Continuous Machine Learning over Streaming Data

Roger Barga, Nina Mishra, Sudipto Guha, Ryan Nienhuis
Amazon Web Service

Kinesis Streaming Services

Robust Random Cut Forrest

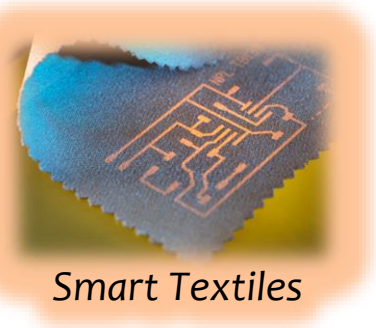
Summary of a dynamic data stream, highly efficient, wide number of use cases…

#Real-time


All data originates in real-time!

127.0.0.1 user-identifier frank [10/Oct/2000:13:55:36 -0700] "GET /apache_pb.gif HTTP/1.0" 200 2326

Common Log Entry



Smart Textiles

Beacons

<R,AMZN,T,G,R1>

NASDAQ OMX Record

"SeattlePublicWater/Kinesis/123/Realtime" – 412309129140

MQTT Record

Health Monitors

Smart Buildings


```
{
  "payerId": "Joe",
  "productCode": "AmazonS3",
  "clientProductCode": "AmazonS3",
  "usageType": "Bandwidth",
  "operation": "PUT",
  "value": "22490",
  "timestamp": "1216674828"
}
```


Metering Record

<165>1 2003-10-11T22:14:15.003Z mymachine.example.com evntslog - ID47 [exampleSDID@32473 iut="3" eventSource="Application" eventID="1011"][examplePriority@32473 class="high"]

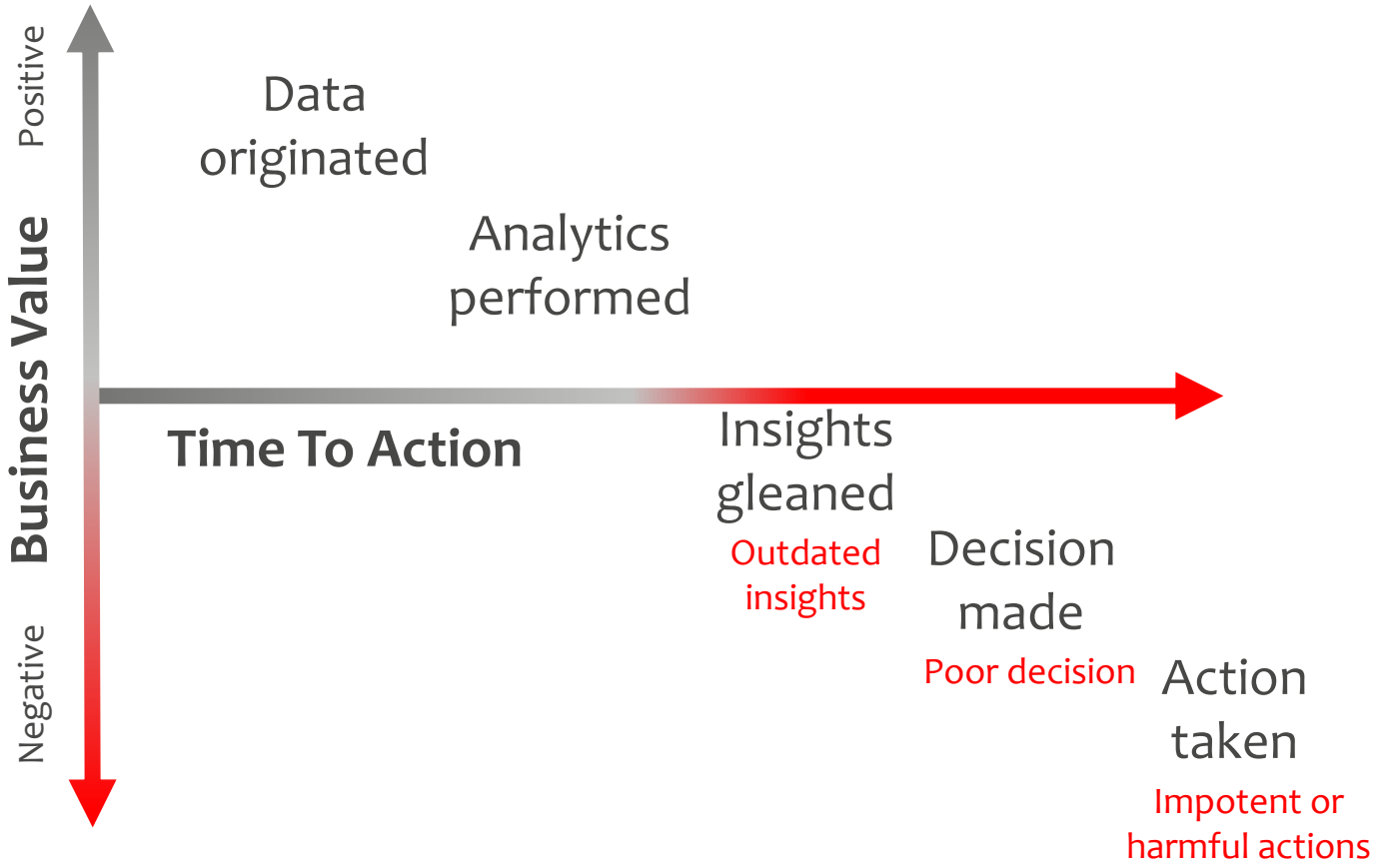
Syslog Entry

But, analytics to gain insights is usually done much, much later.

#WhyWait

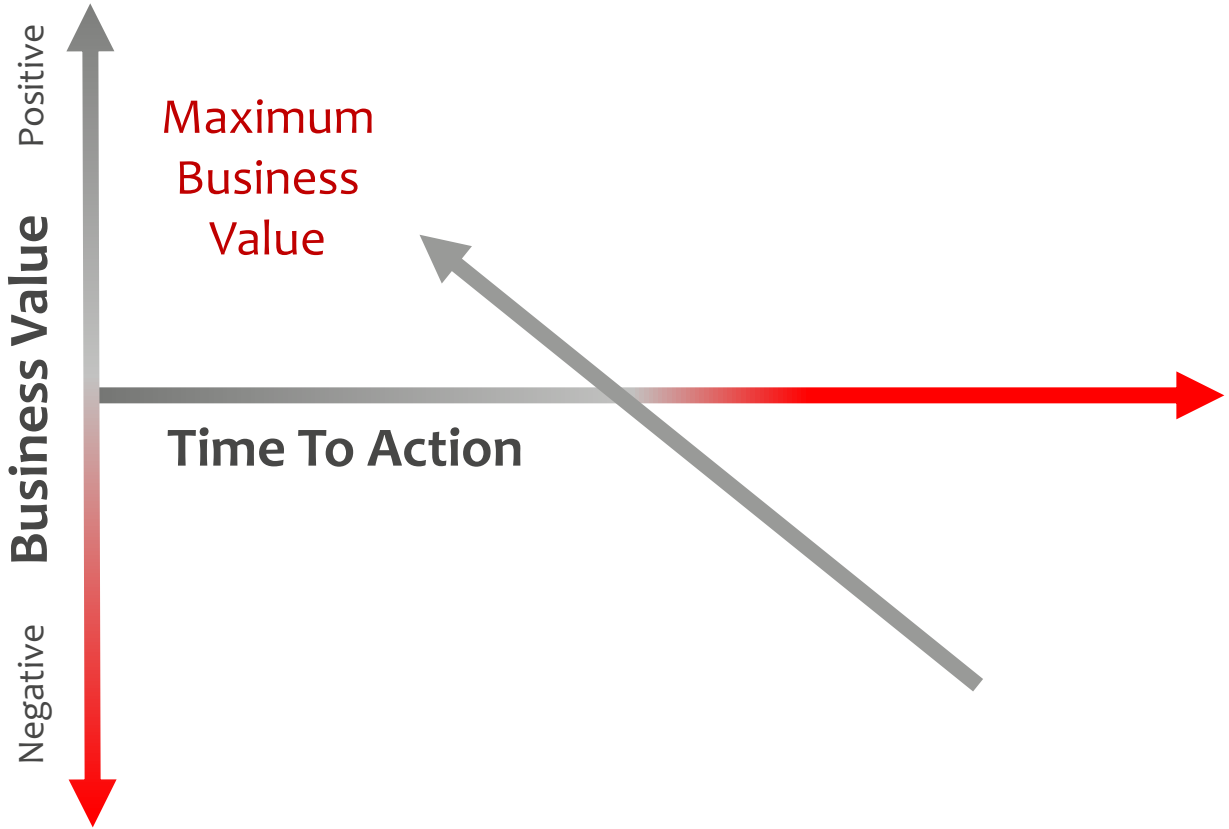

Insights are perishable.

Batch analytics operations take too long



Compress the analytics lifecycle

Maximize the value of data

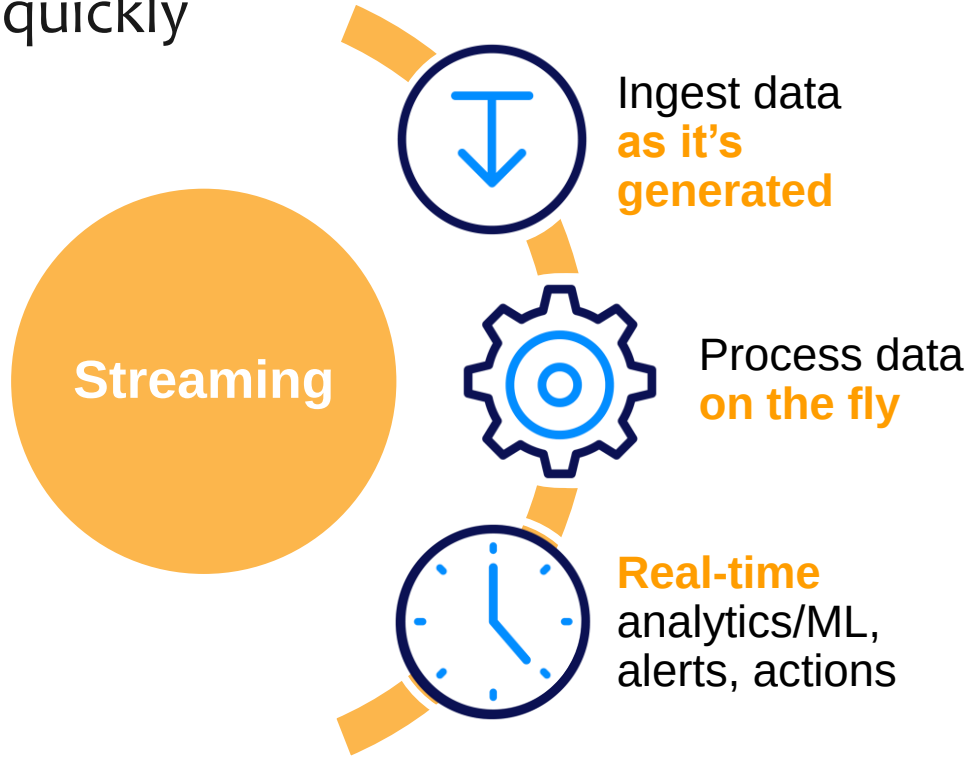



Streaming technology is necessary to detect and act
on real-time perishable insights.

