thatDot
The Streaming Data Company

Streaming Anomaly Detection for Categorical Data

White Paper

Detecting interesting data among a vast sea of otherwise non-interesting data is an important problem with many applications. *thatDot, Inc.* is offering a unique new anomaly detection technology to calculate "novelty scores" in streaming data using our patent-pending technology, developed with funding from DARPA.

With the capabilities provided by *thatDot*, real-time detection of anomalies in categorical data has become possible for the first time. This system uses cutting edge techniques to provide results in a real-time streaming fashion — not batch processing — that automatically uses context to find anomalies and reduce false positives.

Novelty, Anomaly, and Fraud Detection

thatDot has built a system to allow rapid analysis of streaming data, scoring each item to explain how novel or anomalous that piece of data is, given all the previous observations.

This system works on categorical data — which is any form of non-numerical data (names, identifiers, strings, etc.). Typical anomaly detection is a very mathematical process, requiring exclusively numeric data, and delivering results only in a batch. However with *thatDot Anomaly Detection*, each data observation is fed into the system, a score is computed immediately to indicate how unusual that particular observation is given the full set of data observed up to that point.

A major advantage of *thatDot Anomaly Detection* is that it is able to dramatically reduce false positives by using contextual information, learned automatically from the data.

thatDot Anomaly Detection uses the structure created from all observations to understand whether the observation is anomalous, given the surrounding context.

If a user always expects to see similar data from their dataset, when a new observation contains data not originally in the dataset, that is quite anomalous. But on the other hand, if data coming in often includes previously unseen elements, then that the observation is not anomalous. Even though the new data has never been seen before, new data is typical and should not trigger anomaly detection. Using context for categorical data dramatically reduces false positives!

To produce real-time novelty scores, *thatDot* builds and maintains a dynamic graphical model structured by the input data. This network represents a complex model of conditional probabilities for each component of the data observation.

Beyond simply measuring whether an observation is "different" data or not, the system uses context and prior observations to calculate the likelihood of seeing this piece of data, given previous observations and what the system can infer about the structure of the data.

This system requires **no labeling of data and no tuning of hyper-parameters**. Simply start the system, and it works immediately. Once a representative sample of data has been observed, high scores returned for each observation indicate whether the input data represents an unusual observation. If the score is above a desired threshold, the data is treated as an anomaly. The system will also describe where that observation diverged from the previously seen data, and the overall graphical model is available to explain the context of that observation.

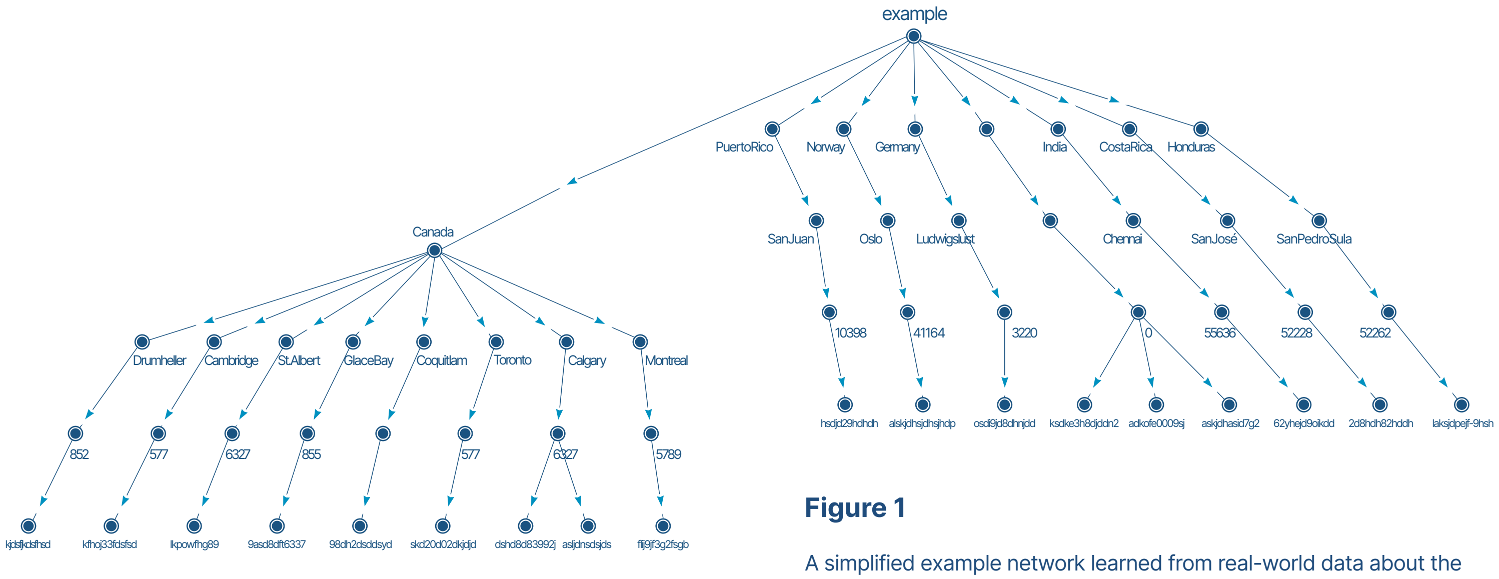


Figure 1

A simplified example network learned from real-world data about the distribution of content among (top-to-bottom):
[Country, City, Provider ID, Content ID].

