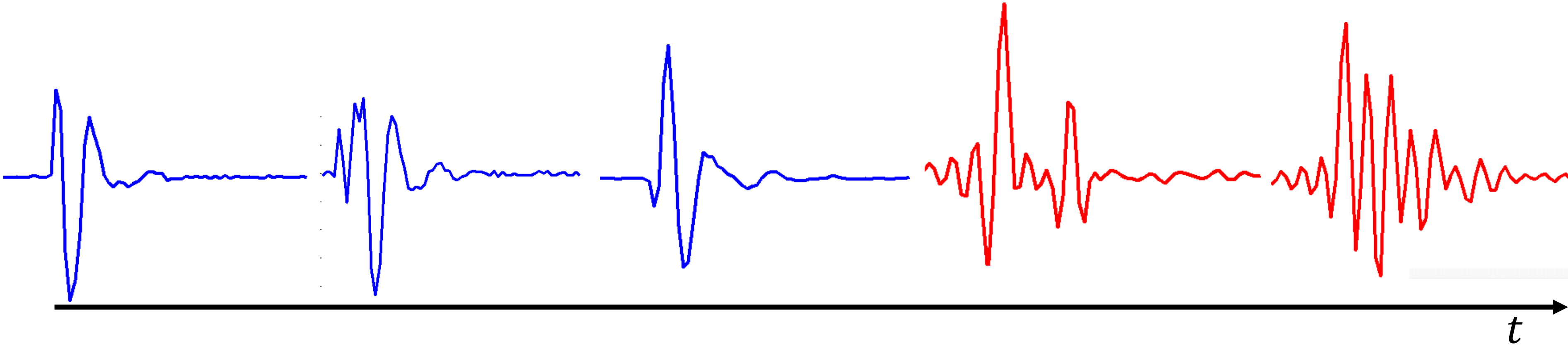
Change Detection in Data Streams: Big Data Challenges

Giacomo Boracchi
DEIB, Politecnico di Milano,
giacomo.boracchi@polimi.it

October 24, 2016
INNS Big Data Conference 2016,
Thessaloniki, Greece

… A CHANGE-DETECTION PROBLEM

Environmental monitoring: a Sensor network for monitoring rockfaces and detect changes waveforms recorded by MEMS sensors in these units.



C. Alippi, G. Boracchi, B. Wohlberg "*Change Detection in Streams of Signals with Sparse Representations*" in Proceedings of IEEE ICASSP 2014, pp 5252 - 5256

… A CHANGE-DETECTION PROBLEM

Learning problems related to **predicting user preferences / interests,** such as:

- Recommendation systems
- Spam / email filtering

Changes arise when users change their own preferences.

Changes have to be detected to update the system accordingly

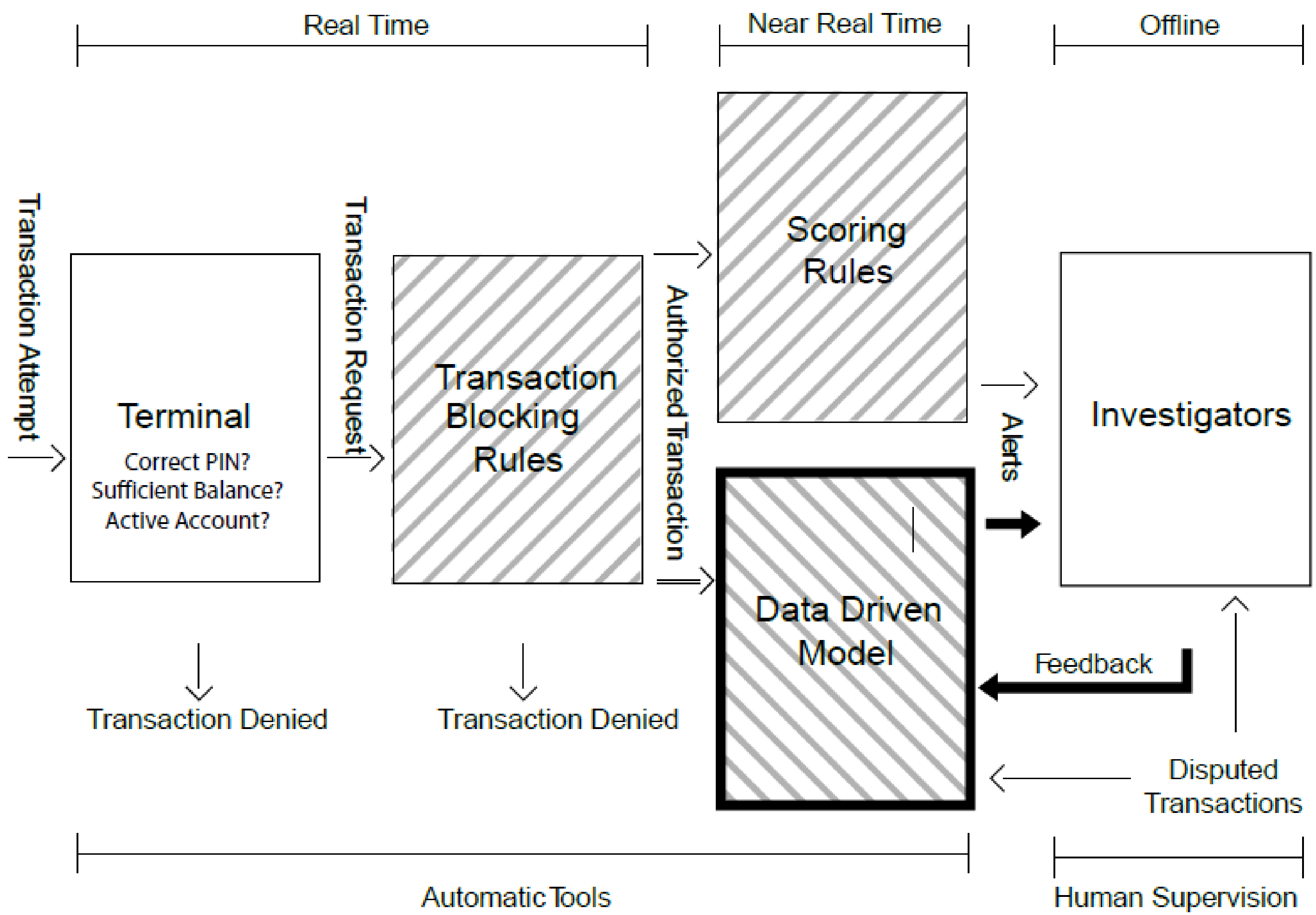
Spam Classification

Alippi, C., Boracchi, G., Roveri, M. *"Just-in-time classifiers for recurrent concepts"*. IEEE TNNLS, 24(4), 620-634 (2013).

Gama, J., Žliobaitė, I., Bifet, A., Pechenizkiy, M., Bouchachia, A. *"A survey on concept drift adaptation"*. ACM Computing Surveys (CSUR), 46(4), 44. (2014)

… AN ANOMALY-DETECTION PROBLEM

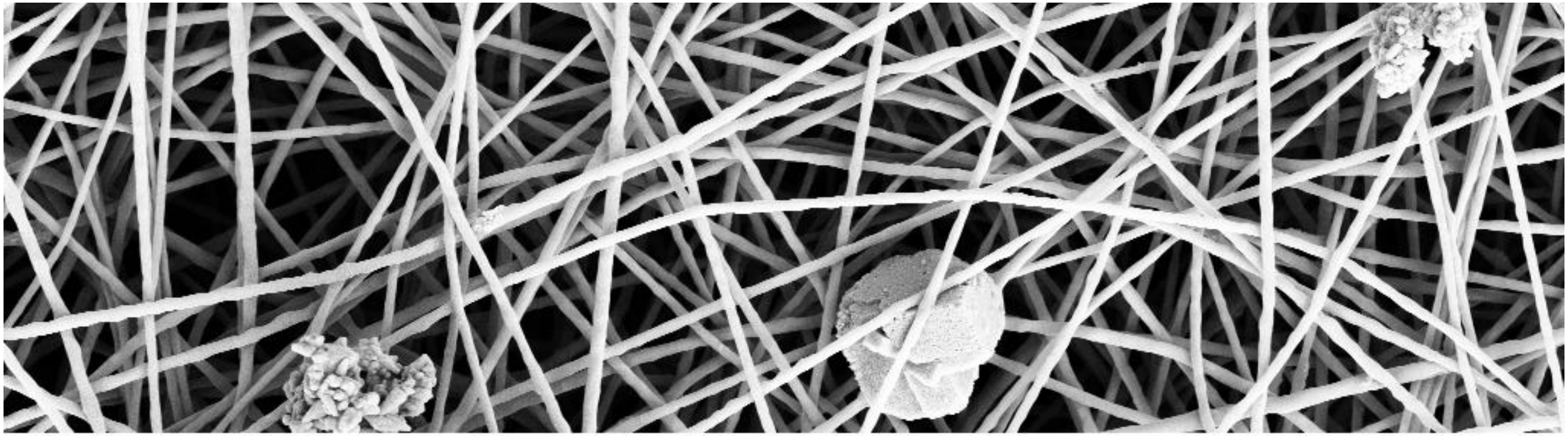
Fraud detection in streams of credit card transactions

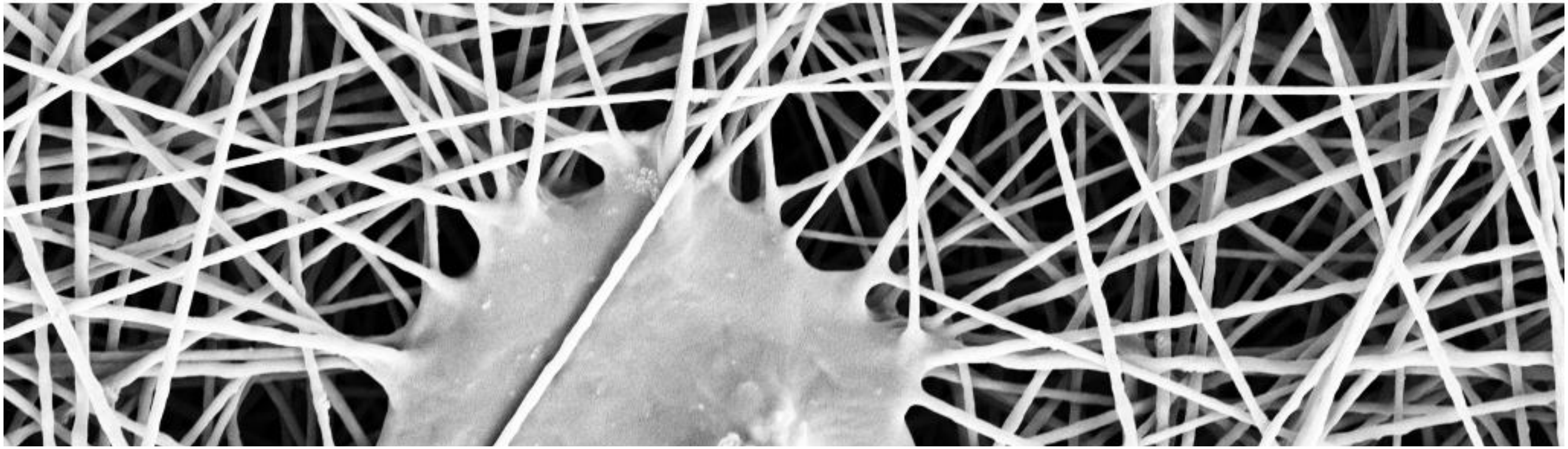


Dal Pozzolo A., Boracchi G., Caelen O., Alippi C. and Bontempi G., *"Credit Card Fraud Detection and Concept-Drift Adaptation with Delayed Supervised Information"*, Proceedings of IJCNN 2015, 8 pages