#Streaming

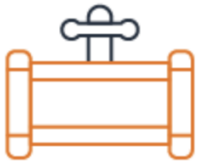
Kinesis Data Streaming Services

Get actionable insights quickly



Streaming with Amazon Kinesis

Easily collect, process, and analyze data and video streams in real time



Kinesis Data Streams

Capture, process, and store data streams

Kinesis Data Firehose

Load data streams into AWS data stores

Kinesis Data Analytics

Analyze data streams with SQL

Kinesis Video Streams

Capture, process, and store video streams

Customer Examples

NETFLIX

Analyze billions of network flows in real-time

Migrated data bus from Kafka to Kinesis

SONOS

1 billion events per week from connected devices

Near-real-time home valuation (Zestimates)

Live clickstream dashboards refreshed under 10s

IoT predictive analytics

100 GB/day clickstreams from 250+ sites

50 billion daily ad impressions, sub-50 ms responses

NORDSTROM

Online stylist processing 10 million events/day

Facilitate communications between 100+ microservices

Amazon Kinesis

Foundational Service Used Across Amazon

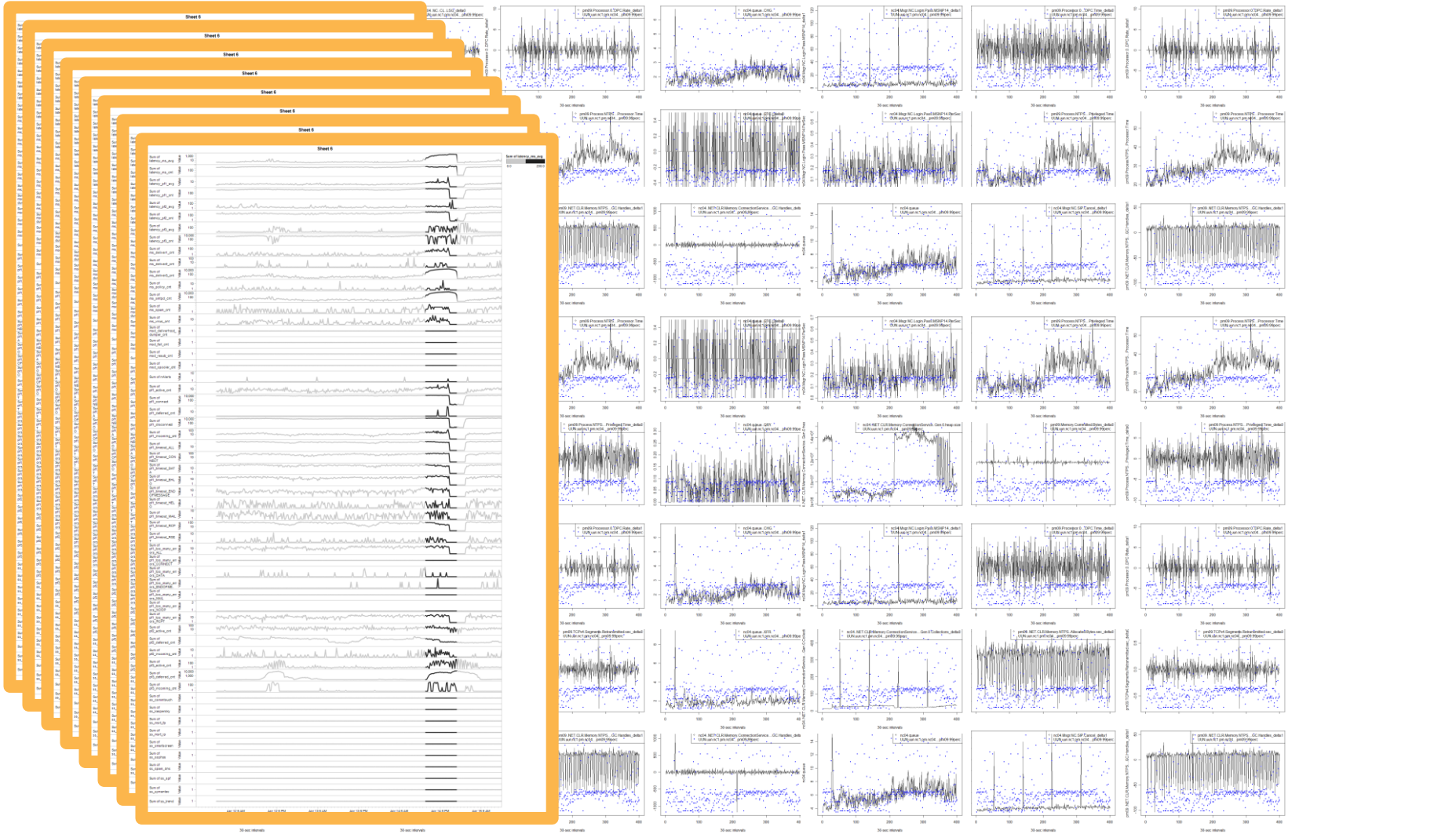
AWS metering

Amazon S3 events

Amazon CloudWatch logs

Amazon.com online catalog

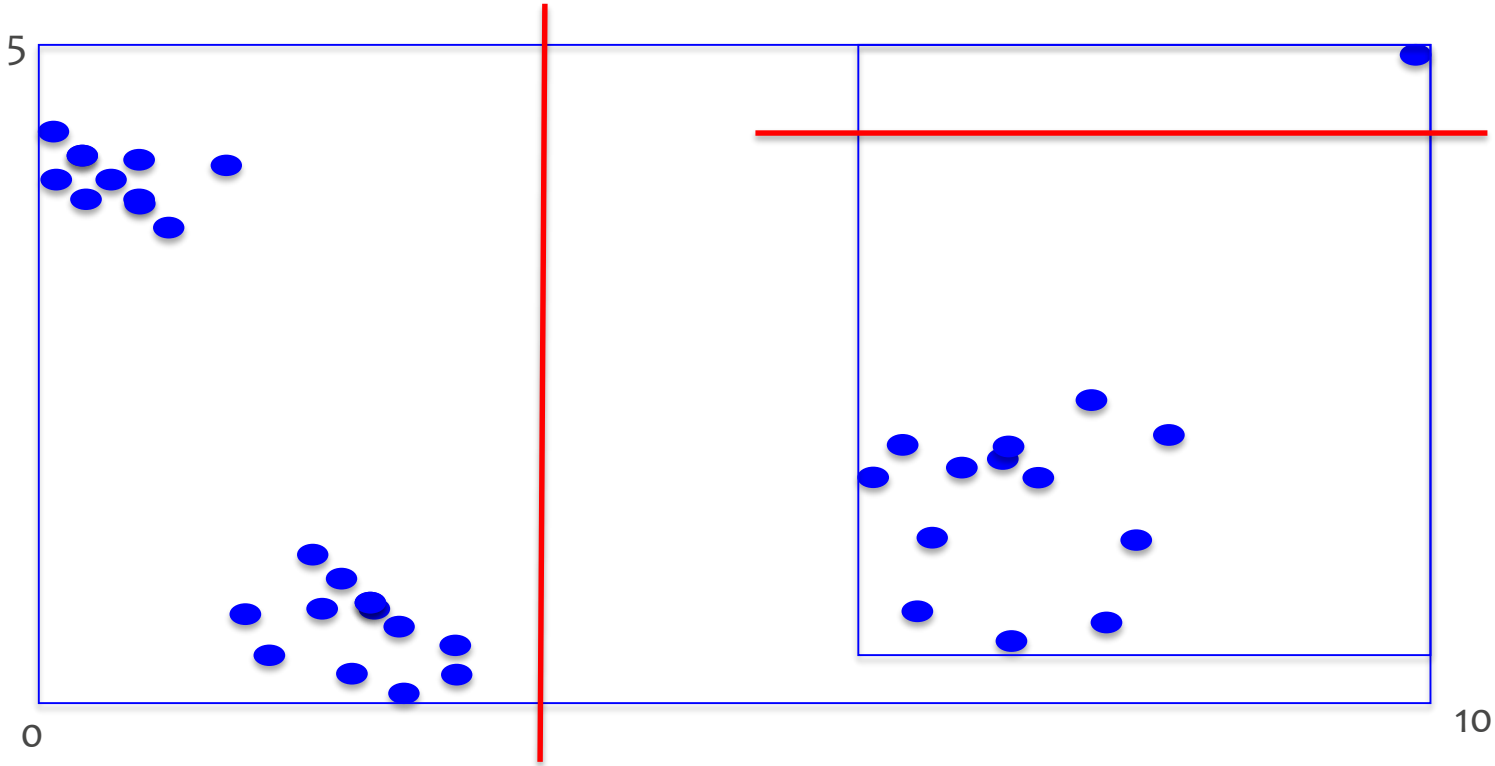
Amazon Go video analytics


Sheet 6

Discover actionable insights in real-time

Anomaly Detection

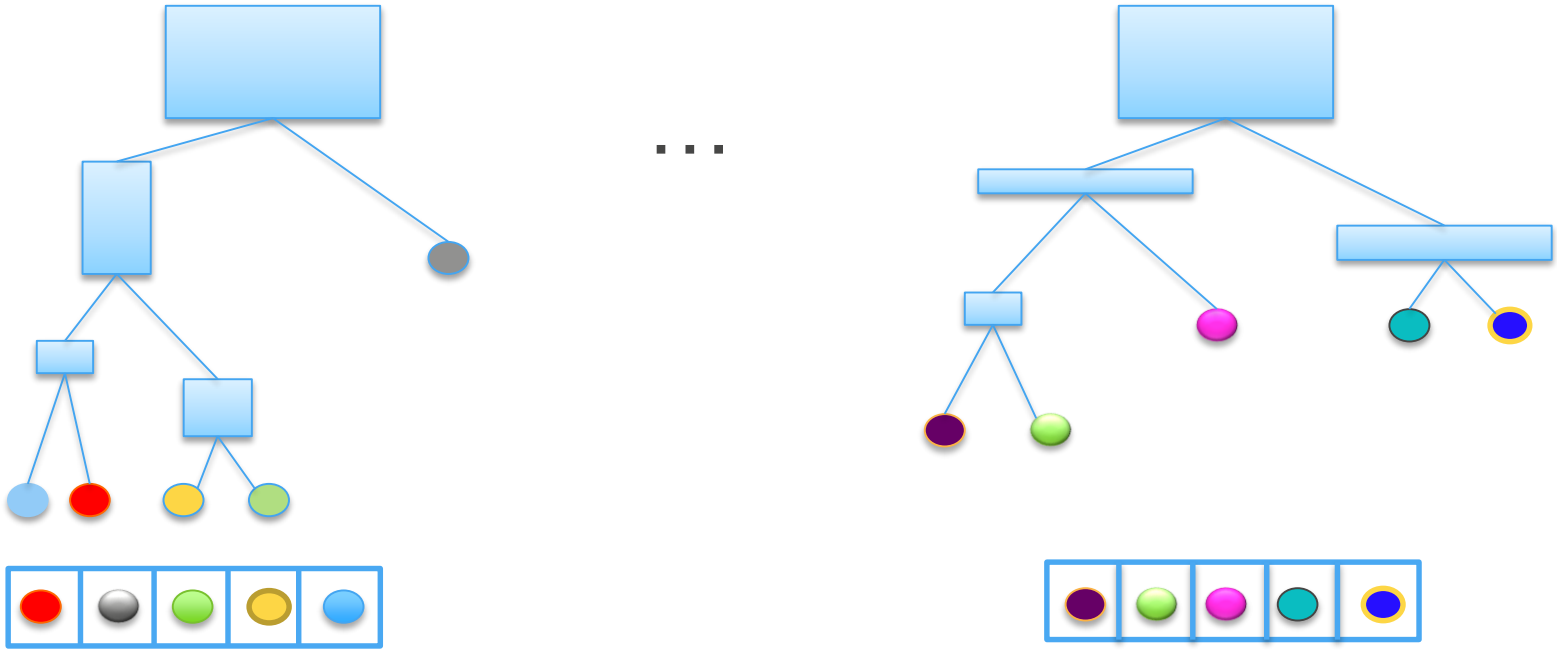
Random Cut Tree



Range-biased Cut

Recurse: The cutting stops when each point is isolated.

Random Cut Forest

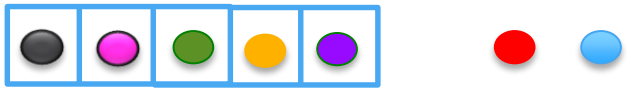


Each tree built on a random sample.

Random Sample of a Stream

Reservoir Sampling [Vitter]

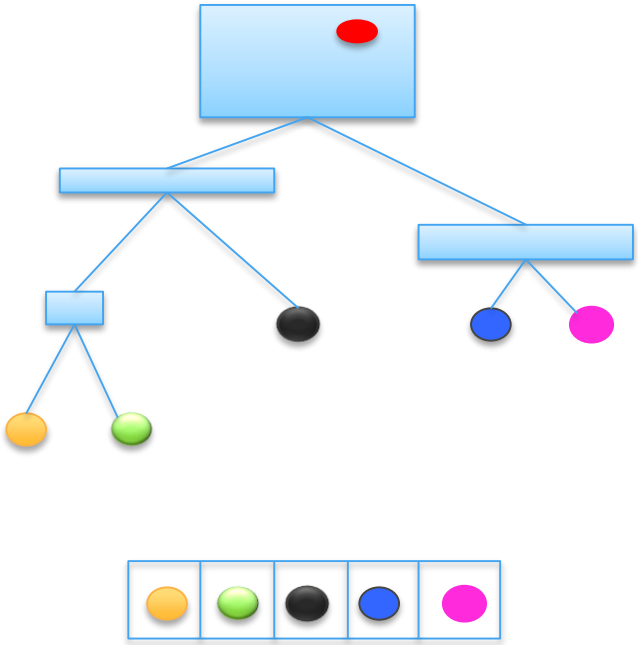
Maintain random sample of 5 points in a stream?



