CRESST

Complete Rare Event Specification using Stochastic Treatment

ODSC Europe, Sep 2020

Walmart Global Tech India

About the Speaker

Debanjana Banerjee,
Data Scientist, Walmart Global Tech India

Walmart Global Tech India

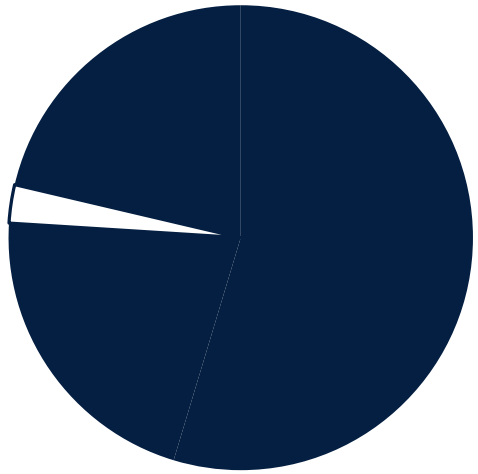
Ritish Menon,
Sr. Data Science Manager,
Walmart Global Tech India

Walmart Global Tech India

The Premise

- What are Rare Events
- Rare Events vs Anomalies
- Challenges
- Getting Started with Rare Events

What are Rare Events?



$P(\bigcirc) \longrightarrow 0$

$P(\bullet) \longrightarrow 1$

Dense Space

Rare Space

- A rare event is one which has numerous junctures to occur in the population space but its true realization is very low.
- In spite of its probability of occurrence being close to 0, its specification can be quite extensive.

  Example: Within the parent rare event of Product Safety (probability of occurrence ~0.01), we have various potential hazard types such as Mechanical Injury, Allergies, etc., each of which rarer still.
- Our objective here, is to study the stochastic aspects associated with such rare events, limited to
  1. Rare Event Forecast
  2. Pattern Deviation Detection in Rare Events

Rare Events vs Anomalies

Rare Events

- Adhere to certain specification
- Allows variability in rate of imbalance
- May pertain to supervised or, unsupervised learning
- Sample independent

Anomalies

- Usually do not adhere to specification
- Variability in rate of imbalance is limited
- Calls for unsupervised learning
- Sample dependent

Identifying Rare Events

Common Challenges

- Fewer Training Examples
- Qualitatively Inexhaustive Training Data
- Unknown/Misreported Rate of Imbalance

Few Methods We Can Use to Identify instances of a Rare Events

- One Class Classifiers
- Semi-Supervised Learners
- Positive-Unlabeled Learners (when you do not have trained labels on the Dense Class)

Best Way to Identify a Rare Event

Case Specific!!

Some links you can check out!

[iCASSTLE (NLP)]

[PU Learning]

[PAC Learning]

Ref Banerjee et al. (2018), Liu et al., towardsdatascience.com

Global Tech India

Rare Event Treatment ***Post*** Identification

1. Forecasting instances (counts) of a Rare Event
2. Detecting Stochastic Deviation in a Rare Event

Rare Event Forecast

Forecasting Case Counts for a Rare Event

- Count Time Series
- How it differs from Classical Time Series
- How we can handle Count Time Series

Walmart Global Tech India

Examples of a Count Time Series

By definition, a count time series is one that enumerates the number of cases of a certain event at a given point in time.

Examples:

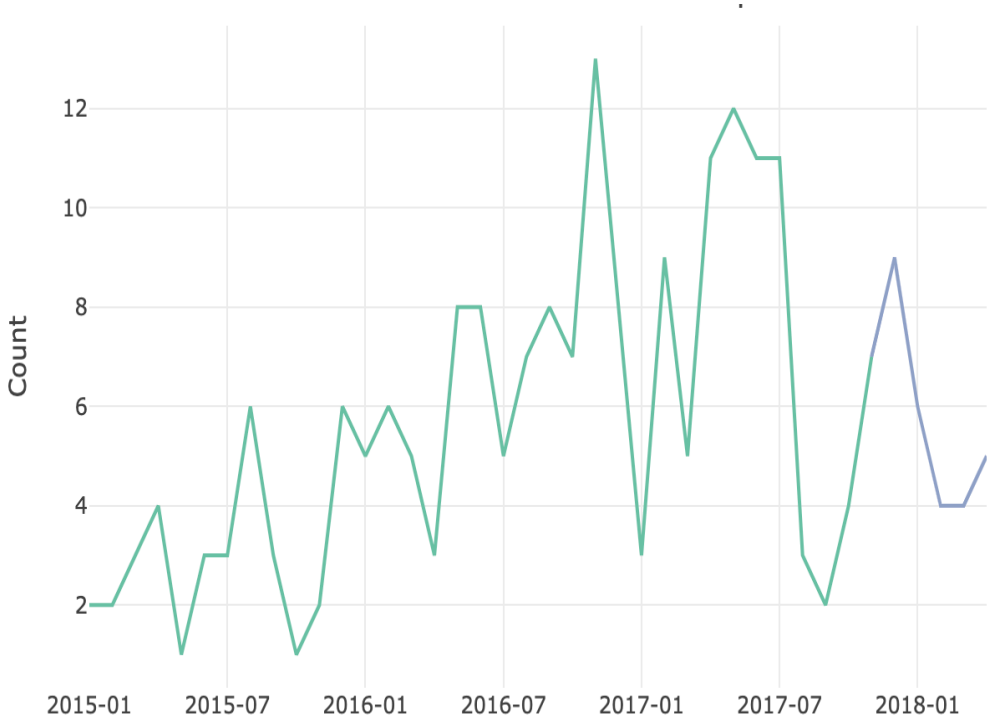
1. Variable A: Count of students graduating in a town in a given year
2. Variable B: Count of players injuring themselves on a field every month

One major difference b/w the two examples above -

The magnitude of Variable A is likely to be in thousands (high), but that of B would be in tens!

