Adaptive Conformal Anomaly Detection for Time-series

Evgeny Burnaev (joint work with Alexander Bernstein, Vlad Ishimtsev and Ivan Nazarov)

Skoltech, Moscow, Russia,
IITP, Moscow, Russia

Outline

Change-point models

- Structure of the observed random process X_t:

$$X_1, X_2, \ldots, X_\theta, X_{\theta+1}, \ldots$$

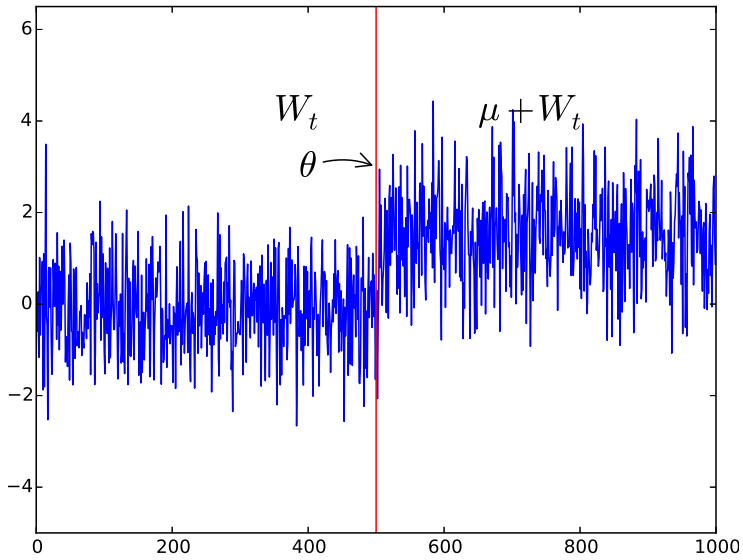
- θ: an unknown change-point time;
- X_t has the distribution $f_\infty(\cdot)$ before the change-point and $f_0(\cdot)$ after the change-point.

Specific example:

$$X_t = \begin{cases} W_t, & 0 \leqslant t < \theta\,, \\ \mu + W_t, & \theta \leqslant t \end{cases}$$

CUSUM statistics: $\gamma_0 = 0$,

$$\gamma_n = \max\left\{0, \gamma_{n-1} + \log \frac{f_0(X_t)}{f_\infty(X_t)}\right\}$$