After making these observations, this network is very informative about differing questions like: Is seeing content distributed to a new city in Canada unusual? (no) —or a new city in Germany? (yes)

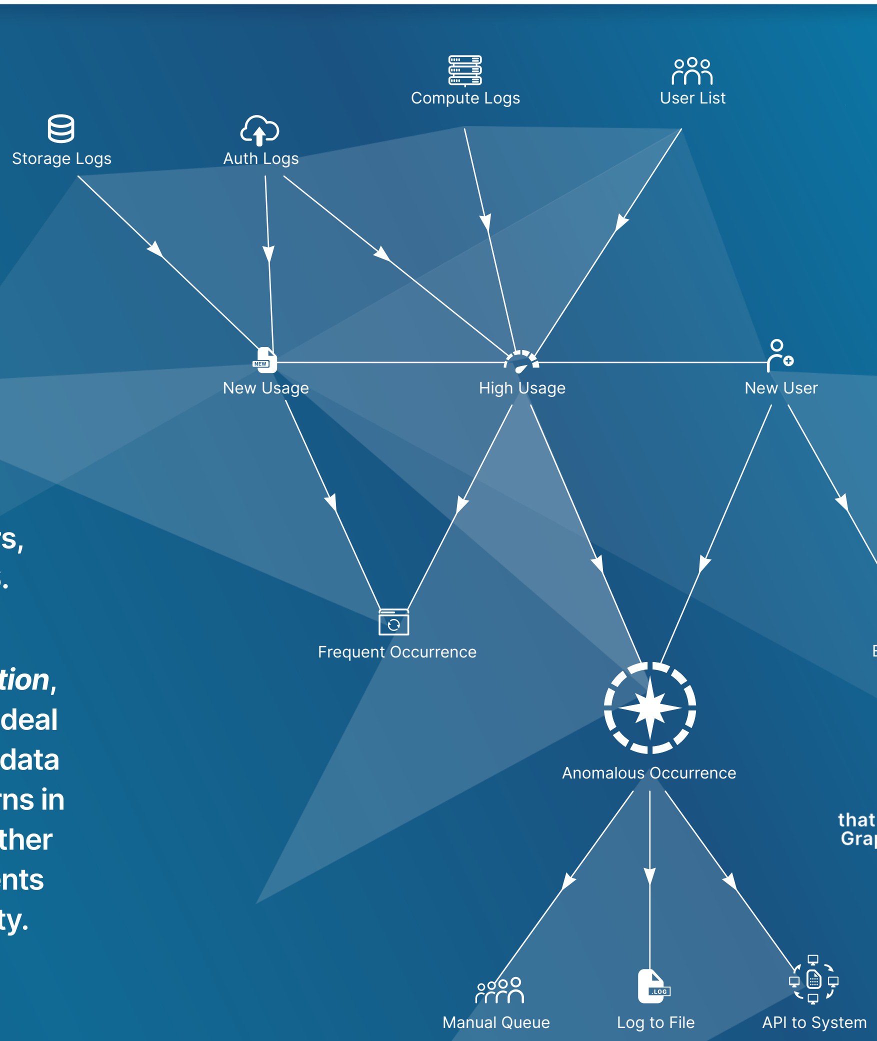
Real-Time Graph Processing to Feed Anomaly Detection

In many real-world applications, data does not always arrive in the form desired for anomaly detection. Many of the most valuable observations are the result of combining and interpreting data from many different sources. In the past, preparing these data sources for consumption was a laborious task requiring weeks, months, or more from data platform engineers building streaming ETL pipelines. When changes to the data or transformation mechanisms are needed, these changes can take a very long time to realize.

thatDot has built a streaming data platform called *thatDot Connect* which streamlines the process of triggering real-time actions (like Anomaly Detection) from complex patterns arising from many different data sources. *thatDot Connect* is the first system to combine a graph data-model and a graph-execution-model, resulting in data transformations that are powerful, flexible, and performed entirely in real-time. Complex data patterns spanning any time period are immediately detected and used to trigger other actions. These actions can trigger data updates to clean or transform data, pass data in to the *Anomaly Detection* sub-system, or they can call out to external systems to trigger other actions.

thatDot Connect has been developed over the last six years, with major funding from the U.S. Government.