Keep  heads with probability $\frac{5}{6}$

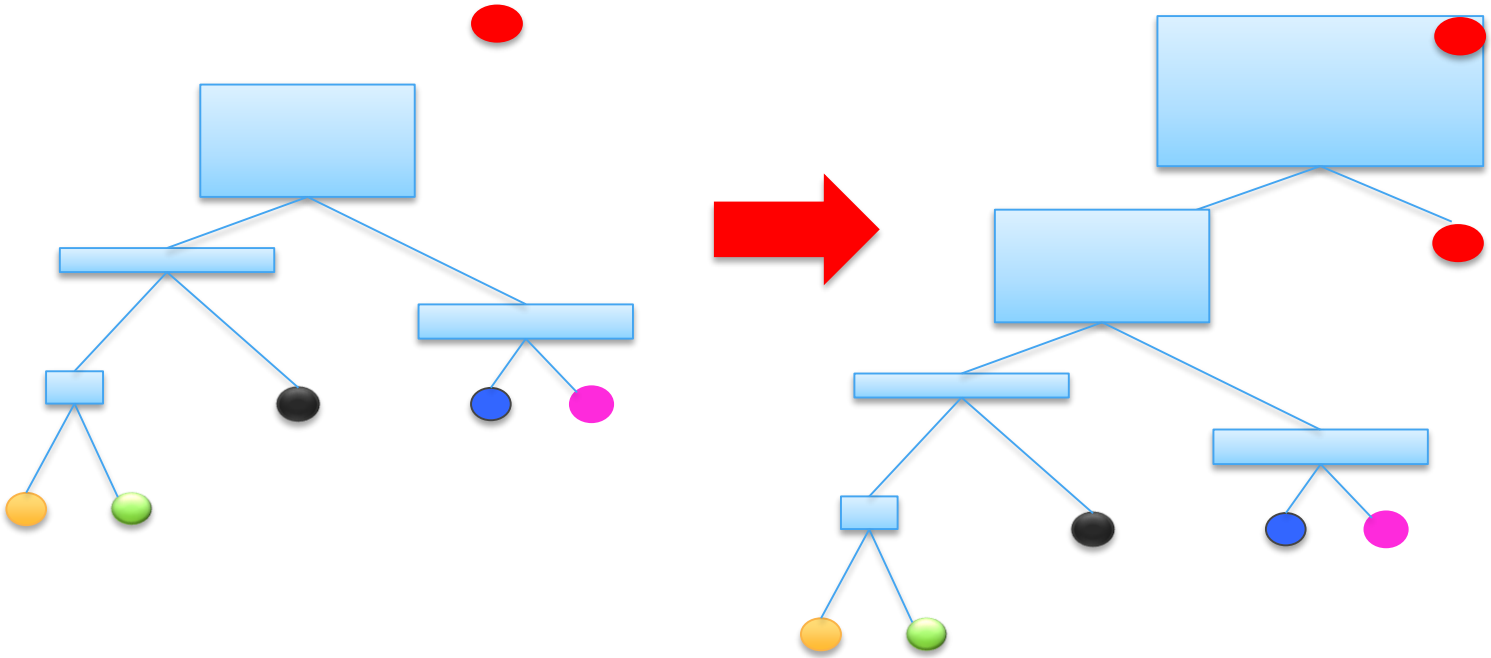
Discard  tails with probability $\frac{1}{6}$

Insert – Case I



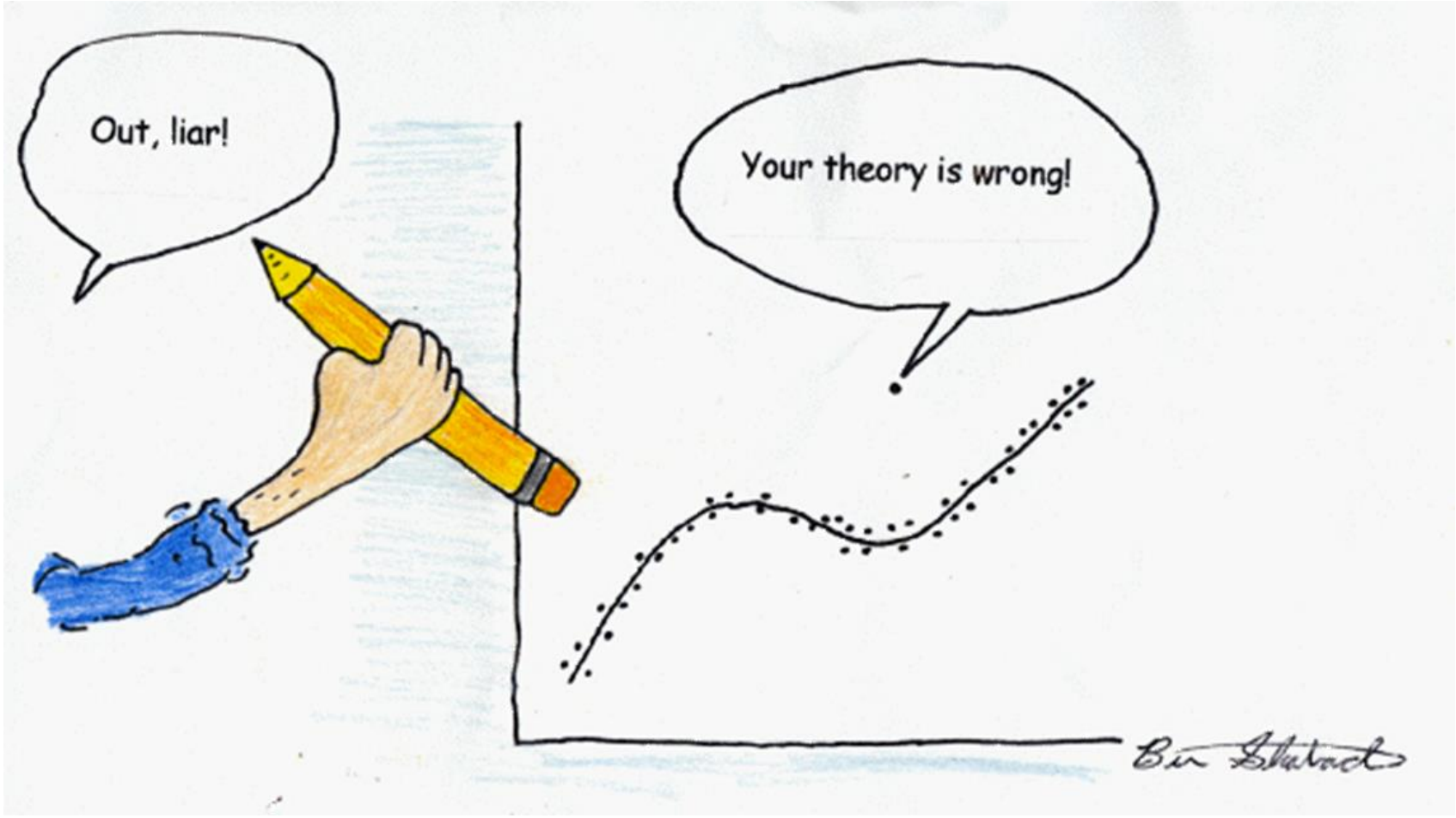
Start with the Root
If the point falls inside the bounding box
  follow the path to the appropriate child

Insert – Case II



Theorem: Insert generates a tree T' ~ T()

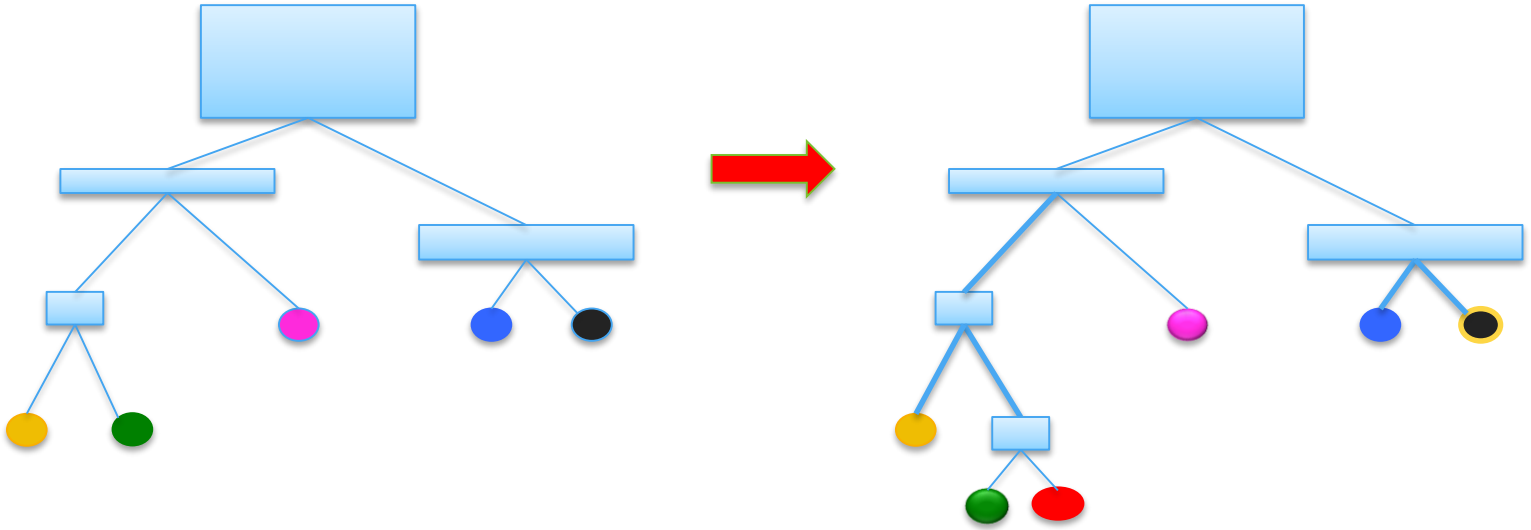
What is an Outlier?



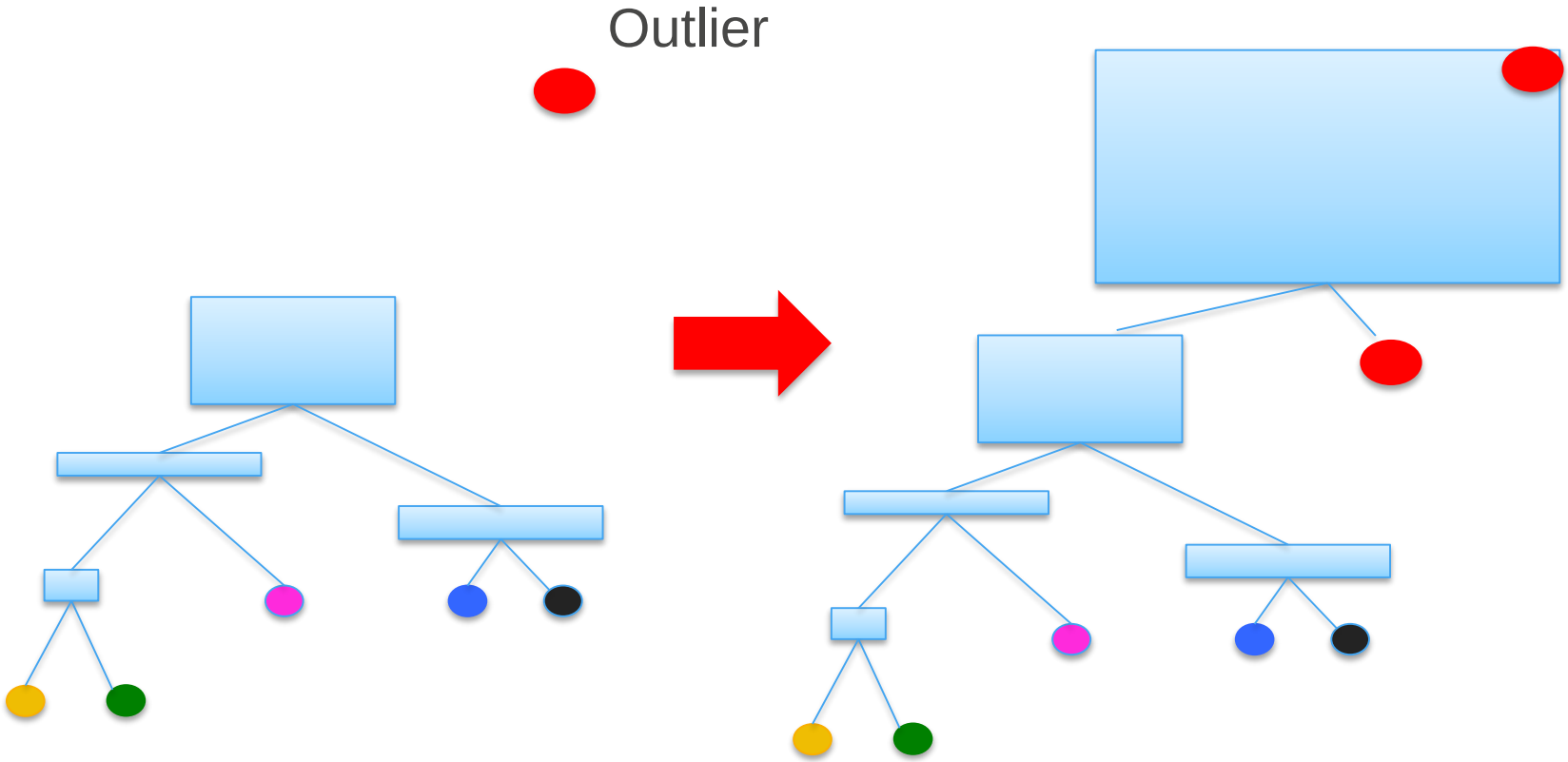
Anomaly Score: Displacement

A point is an *anomaly* if its insertion greatly increases the tree size (= sum of path lengths from root to leaves = description length).