High magnitude count data can be approximated to a normally distributed variable but the same cannot be done for low magnitude count data

Monthly Count of Player Injuries



Global Tech India

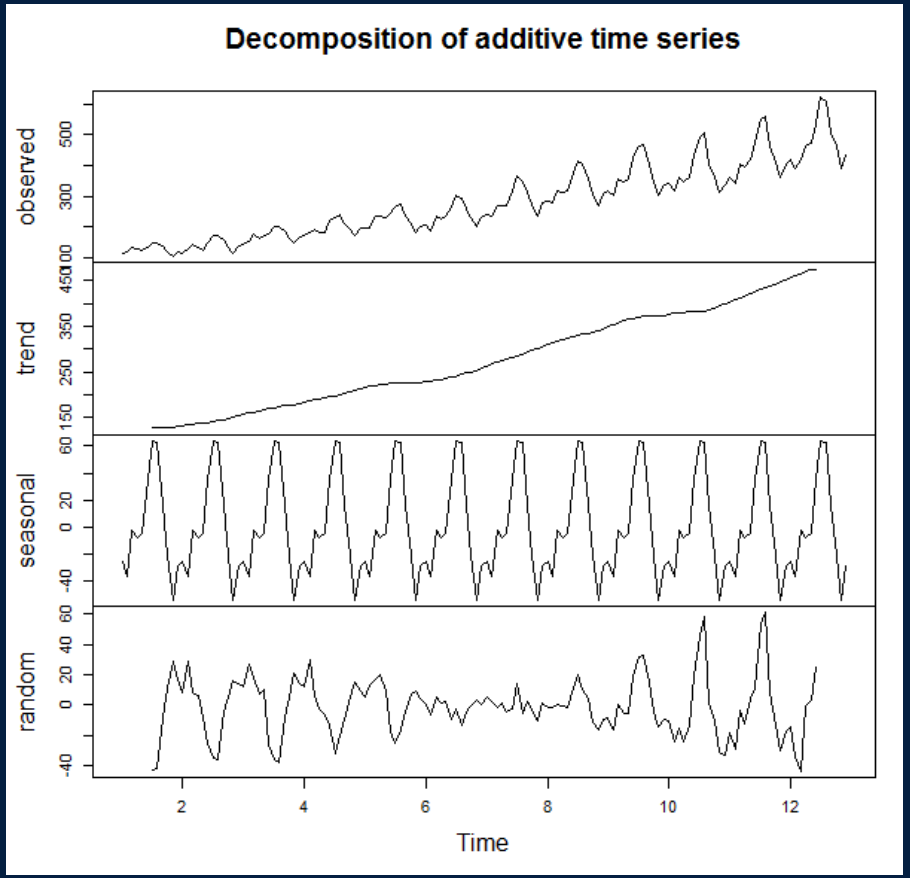
A time series capturing the number of instances of a Rare Event across time is essentially a ***low magnitude Count Time Series***

Classical Time Series Forecast

- Decompose the Time Series into
  - Trend
  - Seasonality
  - Random Component

- Normality Assumption on the Random Component of the Time Series !!

None if these fits a Rare Count Time Series which usually originates from a Poisson Distribution!



What We Want

- To forecast the number of instances of the rare event
- To help in proper planning
- To foresee the necessity of any precautionary measure (particularly if the rare event is of an adverse nature)

Things to Bear in Mind

- We are working with Count Data
- Magnitude of each point of the count time series is very low (due to rarity)
- Poor fit to Normal Distribution
- Better fit to Poisson, Negative Binomial distributions
- Try to maintain inherent data origin

How to Proceed

- Maintain ‘countness’ of data by studying the underlying stochastic process instead of Trend, Seasonality, Randomness individually
- Count Time Series can be handled with Count Processes
- Choose the closest distribution to estimate parameters of the Count Process
- Allow higher freedom in choice of regressors (stochastic or, otherwise)

Poisson Process

An example of Count Process

- Non-Homogeneous Poisson Process
- Predicting Rare Case Counts using tscount

Non-Homogeneous Poisson Process

$\{N_t, t \geq 0\}$ is said to be a NHPP if

- $\{N_t, t \geq 0\}$ has a continuous time space
- $N_0 = 0$
- $P((N_{t+h} - N_t) = 1) = \lambda(t) * h + o(h)$
- $P((N_{t+h} - N_t) > 1) = o(h)$
- $(N_{t1+h} - N_{t1})$ & $(N_{t2+h} - N_{t2})$ are independent $\forall\ h\ if\ t1 \neq t2$
- $N_t \sim P\big(m(t)\big)$

(the Poisson distribution) where $m(t)$ is the mean value function

Mean Value Function & Intensity Function (of a Rare Count Process)

If $\{N_t, t \geq 0\}$ denotes the number of instances of the Rare Event till time t and can be fit to a NHPP with MVF $m(t)$, then

- $m(t) = E(N_t)$
- Intensity Function $\lambda(t) = \frac{d}{dt} m(t)$

Physically, $\lambda(t)$ is the expected number of cases of the Rare Event per unit time i.e.,

the *Hazard Rate*

Count Process | Poisson | R Package Used: tscount

Predicting Equation

Let $\{Y_t;\ t:\ 1(1)n\}$ denote the Count Time Series.

$$Y_t \sim P(m_t)\ ;\ E(Y_t \mid Y_{t-1}, Y_{t-2}, \ldots Y_1, X_t) = m_t$$

Choosing an identity link, the predicting equation for Y_t is given as

$$\widehat{Y_t} = \beta_0 + \sum_{k=1}^{p} (\beta_k * Y_{t-i_k}) + \sum_{l=1}^{q} (\alpha_l * m_{t-j_l}) + \gamma * X_t$$

where $(i_1, i_2, \ldots . i_p)$ & $(j_1, j_2, \ldots . j_q)$ are arbitrary lags of Y_t & m_t respectively.