… AN ANOMALY-DETECTION PROBLEM

Quality Inspection Systems: monitoring the nanofiber production

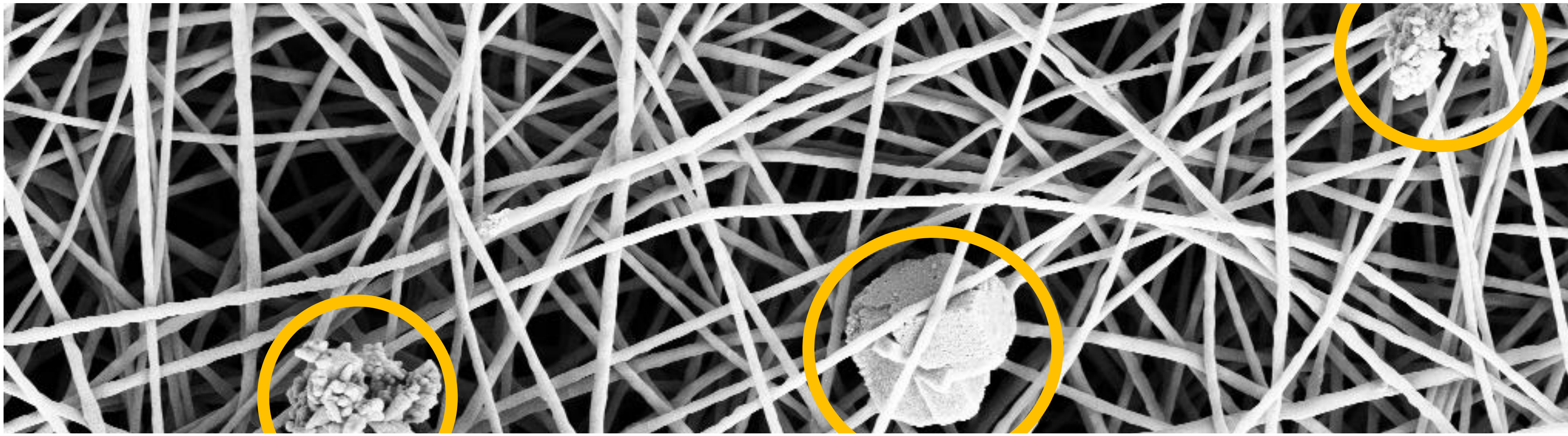


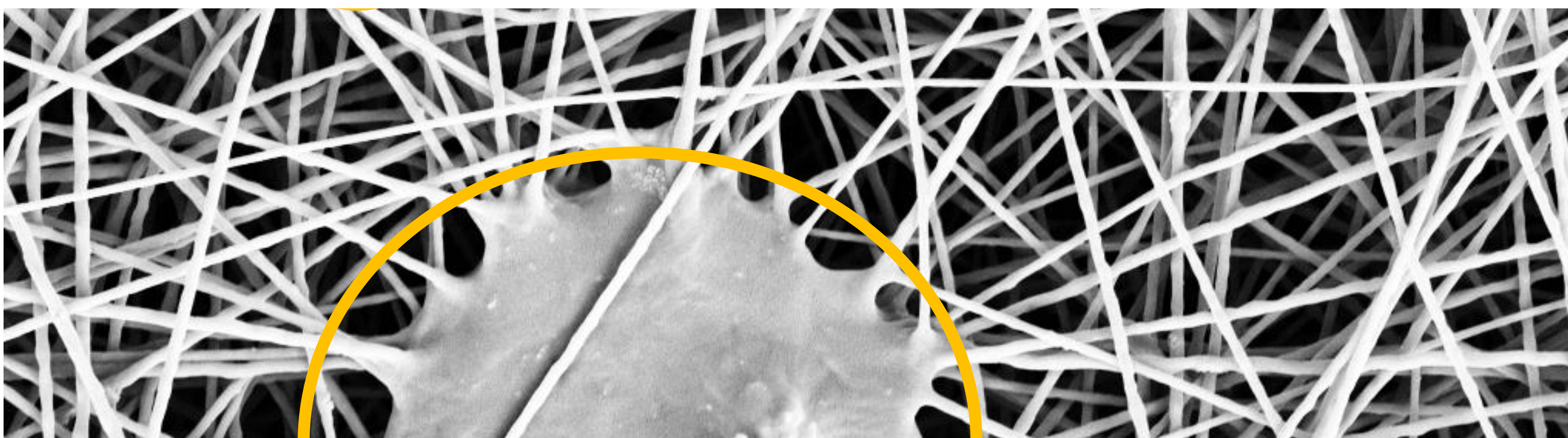


G. Boracchi, D. Carrera and B. Wohlberg *"Novelty Detection in Images by Sparse Representations"* in Proceedings of Intelligent Embedded Systems at SSCI 2014

… AN ANOMALY-DETECTION PROBLEM

Quality Inspection Systems: monitoring the nanofiber production





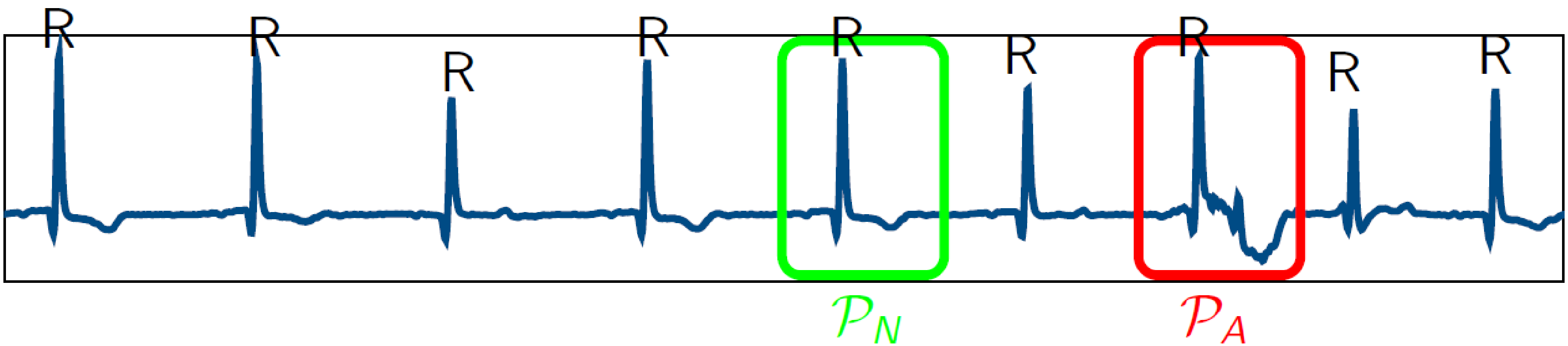
G. Boracchi, D. Carrera and B. Wohlberg *"Novelty Detection in Images by Sparse Representations"* in Proceedings of Intelligent Embedded Systems at SSCI 2014

… AN ANOMALY-DETECTION PROBLEM

Health monitoring / wearable devices:

Automatically analyze EGC tracings to detect arrhythmias or device mispositioning

This is important to provide user-specific monitoring



D. Carrera, B. Rossi, D. Zambon, P. Fragneto, and G. Boracchi "*ECG Monitoring in Wearable Devices by Sparse Models*" in Proceedings of ECML-PKDD 2016, 16 pages

PRESENTATION OUTLINE

- Problem formulation (in a statistical framework)
- Solutions in the ideal conditions
- Solutions when data-distributions are unknown
- Solutions when data are not random variables
- Big data challenges related to change detection
- Detectability Loss
- Conclusions

DISCLAIMER

I am focused on **datastreams**, which do **not have a fixed length** and that have to be **analyzed while data are received**. I am not considering retrospective / offline analysis tools

I am mainly considering **numerical data**. In some cases, extensions apply to categorical or ordinal ones.

I refer to either changes/anomalies according to **my personal experience** in the applications I have considered

For **complete survey** on change/anomaly detection please refer to the very good surveys reported below

V. Chandola, A. Banerjee, V. Kumar. "*Anomaly detection: A survey*". ACM Comput. Surv. 41, 3, Article 15 (July 2009), 58 pages.

Pimentel, M. A., Clifton, D. A., Clifton, L., Tarassenko, L. *"A review of novelty detection"* Signal Processing, 99, 215-249 (2014)

A. Zimek, E. Schubert, H.P. Kriegel. *"A survey on unsupervised outlier detection in high-dimensional numerical data"* Statistical Analysis and Data Mining: The ASA Data Science Journal, 5(5), 2012.

THE PROBLEM FORMULATION

Anomaly / Change Detection Problems in a Statistical Framework

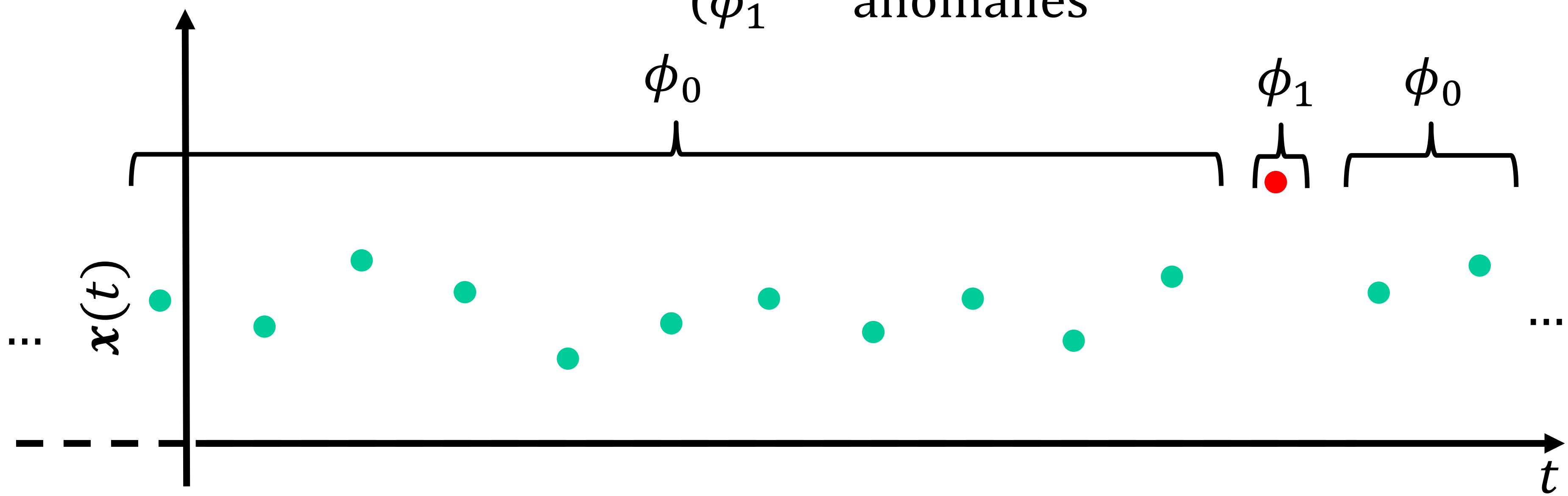
Often, the anomaly-detection problem **boils down to:**