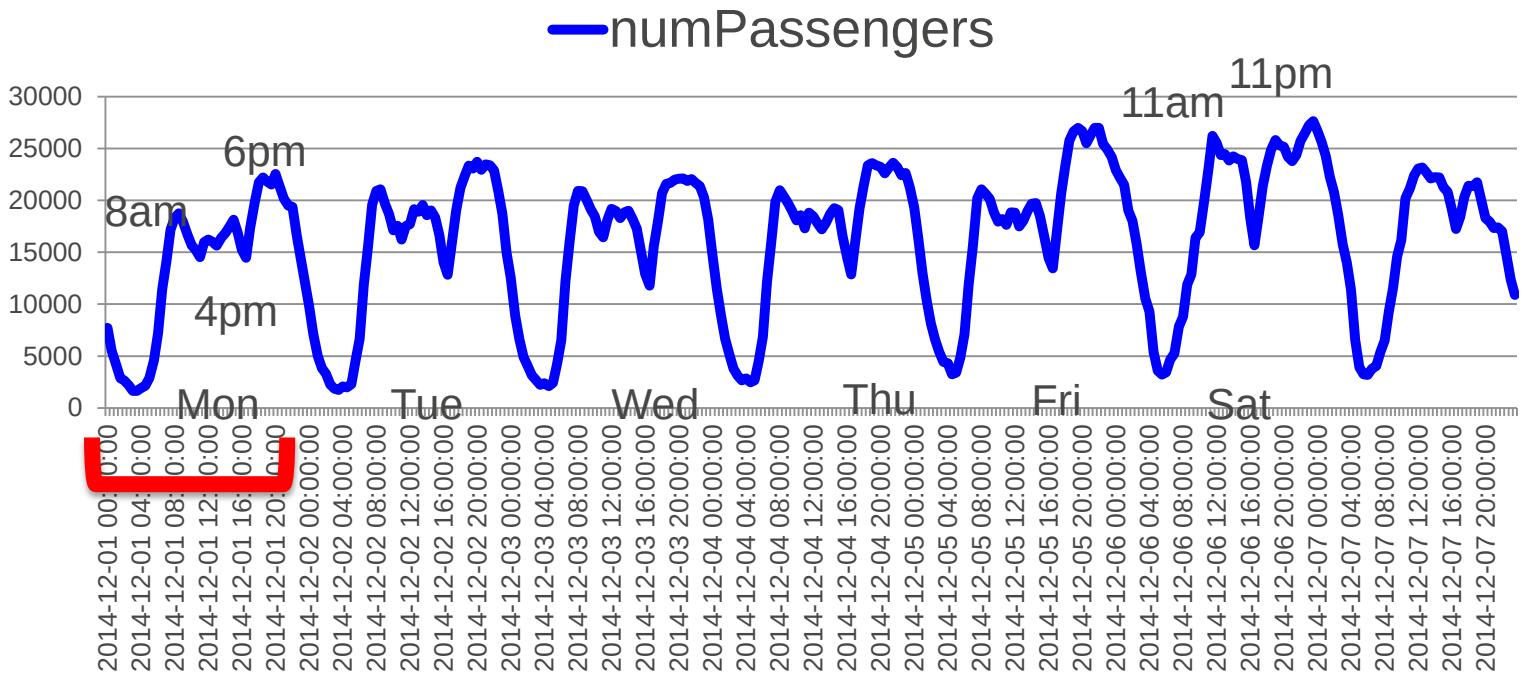
Inlier:



Anomaly Score: Displacement



NYC Taxi Ridership



Date aggregated every 30 minutes,
Shingle Size: 48

[1]Public Data: http://www.nyc.gov/html/tlc/html/about/trip_record_data.shtml

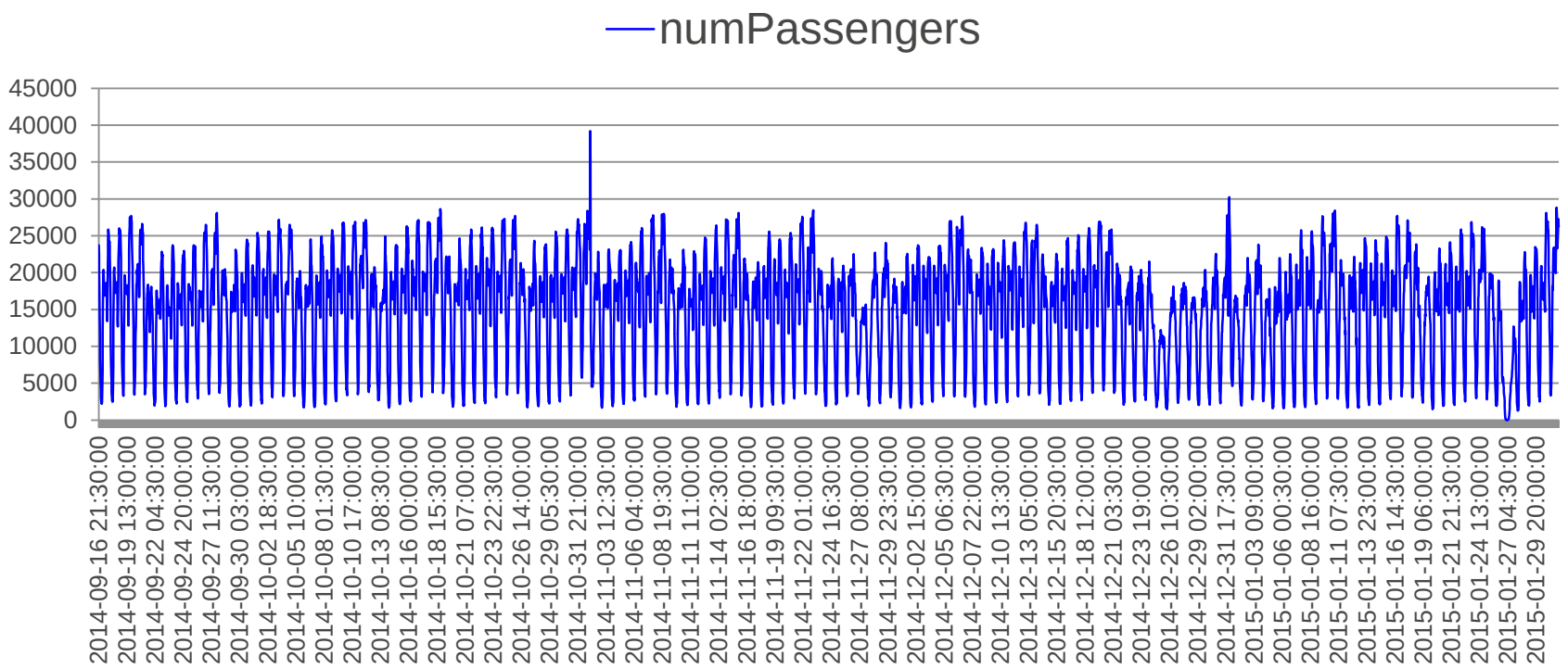
```
-- creates a temporary stream.
CREATE OR REPLACE STREAM "TEMP_STREAM" (
               "passengers"        INTEGER,
          "distance"          DOUBLE,
          "ANOMALY_SCORE"     DOUBLE);

-- creates another stream for application output.
CREATE OR REPLACE STREAM "DESTINATION_SQL_STREAM" (
          "passengers"        INTEGER,
          "distance"          DOUBLE,
          "ANOMALY_SCORE"     DOUBLE);
```

```
-- Compute an anomaly score for each record in the input stream
-- using Random Cut Forest
CREATE OR REPLACE PUMP "STREAM_PUMP" AS
      INSERT INTO "TEMP STREAM"
     SELECT STREAM "passengers", "distance", ANOMALY_SCORE
     FROM TABLE (RANDOM_CUT_FOREST (
          CURSOR(SELECT STREAM * FROM "SOURCE_SQL_STREAM")))
```

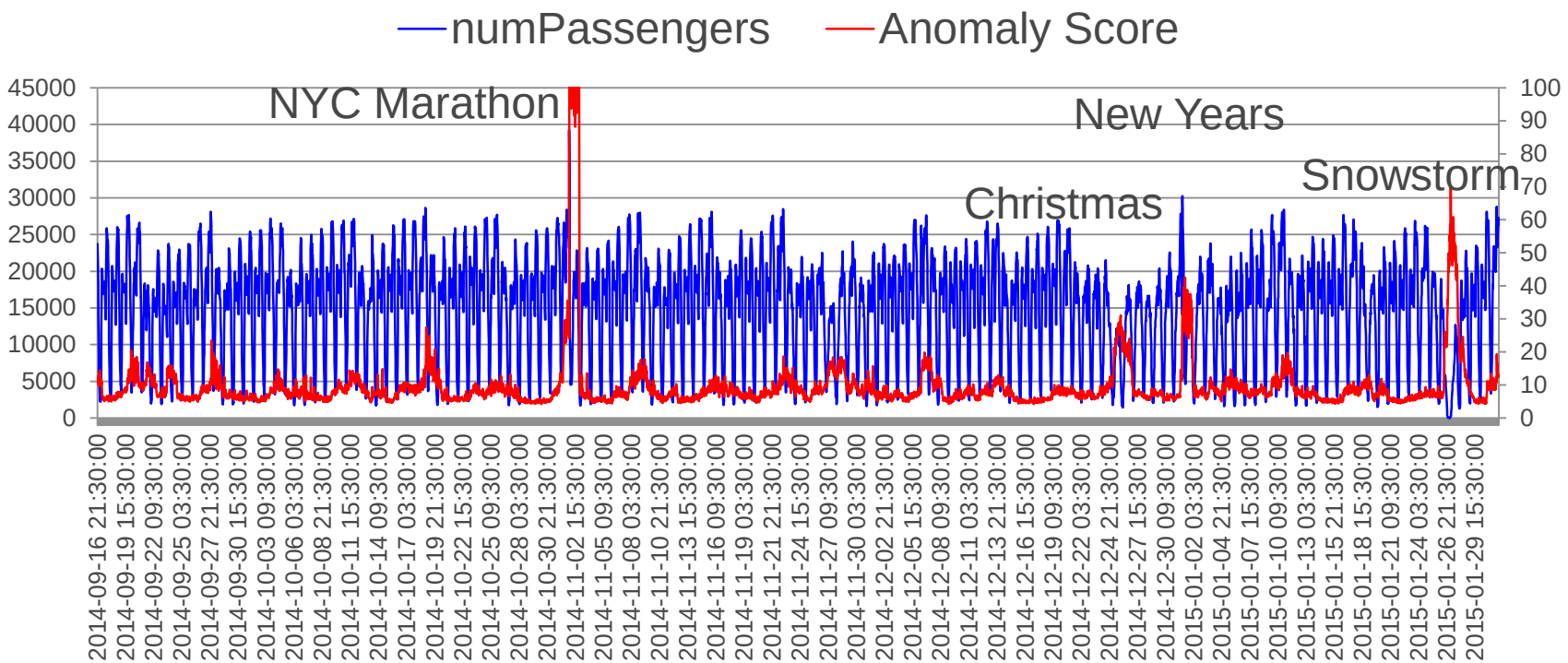
```
-- Sort records by descending anomaly score, insert into output stream
CREATE OR REPLACE PUMP "OUTPUT_PUMP" AS
  INSERT INTO "DESTINATION_SQL_STREAM"
   SELECT STREAM * FROM "TEMP_STREAM"
   ORDER BY FLOOR("TEMP_STREAM".ROWTIME TO SECOND), ANOMALY_SCORE
DESC;
```

SQL to call Random Cut Forest

NYC Taxi Data



NYC Taxi Data



Robust Random Cut Forest Based Anomaly Detection On Streams

Sudipto Guha SUDIPTO@CIS.UPENN.EDU
University of Pennsylvania, Philadelphia, PA 19104.

Nina Mishra NMISHRA@AMAZON.COM
Amazon, Palo Alto, CA 94303.

Gourav Roy GOURAVR@AMAZON.COM
Amazon, Bangalore, India 560055.

Okke Schrijvers OKKES@CS.STANFORD.EDU
Stanford University, Palo Alto, CA 94305.