OPTIMIZATION CRITERION

The regression coefficient as well the time lags to be included are chosen by a trade-off between QIC and MAPE.
QIC is an indicator of model complexity; lower QIC is preferred for better model health.
MAPE is an indicator of model accuracy; lower MAPE indicates higher model accuracy.

QIC usually rises with lower MAPE. All optimization are performed over discrete spaces.

Ref Liboschik et al. (2015)

Pattern Deviation

For a Rare Event Sub-Class

Walmart Global Tech India

Problem Premise

- Parent Rare Event P
- Within P, we have sub-classes A, B, C etc., each representing a specification in P

Example I:

P: Accidents in a factory per day

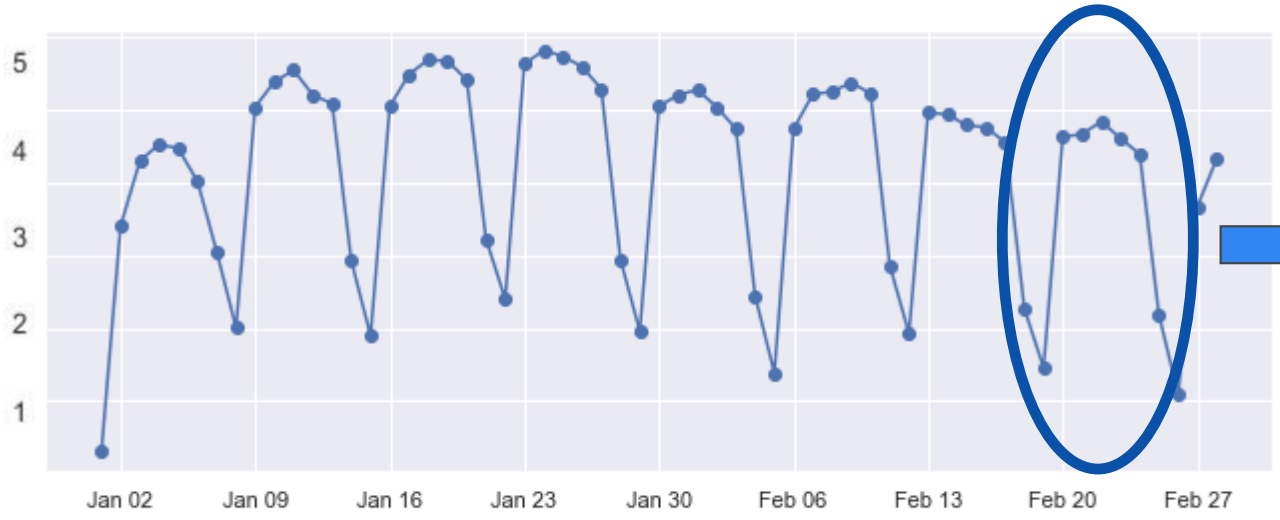
Sub Classes of P: Mechanical Injury, Fire Hazards, etc.

Example II:

P: No. of fraudulent transactions in an eCommerce

Sub Classes of P: Payment Fraud, Offer Fraud, etc.

Visualizing **Subsequences**, Current State and Base State for a Time Series



Subsequence of length 7

Quick Notation:

$T: T^{th}\ subsequence$

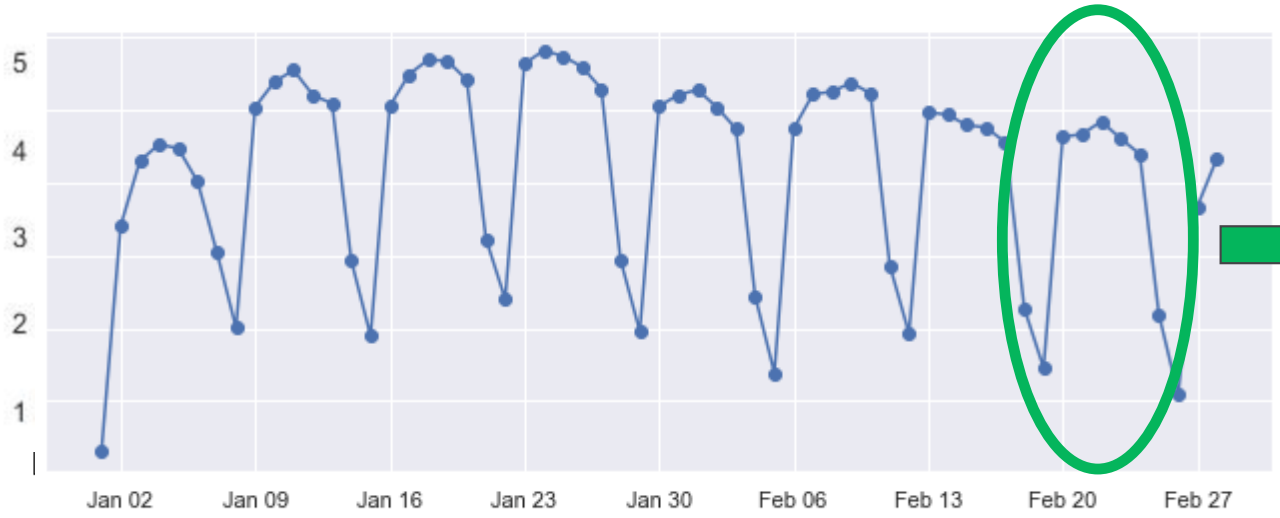
$t:\ t^{th}\ point\ in\ time$

$C_T: a\ subsequence$

We want to identify anomalous subsequences in the time series

Ref Image: Dataquest

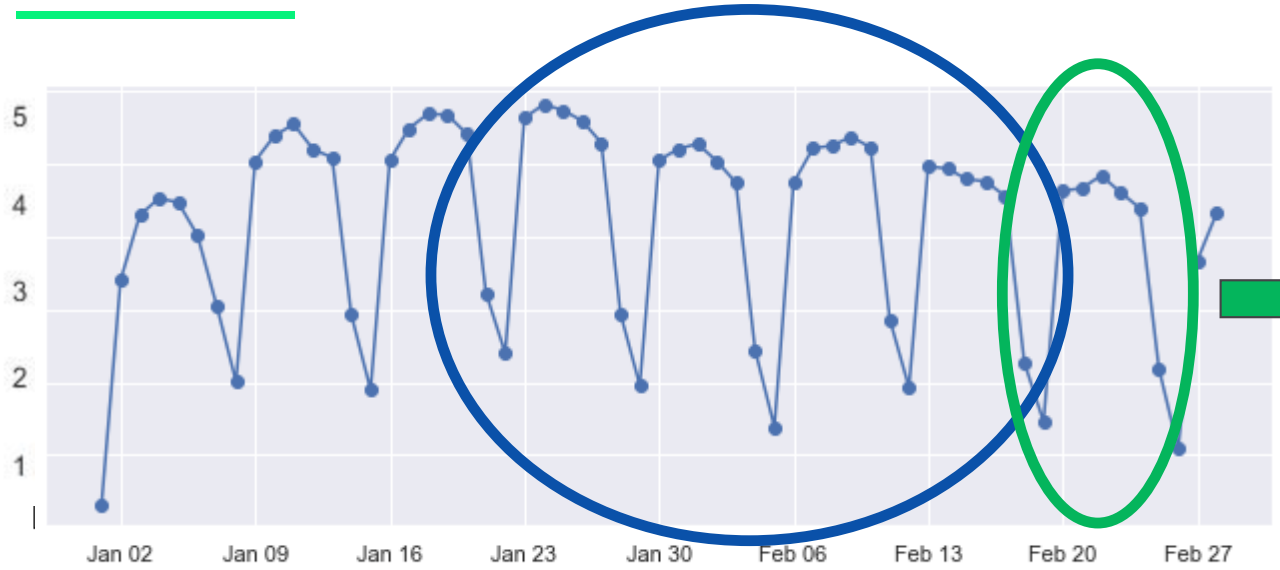
Visualizing Subsequences, **Current State** and Base State for a Time Series



Current State
Subsequence for which we want to obtain the deviation score

Ref Image: Dataquest

Visualizing Subsequences, Current State and **Base State** for a Time Series



Current State
Subsequence for which we want to obtain the deviation score

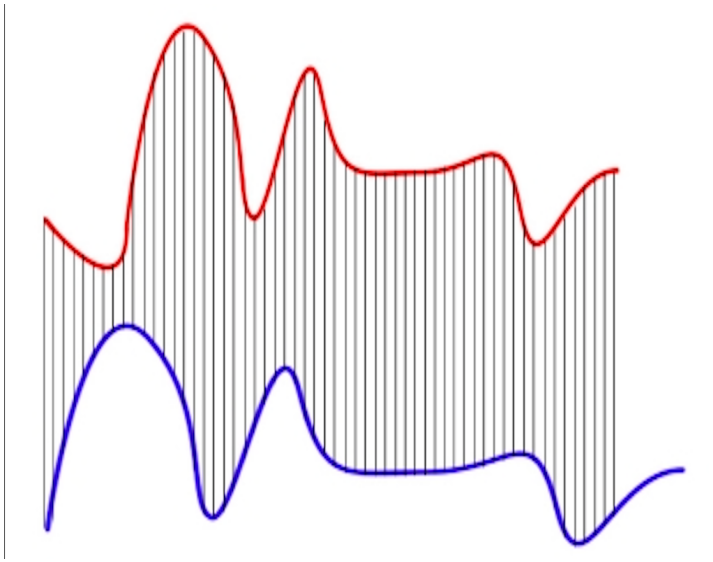
Base State
Subsequence against which we want to compare our Current State

Note, Base State and Current State need not have the same length!

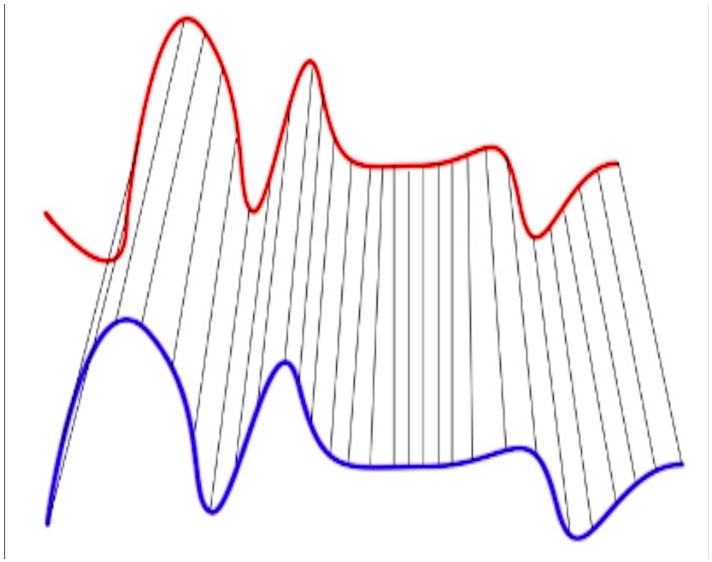
Ref Image: Dataquest

Dynamic Time Warping & DTW Distance (the minimum distance b/w two dynamically warped time series!)