Monitor a datastream

$$\{\boldsymbol{x}(t),\ t = t_0, \dots\},\ \ \boldsymbol{x}(t) \in \mathbb{R}^d$$

where $x(t)$ are realizations of a **random variable having pdf** ϕ_o, and detect those points that are **outliers** i.e.,

$$\boldsymbol{x}(t) \sim \begin{cases} \phi_0 & \text{normal data} \\ \phi_1 & \text{anomalies} \end{cases},$$



ANOMALY-DETECTION IN A STATISTICAL FRAMEWORK

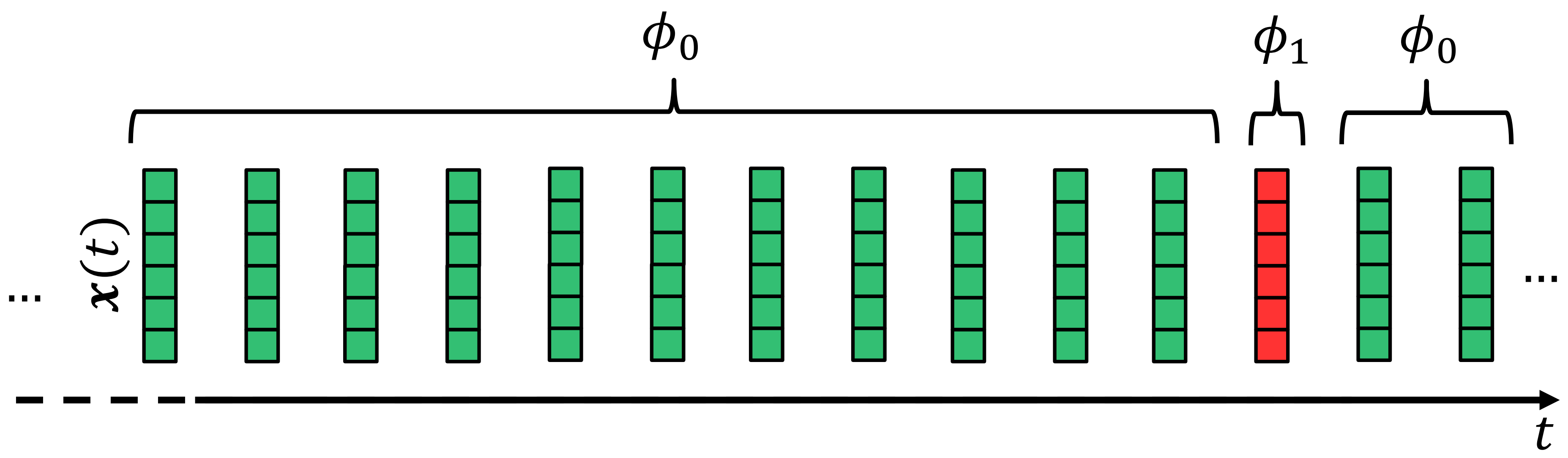
Often, the anomaly-detection problem **boils down to:**

Monitor a datastream

$$\{\boldsymbol{x}(t),\ t = t_0, \dots\},\quad \boldsymbol{x}(t) \in \mathbb{R}^d$$

where $x(t)$ are realizations of a **random variable having pdf** ϕ_o, and detect those points that are **outliers** i.e.,

$$\boldsymbol{x}(t) \sim \begin{cases} \phi_0 & \text{normal data} \\ \phi_1 & \text{anomalies} \end{cases},$$



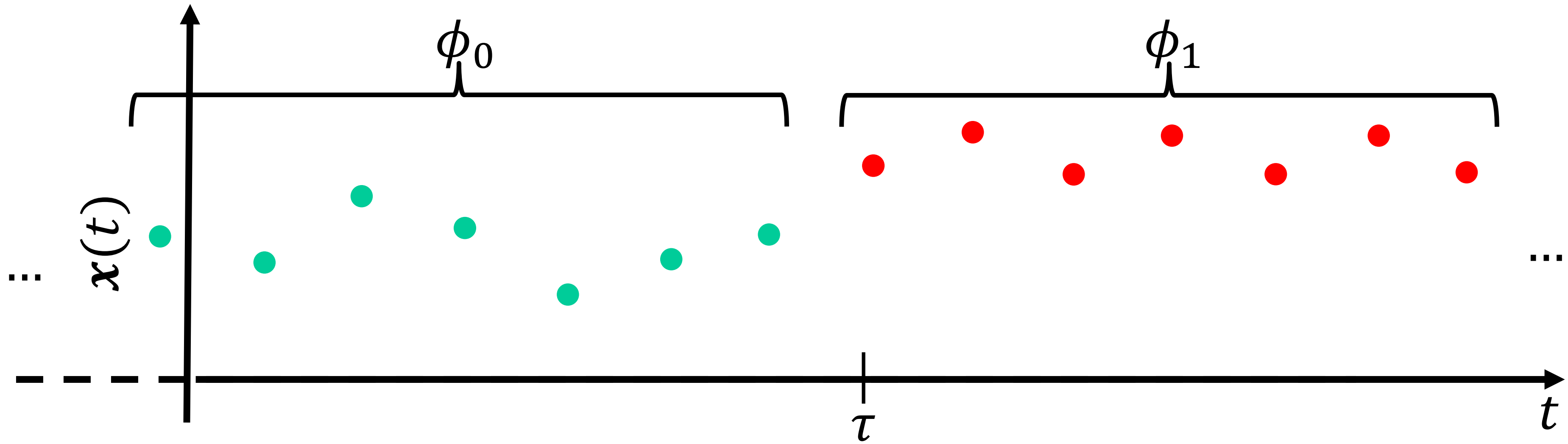
Often, the change-detection problem **boils down to**:

Monitor a **stream** $\{\boldsymbol{x}(t), t = 1, \dots\}$, $\boldsymbol{x}(t) \in \mathbb{R}^d$ of realizations of a **random variable**, and **detect the change-point** τ,

$$\boldsymbol{x}(t) \sim \begin{cases} \phi_0 & t < \tau \\ \phi_1 & t \geq \tau \end{cases},$$

where $\{\boldsymbol{x}(t),\ t < \tau\}$ are i.i.d. and $\phi_0 \neq \phi_1$

We denote such change as: $\phi_o \to \phi_1$



CHANGE-DETECTION IN A STATISTICAL FRAMEWORK

Often, the change-detection problem **boils down to**:

Monitor a **stream** $\{\boldsymbol{x}(t), t = 1, \dots\}$, $\boldsymbol{x}(t) \in \mathbb{R}^d$ of realizations of a **random variable**, and **detect the change-point** τ,

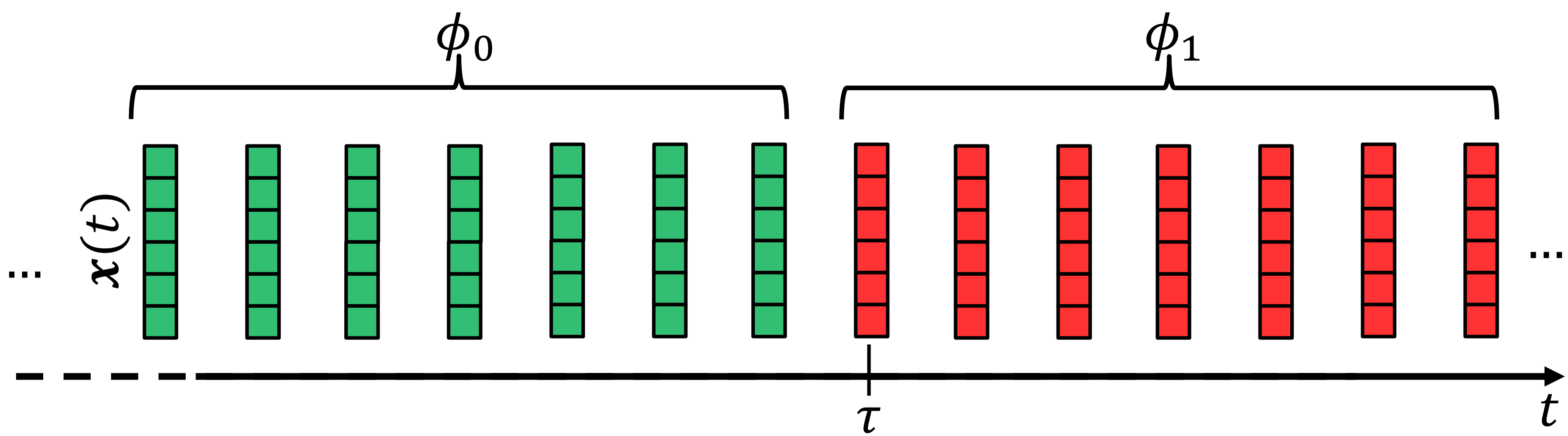
$$\boldsymbol{x}(t) \sim \begin{cases} \phi_0 & t < \tau \\ \phi_1 & t \geq \tau \end{cases},$$

in control state

out of control state

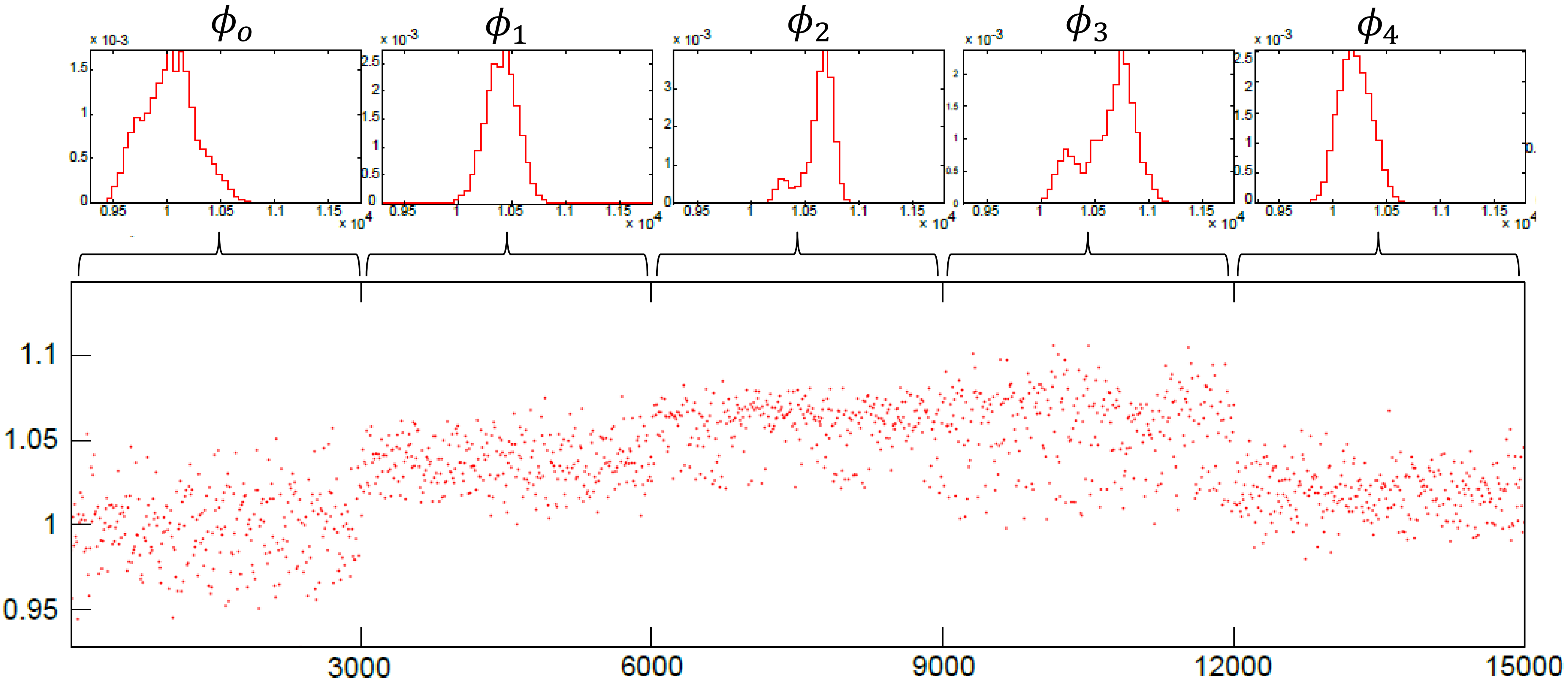
where $\{\boldsymbol{x}(t),\ t < \tau\}$ are i.i.d. and $\phi_0 \neq \phi_1$