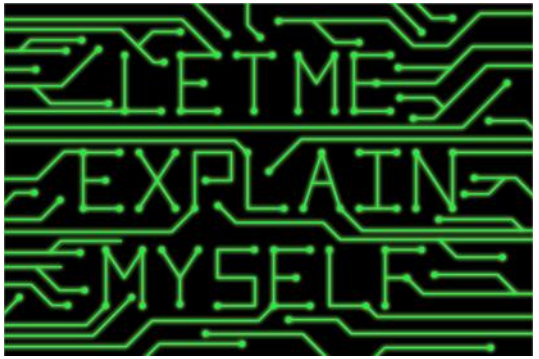
Abstract

In this paper we focus on the anomaly detection problem for dynamic data streams through the lens of random cut forests. We investigate a robust random cut data structure that can be used as a sketch or synopsis of the input stream. We provide a plausible definition of non-parametric anomalies based on the influence of an unseen point on the remainder of the data, i.e., the externality imposed by that point. We show how the sketch can be efficiently updated in a dynamic data stream. We demonstrate the viability of the algorithm on publicly available real data.

a point is data dependent and corresponds to the externality imposed by the point in explaining the remainder of the data. We extend this notion of externality to handle "outlier masking" that often arises from duplicates and near duplicate records. Note that the notion of model complexity has to be amenable to efficient computation in dynamic data streams. This relates question (1) to question (2) which we discuss in greater detail next. However it is worth noting that anomaly detection is not well understood even in the simpler context of static batch processing and (2) remains relevant in the batch setting as well.

For question (2), we explore a randomized approach, akin to (Liu et al., 2012), due in part to the practical success reported in (Emmott et al., 2013). Randomization is a powerful tool and known to be valuable in supervised learn-

Proceedings 33rd International Conference on Machine Learning, New York, NY, USA, 2016.



Attribution and Directionality

Explainable/Transparent/Interpretable ML

"If my time-series data with 30 features yields an unusually high anomaly score. How do I explain why this particular point in the time- series is unusual? [..] Ideally I'm looking for some way to visualize "feature importance" for a specific data point."

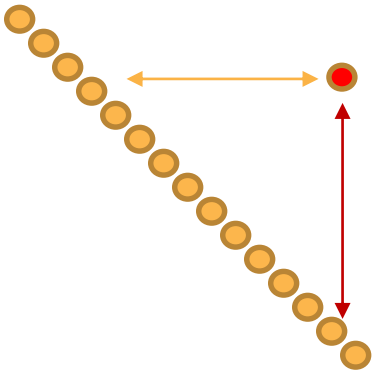
--- Robin Meehan, Inasight.com

What is Attribution?

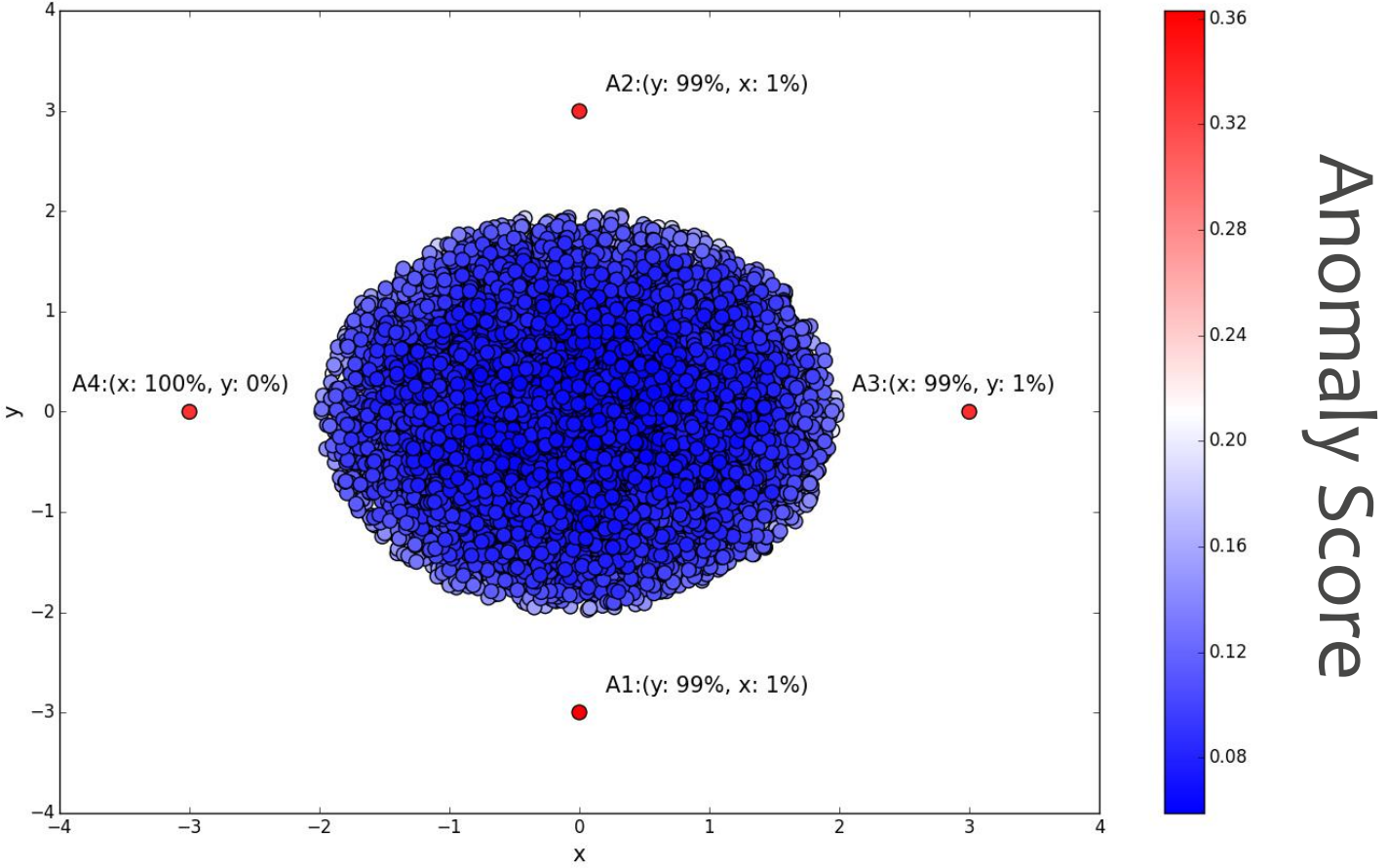
It's the ratio of the "distance" of the anomaly from normal.

(It's a distance in space of repeated patterns in the data.)

$$\Delta^i(p) = \frac{(\text{Score}^{+i}(p) - \text{Score}^{-i}(p))}{\text{Score}^{+i}(p)}$$



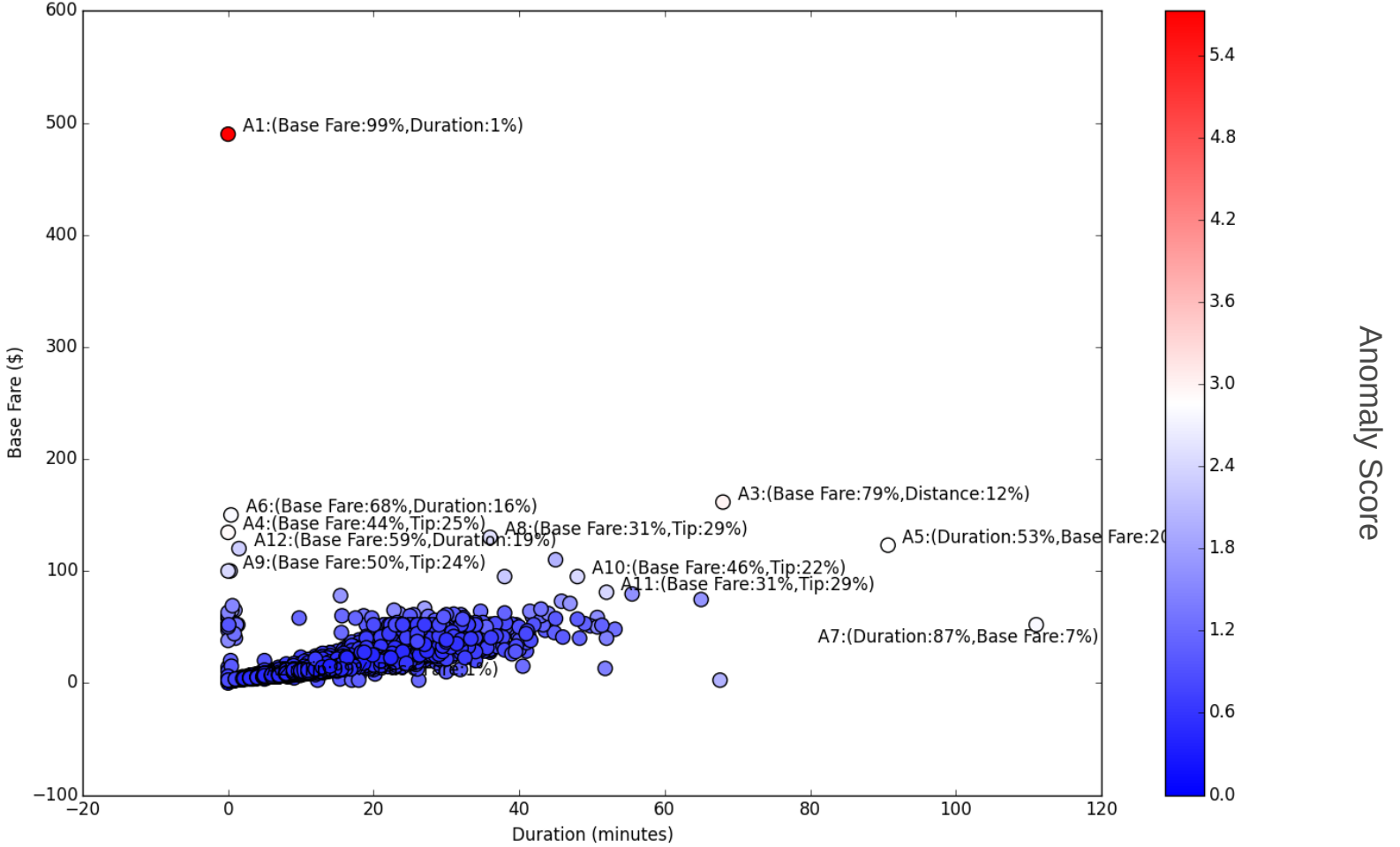
What is Attribution?

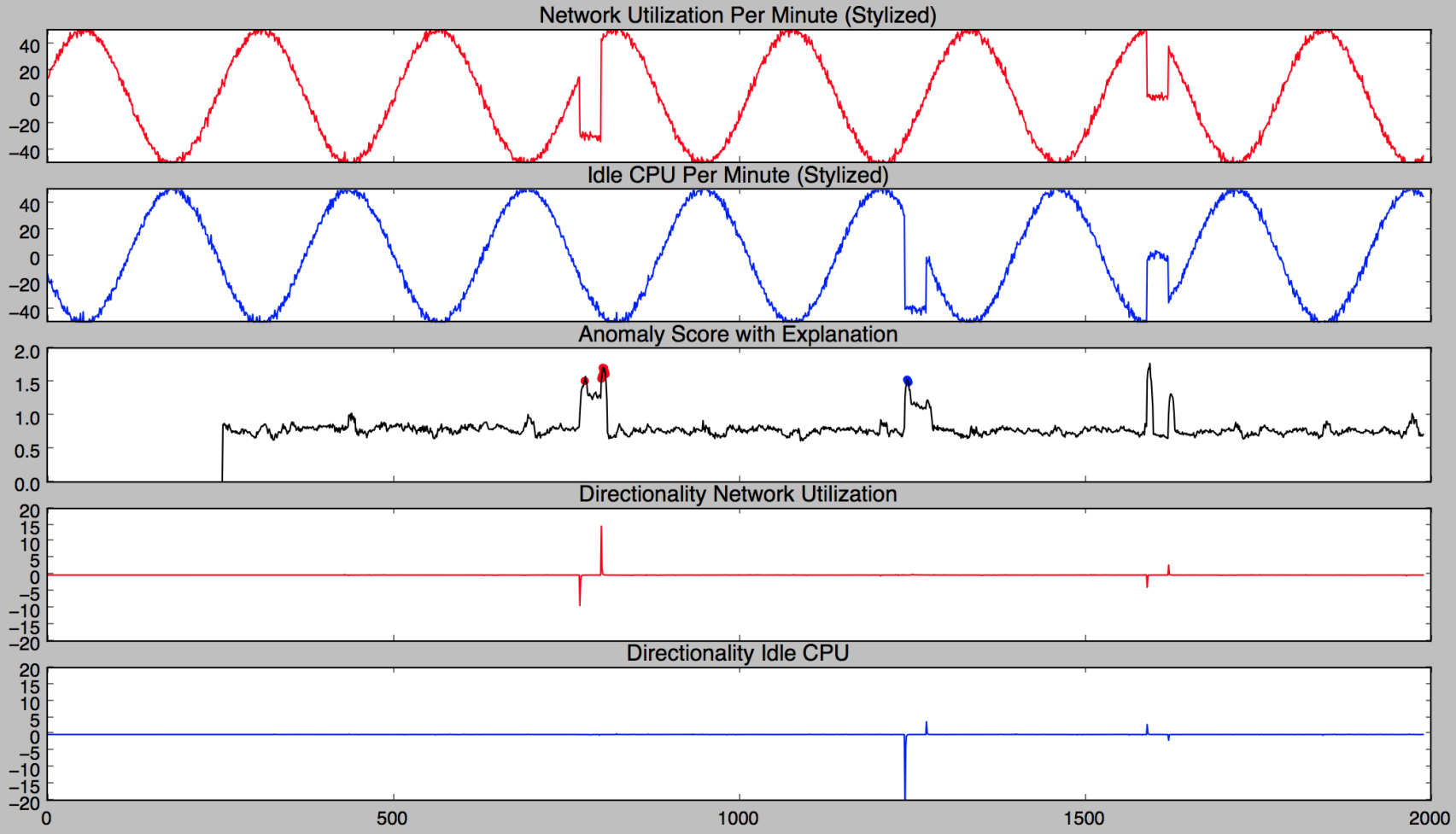


NYC Taxi Ridership Data[1]