Euclidean Map



Dynamically Warped Map

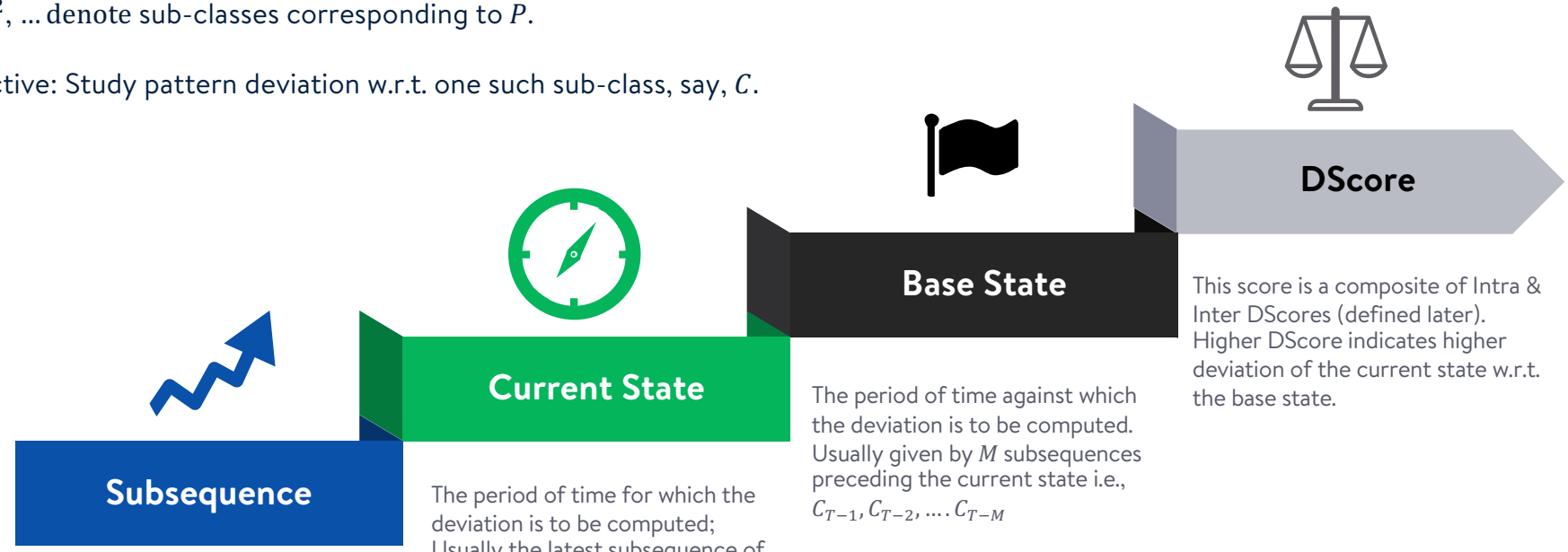


Ref Image: SFL Scientific

Deviation Scores | Concept & Definitions

Let P be a Parent Rare Event.
$C^1, C^2, \ldots$ denote sub-classes corresponding to P.

Objective: Study pattern deviation w.r.t. one such sub-class, say, C.



This score is a composite of Intra & Inter DScores (defined later). Higher DScore indicates higher deviation of the current state w.r.t. the base state.

The period of time against which the deviation is to be computed. Usually given by M subsequences preceding the current state i.e., $C_{T-1}, C_{T-2}, \ldots . C_{T-M}$

The period of time for which the deviation is to be computed; Usually the latest subsequence of C, say, $C_T = \{C_t, C_{t-1}, \ldots C_{t-k}\}$

A subsequence of C, say, C_T (of length k) is defined as $\{C_t, C_{t-1}, \ldots C_{t-j-1}\}$.
In general, the subsequence length may differ but for our purpose, we will always consider $j = 7$ (a week)