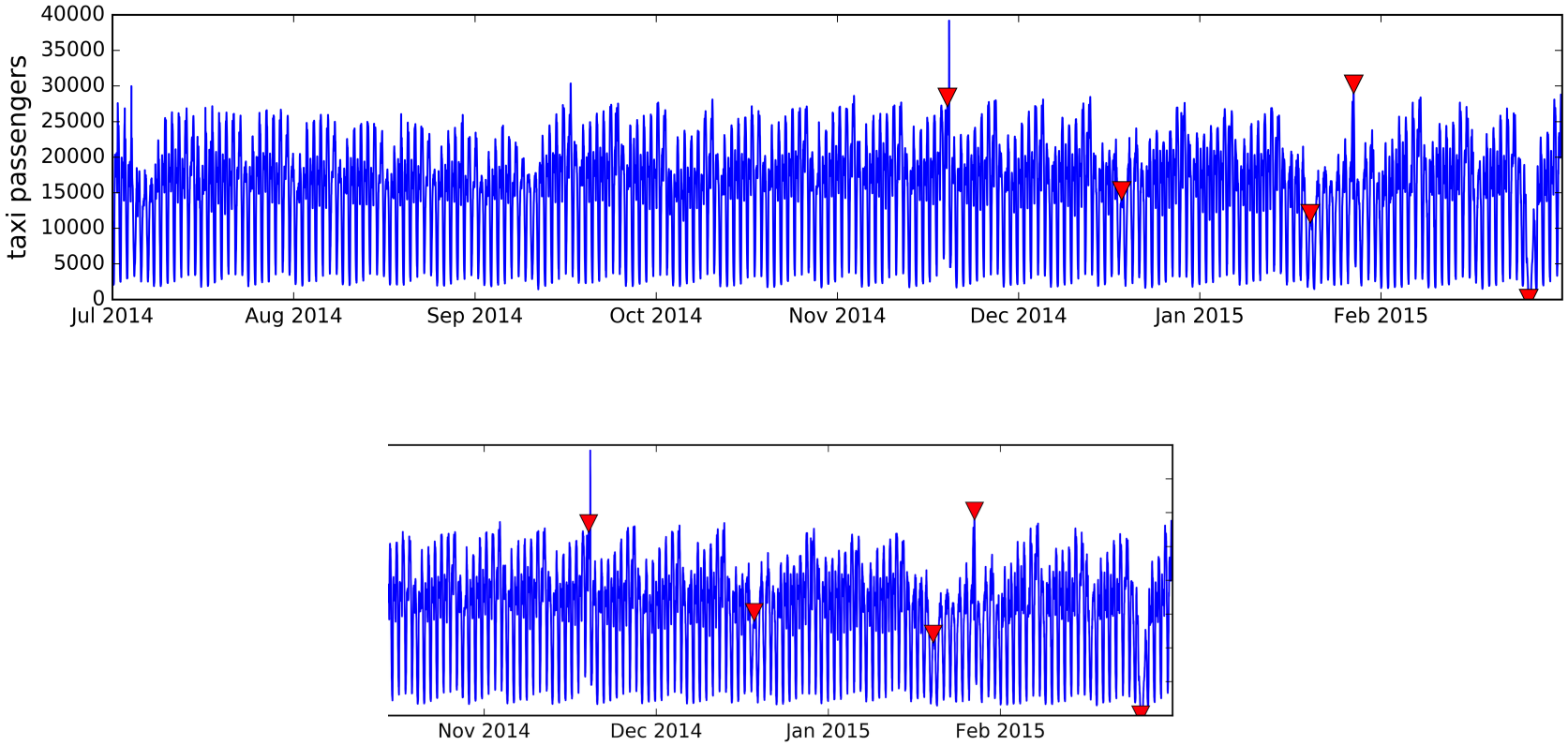


Anomaly Detection

- Important applications: healthcare, security, production, equipment maintenance, internet of things, traffic routing, etc.
- Challenges:
  - Large volumes of time-series data;
  - Complex, intractable or unknown dynamic models;
  - Apparent non-stationary and quasi-periodicity

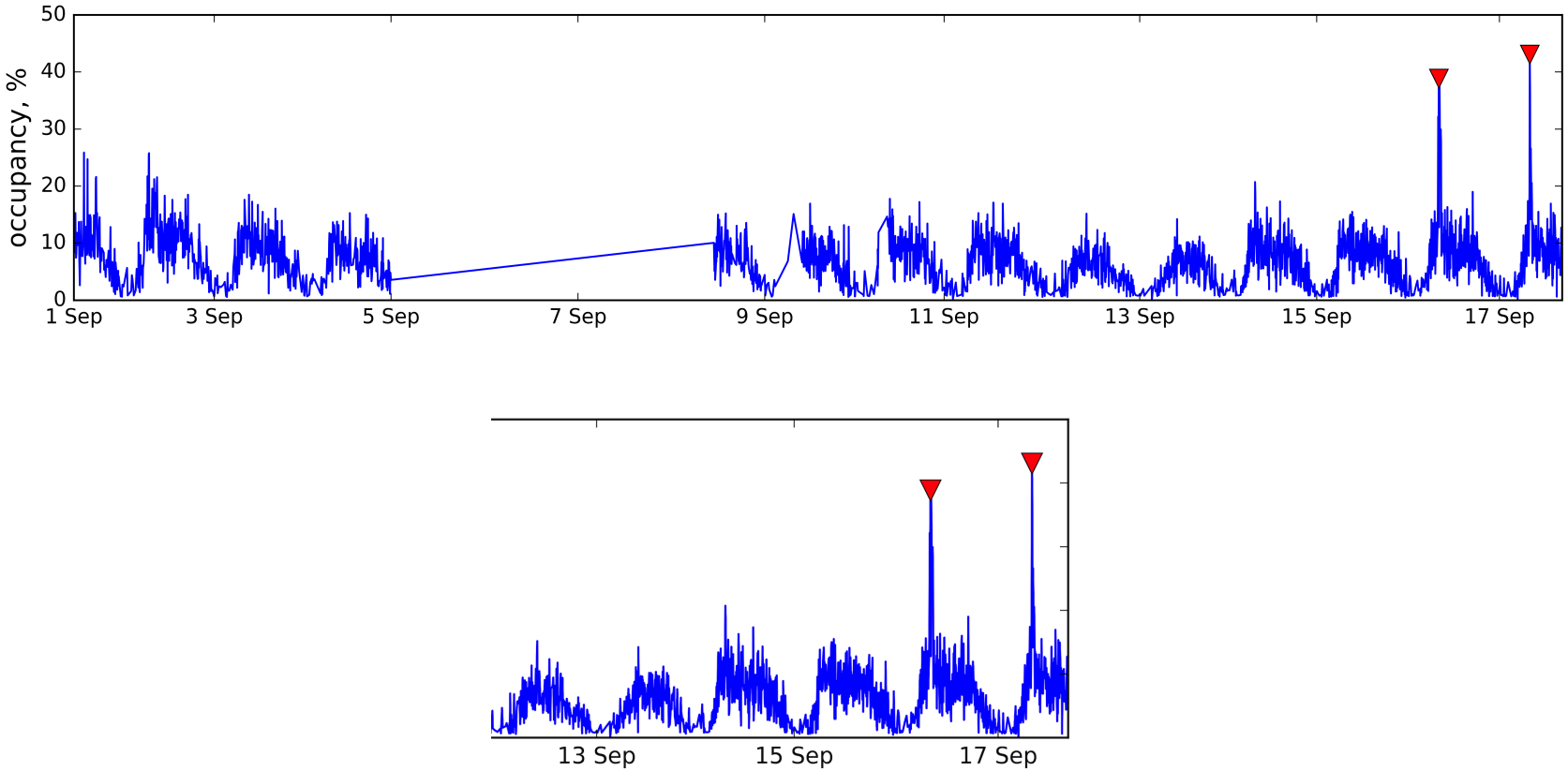
Automatic anomaly detection is critical in today's world where the sheer volume of data makes it impossible to tag outliers manually

