Outlier detection

From statistical analysis to more advanced ML/DL approaches

Raphael Espanha
Data Scientist

October 15th, 2019

Know the unknown..

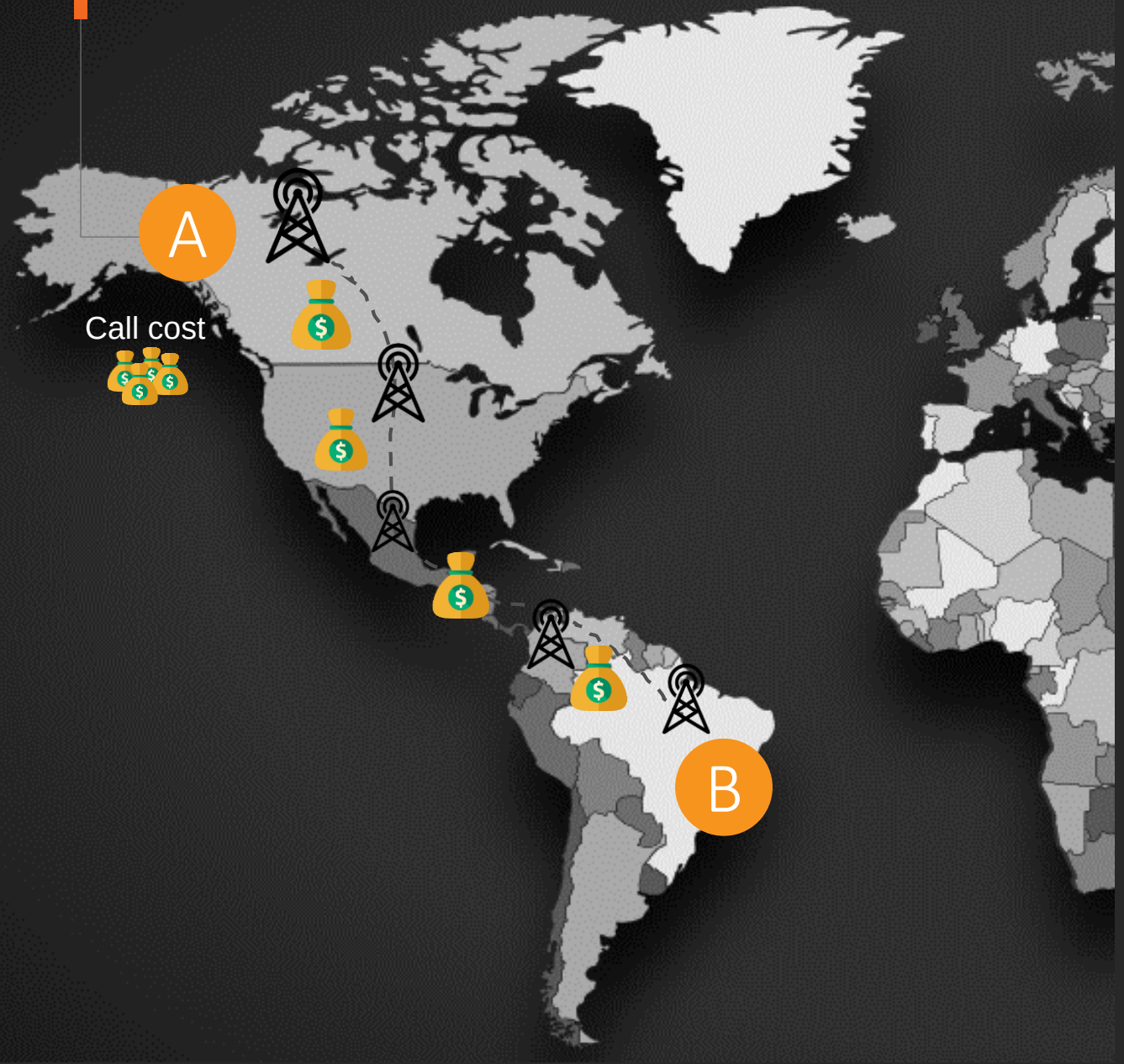
DSPT
DATA SCIENCE PORTUGAL

weDo technologies
A MOBILEUM Company

BYPASS

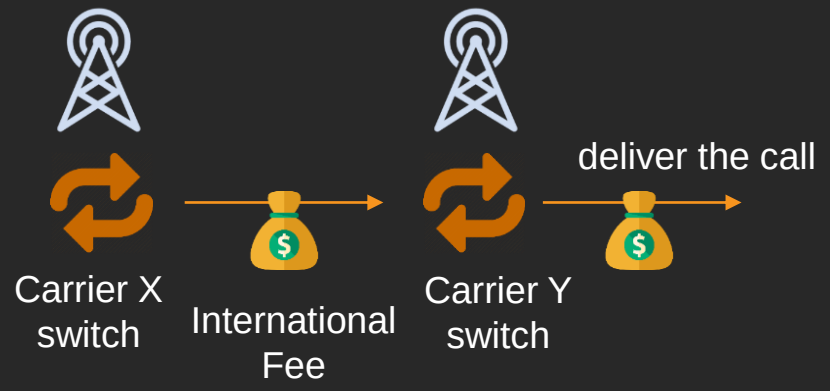
MTR USE CASE

International Call



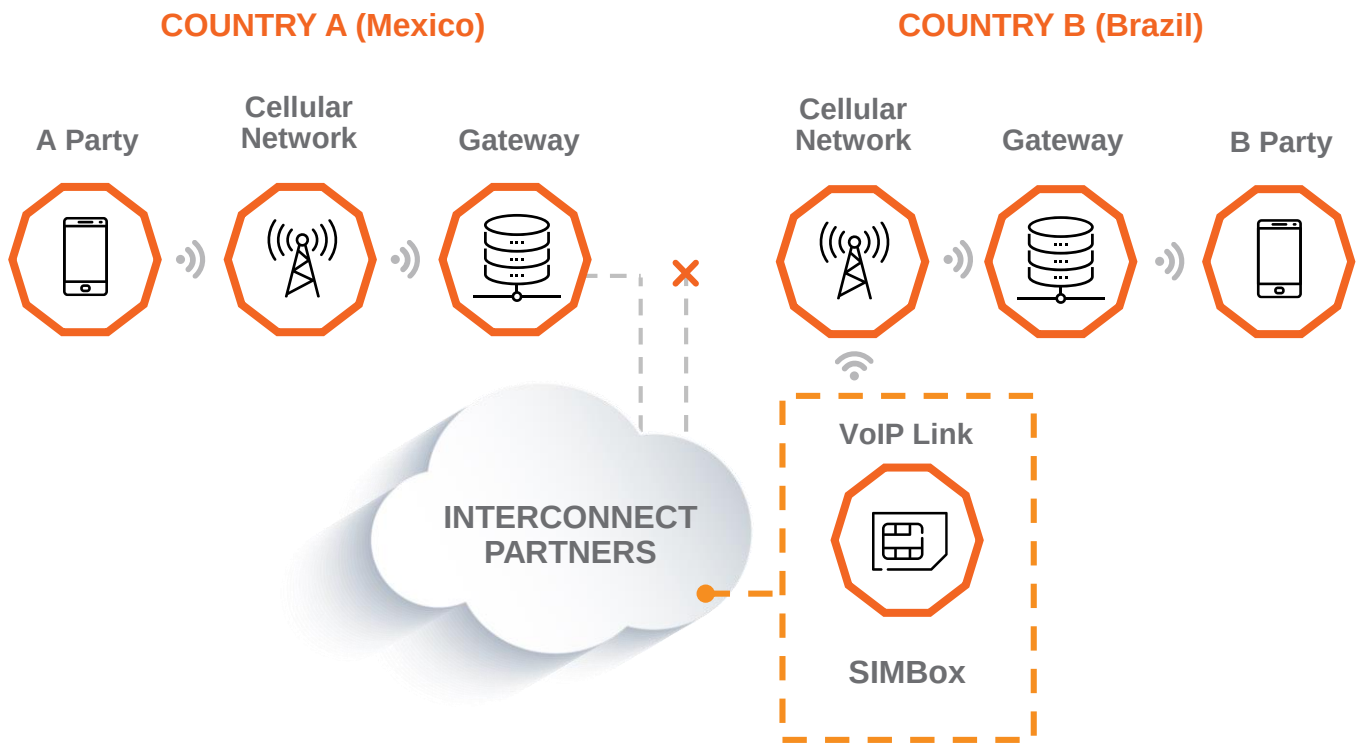
Phone companies have agreements with each other and use the services of carriers