We denote such change as: $\phi_o \to \phi_1$

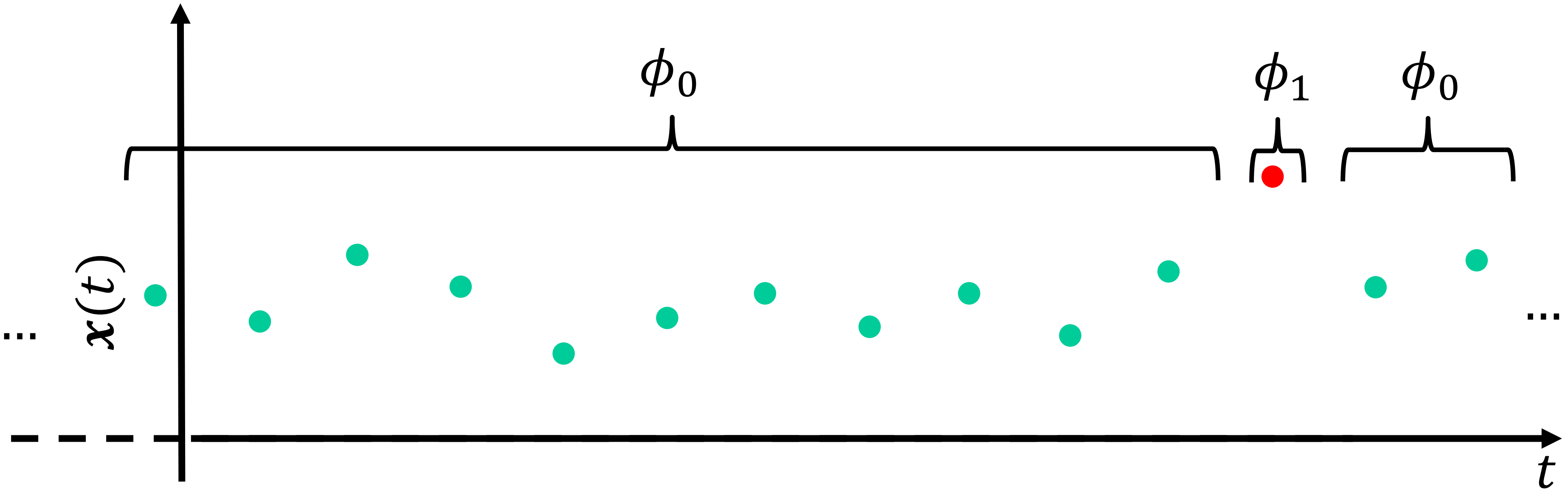


Here are data from an X-ray monitoring apparatus.

There are 4 changes $\phi_o \to \phi_1 \to \phi_2 \to \phi_3 \to \phi_4$ corresponding to different monitoring conditions/materials

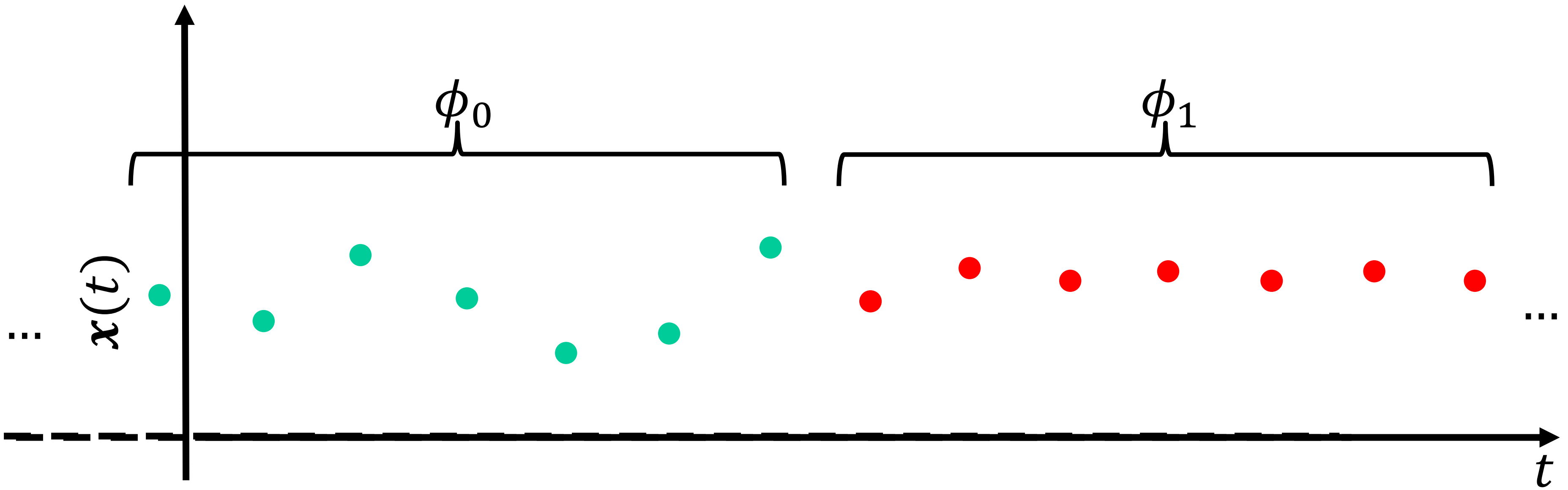


Not all anomalies are due to process changes



PROCESS CHANGES VS ANOMALIES

Not all process changes result in anomalies



THE TWO PROBLEMS

Anomaly-detection problem:

Locate those samples that do not conform the normal ones or a model explaining normal ones

Anomalies in data translate to significant information

Change-detection problem:

Given the previously estimated model, the arrival of new data invites the question: "is yesterday's model capable of explaining today's data?"

Detecting process changes important to understand the monitored phenomenon

V. Chandola, A. Banerjee, V. Kumar. "*Anomaly detection: A survey*". ACM Comput. Surv. 41, 3, Article 15 (2009), 58 pages.

C. J. Chu, M. Stinchcombe, H. White "*Monitoring Structural Change*" Econometrica Vol. 64, No. 5 (Sep., 1996), pp. 1045-1065.

THE TYPICAL SOLUTIONS

Most algorithms are composed of:

- A **statistic** that has a known response to normal data (e.g., the average, the sample variance, the log-likelihood, the confidence of a classifier, an "anomaly score"…)
- A **decision rule** to analyze the statistic (e.g., an adaptive threshold, a confidence region)

Anomaly-detection problem:

Statistics and decision rules are "one-shot", analyzing a set of historical data or each new data (or chunk) independently

Change-detection problem:

Statistics and decision rules are "sequential", as they make a decision considering all the data received so far

SOLUTIONS IN THE IDEAL CONDITIONS

… when ϕ_0 and ϕ_1 are known

ONE-SHOT DETECTOR: NEWMAN PEARSON TEST

Assume data are generated from a parametric distribution ϕ_θ and formulate the following hypothesis test

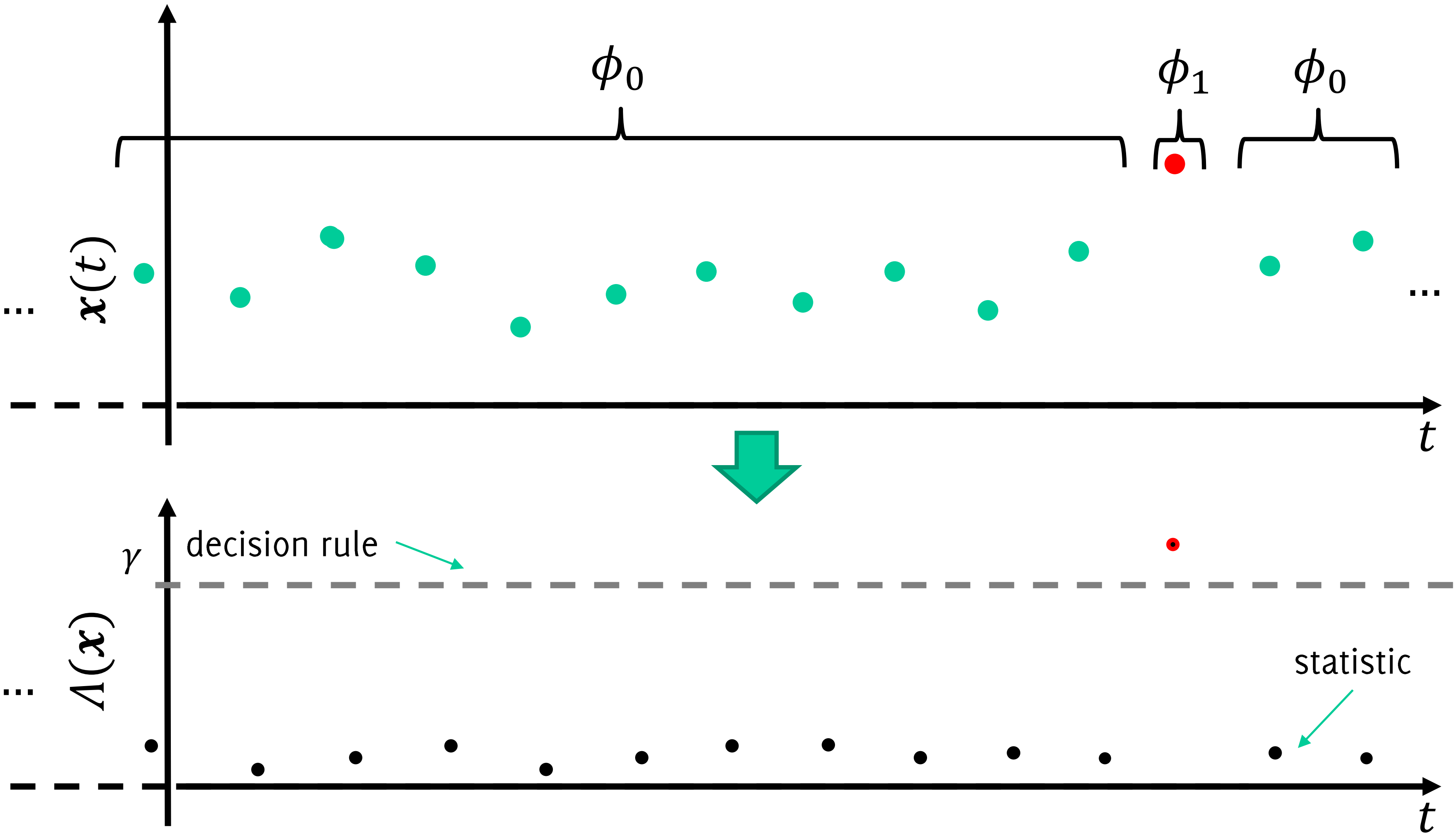
$$H_0: \theta = \theta_0 \text{ vs } H_1: \theta = \theta_1$$

According to the Neumann Pearson lemma the most powerful **statistic** to detect changes is the **log-likelihood ratio**

$$\Lambda(x) = \frac{\phi_1(x)}{\phi_0(x)}$$

and the **detection rule** being $\Lambda(x) > \gamma$, where γ is set to **control the false alarm rate** (type I errors of the test).