Pickup Time	Dropoff Time	Distance	Base Fare	Surcharge	Tax	Tip	Tolls	Total
7/1/14 1:43	7/1/14 1:51	2.59	9.5	0.5	0.5	0	0	10.5
7/1/14 1:33	7/1/14 1:47	2.38	12	0.5	0.5	0	0	13
7/1/14 1:37	7/1/14 1:50	2.87	11.5	0.5	0.5	0	0	12.5
7/1/14 1:35	7/1/14 1:50	4.68	16	0.5	0.5	4.95	0	21.95
7/1/14 1:25	7/1/14 1:49	6.72	23	0.5	0.5	0	0	24
7/1/14 1:30	7/1/14 1:50	5.04	18.5	0.5	0.5	0	0	19.5
7/1/14 0:17	7/1/14 0:24	2.53	9	0.5	0.5	0	0	10
7/1/14 0:06	7/1/14 0:22	2.48	12.5	0.5	0.5	2.6	0	16.1
7/1/14 0:17	7/1/14 0:24	1.81	7.5	0.5	0.5	0	0	8.5
7/1/14 1:38	7/1/14 1:50	6.26	19	0.5	0.5	1	0	21

[1]Public Data: http://www.nyc.gov/html/tlc/html/about/trip_record_data.shtml


A1:(Base Fare:99%,Duration:1%)
A6:(Base Fare:68%,Duration:16%)
A4:(Base Fare:44%,Tip:25%)
A12:(Base Fare:59%,Duration:19%)
A9:(Base Fare:50%,Tip:24%)
A8:(Base Fare:31%,Tip:29%)
A3:(Base Fare:79%,Distance:12%)
A5:(Duration:53%,Base Fare:2
A10:(Base Fare:46%,Tip:22%)
A11:(Base Fare:31%,Tip:29%)
A7:(Duration:87%,Base Fare:7%)
Base Fare ($)
Duration (minutes)
Anomaly Score
600
500
400
300
200
100
0
−100
−20
0
20
40
60
80
100
120
5.4
4.8
4.2
3.6
3.0
2.4
1.8
1.2
0.6
0.0


Network Utilization Per Minute (Stylized)
40
20
0
−20
−40
Idle CPU Per Minute (Stylized)
40
20
0
−20
−40
Anomaly Score with Explanation
2.0
1.5
1.0
0.5
0.0
Directionality Network Utilization
20
15
10
5
0
−5
−10
−15
−20
Directionality Idle CPU
20
15
10
5
0
−5
−10
−15
−20
0
500
1000
1500
2000

The Moving Example

A Fan/Turbine

1000 pts in each blade

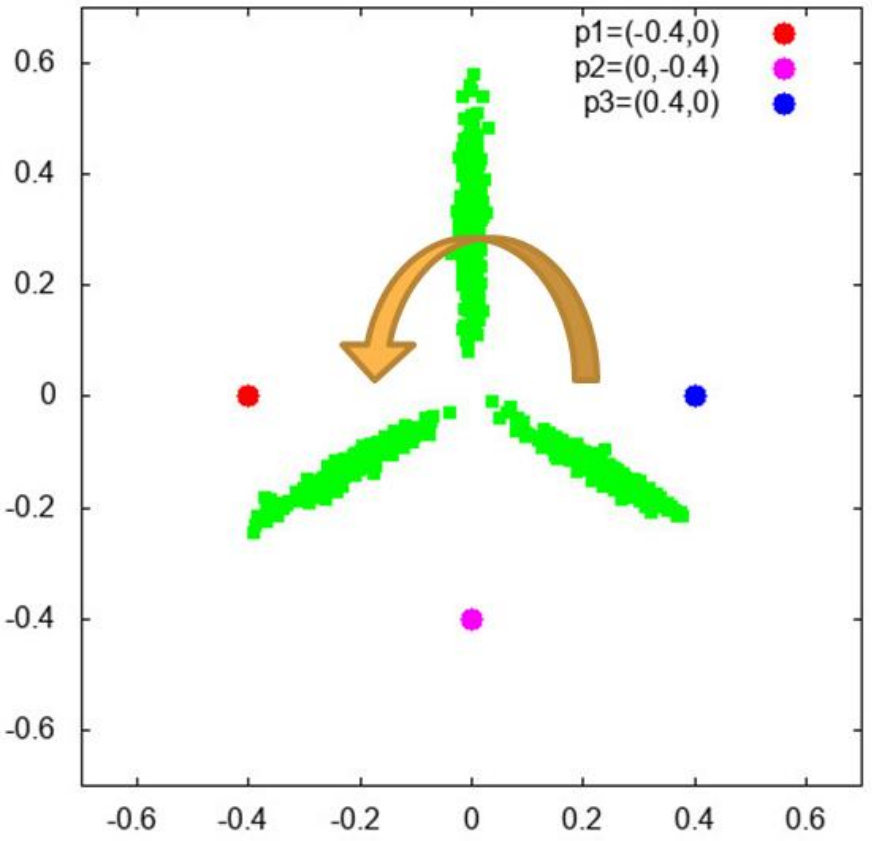
Gaussian, for simplicity

Blades designed unequal

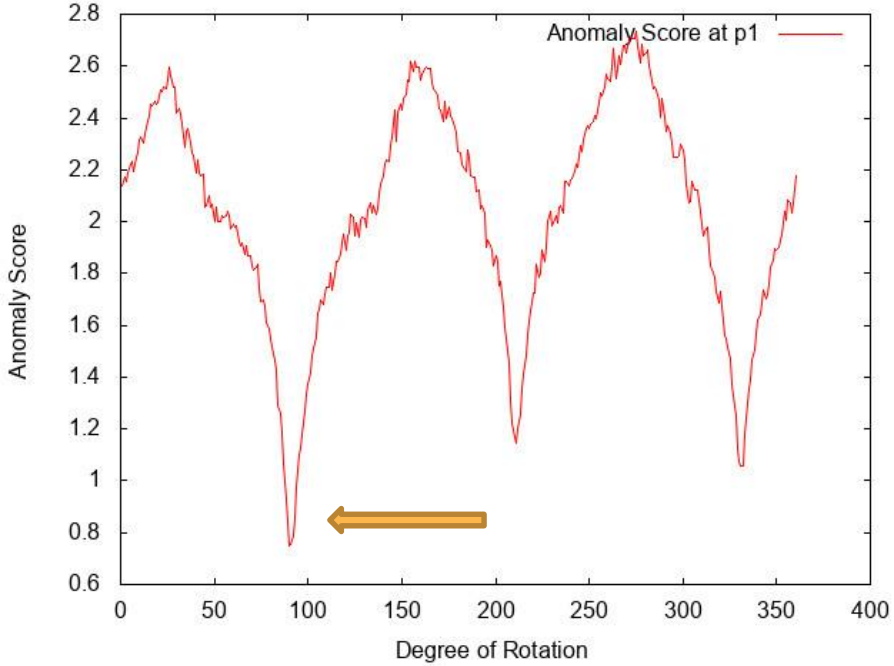
Rotate counterclockwise

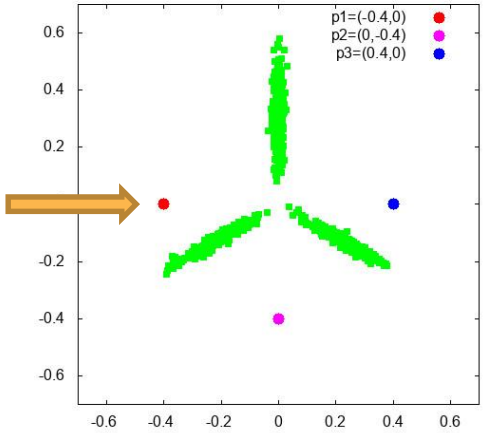
3 special “query” points

100 trees, 256 points each



Anomaly Score at P1

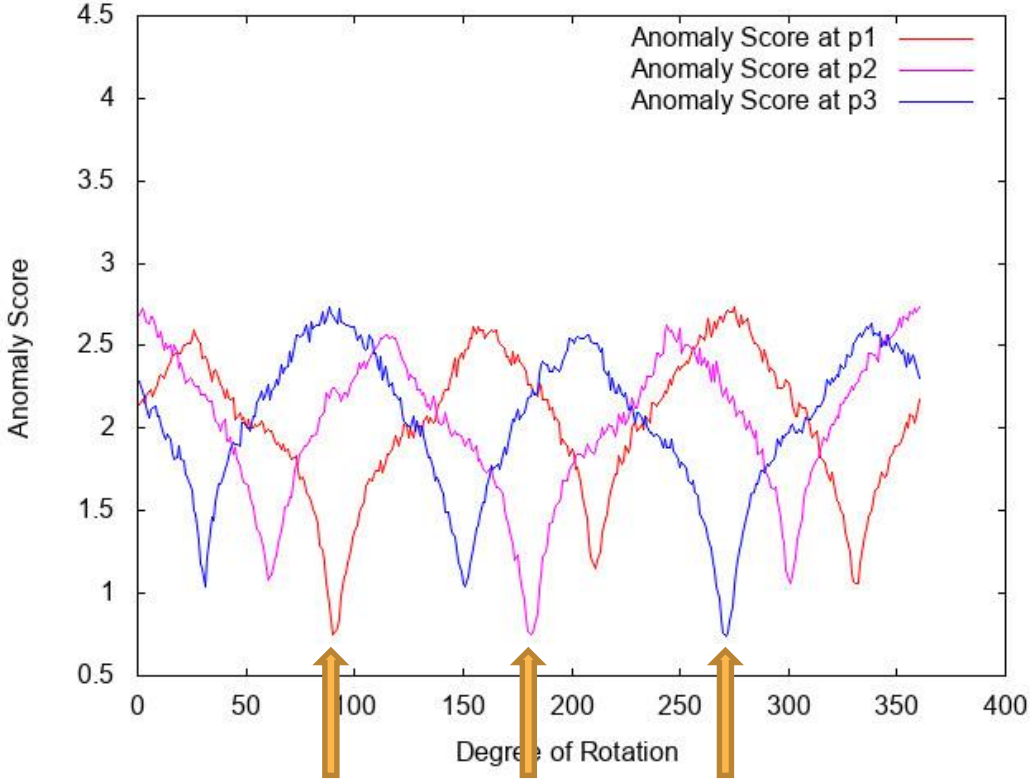


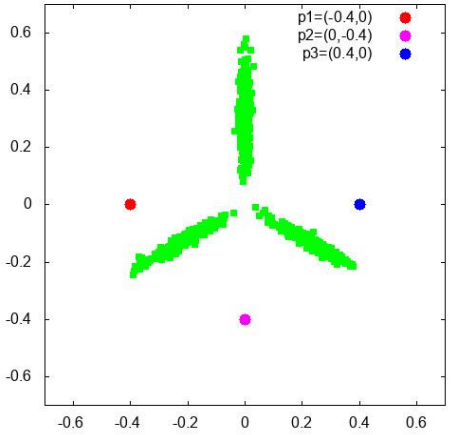


Blade overhead = Not an anomaly

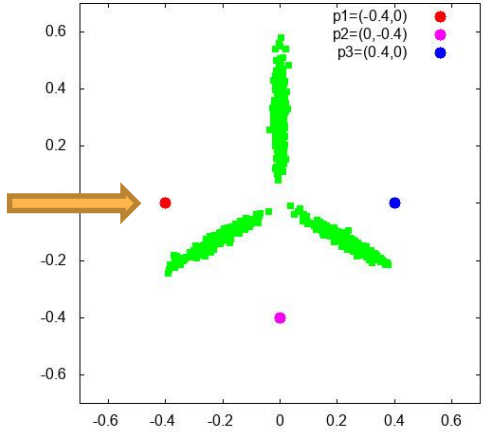
What is going on at 90 degrees?