By paying interconnection fees



weDO technologies
A MOBILEUM Company

Bypass



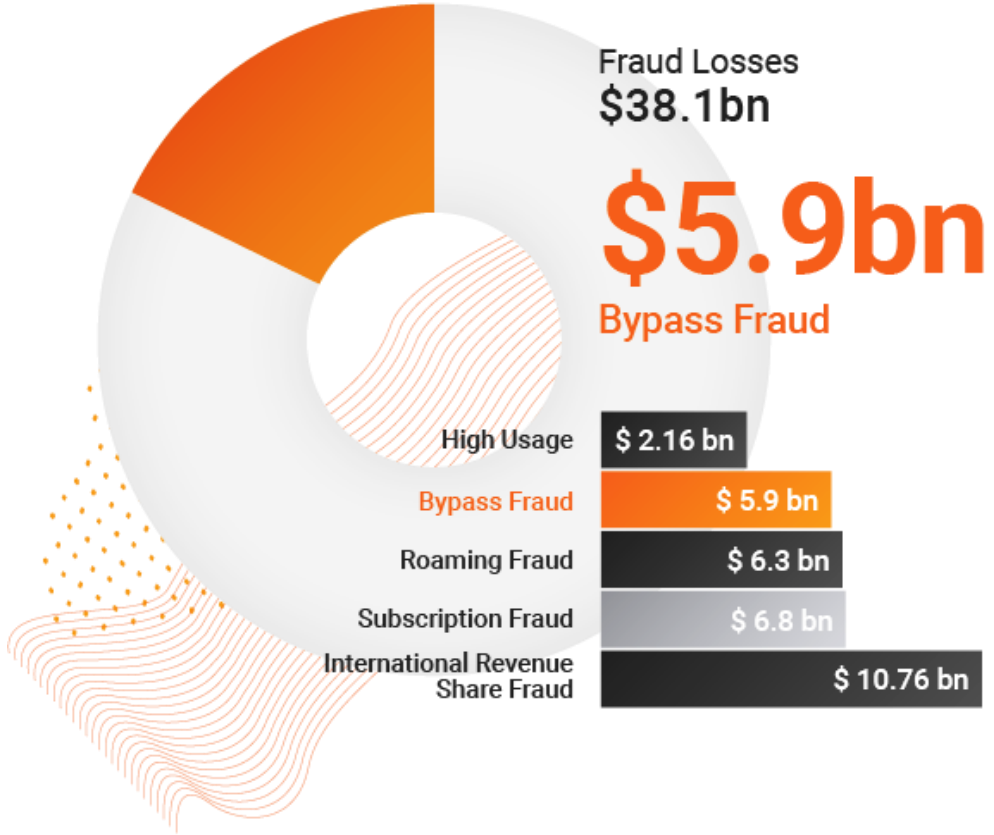
Fraudster profit = -

International Fee to forward the call

Local call cost

weDO technologies
A MOBILEUM Company

Bypass



Financial Impact

- Loss of revenue for legitimate operators

Image Impact

- No call completion
- Odd or no calling number
- Bad call quality

weDO technologies
A MOBILEUM Company

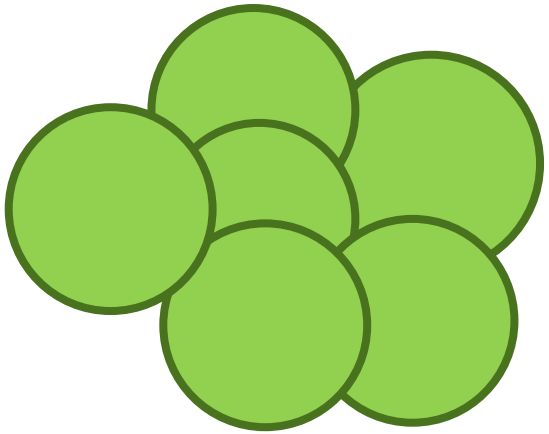
GOAL

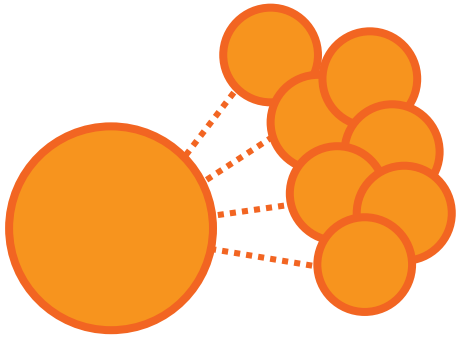
Goal – Detect abnormal ranges

a_number
ABCDE
ABCDF
ABCDG

range_1
ABCD

Normal ranges



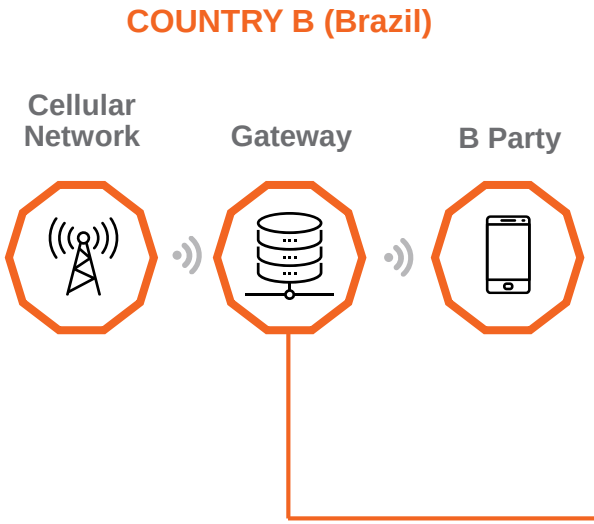


A numbers

Range with abnormal behavior

DATA

Data – Signaling events



Signaling data from august 2018

a_number	b_number	time
ZNFNNUULP	BNFVZNOPOIUFP	20180806211322
ZIIIPZIUZ	BNFVZNFROIRFO	20180816191720
IIFLIZPUOFL	BNFVZNLFVNOIO	20180820154021
LLRRNRUIUZUI	BNFVZNNIUPZVP	20180814122533
ZNFZUPPPR	BURVDUZNZUUZFOLIFNFP	20180815204314
VNNVVFRLZ	BNFVZNOLOUVNV	20180806160006
IOVFFNVOOLNF	BURVZNNIUFRUV	20180804105707
LLRIZIILIPFR	BNFVZNNIUZIUO	20180814012423
ZNUPOZROO	BNFVZNNIUFOPL	20180813171510
ZNOZNULOZ	BNFVZNVOLLLNO	20180815141804

a_number - origin number
b_number - destiny number
time: call timestamp

weDO technologies
A MOBILEUM Company

Data Cleansing

Client requested filters were applied:

- A numbers from starting with L or I
- A number length > 9
- Null values

a_number	b_number	time
ZNFNNUULP	BNFVZNOPOIUFP	20180806211322
ZIIIPZIUZ	BNFVZNFROIRFO	20180816191720
IIFLIZPUOFL	BNFVZNLFVNOIO	20180820154021
LLRRNRUIUZUI	BNFVZNNIUPZVP	20180814122533
ZNFZUPPPR	BURVDUZNZUUZFOLIFNFP	20180815204314
VNNVVFRLZ	BNFVZNOLOUVNV	20180806160006
IOVFFNVOOLNF	BURVZNNIUFRUV	20180804105707
LLRIZIILIPFR	BNFVZNNIUZIUO	20180814012423
ZNUPOZROO	BNFVZNNIUFOPL	20180813171510
ZNOZNULOZ	BNFVZNVOLLLNO	20180815141804