Outliers can be detected by a threshold on $\Lambda(\boldsymbol{x})$



THE CUSUM TEST ON THE LIKELIHOOD RATIO

CUSUM involves the calculation of a **CU**mulative **SUM,** which makes it a sequential monitoring scheme.

It can be applied to the log-likelihood ratio:

$$\log(\Lambda(x)) = \log\left(\frac{\phi_1(x)}{\phi_0(x)}\right) = \begin{cases} < 0 & \text{when} \quad \phi_0(x) > \phi_1(x) \\ > 0 & \text{otherwise} \end{cases}$$

The CUSUM statistic is:

$$S(t) = \max\left(0, \mathrm{S}(\mathrm{t}-1) + \log\left(\Lambda(x(t))\right)\right)$$

And the decision rule is

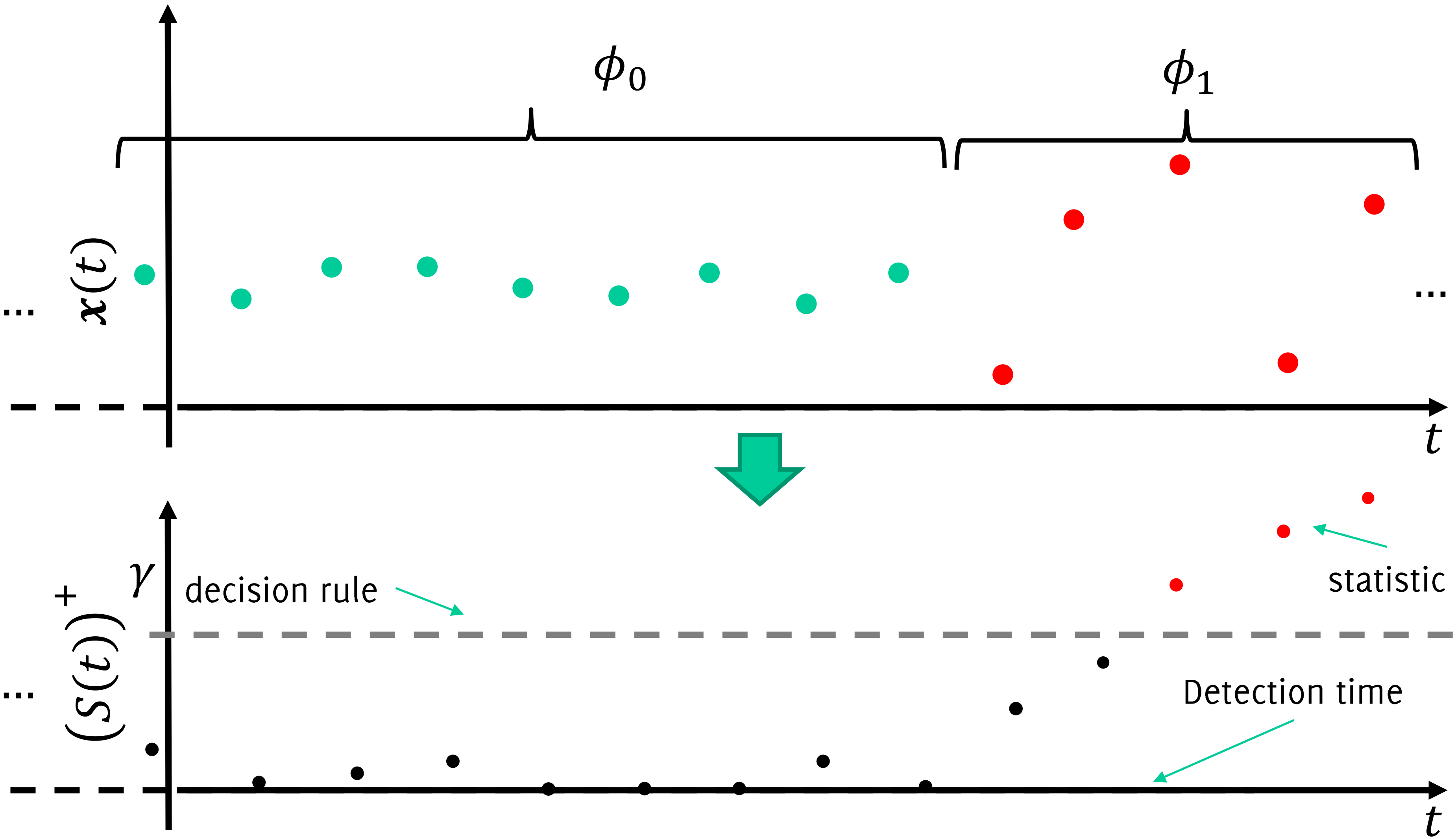
$$S(t) > \ \gamma$$

M. Basseville, I. V. Nikiforov, "*Detection of Abrupt Changes - Theory and Application*", Prentice-Hall, Inc. April 1993

Page, E. S. "*Continuous Inspection Scheme*". Biometrika. 41 (1/2): 100–115 (June 1954).

ONE-SHOT DETECTOR: NEWMAN PEARSON TEST

Outliers can be detected by a threshold on $\Lambda(\boldsymbol{x})$



ANOMALY-DETECTION WHEN ϕ_0 AND ϕ_1 ARE UNKNOWN

DATA DISTRIBUTION IS UNKNOWN

Most often, **only a training set** TR is **provided**:

There are three scenarios:

- **Supervised:** Both normal and anomalous training data are provided in TR.
- **Semi-Supervised:** Only normal training data are provided, i.e. no anomalies in TR.
- **Unsupervised:** TR is provided without label.

V. Chandola, A. Banerjee, V. Kumar. "*Anomaly detection: A survey*". ACM Comput. Surv. 41, 3, Article 15 (2009), 58 pages.

SUPERVISED ANOMALY DETECTION - SOLUTIONS

In **supervised methods** the training data are divided in normal (+) and anomalous (−) ones:

$$TR = \{(\boldsymbol{x}(t), y(t)), t < t_0, x \in \mathbb{R}^d, y \in \{+, -\}\}$$

Solution:

- Use a classifier to distinguish normal vs anomalous data

During training:

- Learn a classifier $\mathcal{K}$ from TR.

During testing:

- compute the classifier output $\mathcal{K}(\boldsymbol{x})$ or set a threshold on the posterior $p_{\mathcal{K}}(-|\boldsymbol{x})$

SUPERVISED ANOMALY DETECTION - CHALLENGES

The **problem is challenging** because of:

- **Class Imbalance:** Normal data far outnumber anomalies
- **Concept Drift:** Anomalies might **evolve** over time
- **Selection Bias:** Training samples are typically selected through a **biased procedure**

This is **what typically happens in fraud detection**:

- Frauds are typically less than 1% of transactions
- New Fraudulent strategies are always devised
- Supervised samples are provided in the form of feedbacks for the alerted transactions

Dal Pozzolo A., Boracchi G., Caelen O., Alippi C. and Bontempi G., *"Credit Card Fraud Detection and Concept-Drift Adaptation with Delayed Supervised Information"*, Proceedings of IJCNN 2015, 8 pages

SEMI-SUPERVISED ANOMALY DETECTION

In semi-supervised methods the TR composed of normal data

$$TR = \{x(t), t < t_0, x \sim \phi_0\}$$

Very practical assumptions:

- **Normal data** are often **easy to gather**
- **Anomalous data** are **difficult/costly** to gather and it would be difficult to have a representative training set
- **Anomalies might also evolve** over time

All in all, it is often **safer** to **detect any data departing from** the **normal conditions**

Semi-supervised anomaly-detection methods are also referred to as **novelty-detection methods**

V. Chandola, A. Banerjee, V. Kumar. "*Anomaly detection: A survey*". ACM Comput. Surv. 41, 3, Article 15 (2009), 58 pages.

Pimentel, M. A., Clifton, D. A., Clifton, L., Tarassenko, L. "*A review of novelty detection*" Signal Processing, 99, 215-249 (2014)

DENSITY-BASED METHODS

Density-Based Methods: ***Normal*** *data occur in* ***high probability regions*** *of a stochastic model, while* ***anomalies*** *occur in the* ***low probability regions*** *of the model*

During training: $\hat{\phi}_0$ can be **estimated** from the training set

$$TR = \{x(t), t < t_0, x \sim \phi_0\}$$

- parametric models (e.g., Gaussian mixture models)
- nonparametric models (e.g. KDE, histograms)

During testing:

- Anomalies are detected as data having $\hat{\phi}_0(\boldsymbol{x}) < \eta$

V. Chandola, A. Banerjee, V. Kumar. "*Anomaly detection: A survey*". ACM Comput. Surv. 41, 3, Article 15 (2009), 58 pages.

DOMAIN-BASED METHODS

Domain-based methods: ***Estimate a boundary around normal*** *data, rather than the density of normal data.*

A **drawback of density-estimation methods** is that they are meant to be accurate in high-density regions, while anomalies live in low-density ones.

One-Class SVM are domain-based methods defined by **the normal samples at the periphery of the distribution.**

Schölkopf, B., Williamson, R. C., Smola, A. J., Shawe-Taylor, J., Platt, J. C. "*Support Vector Method for Novelty Detection*". In NIPS 1999 (Vol. 12, pp. 582-588).

Tax, D. M., Duin, R. P. "*Support vector domain description*". Pattern recognition letters, 20(11), 1191-1199 (1999)

Pimentel, M. A., Clifton, D. A., Clifton, L., Tarassenko, L. *“A review of novelty detection”* Signal Processing, 99, 215-249 (2014)

UNSUPERVISED ANOMALY-DETECTION

The training set TR might contain **both normal and anomalous data**. However, **no labels** are provided

$$TR = \{x(t), t < t_0 \}$$

Underlying assumption: *Anomalies are rare* *w.r.t. normal data* TR

Remarks:

- Density/Domain based methods that are robust to outliers can be applied in an unsupervised scenario
- Unsupervised methods can be improved whenever labels are available

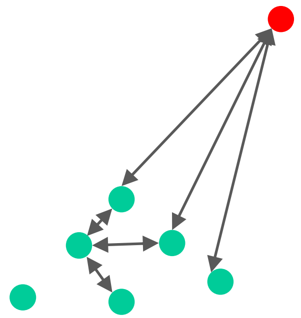
DISTANCE-BASED METHODS

Distance-based methods: ***Normal*** *data instances occur in* ***dense neighborhoods****, while* ***anomalies*** *occur* ***far from their closest neighbors****.*

A critical aspect is the **choice of the similarity measure** to use.