Example: NAB benchmark



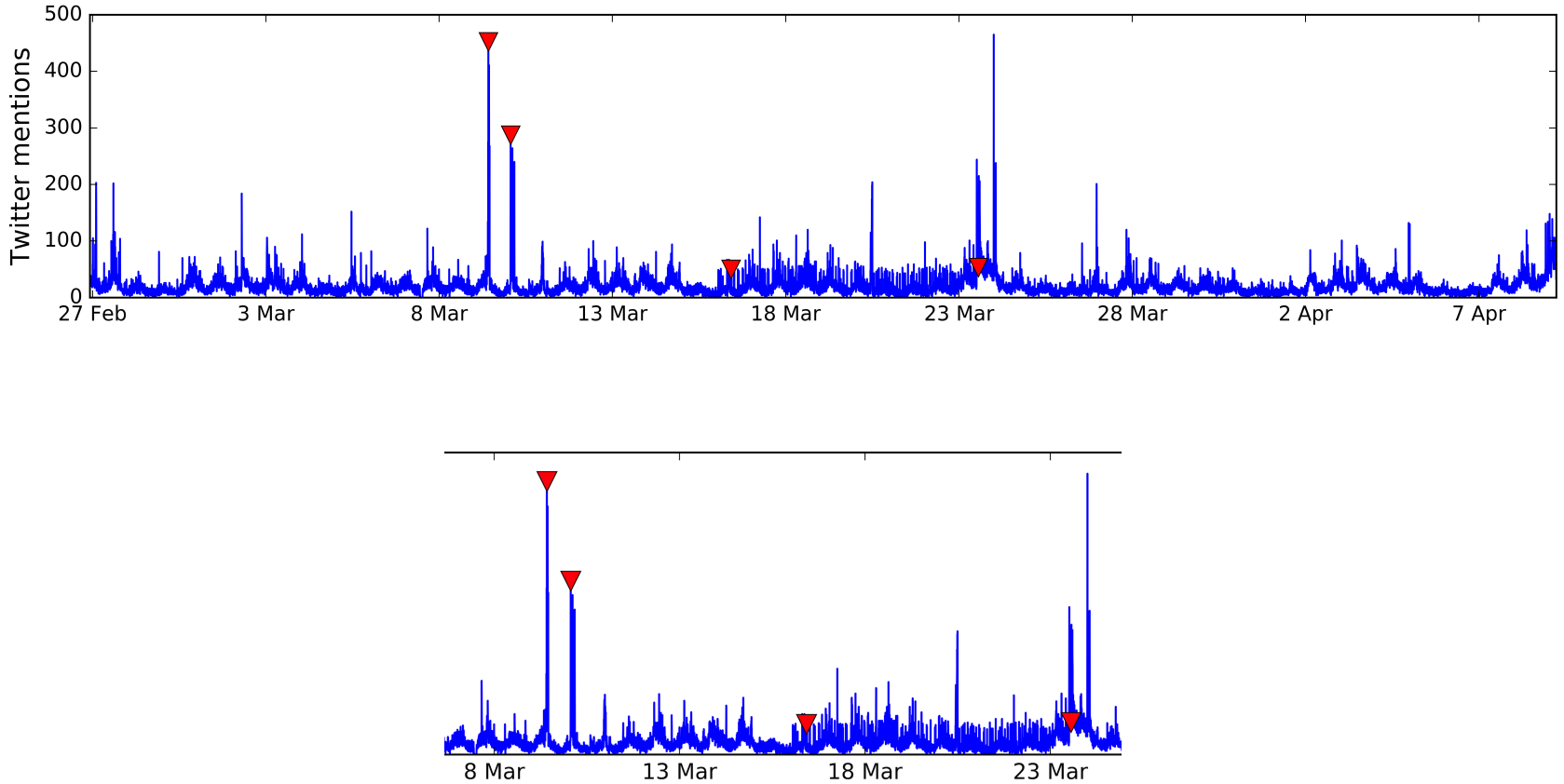
NYC taxi passengers (aggregated into 30-min buckets), July 2014 — March 2015. Anomalies – NYC marathon, Thanksgiving day, Christmas, New Years day, and a snow storm. Source: NYC Taxi and Limousine Commission