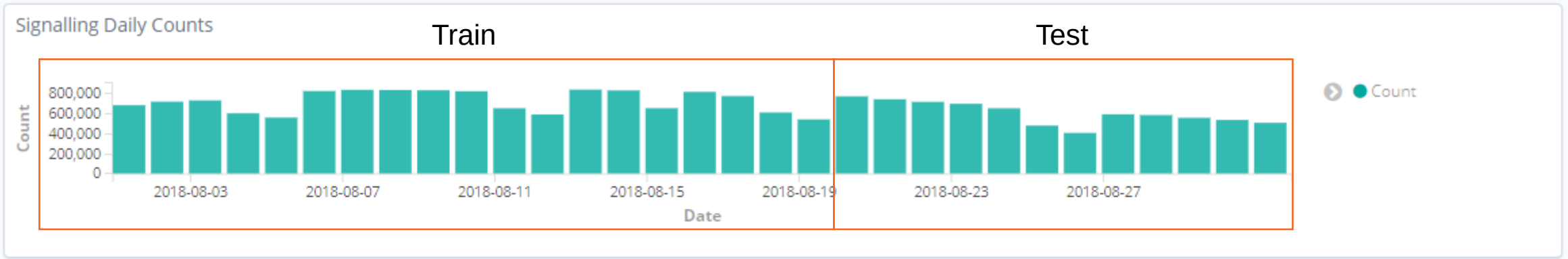
33 929 956 events

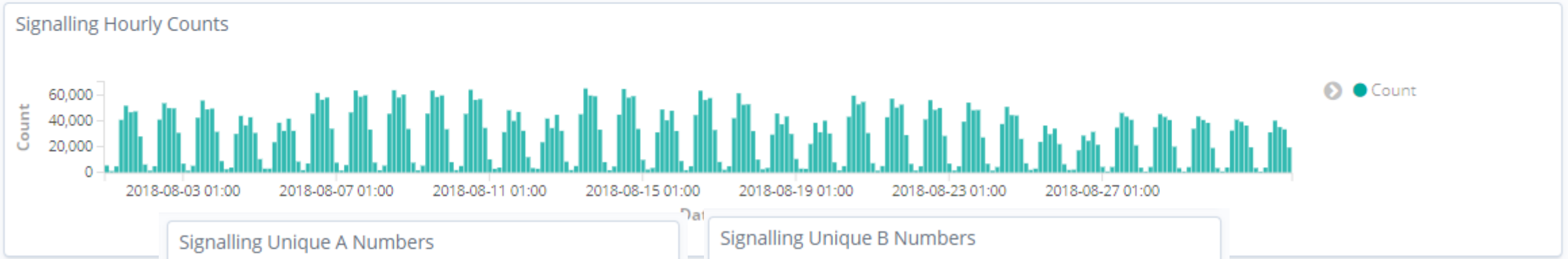
a_number	b_number	time
IIFVNZFFNZR	BNFVZNNIUZIZL	20180821201213
IOVFZNPZZLUZ	BURVZNNIUFUVR	20180805114643
ILFNFVNPURP	BNFVZNNIUOILR	20180824190610
LLRPILRNPLUU	BNFVZNNIUZRNZ	20180820130935
IOVFVNVRPZLU	BURVZNNIUFIPU	20180805202844
LVUFUOLNRFLZ	BNFVZNNIULIVI	20180802125215
LURFVPIVLVN	BURVZVULLUPRU	20180815202426
IPUZZUOIIRVP	BURVZVUIIVPNU	20180803114237
IIFNUNRLPZV	BNFVZNNIURNOO	20180806085534
IIFUIPPNVNU	BNFVZNNIURLIU	20180816120950

