyDetection
      config:
        detector:
          type: 'ZScore'
          config:
            sub_stat_tracker:
              type: 'IncSlidingMeanTracker'
              config:
                window_size: 500
            stdev_tracker:
              type: 'IncSlidingStdevTracker'
              config:
                window_size: 500
    - type: PyMap
      config:
        fn: "lambda x: (x[1].predictions[0].label)"
```

Upcoming Features (cont.)

- Anomaly Detection on Time Series*



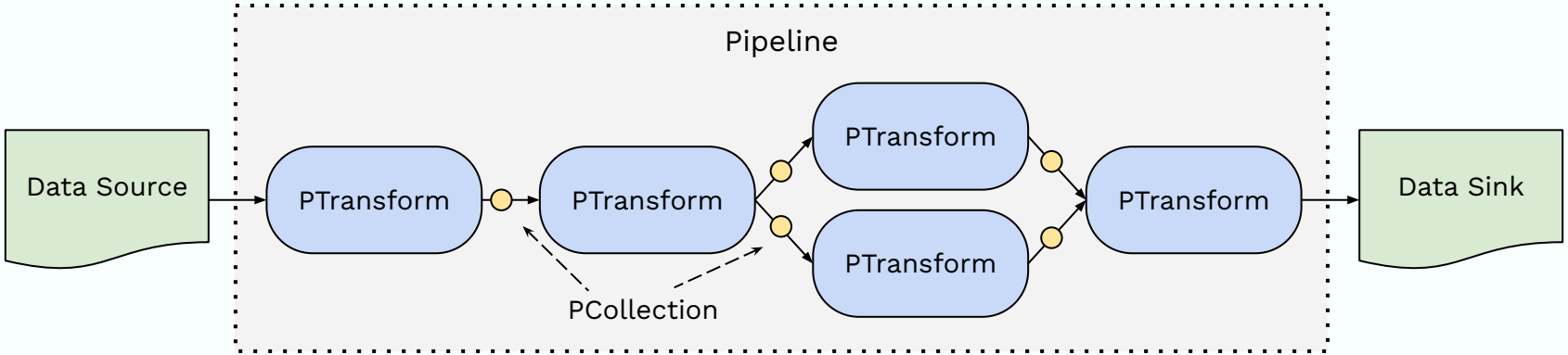
* In collaboration with Google 2025 summer intern, Ashok Devireddy.

Links

- Design Doc
  https://docs.google.com/document/d/1tE8lz9U_vjlNn2H7t-GRrs3vfhQ5UuCgWiHXCRHRPns/edit?usp=sharing
- Source Code
  https://github.com/apache/beam/tree/master/sdks/python/apache_beam/ml/anomaly
- Python Doc
  https://beam.apache.org/releases/pydoc/current/apache_beam.ml.anomaly.html
- Colabs
  https://github.com/apache/beam/tree/master/examples/notebooks/beam-ml/anomaly_detection

One-Pager: Beam Basics

- A framework to unify the **batch** (bounded) and **streaming** (unbounded) processing.
- Key Concepts
  - **PCollection** - a representation of data for parallel processing
  - **PTransform** - a representation of computation to transform data
  - **Pipeline** - a Directed Acyclic Graph (DAG) of PTransforms



QUESTIONS?

Shunping Huang
shunping@google.com

BEAM SUMMIT NYC 2025