Example: NAB benchmark



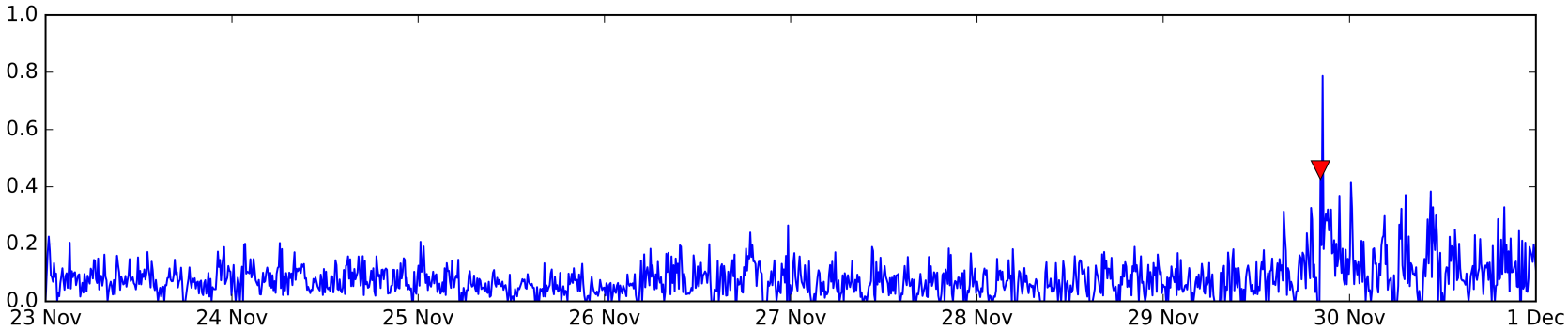
Traffic data from the Twin Cities Metro area in Minnesota, 1-18 September 2015. Occupancy: the percentage of time, during the 30 second sample period, that the detector sensed a vehicle.
Source: Minnesota Department of Transportation

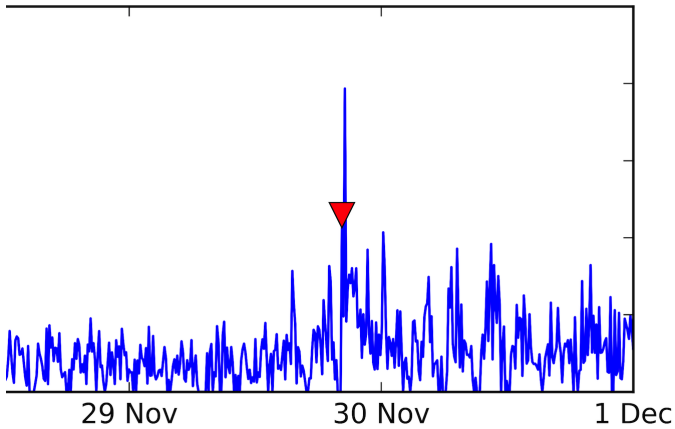
Example: NAB benchmark



A collection of Google mentions in Twitter. The metric value represents the number of mentions for a given ticker symbol every 5 minutes from February, 27 to April, 22 2015

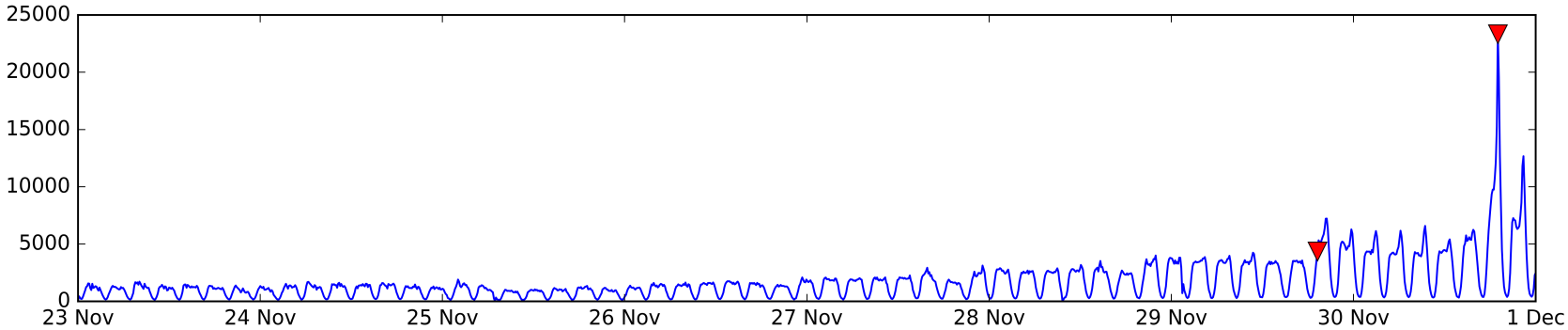
Example: Yahoo! benchmark

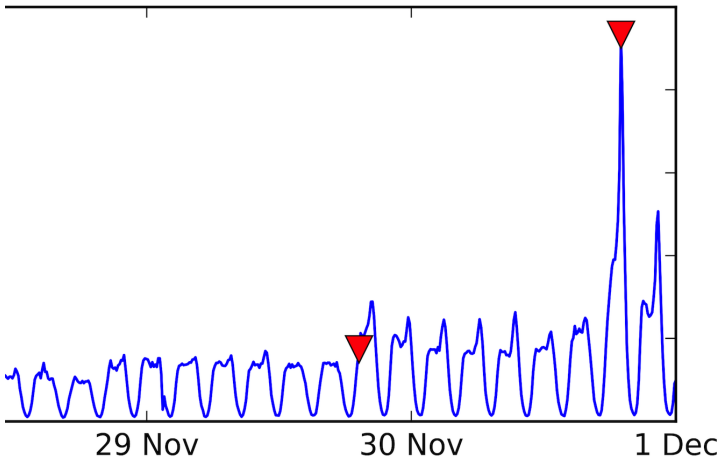




Real production traffic to some of the Yahoo! properties from 23.11.2014 to 02.11.2014