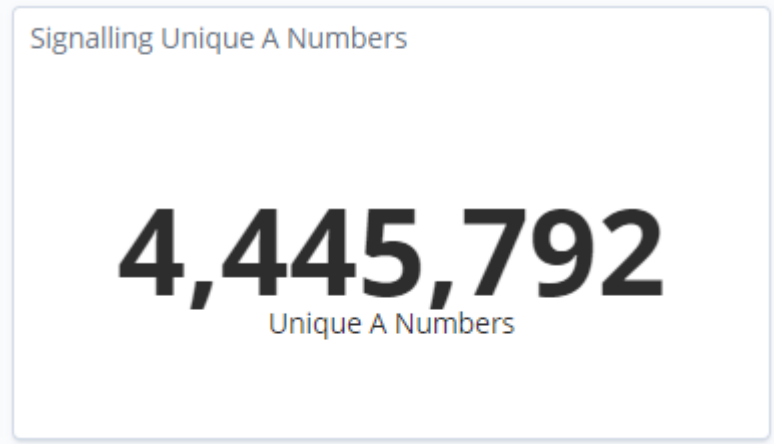
20 636 037 events

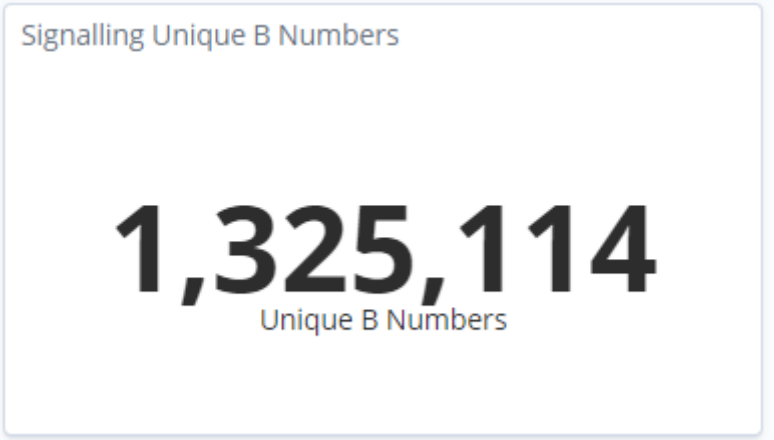
weDo technologies
A MOBILEUM Company

Data Analysis – Time and distinct values







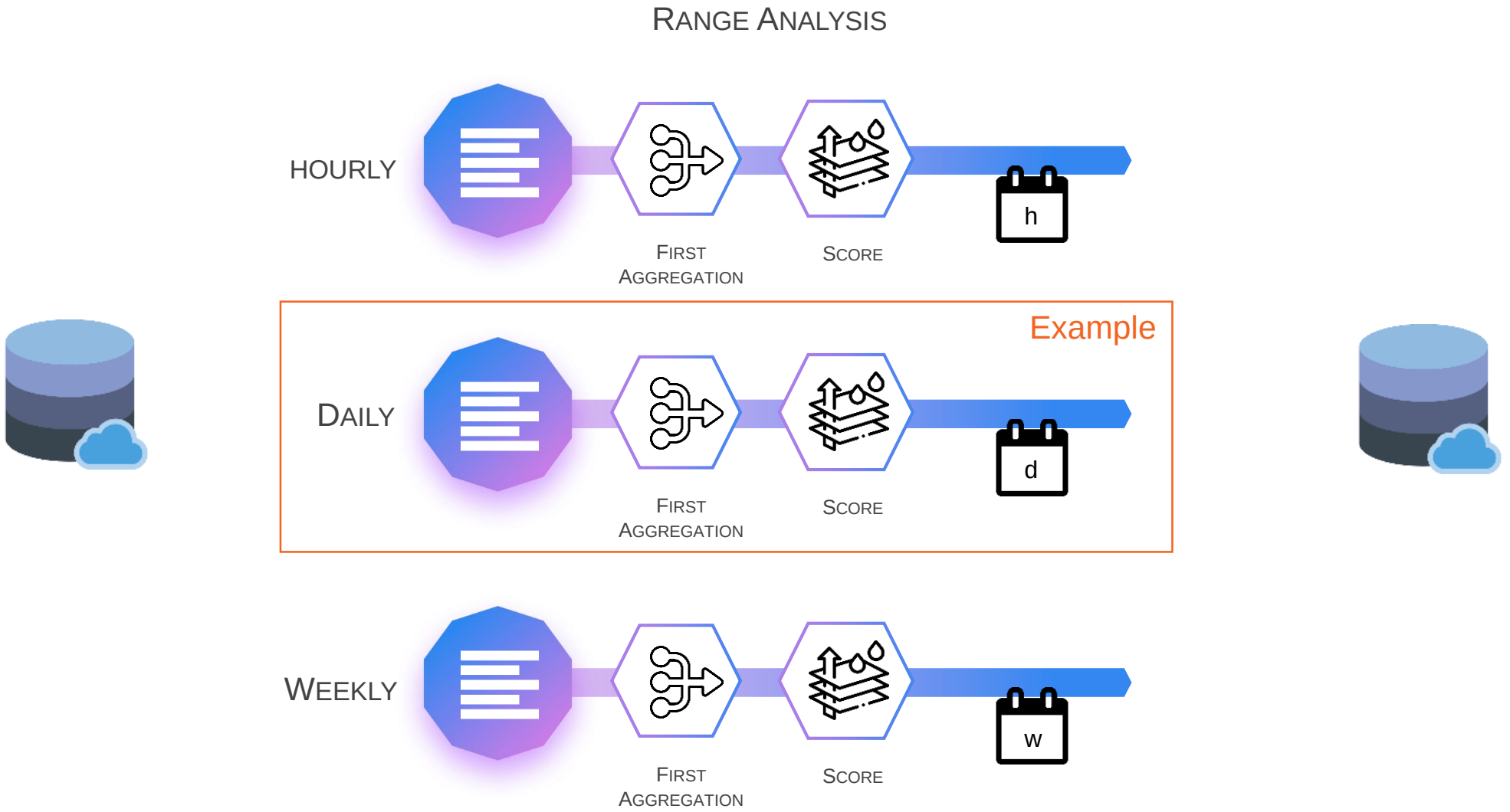


wedo technologies
A MOBILEUM Company

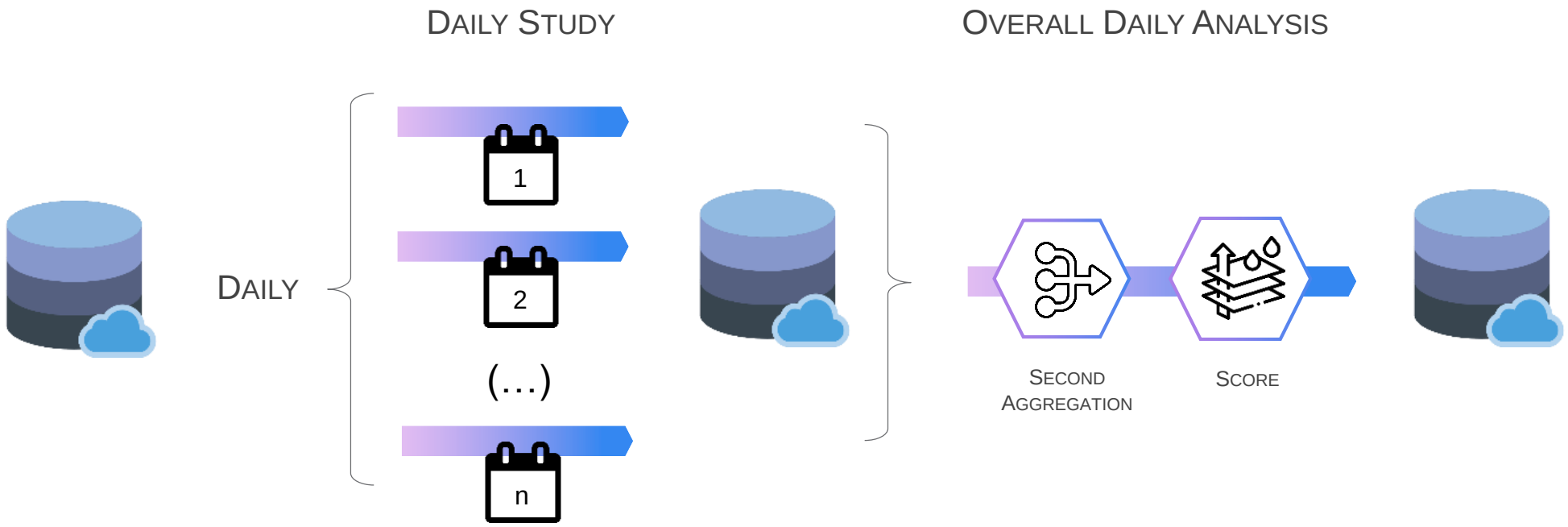
Outlier Detection
Range Numbers

Statistical approach

Statistical Approach – Multiple time context analysis



Statistical Approach – Daily analysis overview

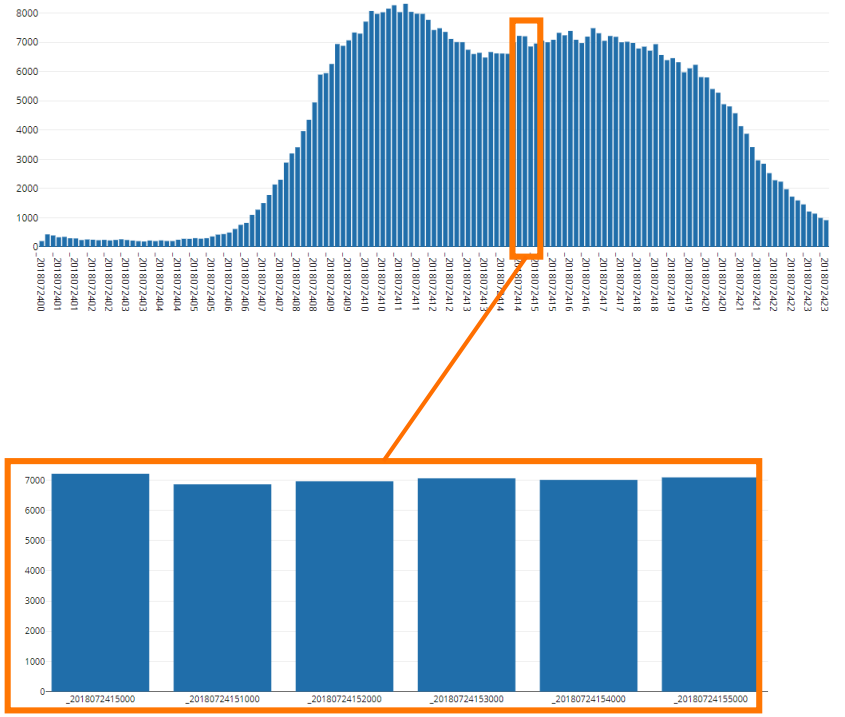


Statistical Approach – Preprocessing - Time binning

Creating intervals of 10 min

10:43	10:40
10:45	10:40
10:47	10:40
10:51	10:50
10:55	10:50
10:59	10:50

time	interval_bin
20180819160732	201808191600
20180806114005	201808061140
20180824092934	201808240920
20180822155336	201808221550
20180807131051	201808071310
20180802085920	201808020850
20180820100427	201808201000
20180825164352	201808251640
20180812123831	201808121230
20180815173523	201808151730
20180824152437	201808241520



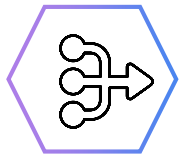
Equal-width Binning

wedo technologies
A MOBILEUM Company

Statistical Approach - First aggregation (daily analysis example)

Group by a_number

a_number	b_number	time	interval_bin
LLRPPNFVOZPL	BNFVZNNIUFNFP	20180801143033	201808011430
IIFOUVZLLUO	BNFVZNNIULZUI	20180801182655	201808011820
IIFRRONLVNU	BNFVZNNIUORRR	20180801193622	201808011930
LLROILIOPRIU	BURVZNLRZNIZO	20180801133657	201808011330
INFIINRFOOU	BNFVZNNIUZUPZ	20180801174707	201808011740
LZLNRNRZFU	BNFVZNNIUOZVO	20180801115318	201808011150
LLROFPNLIZOO	BNFVZNNIUPLZU	20180801190527	201808011900
IIFVRIFOUNF	BNFVZNNIUFFVZ	20180801210050	201808012100
ILFUPUZORFR	BURVZNIPIONZV	20180801101755	201808011010
IIRFZZUFZNV	BNFVZNZIPPIVR	20180801123153	201808011230
IVLZNVIPVFF	BNFVZNNIUPVOV	20180801124409	201808011240
IIFNFVOIZZO	BURVZNFZFVORU	20180801201836	201808012010
IIFUZZPPUFP	BNFVZNNIUOLRU	20180801171554	201808011710

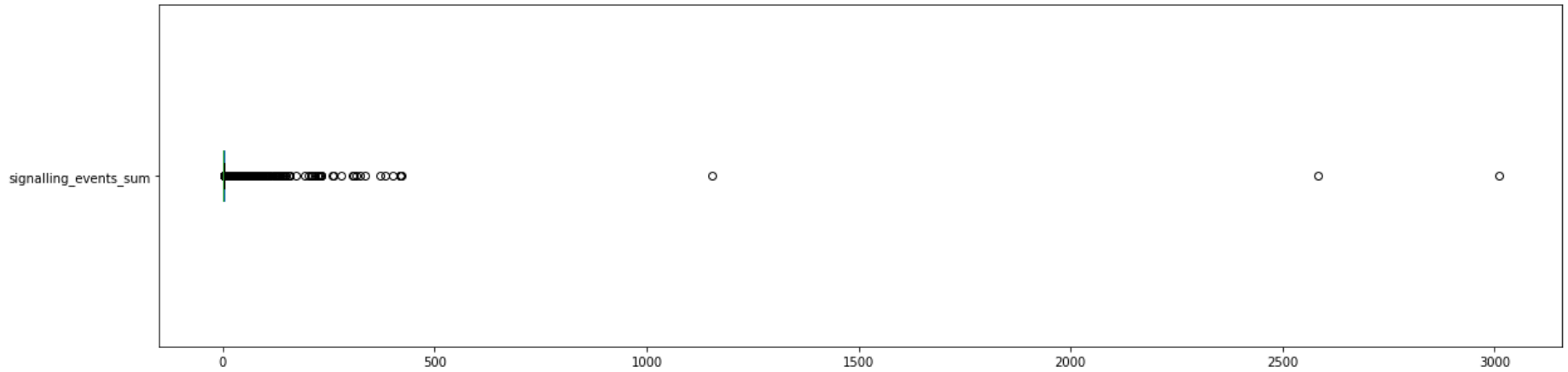


a_number	distinct_b_numbers	distinct_interval_bin	signalling_events
IIFZOVRLUOI	1	1	1
IIFNOOZPZRN	1	1	1
LLVURNFFZUPI	1	1	1
LFRIURNFPIV	1	1	1
LLRPFRINZUVU	1	1	1
IIFVVZIOLZF	1	1	1
IIFFNNUVIIP	4	4	4
IZULVVPIVPZU	1	1	1
LLVURPIINZII	1	1	1
ILFRZNRPRZR	1	1	1
LLRLZILUPZFP	2	1	2
IIFNZIVIVVN	1	1	1
LRZORIULII	1	1	1