All 3 Blades

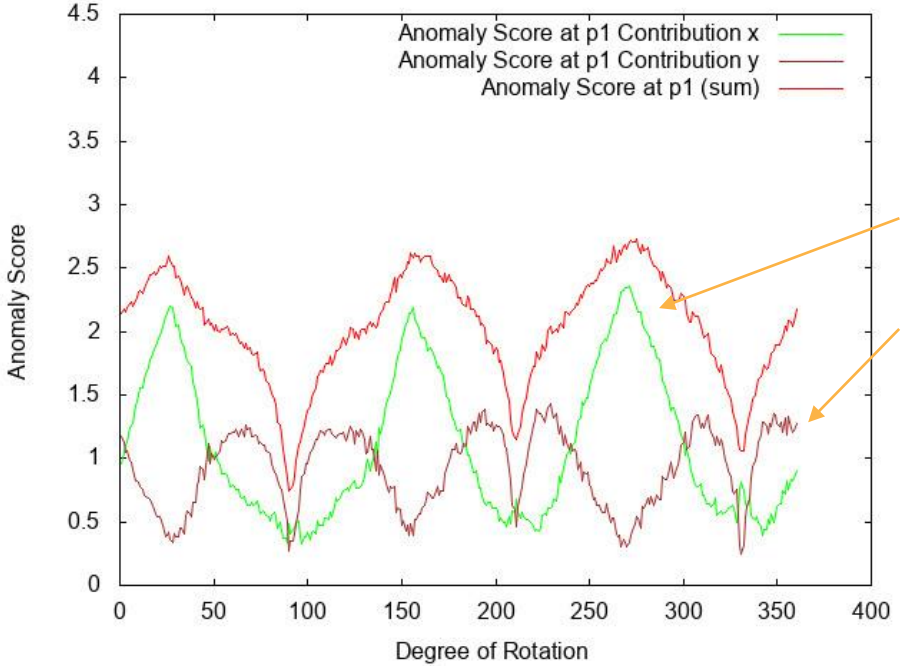

Anomaly Score at p1
Anomaly Score at p2
Anomaly Score at p3
Anomaly Score
Degree of Rotation
4.5
4
3.5
3
2.5
2
1.5
1
0.5
0
50
100
150
200
250
300
350
400


p1=(-0.4,0)
p2=(0,-0.4)
p3=(0.4,0)
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.6
-0.4
-0.2
0
0.2
0.4
0.6

Transparent Attributions



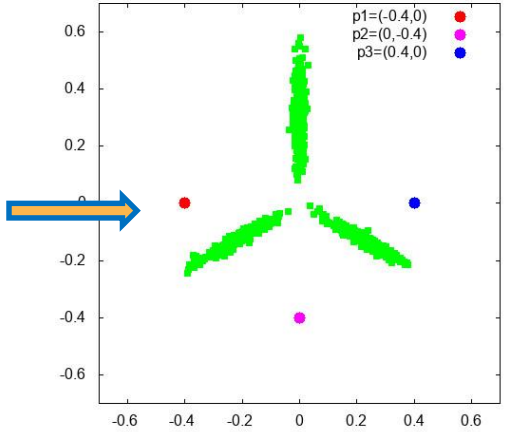
x coordinate's contribution for p1?

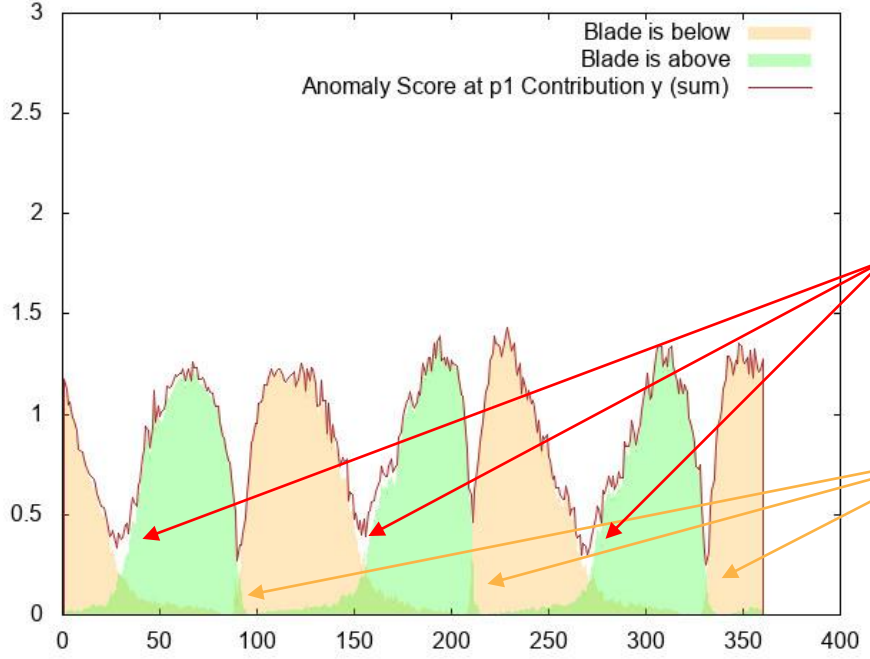


p1 is far away in x-coord most of the time

But what is happening to y?

Directionality





Slowly rotating away
Total score remains high

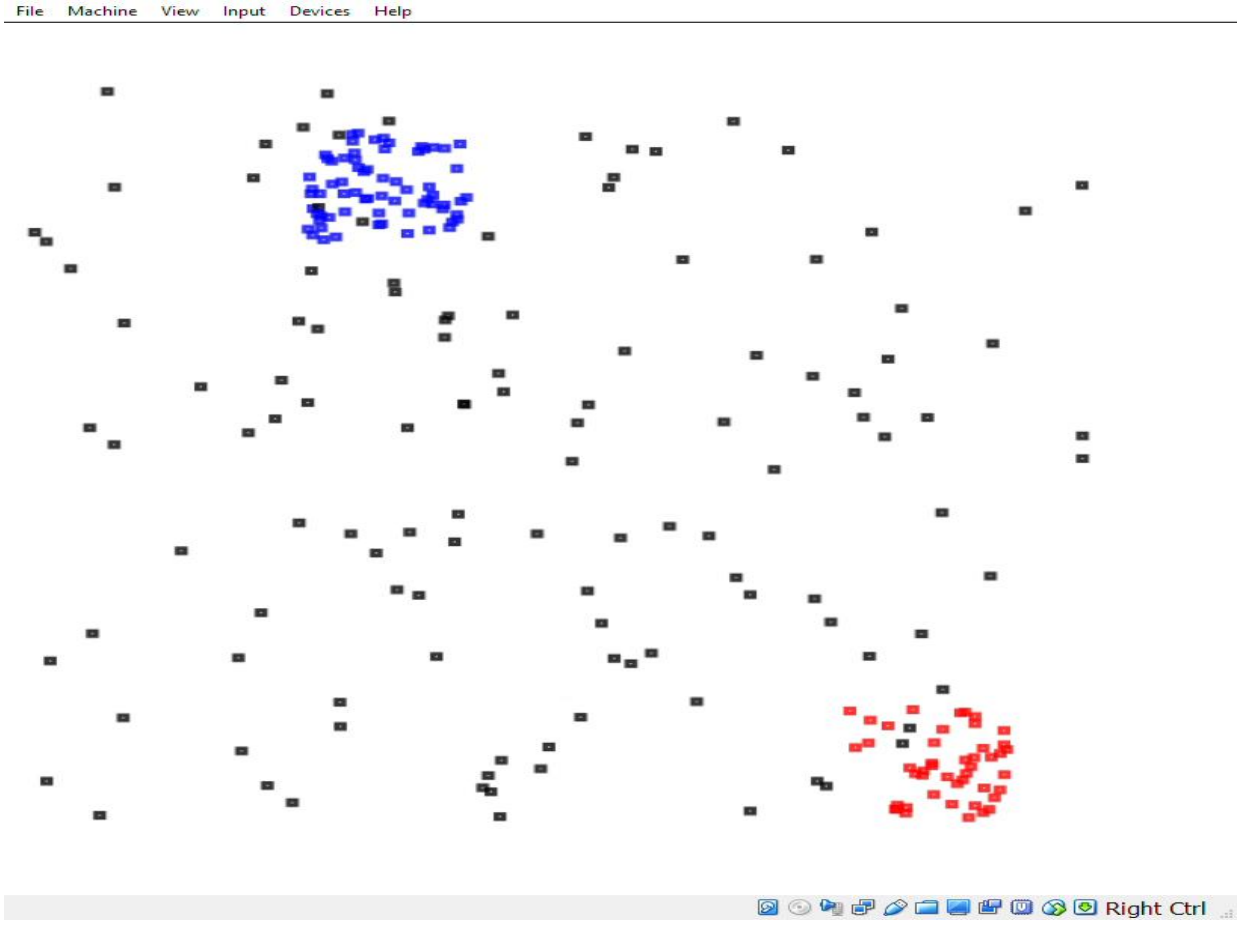
Sharp transition when the blade moves from above to below at p1!
Total score plummets.