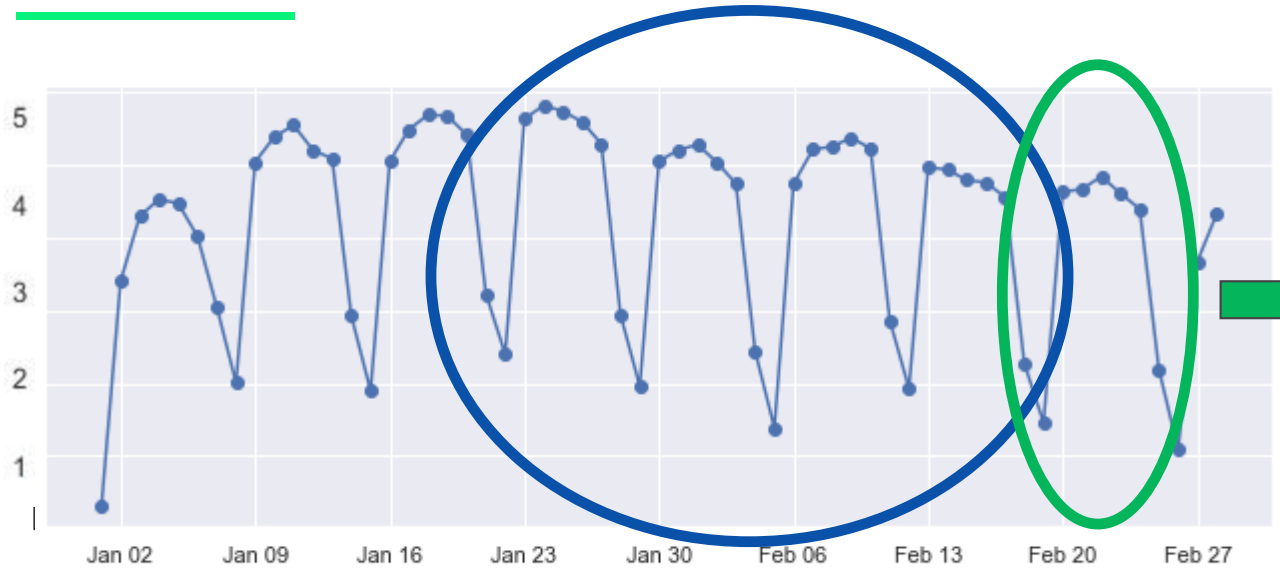
Deviation Score

DScore

Identifies if a subsequence is deviating

1. W.r.t. itself (Intra DScore)
2. W.r.t. its cluster or, sister classes (Inter DScore)

Visualizing the **Intra DScore**

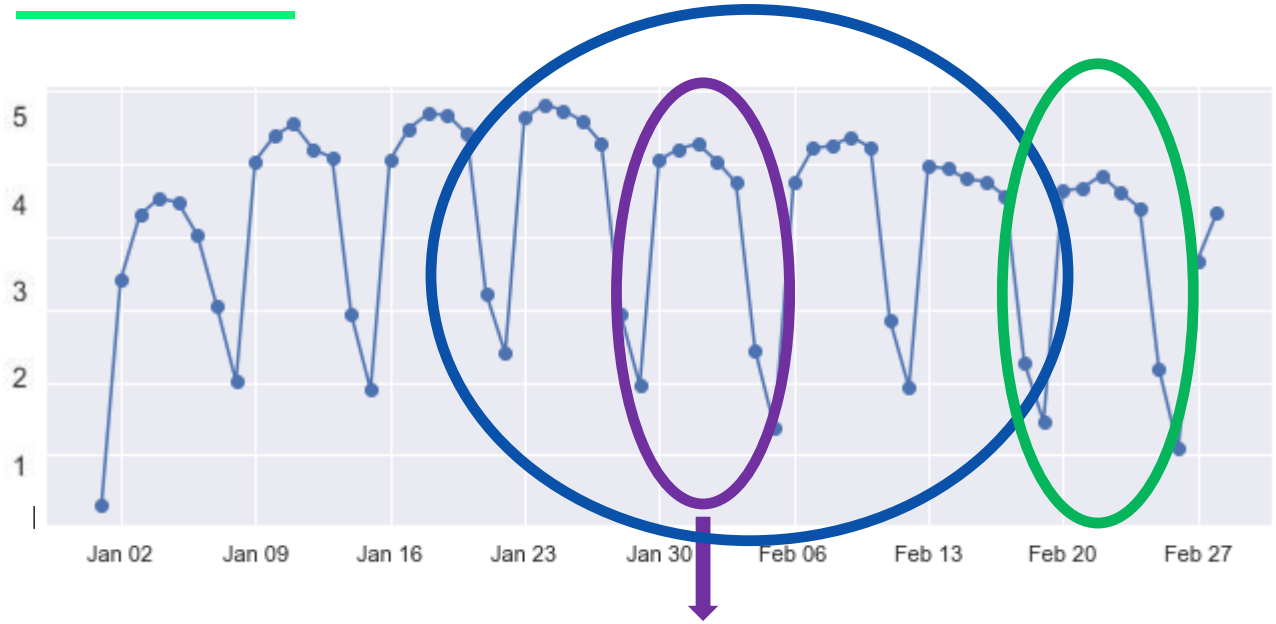


Current State
Subsequence for which we want to obtain the deviation score

Base State
State against which we want to compare our Current State

Ref Image: Dataquest

Visualizing the **Intra DScore**



Nearest Neighbor to the Current State
Based on DTW distance

Intra DScore is physically the normalized DTW distance b/w the Current State and its NN in the base state!

Intuition: If you are not close enough to our closest neighbor, you are an anomaly!

Ref Image: Dataquest

Intra Dscore | Deviation w.r.t. itself

BASE STATE

We are comparing C_T against its immediate past. Depending on use case, the number of preceding subsequences chosen is M_1. For longer base states, M_1 could be larger.

C_T is thus compared against $C_{T-1}, C_{T-2}, \dots \dots C_{T-M_1}$

DISCORDS

A subsequence C_T is said to be in discord with its base state iff $DTW\left(C_T, NN(C_T)\right) > A$, A being a pre-determined threshold.

INTRA DSCORE

The Intra Dscore of C_T w.r.t. its base state is given as

$$IntraD(C,T) = \frac{DTW\left(C_T, NN(C_T)\right) - DTW_{min}}{DTW_{max} - DTW_{min}}$$