wedo technologies
A MOBILEUM Company

Statistical Approach – dropping low event ranges

Assuming that a numbers and ranges with very low daily events are not fraudsters



Using Tukey's fences

$$\underbrace{[Q_1 - k(Q_3 - Q_1)}_{\varnothing}, \underbrace{Q_3 + k(Q_3 - Q_1)]}_{5\ (k=3)}$$

Removing a_numbers with less than 5 events in one day:

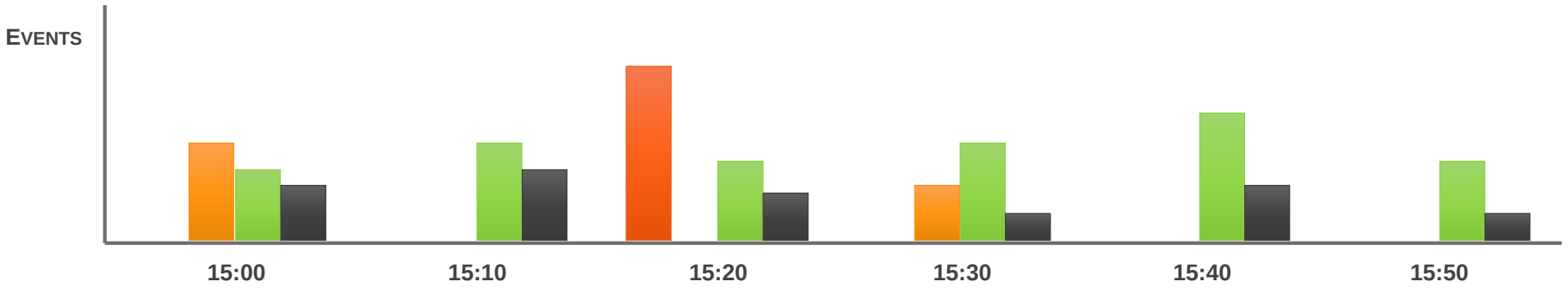
✓ Greatly improves algorithm performance

328 368 ranges to 15 490 (~4%)

wedo technologies
A MOBILEUM Company

Statistical Approach – Event's intensity

a_number	distinct_b_numbers	distinct_interval_bins	signalling_events	event_intensity_score
IFIULVNPLOR	16	1	20	0.9328
IFRUIRFUVIL	22	2	30	0.9036
IFRUZURNUIV	86	6	100	0.3165
LZVNNPPVFVFNZ	32	6	39	0.2633



weDO technologies
A MOBILEUM Company

Statistical Approach – Coefficient of Variation (CV)

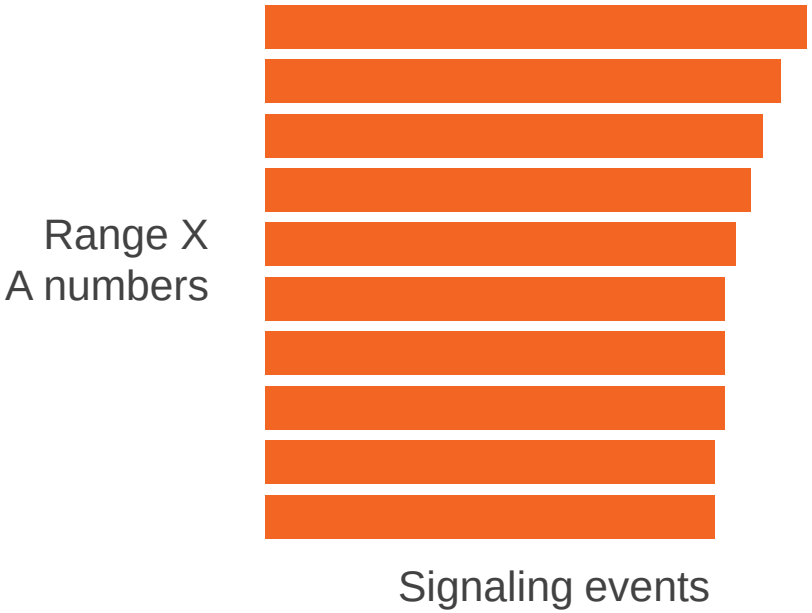
Also known as relative standard deviation (RSD)

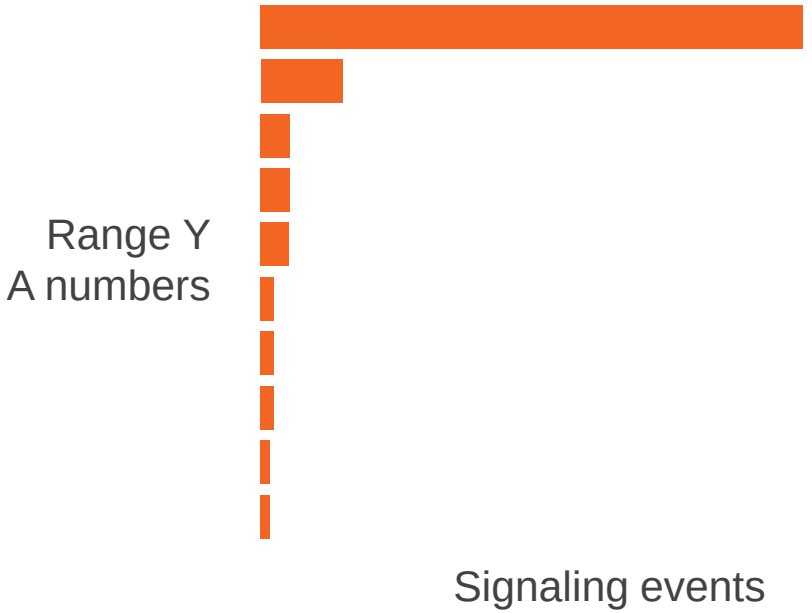
Standardized measure of dispersion of a **frequency distribution**