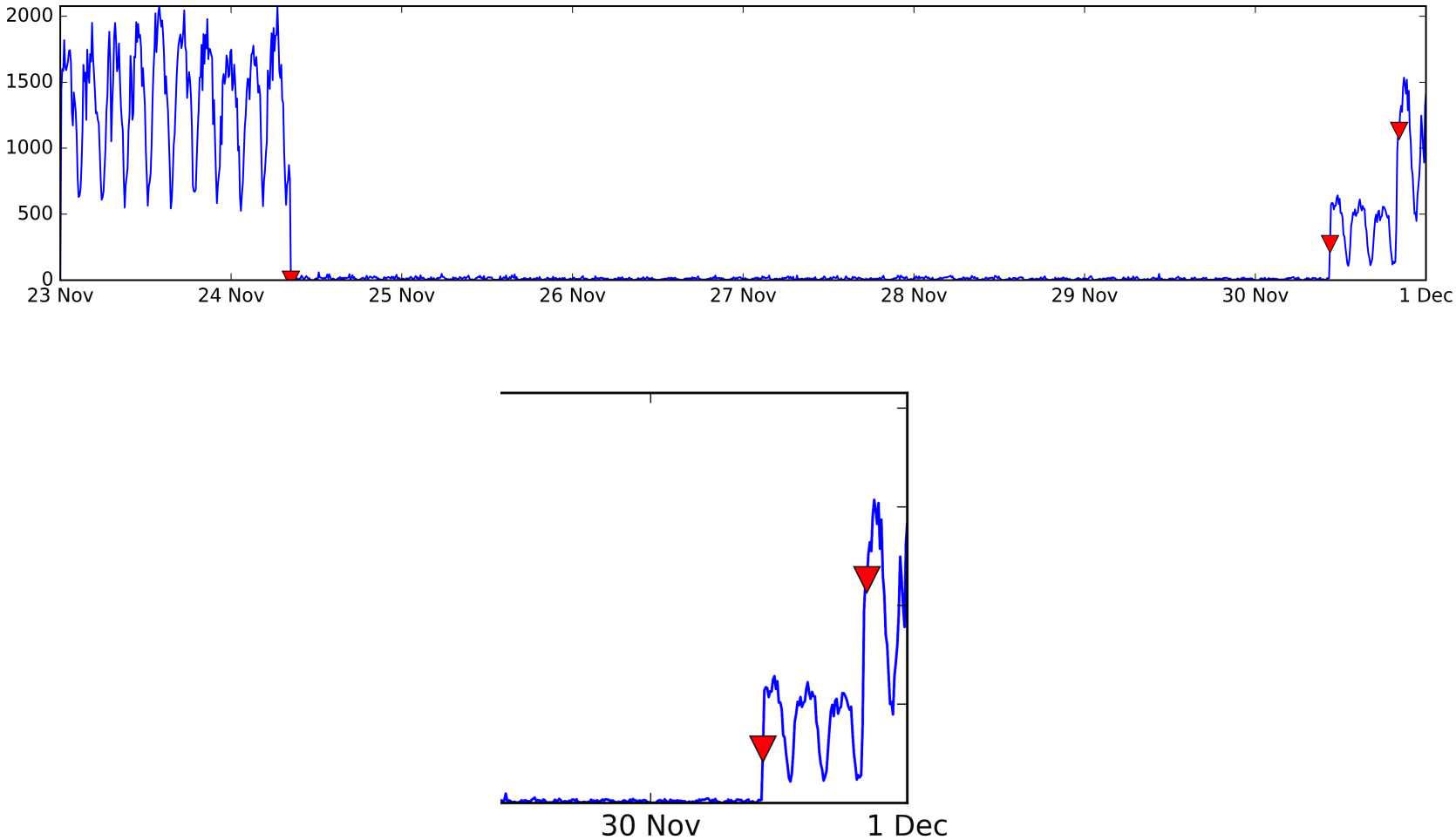
Example: Yahoo! benchmark





Real production traffic to some of the Yahoo! properties from 23.11.2014 to 02.11.2014

Example: Yahoo! benchmark



Real production traffic to some of the Yahoo! properties from 23.11.2014 to 02.11.2014

Change-point models: difficulties

Data:

- Change-point methods require strong pre-, and post-change probabilistic assumptions;
- Observations are assumed stationary
- Real data model may be significantly different

Change-points:

- Single or multiple change-points of similar or different nature;
- Clustering of change-points.