Anomalies are detected by **monitoring**:

- **distance** between each data and its $\boldsymbol{k}$ **−nearest neighbor**



V. Chandola, A. Banerjee, V. Kumar. "*Anomaly detection: A survey*". ACM Comput. Surv. 41, 3, Article 15 (2009), 58 pages.

Zhao, M., Saligrama, V. (2009). Anomaly detection with score functions based on nearest neighbor graphs. In Advances in neural information processing systems (pp. 2250-2258).

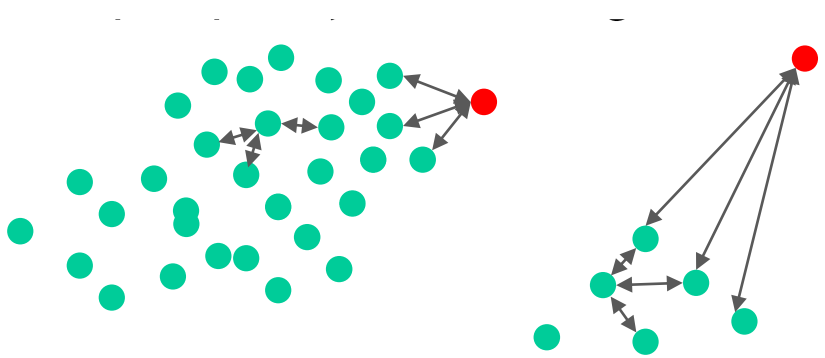
DISTANCE-BASED METHODS

Distance-based methods: *Normal* *data instances occur in* ***dense neighborhoods****, while* ***anomalies*** *occur* ***far from their closest neighbors***.

A critical aspect is the **choice of the similarity measure** to use.

Anomalies are detected by **monitoring**:

- **distance** between each data and its $\boldsymbol{k}$ **−nearest neighbor**
- the **density** of each data **relatively to its neighbors**



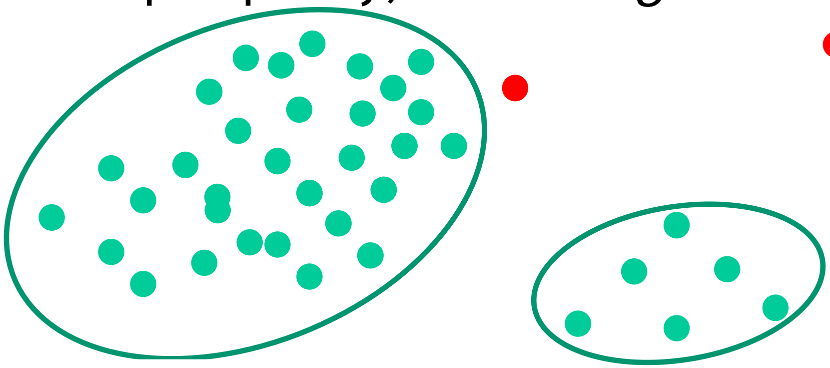
V. Chandola, A. Banerjee, V. Kumar. "*Anomaly detection: A survey*". ACM Comput. Surv. 41, 3, Article 15 (2009), 58 pages.

Distance-based methods: ***Normal*** *data instances occur in* ***dense neighborhoods****, while* ***anomalies*** *occur* ***far from their closest neighbors****.*

A critical aspect is the **choice of the similarity measure** to use.

Anomalies are detected by **monitoring**:

- **distance** between each data and its k **−nearest neighbor**
- the **density** of each data **relatively to its neighbors**
- whether they do not belong to **clusters**, or are at the cluster periphery, or belong to small and sparse clusters



V. Chandola, A. Banerjee, V. Kumar. "*Anomaly detection: A survey*". ACM Comput. Surv. 41, 3, Article 15 (2009), 58 pages.

CHANGE-DETECTION WHEN ϕ_0 AND ϕ_1 ARE UNKNOWN

PARAMETRIC SETTINGS

Parametric settings:

ϕ_0 and ϕ_1 are known up to their parameters, thus the change $\phi_0 \rightarrow \phi_1$ corresponds to a change $\theta_0 \rightarrow \theta_1$

Change-Point Methods (CPM) are **sequential** monitoring schemes that **extend** traditional **parametric hypothesis tests**

These assumptions typically hold in **quality control** applications, but not in applications **where** the **change is unpredictable** (e.g. it is not known which parameter will be affected)

Hawkins, D. M., and Zamba, K. D. *"Statistical process control for shifts in mean or variance using a changepoint formulation"* Technometrics 2005

Ross, G. J. "Sequential change detection in the presence of unknown parameters". Statistics and Computing, 24(6), 1017-1030, 2014

NONPARAMETRIC SETTINGS

Both ϕ_0 and ϕ_1 are unknown, thus **the change** $\phi_0 \to \phi_1$ is **completely unpredictable**

Typical statistics:

- **Nonparametric statistics**, like the Mann-Whitney, Mood, Lepage, Cramer von Mises, Kolmogorov-Smirnov
- **Feature-extraction** to bring stationary data to some known distribution (e.g. the Box-Cox Transform)

Ross, G. J., Tasoulis, D. K., Adams, N. M. *"Nonparametric monitoring of data streams for changes in location and scale"* Technometrics, 53(4), 379-389, 2012.

Alippi, C., Boracchi, G., Roveri, M. *"Change detection tests using the ICI rule"* Proceedings of IJCNN 2010 (pp. 1-7).

NONPARAMETRIC SETTINGS

Both ϕ_0 and ϕ_1 are unknown, thus **the change** $\phi_0 \rightarrow \phi_1$ is **completely unpredictable**

Typical decision rules like:

- **CPM** which can control the ARL_0
- **CUSUM** to detect changes in the expectation of the statistic
- **ICI rule** or other critieria to yield a sequential decision

Unfortunately **most** nonparametric statistics and the decision rules **do not natively apply to multivariate data.**

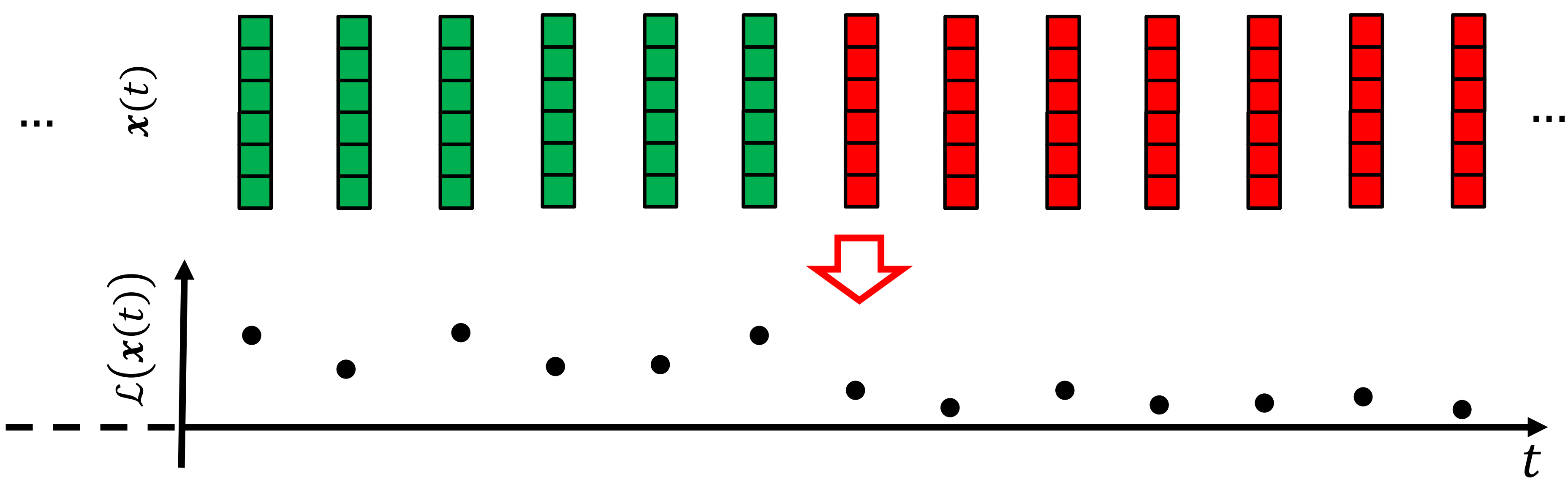
Ross, G. J., Tasoulis, D. K., Adams, N. M. "*Nonparametric monitoring of data streams for changes in location and scale*" Technometrics, 53(4), 379-389, 2012.

Alippi, C., Boracchi, G., Roveri, M. *“Change detection tests using the ICI rule”* Proceedings of IJCNN 2010 (pp. 1-7).

Tartakovsky, A. G., Veeravalli, V. V. "*Change-point detection in multichannel and distributed systems*". Applied Sequential Methodologies: Real-World Examples with Data Analysis, 173, 339-370, 2004

A typical approach is to reduce data dimension by monitoring the log-likelihood of normal data (as in density-based methods)

1. During training, estimate $\hat{\phi}_0$ from TR
2. During testing, compute
$$\mathcal{L}\big(\boldsymbol{x}(t)\big) = \log(\hat{\phi}_0(\boldsymbol{x}(t)))$$
3. Monitor $\{\mathcal{L}\big(\boldsymbol{x}(t)\big),\ t = 1, \dots\}$



MONITORING THE LOG-LIKELIHOOD

A typical approach is to reduce data dimension by monitoring the log-likelihood of normal data (as in density-based methods)

1. During training, estimate $\hat{\phi}_0$ from TR
2. During testing, compute
$$\mathcal{L}\big(\boldsymbol{x}(t)\big) = \log(\hat{\phi}_0(\boldsymbol{x}(t)))$$
3. Monitor $\{\mathcal{L}\big(\boldsymbol{x}(t)\big),\ t = 1, \dots\}$

This is quite a popular approach in sequential monitoring and in anomaly detection

L. I. Kuncheva, *"Change detection in streaming multivariate data using likelihood detectors,"* IEEE Transactions on Knowledge and Data Engineering, vol. 25, no. 5, 2013.

X. Song, M. Wu, C. Jermaine, and S. Ranka, *"Statistical change detection for multidimensional data,"* in Proceedings of International Conference on Knowledge Discovery and Data Mining (KDD), 2007.

J. H. Sullivan and W. H. Woodall, *"Change-point detection of mean vector or covariance matrix shifts using multivariate individual observations,"* IIE transactions, vol. 32, no. 6, 2000.

C. Alippi, G. Boracchi, D. Carrera, M. Roveri, "*Change Detection in Multivariate Datastreams: Likelihood and Detectability Loss*" IJCAI 2016, New York, USA, July 9 - 13

HIERARCHICAL CHANGE-DETECTION TESTS

In nonparametric sequential monitoring it is convenient to

- **online sequential CDTs** for detection purposes
- **offline hypothesis tests** for validation purposes.