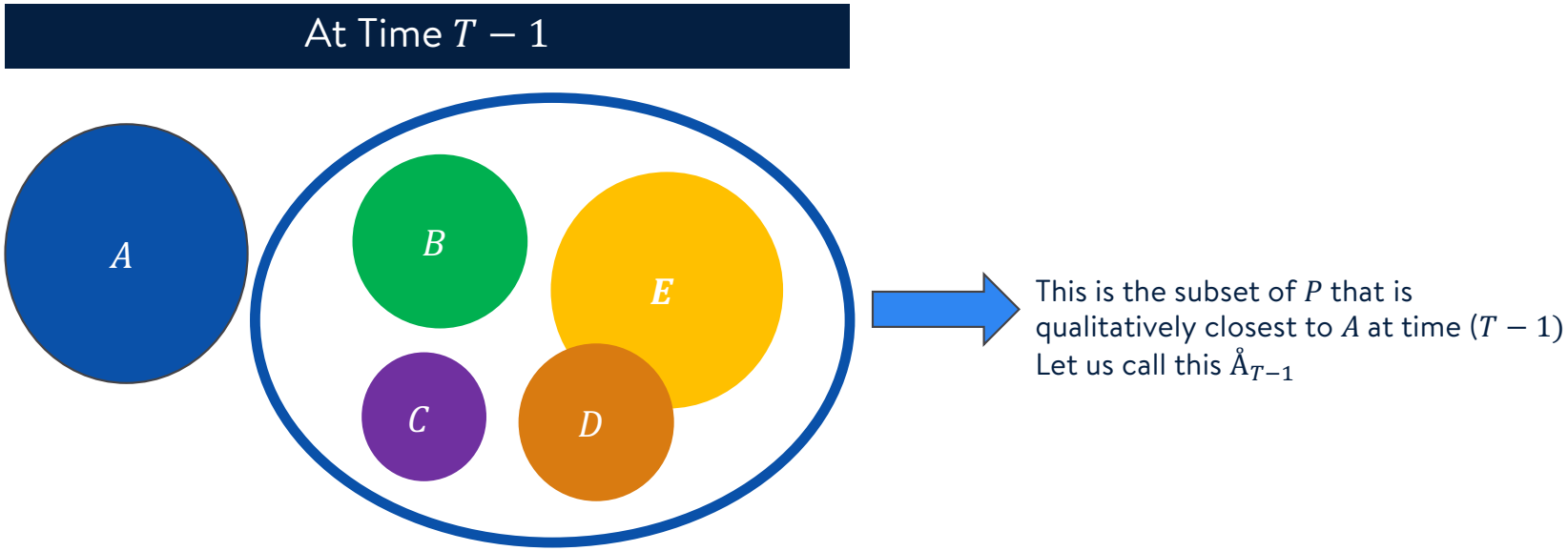
$$where\ DTW_{min} = \min\left(DTW\left(C_{T'}, NN(C_{T'})\right)\right) \&\ DTW_{max} = \max\left(DTW\left(C_{T'}, NN(C_{T'})\right)\right)$$
$$T' \in \{C_T, C_{T-1}, C_{T-2}, \dots \dots C_{T-M_1}\}$$

** DTW indicates DTW distance (Mininum distance between dynamically warped set of points)

Visualizing the Inter Dscore:
Relevant Sister Classes

Say, the Parent Rare Event is called P

The Rare Event Sub-Class for which you are doing the analysis is called A



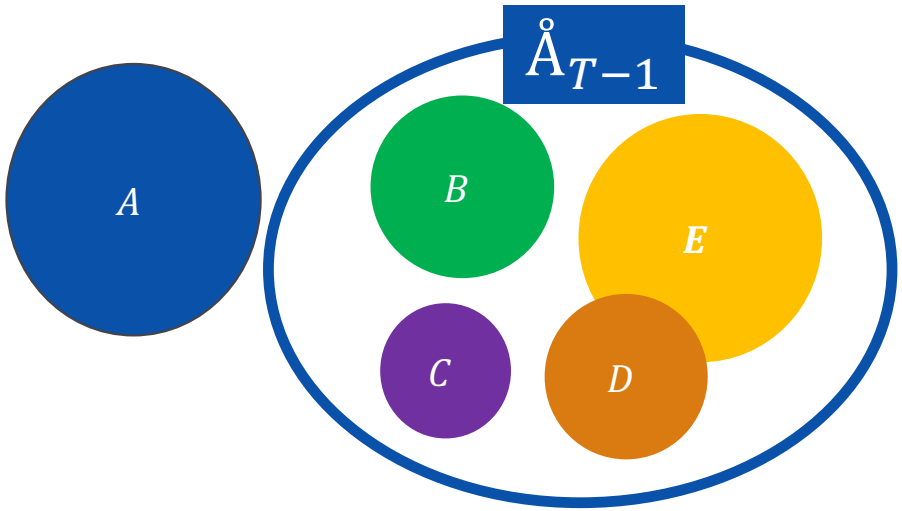
Visualizing the **Inter DScore**

A is believed to have deviated from its Sister Classes if
$|\rho_T - \rho_{T-1}| > \epsilon$

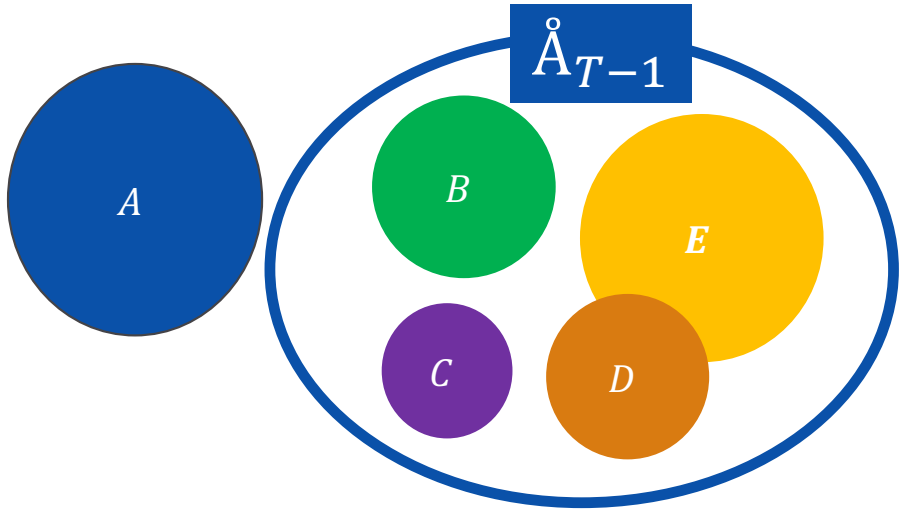
Say, the Parent Rare Event is called P

The Rare Event Sub-Class for which you are doing the analysis is called A

At Time $T-1$: $Corr\left(A, \mathring{A}_{T-1}\right) = \rho_{T-1}$



At Time T: $Corr\left(A, \mathring{A}_{T-1}\right) = \rho_T$



Inter Dscore | Deviation w.r.t. Sister Classes (how the belongingness is changing)

SISTER CLASSES

Since C_T is known to be a sub-class of a parent rare event, we want to study the relationship between C_T and the sub-classes closest to C_T and check how the relationship is mutating over time.