Hotspots on a Stream

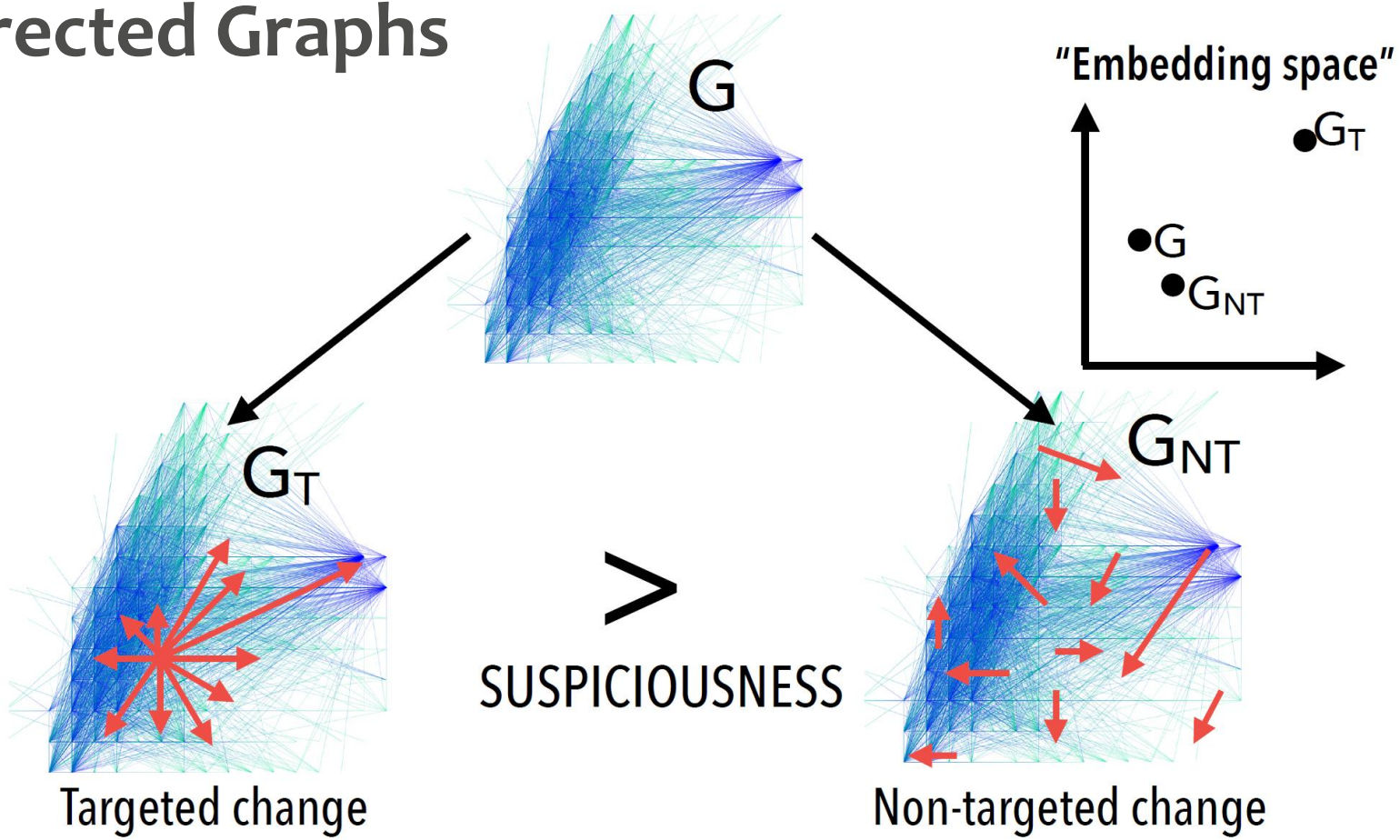
lyft

UBER

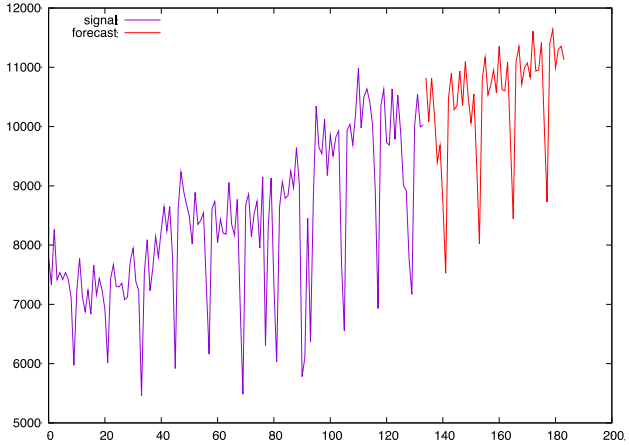
Hotspots on a Stream

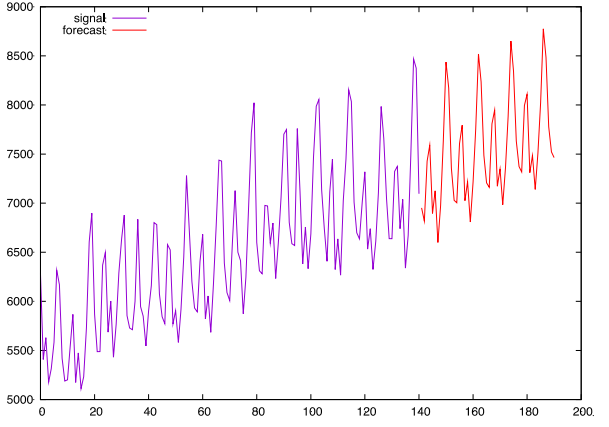


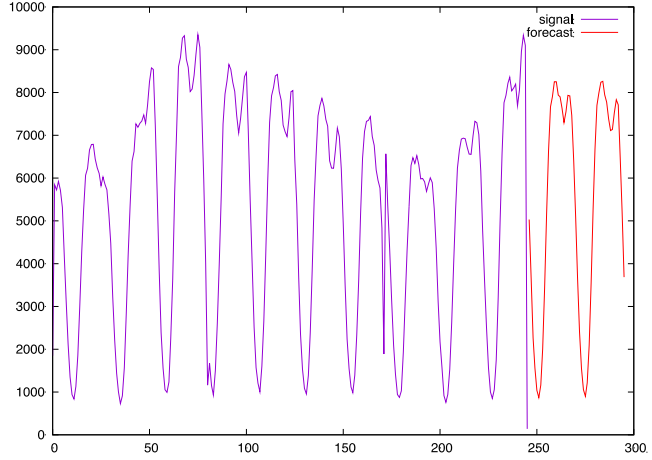
Detecting Anomalies in Directed Graphs

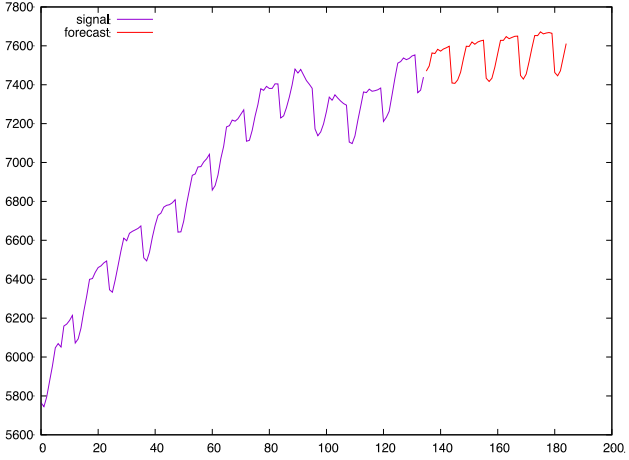


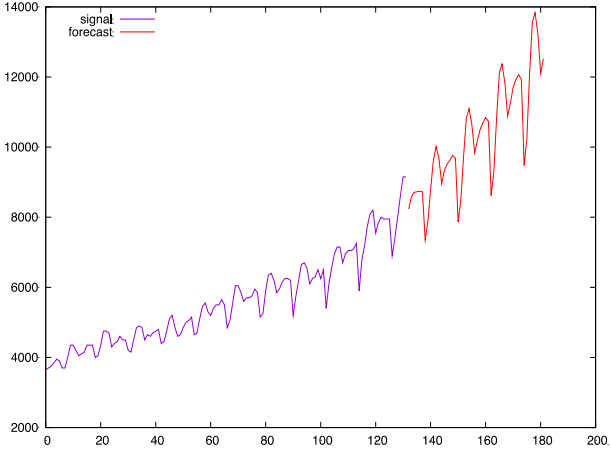
Time Series Forecasting











False Alarms

Anomaly Detection with user feedback

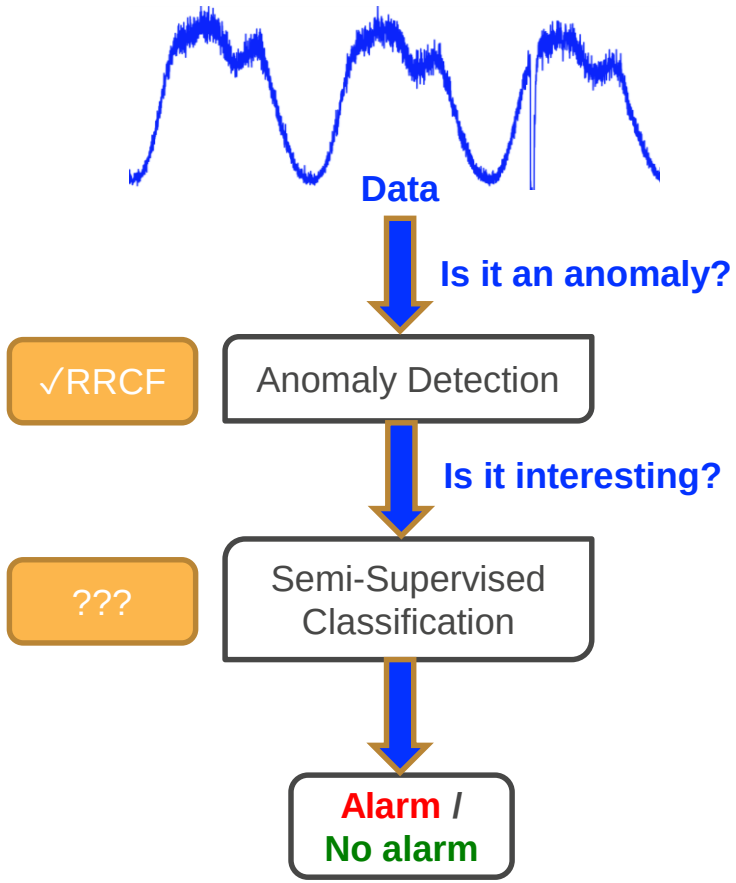
Alarm fatigue: Personnel become desensitized

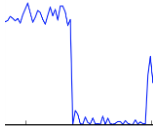
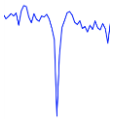
560 alarm related deaths during 2005—2008 (FDA data)

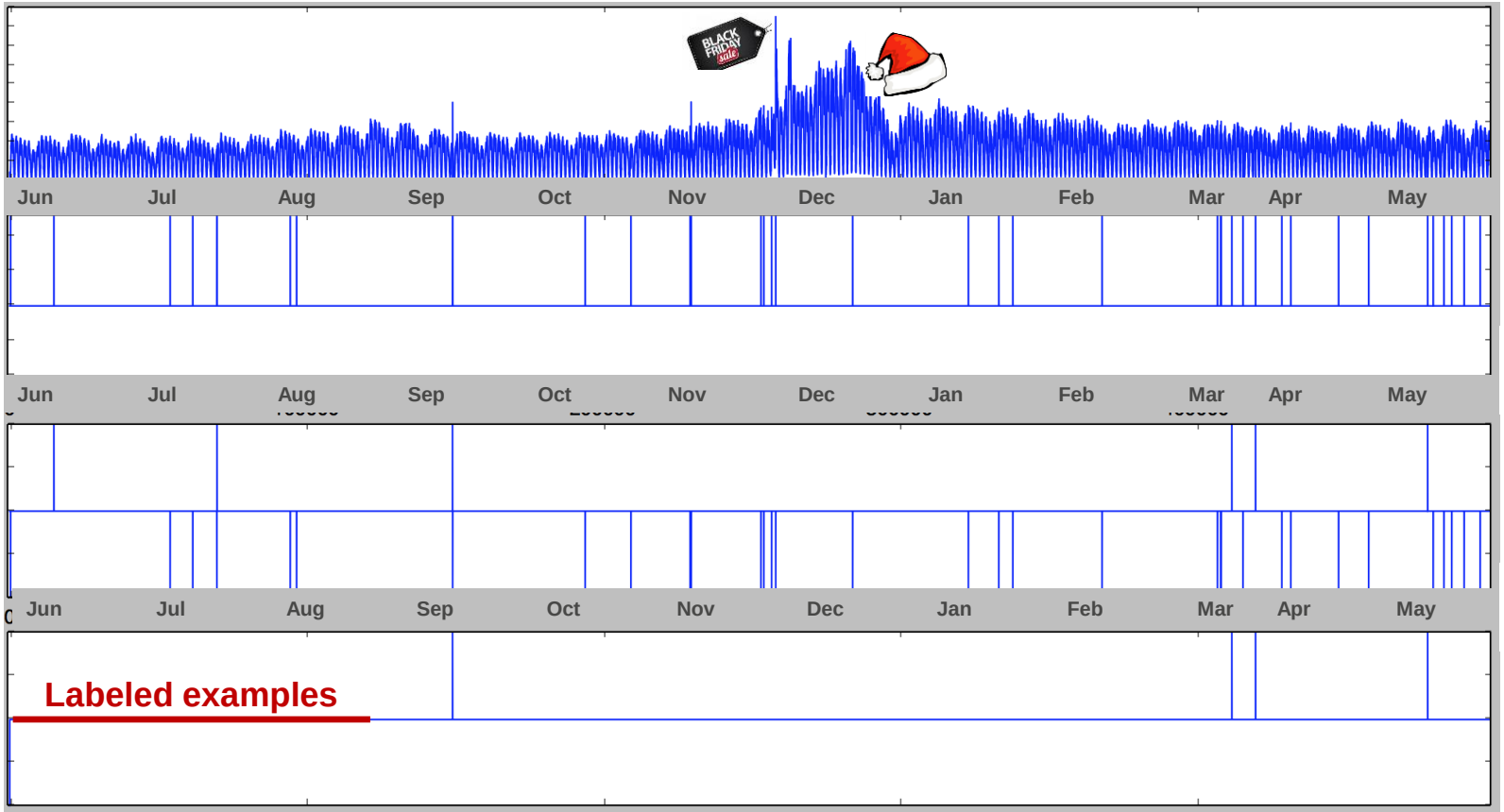
""Alarms sounded 1 hour before the nurse discovered he was unresponsive. He eventually died. An investigation found the alarm volume had been turned off."

System View

Streaming semi-supervised learning



Orders Data:  vs. 



Robust Random Cut Forrest

Summary of a dynamic data stream, efficient, number of use cases…

Anomaly Detection

Attribution and Directionality

Hotspot Detection

Classification

Forecasting

Missing Value Imputation

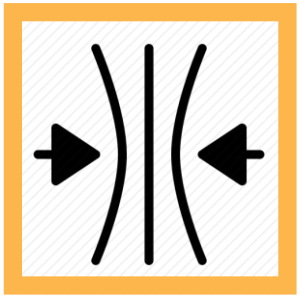
Anomaly Detection in Streaming Directed Graph

Amazon Kinesis Data Analytics

The *easiest* way to use machine learning!

Unsupervised

Online



Adaptive

Real-time

Available Now

- Anomaly Detection
- Anomaly Detection with explanations
- Hotspot Detection (releasing soon!)

Coming Soon

- Classification
- Time Series Forecasting
- Missing Value Imputation

Contributors To This Project

Roger Barga, Kapil Chhabra, Charles Elkan,
Dhivya Eswaran, Christos Faloutsos, Praveen Gattu,
Gaurav Ghare, Sudipto Guha,
Shiva Kasiviswanathan, Nina Mishra,
Morteza Monemizadeh, Lauren Moos,
Yonatan Naamad, Ryan Nienhuis, Gourav Roy,
Okke Schrijvers, Joshua Tokle, and
Tal Wagner