$$c_v = \frac{\sigma}{\mu}$$

Requires group by range first

Compute **a rotation score** based on CV computed from events





Range X rotation score >> Range Y rotation score

weDo technologies
A MOBILEUM Company

Statistical Approach – Map back possible a numbers level up

Range Y
A numbers

Range Y should not be marked as a potential fraudulent range

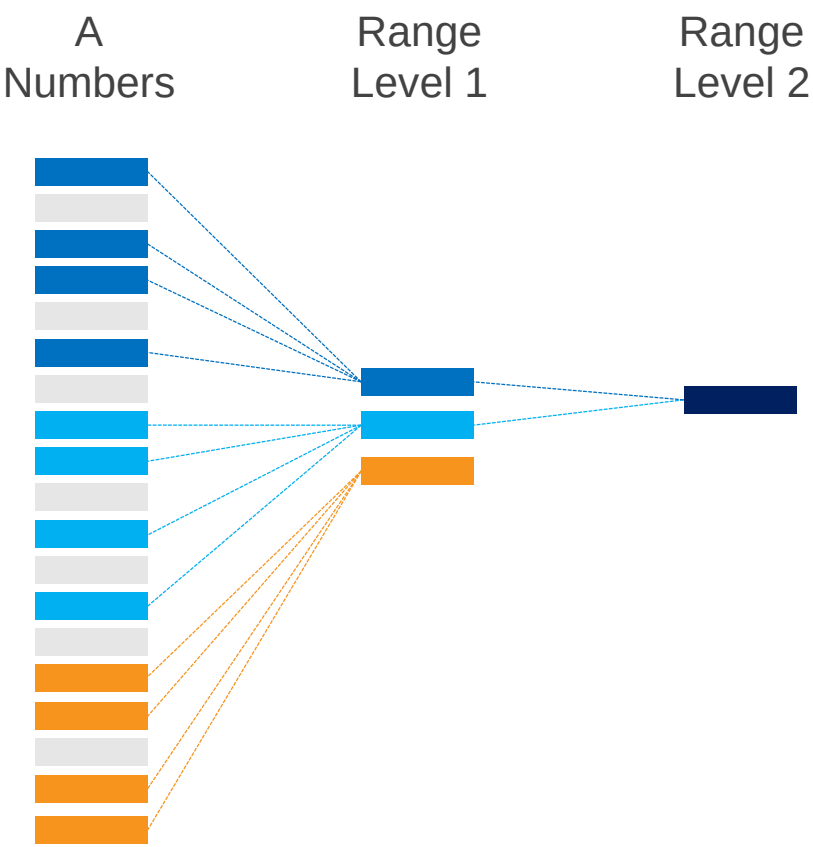
Possible False Positives

However, its first a number may have a fraudster potential

Range X
A numbers

SIMBox

Using a SIMBox with a sequence of SIM cards can result in a high a number rotation



weDo technologies
A MOBILEUM Company

Statistical Approach – Overall Analysis and Profiles

Rotation score

Hour consistency score

Event intensity score

B dispersion score

A/B ratio score

A number
level 1 {0,1,2}
High profile score
Rotator profile score

Day
20181201
analysis

Daily Analysis
Finished

Overall Analysis

Day 20181201
analysis

(…)

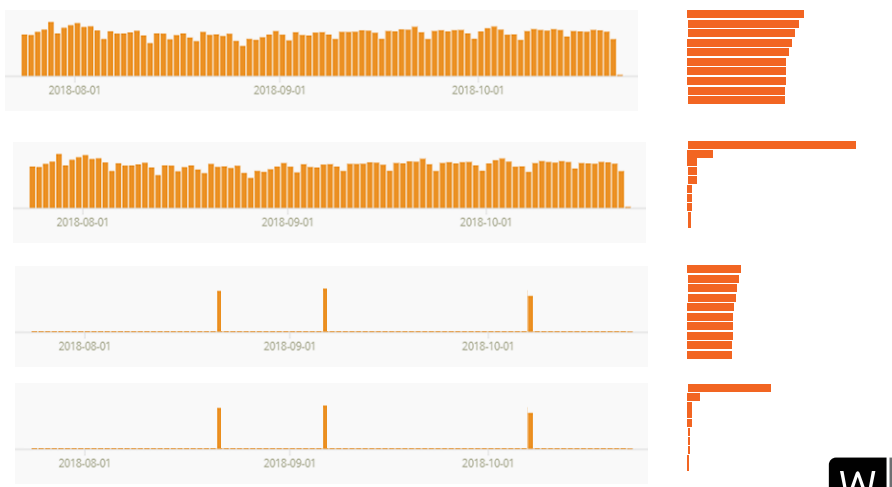
Day X
analysis

Global Rotators

Global High

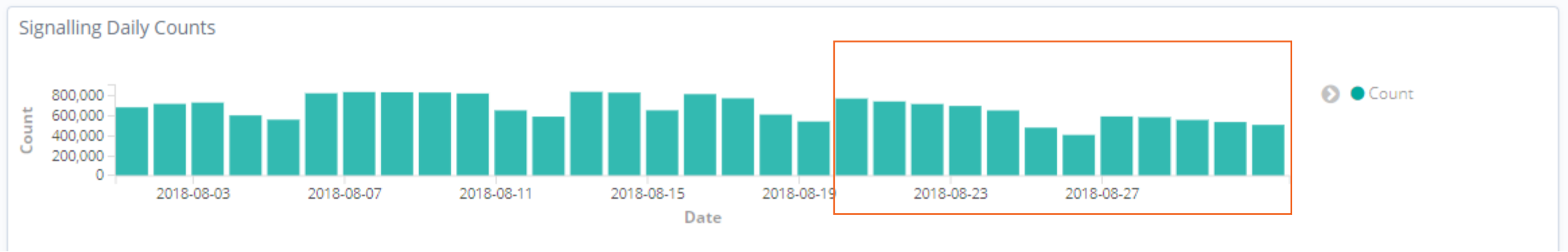
Local Rotators

Local High



wedo technologies
A MOBILEUM Company

Statistical Approach – Results

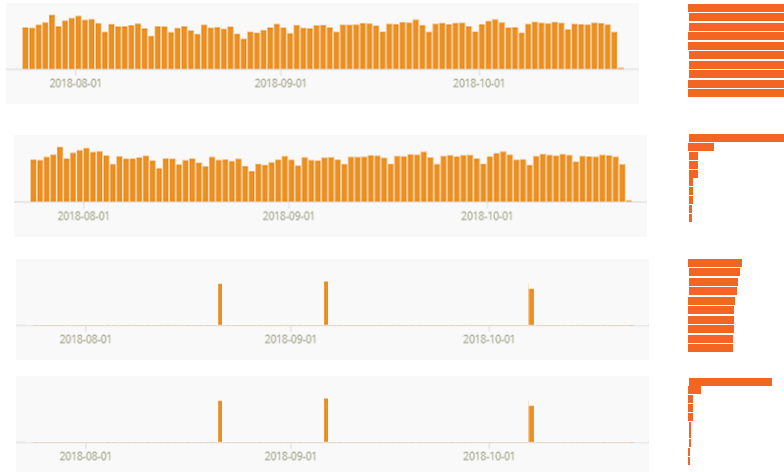


Global Rotators

Global High

Local Rotators

Local High



Tukey's fences
K=3

383 cases

19 / 21 known bypass (90%)

0 / 52 known FP (0%)

365 unclassified

weDo technologies
A MOBILEUM Company

Outlier Detection
Range Numbers

Machine Learning - PCA

Principal Component Analysis

Similar data preparation steps

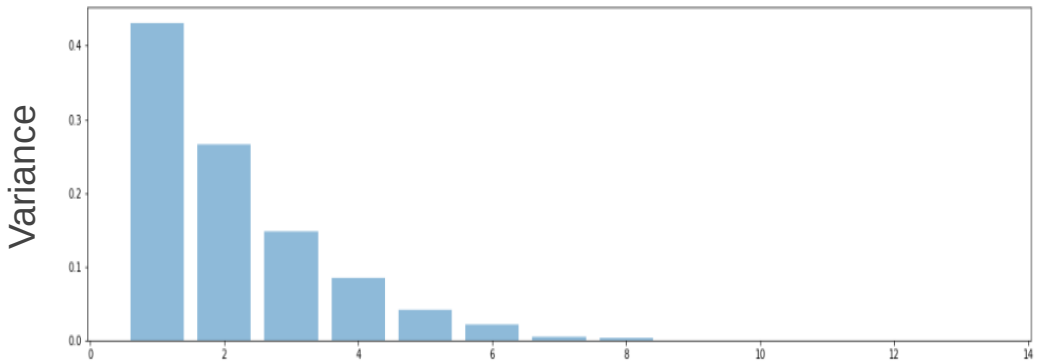
- Range
- Range Level
- Signaling Events CV
- AB Ratio
- Rotation Score
- Interval consistency
- (…)

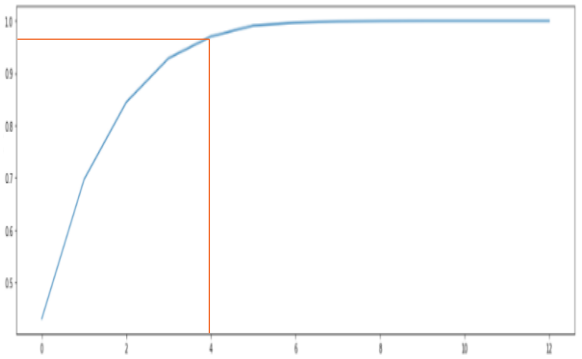
standardized scale (z)

$$z = \frac{x - \mu}{\sigma}$$

Principal Component Analysis is a **linear transformation** of data that reduces its dimension into **Principal Components** (PCs)

Principal Components are new variables that are constructed as linear combinations of the initial variable