Sub-class S is said to be a **Relevant Sister Class** of C at time T iff $Corr(C_T, S_T) > B$, B being a pre-determined threshold.

Note that, the same sub-classes will not remain the closest with C over time.
For our purpose, we will compare the base and the current relationship between C and its relevant sister classes for the **base state**.

BASE STATE

Depending on use case, the number of preceding subsequences chosen is M_2. For longer base states, M_2 could be larger.
For our purpose, M_2 is fixed at 1.

Hence, we will be comparing $Corr(C_T, S_T)$ & $Corr(C_{T-1}, S_{T-1})$ $\forall\ S\ \in \{the\ relevant\ sister\ classes\ of\ C\ for\ base\ time\ period\ T-1\}$

INTER DSCORE

The Inter Dscore of C_T w.r.t. its base state is given as

$$InterD(C,T) = \frac{\sum_1^K |Corr(C_T, S_T^k) - Corr(C_{T-1}, S_{T-1}^k)|}{2 * K}$$

$$where\ S^1, S^2, \ldots S^K\ are\ the\ K\ relevent\ sister\ classes\ of\ C\ for\ base\ state\ T-1$$

Final Deviation Score | Thresholding

The final DScore for C_T is given as

$$DScore(C,T) = IntraD(C,T) + InterD(C,T)$$

THRESHOLDING

C_T is believed to have deviated significantly w.r.t. its base state iff

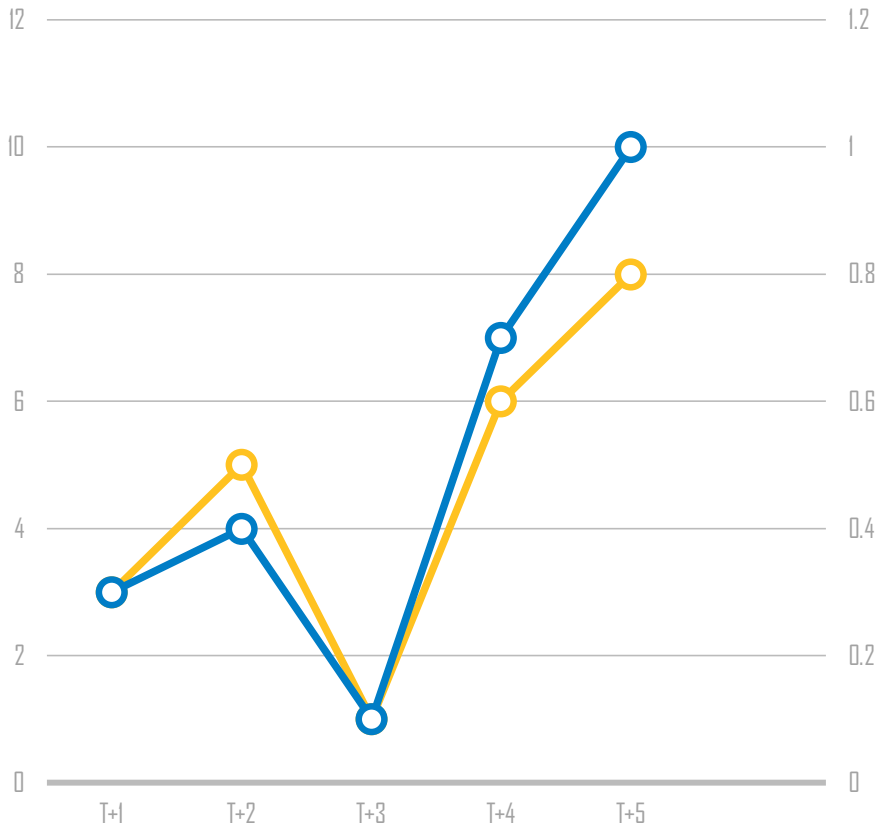
$$DScore(C,T) > \boldsymbol{D^*}$$

COMPUTATION OF $D*$

- Based on DScores of previously anomalous subsequences
- Based on comparing the DScores of the current state to the DScores of subsequence in the base state

Experiments

Forecast | Five Point Prediction



Objective: Predict the number of cases for rare event A for five future weeks based on simulated data

Training Data: **Simulated** cases of rare event A for 156 weeks with parameters from a previously learnt Poisson distribution
Note that, for most weeks, the net number of cases for event A is zero. Hence, the originating time series is highly sparse.

No. of Instances

Week	Actual	Predicted
T+1	3	3
T+2	5	4
T+3	1	1
T+4	6	7
T+5	8	10
Test MAPE	17.3%	

Global Tech India

Pattern Deviation | Case Specific Interpretation

Test Time Series: No. of instances of event A + median word embedding originating from simulations of A

Objective: Detect weeks undergoing significant deviation

The interpretation of Pattern Deviation is highly case specific, since it is both qualitative and quantitative in nature.

For our experiment, we started with 312 weeks' worth of data to detect pattern deviations and the algorithm was able to flag 10 Weeks (3.2%) as anomalous from the huge dataset.
The universal benefit lies in the data reduction, since the end user now needs to deep dive into only 3.2% of the whole data as opposed to the entire bulk.

For our purpose, we were also able to detect a mutating nature of event A w.r.t. its relevant sister classes.

Flagged Weeks

Week	DScore
T-5	1.47
T-29	1.47
T- 54	1.46
T-93	1.46
T-135	1.25
T-161	1.48
T-165	1.31
T-220	1.26
T-265	1.24
T-299	1.34
$D* = 1.21$	

Ritish Menon
Sr. Data Science Manager, Walmart Global Tech India
ritish.menon@walmart.com
Link To LinkedIn

Debanjana Banerjee
Data Scientist, Walmart Global Tech India
debanjana.banerjee@walmart.com
Link To LinkedIn

Thank You!

CRESST

Global Tech India