Anomaly detection challenge

- large volumes of data;
- non-stationary and/or quasi-periodic dynamics;
- intractable or unknown data distributions;
- lack of anomaly occurrence model

The challenges justify the need for a model-free adaptive and computationally efficient methods

Detector requirements

A good online anomaly detector should:

- detect as many anomalies as possible;
- detect anomalies as soon as possible, ideally before the anomaly becomes apparent;
- trigger as few as possible false alarms;
- must be causal

NAB: datasets

Numenta Anomaly Benchmark for evaluating algorithms for online anomaly detection:

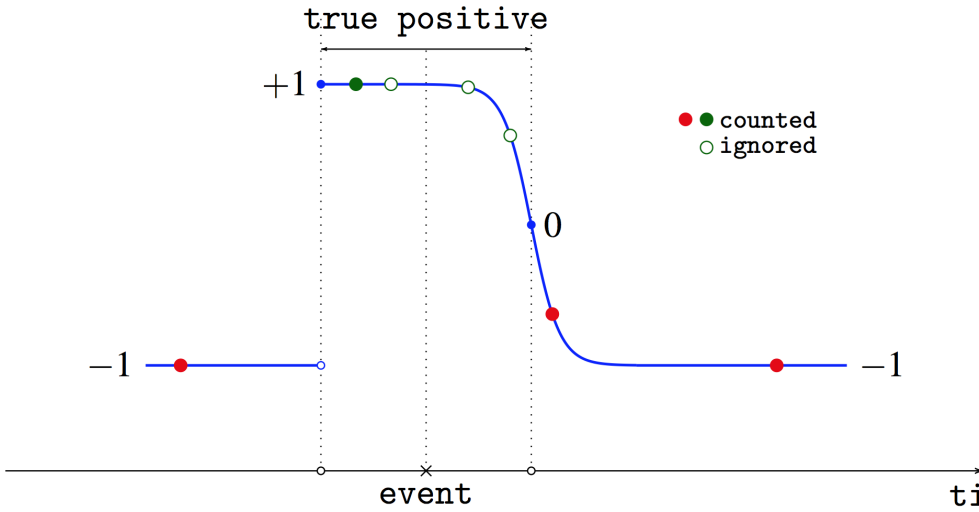
- ~ 58 (Numenta dataset) ~ 100 (Yahoo dataset) manually labeled real-world and artificial time-series;
- 1000-22000 data instances per series
- real world data: metrics ranging from IT metrics (network traffic, CPU utilization) to sensors on industrial machines to social media chatter;
- several anomaly-free time-series.

Precision and recall do not penalize late or reward early detection: NAB uses performance metrics which explicitly incorporate time

NAB: detector performance score

Scoring mechanism:

- anomaly windows centered around a ground truth label;
- only the earliest detection within a window is a TP, others are ignored;
- time-based weights:
  - favours earlier TP detections;
  - softly penalizes FP detections;
- FN is the number of windows without any detection



Current work

We propose model-free multivariate anomaly detection methods and their adaptation to time-series data:

- performs well on the Anomaly Detection benchmark;
- provides a confidence measure of the detection;
- is computationally efficient;
- adapts to possible non-stationarity and quasi-periodicity

Conformal Anomaly Detection

Let us consider a sequence of multidimensional observations $\mathbf{x}_i \in \mathbb{R}^d$, $i = 1, 2, \ldots$

Conformal Anomaly Detection — a distribution-free method to assign confidence score to observations

- Non-Conformity Measure (NCM) $A(X_{:m}, \mathbf{x}_{m+1})$ quantifies how $\mathbf{x}_{m+1}$ is relative to $X_{:m} = (\mathbf{x}_i)_{i=1}^m$;
- Non-conformity scores $\{\alpha_i\}_{i=1}^{m+1} = \{A(X_{:m,-i}, \mathbf{x}_i)\}_{i=1}^{m+1}$ are used to get the empirical p-value

$$p(\mathbf{x}_{m+1}, X_{:m}, A) = \frac{1}{m+1}\big|\{i : \alpha_i \geq \alpha_{m+1}\}\big| \qquad \text{(CPv)}$$

The more abnormal $\mathbf{x}_{m+1}$ is, the lower its (CPv) is

Conformal Anomaly Detection

Conservative coverage guarantees in online learning setting:

- i.i.d. data $\mathbf{x}_m \sim D$ is fed into the conformal procedure one observations at a time,
- then for any NCM A and all $m \geq 1$ we have

$$\mathbb{P}_{X \sim D^{m+1}} \left(p(\mathbf{x}_{m+1}, X_{-(m+1)}, A) < \epsilon \right) \leq \epsilon$$

Choice of the NCM affects

- the tightness of the guarantee,
- volume of computations

As **NCM** we use k Nearest Neighbors statistics

k Nearest Neighbours Anomaly Detector

Assign the abnormality score to $\mathbf{x}$ based on the average neighbourhood proximity:

$$\mathrm{NN}(\mathbf{x}; k) = k^{-1} \sum_{\mathbf{y} \in N_k(\mathbf{x})} d(\mathbf{x}, \mathbf{y}),$$

where $N_k(\mathbf{x})$ are k-nearest neighbors of $\mathbf{x}$ (excluding $\mathbf{x}$) from a given dataset