Alippi, C., Boracchi, G., Roveri, M. "Hierarchical Change-Detection Tests" TNNLS 2016 pp 1- 13

HIERARCHICAL CHANGE-DETECTION TESTS

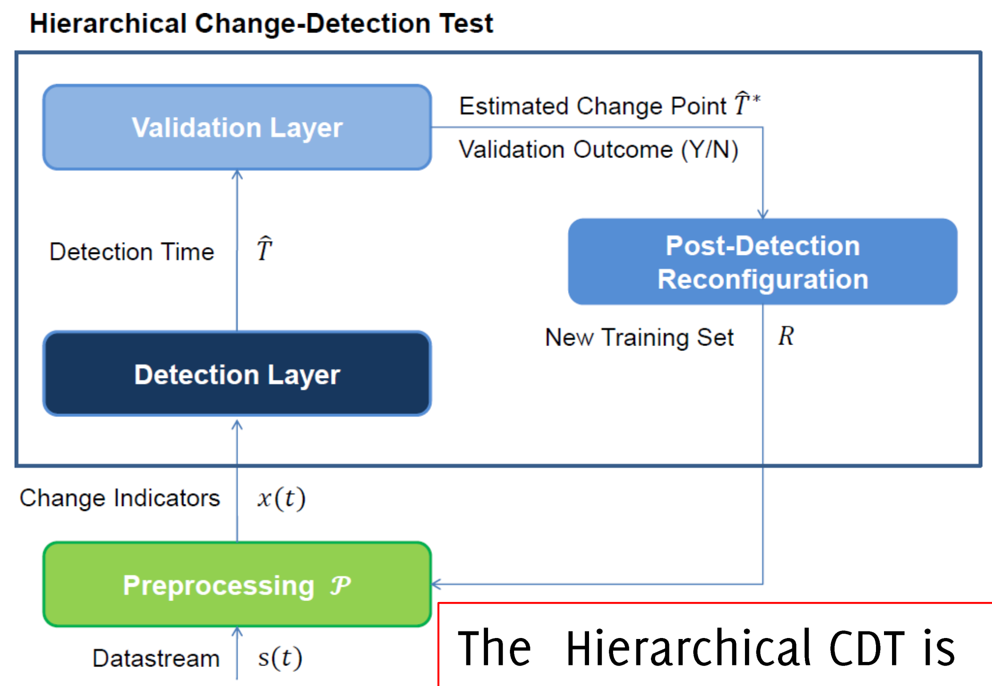
In nonparametric sequential monitoring it is convenient to

- **online sequential CDTs** for detection purposes
- **offline hypothesis tests** for validation purposes.

This results in two-layered (hierarchical) CDTs

Offline HT is activated to validate any detection

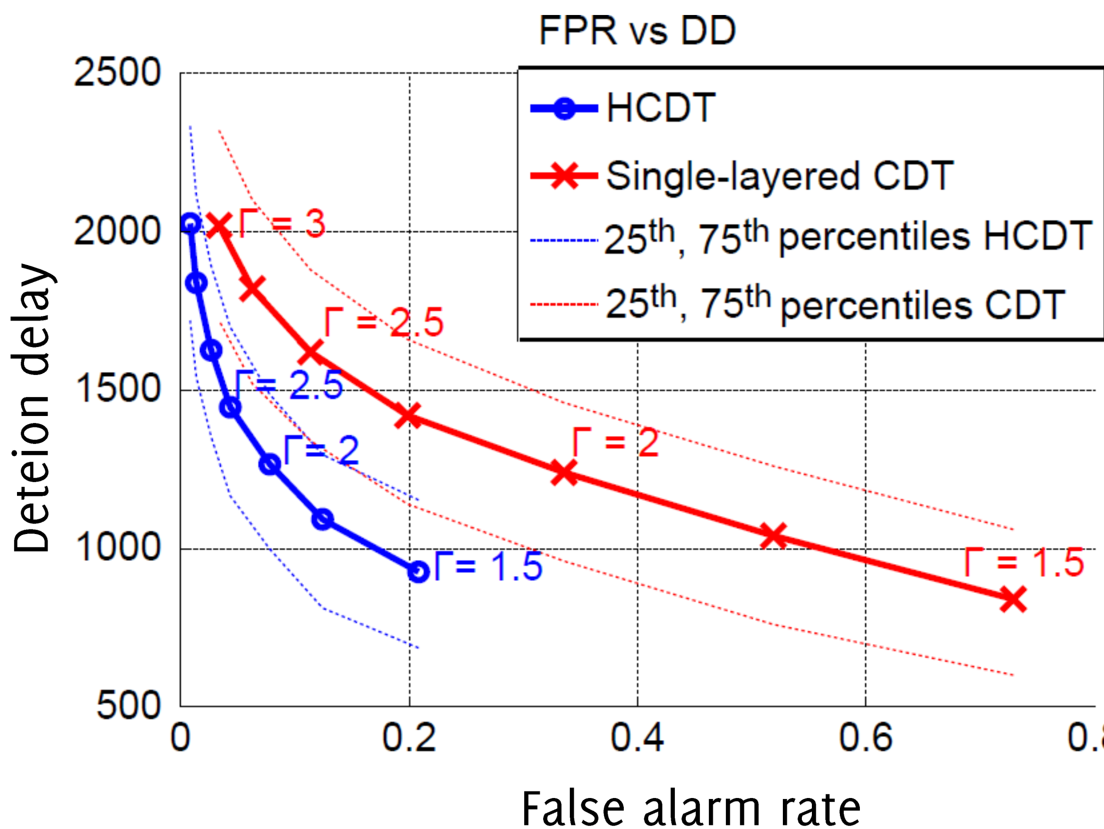
Online CDT detects process changes in the input datastream



The Hierarchical CDT is automatically reconfigured

Alippi, C., Boracchi, G., Roveri, M. "Hierarchical Change-Detection Tests" TNNLS 2016 pp 1- 13

Hierarchical CDTs can achieve a far **more advantageous trade-off** between false-positive rate and detection delay **than their single-layered**, more traditional, **counterpart.**



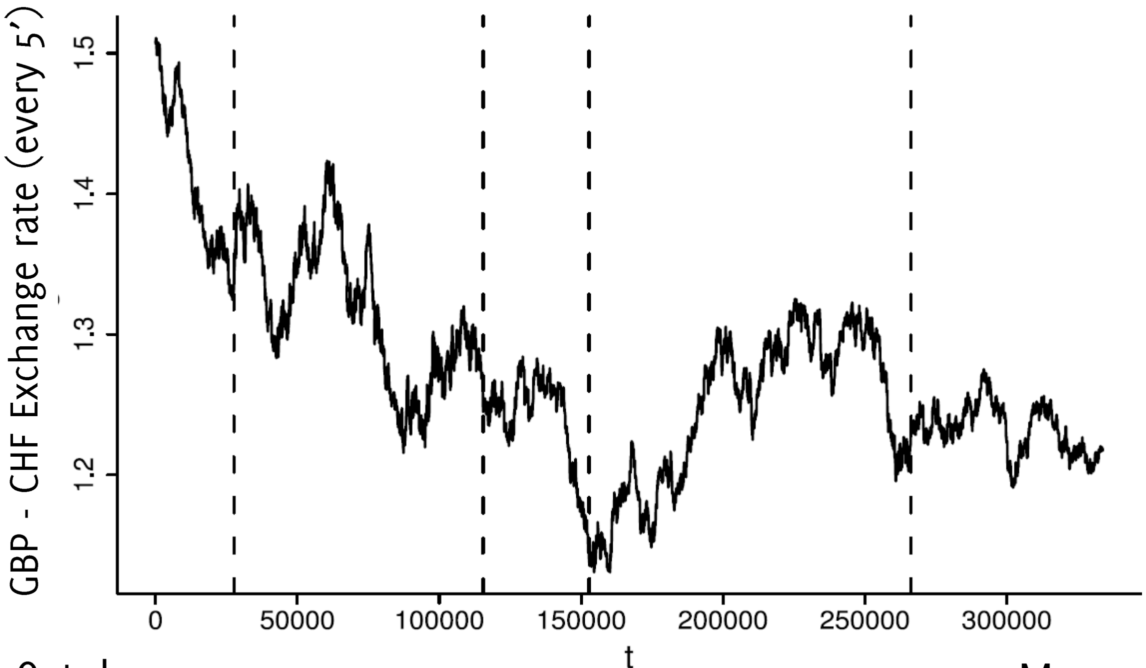
Alippi, C., Boracchi, G., Roveri, M. *"Hierarchical Change-Detection Tests"* TNNLS 2016 pp 1- 13

CHANGE/ANOMALY DETECTION OUT OF THE RANDOM VARIABLE WORLD

… monitoring signals, images, …

THE LIMITATIONS OF THE RANDOM VARIABLE MODEL

Often **data are in the form of time series**, and are not i.i.d



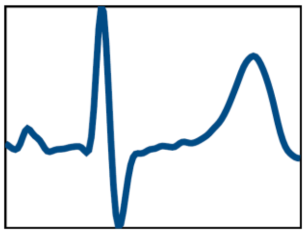
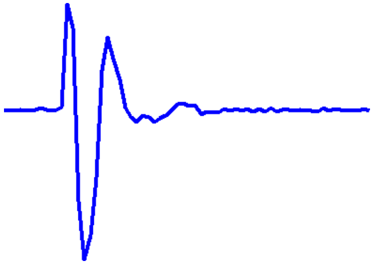
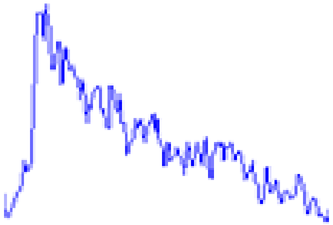
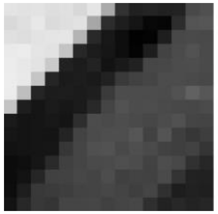
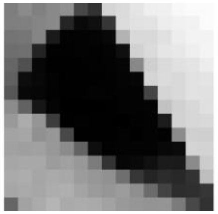
October 2002 May 2007

Data are **clearly correlated over time**

Changes in the data **correlation** are the most **important** ones

THE LIMITATIONS OF THE RANDOM VARIABLE MODEL

Random variable model **does not apply on signal / images**

Stacking each signal in a vector is not convenient:

- **Data dimension** becomes **huge**
- **Correlation among components is difficult to model**

Often **normal data** exhibit some form **of structure**. Thus,

- Normal data live in a **low-dimensional space**
- **Dimensionality reduction can be applied**

We are interested in **changes/anomalies affecting structures**

… OUT OF RANDOM VARIABLE WORLD

Typical approach: *Fit a statistical model to the observation to **model dependence**, **apply change-detection** on the independent **residuals**.*

The change/anomaly detection methods will tell whether **incoming data fit or not the normal model**

This can be done by

- **Detrending/Filtering**: remove the deterministic and correlated components of the data
- **Feature extraction**: meaningful indicators to be monitored which have a known / controlled response to normal data

FEATURE EXTRACTION FOR CHANGE-ANOMALY DETECTION

Features can be either:

- **Expert-driven**, or manually crafted
- **Data-driven,** or learned from data

And can be used in one-shot/sequential monitoring schemes

V. Chandola, A. Banerjee, V. Kumar. "*Anomaly detection: A survey*". ACM Comput. Surv. 41, 3, Article 15 (2009), 58 pages.