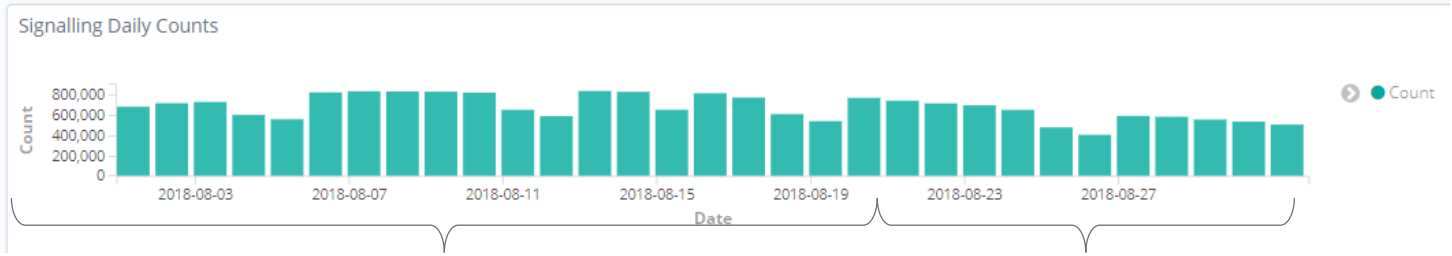




PCs	Retained Information (%)
1	43
2	70
3	84
4	93

wedo technologies
A MOBILEUM Company

Principal Component Analysis - Results



Normal samples → Fit

Predict / transform →

Sample X input − Weighted eigenvector = Error vector

mean

Error Value

Eigenvectors
1 x dimension size

1 2 3 4

Eigenvalues λ1 λ2 λ3 λ4

Σ

Weighted eigenvector

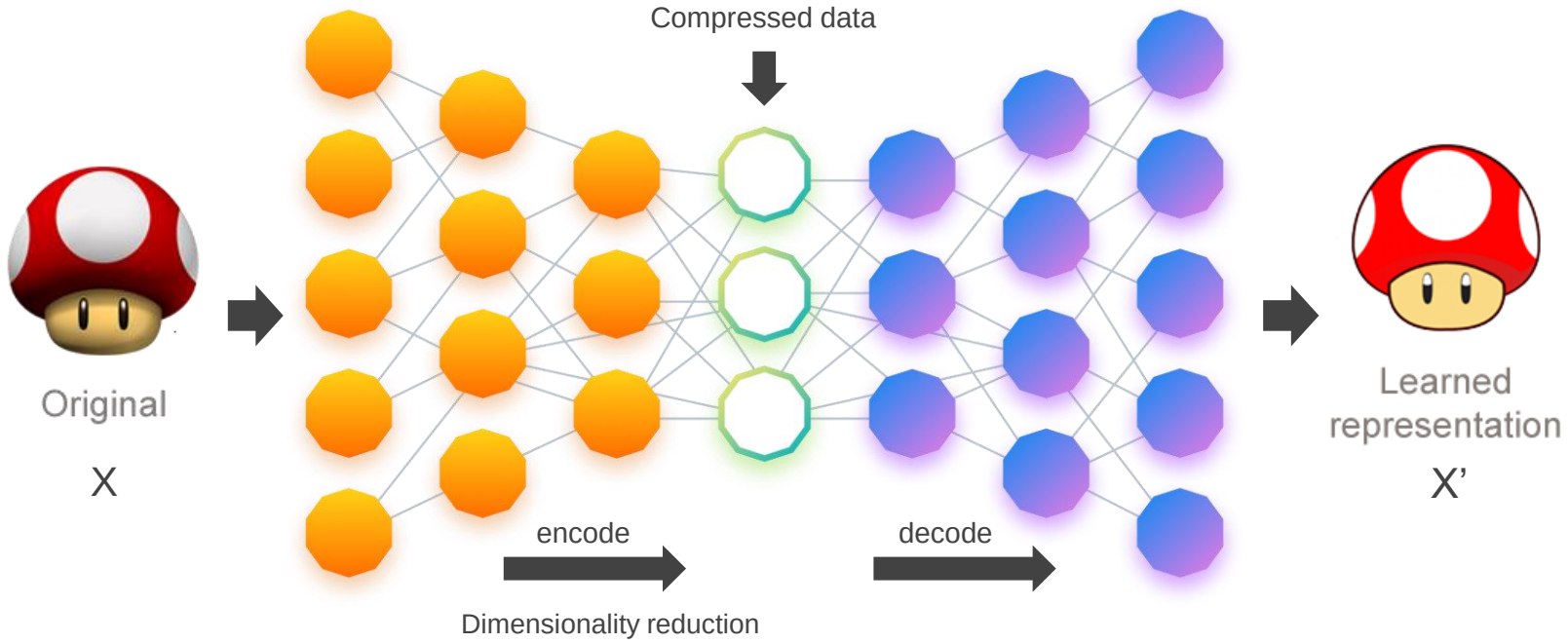
Tukey (k=3) 62 713 cases	Q (0.99) 2568 cases
17 / 21 known bypass (81%)	13 / 21 known bypass (62%)
2 / 52 known FP (2%)	0 / 52 known FP (0%)

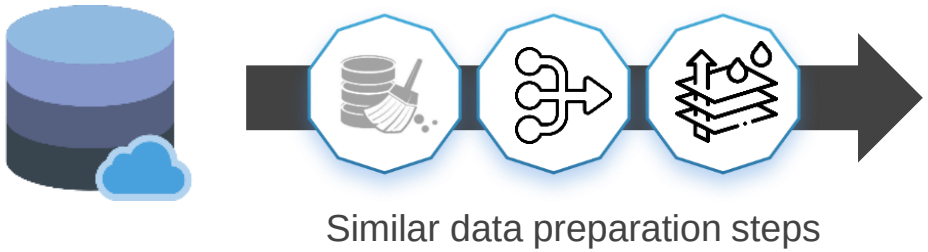
weDO technologies
A MOBILEUM Company

Outlier Detection
Range Numbers

Deep Learning - Autoencoders

Autoencoders





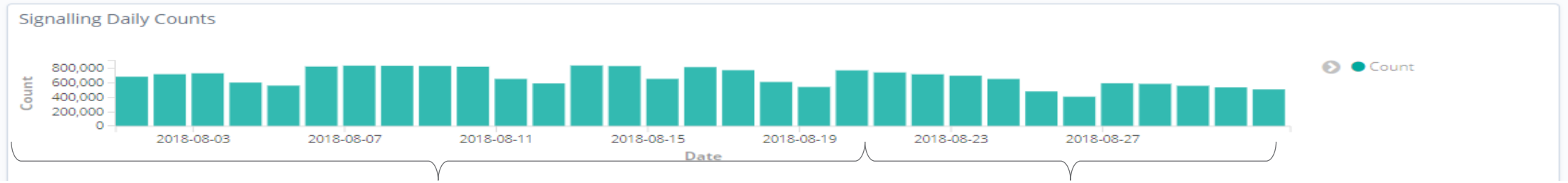
Similar data preparation steps

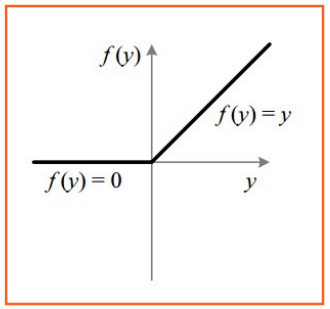
- Range
- Range Level
- Signaling Events CV
- AB Ratio
- Rotation Score
- Interval consistency
- (…)

standardized scale (z)