As $d(\mathbf{x}, \mathbf{y})$ we use

$$d(\mathbf{x}, \mathbf{y}) = \sqrt{(\mathbf{x} - \mathbf{y})' \hat{\mathbf{\Sigma}}^{-1} (\mathbf{x} - \mathbf{y})},$$

where $\hat{\mathbf{\Sigma}}^{-1}$ is an estimate of the covariance matrix

- every observation with $\mathrm{NN}(\mathbf{x}; k) \geq \epsilon$ is an anomaly;
- sensitive to k;
- the score $\mathrm{NN}(\mathbf{x}; k)$ is poorly interpretable

Lazy Drifting Conformal Detector

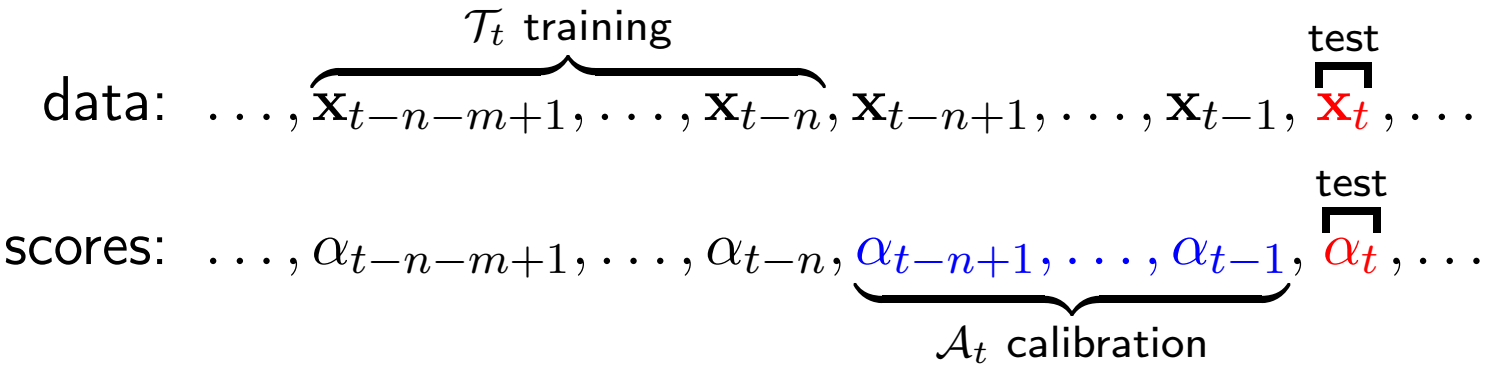
CAD is inefficient: all non-conformity scores are recomputed for each new observation

We propose a lazy procedure LDCD, which

- does less computations;
- adapts to non-stationarity by tracking the data;
- produces conformal confidence scores.

Lazy Drifting Conformal Detector

At each step t LDCD with the non-conformity measure A over $(\mathbf{x}_s)_{s\geq 1}$:



- get the score α_t by applying A, trained over $\mathcal{T}_t$, to $\mathbf{x}_t$;
- compute the p-value of α_t in the calibration **queue** $\mathcal{A}_t$;
- $\mathcal{A}_{t+1} \leftarrow \mathcal{A}_t$: push α_t into $\mathcal{A}_t$ end evict α_{t-n+1};

Time-delay embedding

Let $X = (x_t)_{t \geq 1}$ is a univariate time-series;

Use a sliding historical window of width L:

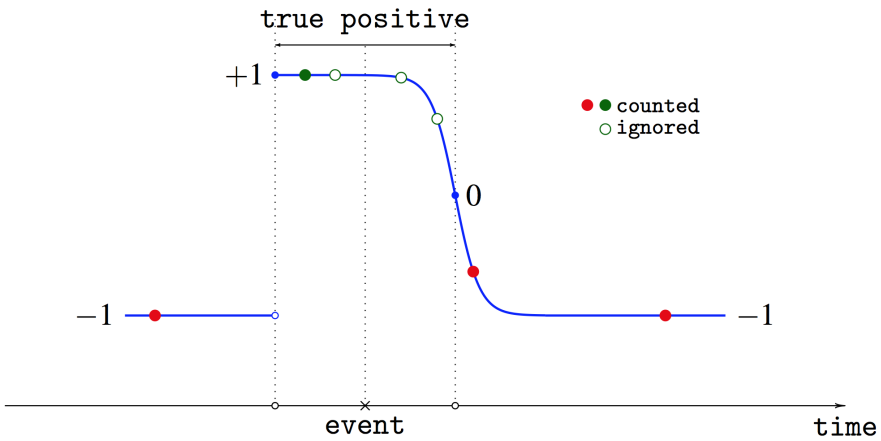
$$\dots, x_{t-L-1}, \overbrace{x_{t-L}, \underbrace{x_{t-L+1}, \dots, x_{t-1}}_{\mathbf{x}_{t+1}}, x_t}^{\mathbf{x}_t}, x_{t+1}, \dots,$$

$\mathbf{x}_t$ are the most recent L observations before t (not including x_t)