Combined with _Anomaly Detection_, _thatDot Connect_ becomes the ideal platform for unifying streaming data sources, finding complex patterns in real-time, and determining whether the corresponding data represents fraud or other anomalous activity.



thatDot Connect Graph Structure

Deployment and Usage

thatDot Connect and Anomaly Detection are deployed to either a single server or to an entire cluster—in the cloud or on premise, and then available for user interaction via web APIs. The core technology runs on the JVM and can be deployed as a *.jar* in virtually any environment.

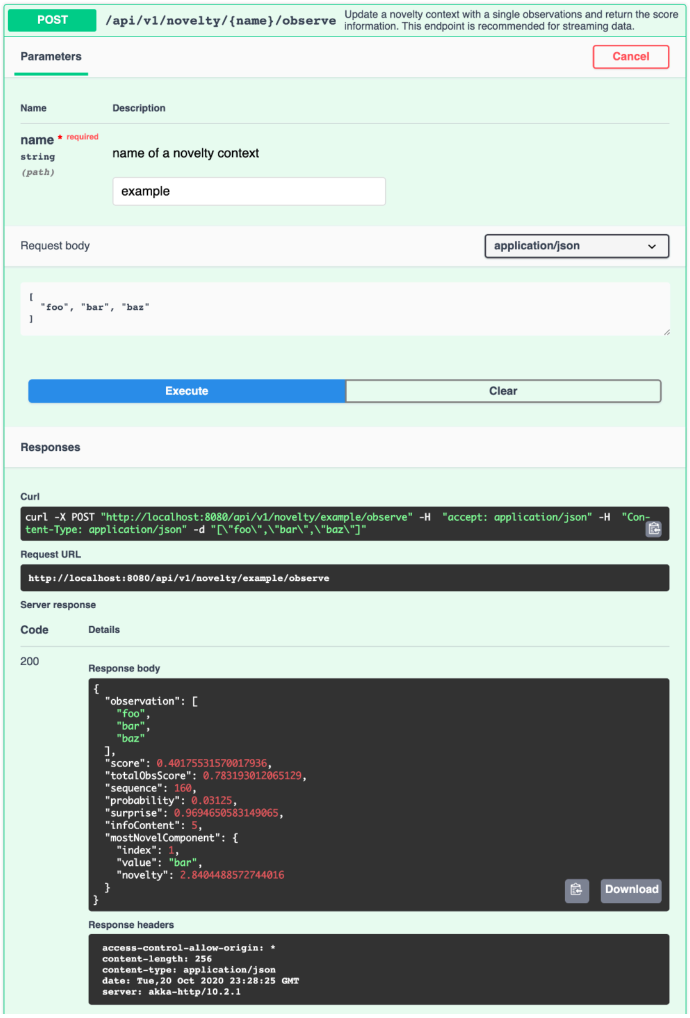
To use *Anomaly Detection* the only interaction needed is to call a single API endpoint, passing in an array of strings as the observation. A JSON object is returned indicating the score for that particular observation and any additional relevant data—like the point at which that observation deviated from prior data seen.

thatDot

Review The Docs

amazon web services

Download The App



`/api/v1/novelty/{name}/observe` Update a novelty context with a single observations and return the score information. This endpoint is recommended for streaming data.

Name	Description
name * required string *(path)*	name of a novelty context example

Request body: application/json

```
[
  "foo", "bar", "baz"
]
```

Curl

```
curl -X POST "http://localhost:8080/api/v1/novelty/example/observe" -H  "accept: application/json" -H  "Con-
tent-Type: application/json" -d "[\"foo\",\"bar\",\"baz\"]"
```

Request URL

```
http://localhost:8080/api/v1/novelty/example/observe
```

Server response

Code	Details
200	Response body

```
{
  "observation": [
    "foo",
    "bar",
    "baz"
  ],
  "score": 0.401755315700179 36,
  "totalObsScore": 0.783193012065129,
  "sequence": 160,
  "probability": 0.03125,
  "surprise": 0.9694650583149065,
  "infoContent": 5,
  "mostNovelComponent": {
    "index": 1,
    "value": "bar",
    "novelty": 2.8404488572744016
  }
}
```

Response headers

```
access-control-allow-origin: *
content-length: 256
content-type: application/json
date: Tue,20 Oct 2020 23:28:25 GMT
server: akka-http/10.2.1
```

Figure 2

An example response returned in our interactive API documentation from a single call to the *Anomaly Detection* system.

The richer capabilities of ***thatDot Connect*** are also available in a similar fashion. REST APIs control the definition of stream ingest sources, standing queries, and other capabilities of that system.

The system can be scaled horizontally to support data volume of any size. Cluster deployment can be performed via Kubernetes, various cloud deployment systems (e.g. AWS CloudFormation) or from manual configuration in any environment.