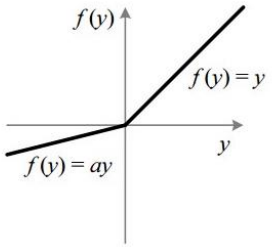
$$z = \frac{x - \mu}{\sigma}$$

Autoencoders - Results





ReLU



Leaky ReLU

Fit

13-8-2-8-13

Optimizer	Adam
Loss	MSE
Act. Function	ReLU
Epochs	50

Predict

MSE

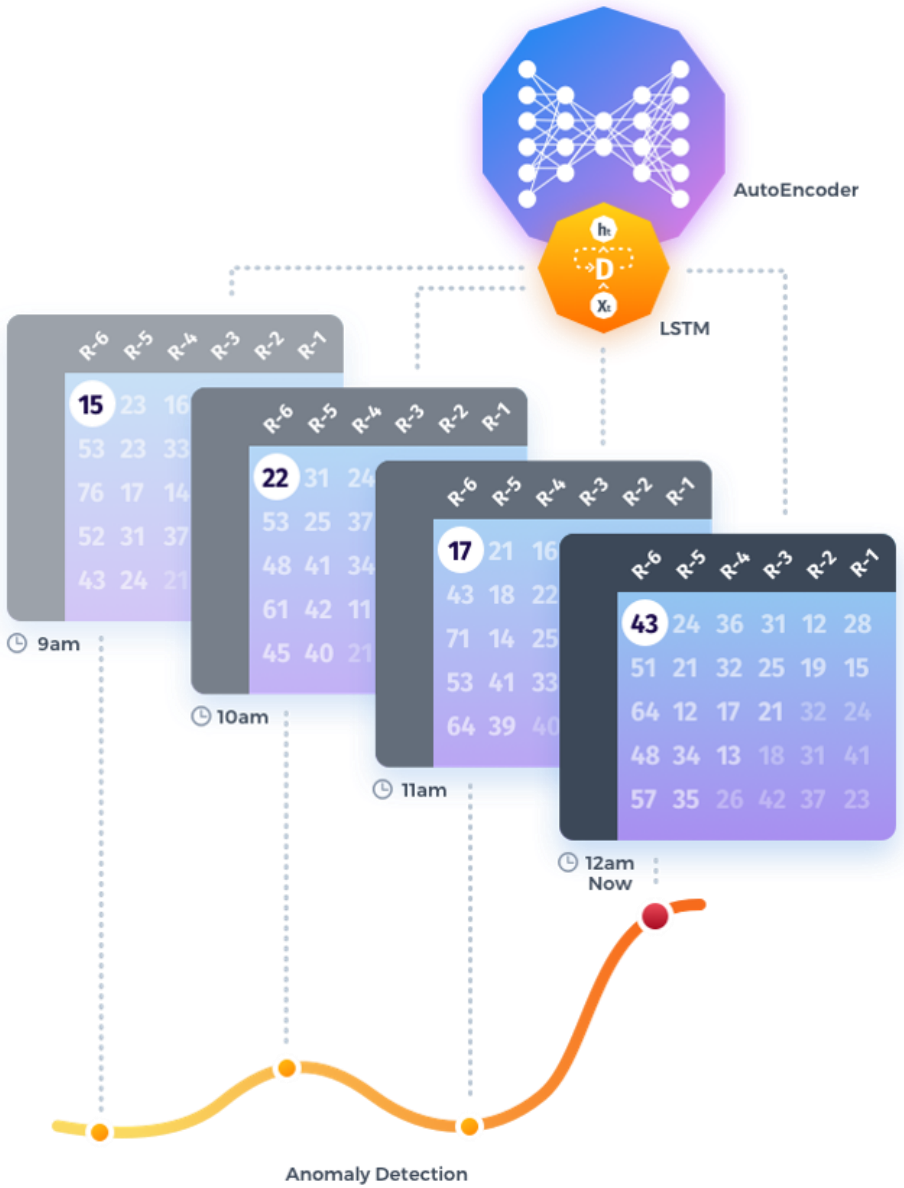
Tukey (k=3) 83 459 cases	Q(0.99) 6807 cases
20 / 21 known bypass (95%)	16 / 21 known bypass (76%)
2 / 52 known FP (2%)	0 / 52 known FP (0%)

weDo technologies
A MOBILEUM Company

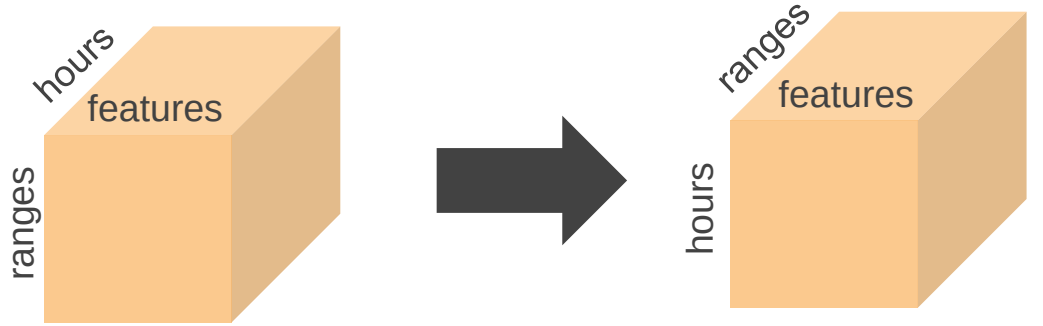
Outlier Detection
Range Numbers

Deep Learning – LSTM + AE

Stacked LSTM + Autoencoders

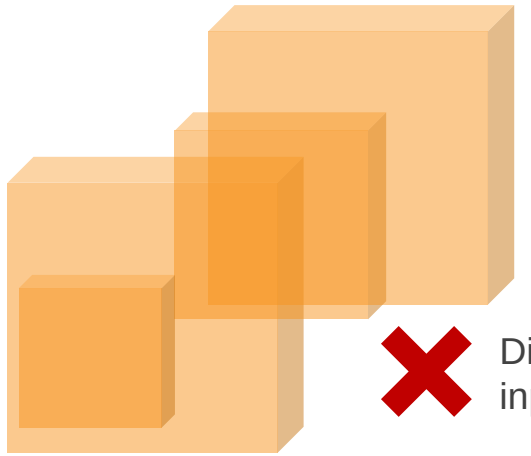


3D Approach



Sequence of hours
(scaled)

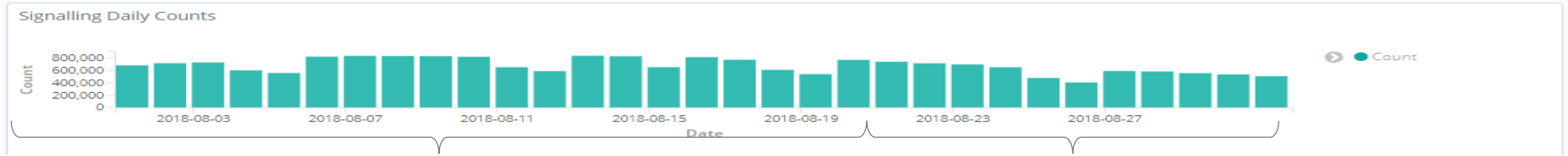
Sequence of ranges
(scaled)



Different
input shape

weDo technologies
A MOBILEUM Company

Stacked LSTM + Autoencoders - Results



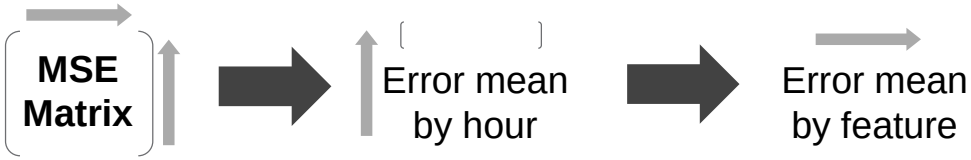
Fit

Example in Keras

```
inputs = Input(shape=(X.shape[1], X.shape[2]))
L1 = LSTM(16, activation='relu', return_sequences=True,
          kernel_regularizer=regularizers.l2(0.00))(inputs)
L2 = LSTM(4, activation='relu', return_sequences=False)(L1)
L3 = RepeatVector(X.shape[1])(L2)
L4 = LSTM(4, activation='relu', return_sequences=True)(L3)
L5 = LSTM(16, activation='relu', return_sequences=True)(L4)
output = TimeDistributed(Dense(X.shape[2]))(L5)
model = Model(inputs=inputs, outputs=output)
```

Optimizer	Adam
Loss	MSE
Act. Function	ReLU
Epochs	50

Predict



Tukey (k=3) 112 713 cases	Q(0.99) 10 235 cases
21 / 21 known bypass (100%)	20 / 21 known bypass (95%)
10 / 52 known FP (20%)	8 / 52 known FP (15%)

weDo technologies
A MOBILEUM Company

Conclusions

Conclusions

Statistical with scoring approach

- Fast (18 secs for 10 days)
- Good results with **low cases**
- **Keeps the history** of daily studies to help end user to understand what is going on
- Simple and **easy to explain**
- Overfitting can occur

PCA

- Fast
- Good results but only with a **high number of cases**
- Not easy to explain results

Autoencoders

- Can learn new feature representation
- Relatively simple to understand
- Hyper parameter tuning
- Good results but only with a **high number of cases**
- Not so easy to explain results (feature importance)

LSTM + Autoencoders

- Can learn new feature representation
- Sequence learning capability
- Good results require **very high number of cases (high FPs)**
- Very high computational cost (fit)
- Complex
- Hard to explain results

THANK YOU

raphael.espanha@wedotechnologies.com

Know the unknown...

Q & A