Procedure:

1. embed the time series $X = (x_t)_{t \geq 1}$ in the L-dimensional space;
2. apply the LDCD to the stream $\tilde{X} = (\mathbf{x}_t)_{t \geq L+n+m+1}$ of multivariate observations

Scoring example for a sample anomaly window



Metric	A_{TP}	A_{FP}	A_{TN}	A_{FN}
Standard	1.0	-0.11	1.0	-1.0
LowFP	1.0	-0.22	1.0	-1.0
LowFN	1.0	-0.11	1.0	-2.0

- The first point is an FP preceding the anomaly window
- Only one detection counts the earliest TP for the score
- There are two FPs after the window. TNs make no score contributions

- NAB score for this example would calculate as:

  $-1.0A_{FP} + 0.9999A_{TP} - 0.8093A_{FP} - 1.0A_{FP} = 0.6909$ for the standard application profile

Runtime complexity

method	Prediction on series $(\mathbf{x}_s)_{s=1-n}^{T}$ Scores	Pv
LDCD	$Tc_A(n)$	Tm
ICAD (sliding)	$Tc_A(n)$	Tm
ICAD (online)	$Tc_A(n)$	$T \log T$
CAD	$\sum_{t=1}^{T}(t+n)c_A(t+n-1)$	$nT + \frac{1}{2}T(T+1)$

Table : Worst case runtime complexity of conformal procedures on $(\mathbf{x}_s)_{s=1-n}^{T}$, n is the length of the train sample.

Comparison: Results for Yahoo! S5 dataset

Detector	LowFN	LowFP	Standard
27-NN $l = 19$ LDCD w. pruning	68.8	56.9	64.3
1-NN $l = 1$ LDCD w. pruning	53.8	36.2	46.9
relativeEntropy	52.5	40.7	48.0
Numenta	44.4	37.5	41.0
Numenta™	42.5	36.6	39.4
bayesChangePt	43.6	17.6	35.7
windowedGaussian	40.7	25.8	31.1
skyline	28.9	18.0	23.6
Random ($p_t \sim \mathcal{U}[0,1]$)	47.2	1.2	29.9

Comparison: Numenta dataset

Detector	LowFN	LowFP	Standard
27-NN $l = 19$ LDCD w. pruning	64.1	42.6	56.8
1-NN $l = 1$ LDCD w. pruning	62.7	30.7	53.5
Numenta	74.3	63.1	70.1
Numenta™	69.2	56.7	64.6
relativeEntropy	58.8	47.6	54.6
windowedGaussian	47.4	20.9	39.6
skyline	44.5	27.1	35.7
bayesChangePt	32.3	3.2	17.7
Random ($p_t \sim \mathcal{U}[0,1]$)	25.9	5.8	16.8

Conclusions

We propose model-free anomaly detection methods for time-series data:

- performs well on the Anomaly Detection benchmark;
- provides a confidence measure of the detection;
- is computationally efficient;
- adapts to possible non-stationarity and quasi-periodicity;
- **third place in a competition**

Dynamic range

- To measure effect of conformal p-values on the performance: test a basic k-NN detector with a heuristic rule to assign confidence
- The score of the t-th observation

$$\alpha_t = \mathsf{NN}(\mathbf{x}_t; k)\,.$$

- The score is dynamically normalized to a value within the $[0, 1]$ range

$$\mathsf{Pv}_t = \frac{\max_{i=0}^{m} \alpha_{t-i} - \alpha_t}{\max_{i=0}^{m} \alpha_{t-i} - \min_{i=0}^{m} \alpha_{t-i}}\,.$$

- The value $p_t = 1 - \mathsf{Pv}_t$ is the conformal abnormality score returned by each detector for the observation x_t

Conformal vs. Dynamic range: $(k, l) = (27, 19)$

Corpus	p-value	LowFN	LowFP	Standard
Numenta	DynR	-9.6	-185.7	-54.9
	LDCD	4.3	-143.8	-34.0
	DynR w. pruning	63.0	36.2	54.9
	LDCD w. pruning	64.1	42.6	56.8
Yahoo!	DynR	50.0	0.3	36.1
	LDCD	50.1	0.4	36.1
	DynR w. pruning	68.2	56.4	63.8
	LDCD w. pruning	68.8	56.9	64.3

NAB scores of the k-NN detector $(27, 19)$ on the Numenta and Yahoo! S5 corpora

Conformal vs. Dynamic range: $(k, l) = (1, 1)$

Corpus	*p*-value	LowFN	LowFP	Standard
Numenta	DynR	-167.0	-658.4	-291.0
	LDCD	62.3	34.8	53.8
	DynR w. pruning	52.2	4.2	39.0
	LDCD w. pruning	62.7	30.7	53.5
Yahoo!	DynR	30.8	-20.7	16.9
	LDCD	47.7	21.5	37.6
	DynR w. pruning	50.6	35.2	44.8
	LDCD w. pruning	53.8	36.2	46.9

NAB scores of the k-NN detector $(1, 1)$ on the Numenta and Yahoo! S5 corpora