FEATURE EXTRACTION FOR CHANGE-ANOMALY DETECTION

Features can be either:

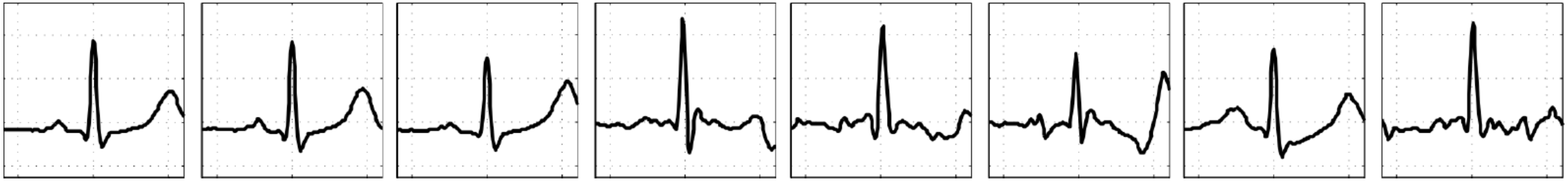
- **Expert-driven**, or manually crafted
- **Data-driven,** or learned from data

And can be used in one-shot/sequential monitoring schemes

Data-driven features are typically **obtained form a model** $\mathcal{M}$ which **represents normal data**

- **During training:** learn the model $\mathcal{M}$ from TR
- **During testing:** assess whether $\boldsymbol{x}$ conforms or not $\mathcal{M}$

Dictionary learned from normal ECG signal (sparse representations)



V. Chandola, A. Banerjee, V. Kumar. "*Anomaly detection: A survey*". ACM Comput. Surv. 41, 3, Article 15 (2009), 58 pages.

RECONSTRUCTION-BASED APPROACHES

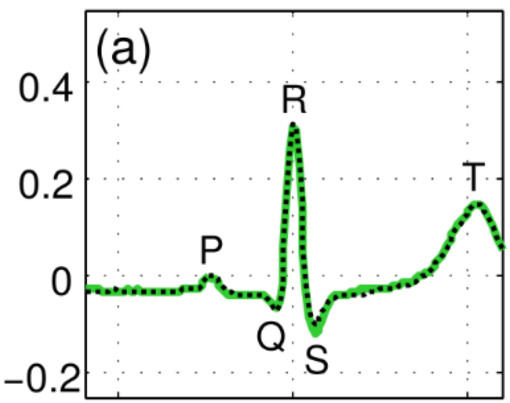
The most widely adopted features are the **residuals**, which involve computing α, the coefficients of the representation of $\boldsymbol{x}$ w.r.t $\mathcal{M}$

$$r(t) = \|\boldsymbol{x} - \mathcal{M}(\alpha)\|_2$$

Very popular models are: autoregressive models, neural networks (auto-encoders), sparse representations

Example of reconstruction based on sparse representations

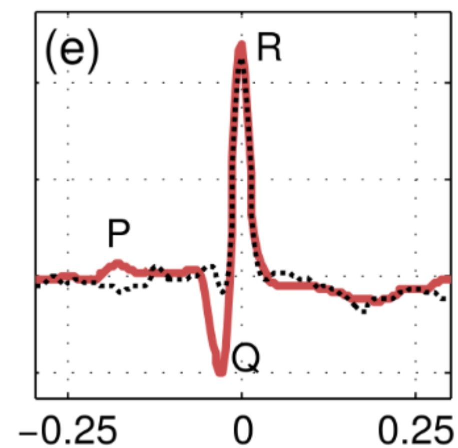
Normal data: good reconstruction



dotted line: $\mathcal{M}\boldsymbol{\alpha}$

solid line : $\boldsymbol{x}$

Anomalous data: poor reconstruction



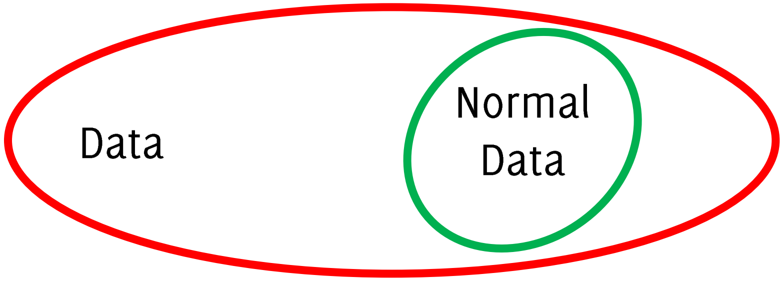
V. Chandola, A. Banerjee, V. Kumar. "*Anomaly detection: A survey*". ACM Comput. Surv. 41, 3, Article 15 (2009), 58 pages.

C. Alippi, G. Boracchi, B. Wohlberg, *“Change Detection in Streams of Signals with Sparse Representations “* IEEE ICASSP 2014

SUBSPACE METHODS

Learn a model describing normal data and project test data into it.

- PCA / Robust PCA / kernel PCA : learn a linear subspace where normal data live
- Sparse representations: learn a union of low-dimensional subspaces where normal data live
- Kernel methods



V. Chandola, A. Banerjee, V. Kumar. "*Anomaly detection: A survey*". ACM Comput. Surv. 41, 3, Article 15 (2009), 58 pages.

C. Alippi, G. Boracchi, B. Wohlberg, *"Change Detection in Streams of Signals with Sparse Representations "* IEEE ICASSP 2014

BIG DATA CHALLENGES

When performing change/anomaly in the random-variable word

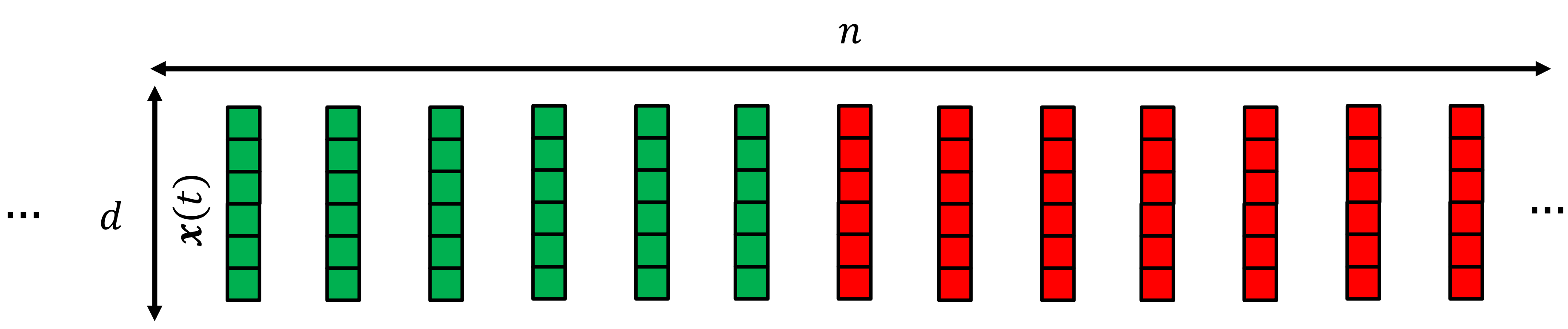
WHEN DATA DIMENSION GROWS

When n (or the data throughput) grows:

- **Memory issues:** not feasible to store all the data in memory
- **Computational issues**: algorithms should be $\mathcal{O}(1)$, and single-pass
- **Having a lot of training samples** is good!

Thus, there is need for

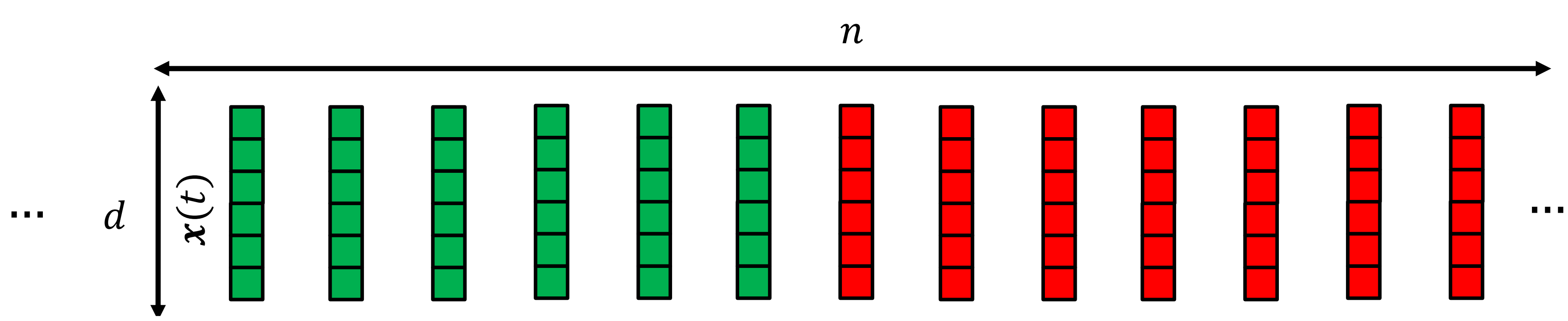
- approximated statistics
- Incremental formulas, dataset pruning



WHEN DATA DIMENSION GROWS

When d grows:

- **Memory issues:** not feasible to store many data in memory
- Difficult to **find a model $\hat{\phi}_0$, many training samples** needed
- Number of irrelevant component might increase
- Distance-based methods are difficult to tune
- Combinatorial growth of the number of subspaces
- Data-visualization issues
- **Detectability loss**



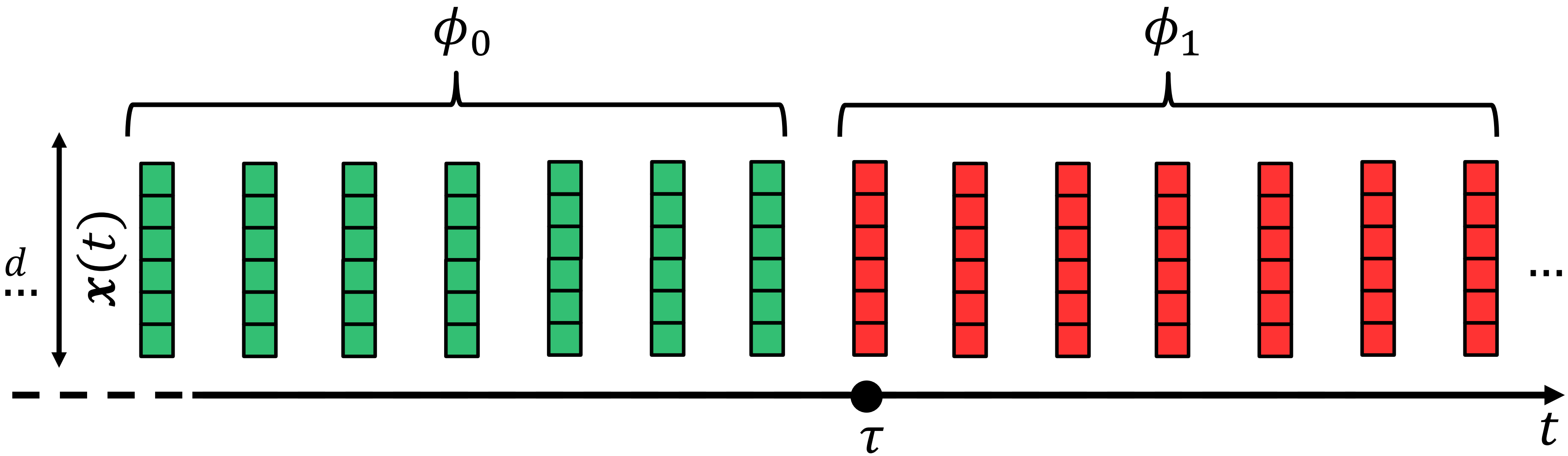
A. Zimek, E. Schubert, H.P. Kriegel. *"A survey on unsupervised outlier detection in high-dimensional numerical data"* Statistical Analysis and Data Mining: The ASA Data Science Journal, 5(5), 2012.

DETECTABILITY LOSS IN HIGH-DIMENSIONAL DATA

How data dimension affects monitoring the Log-likelihood

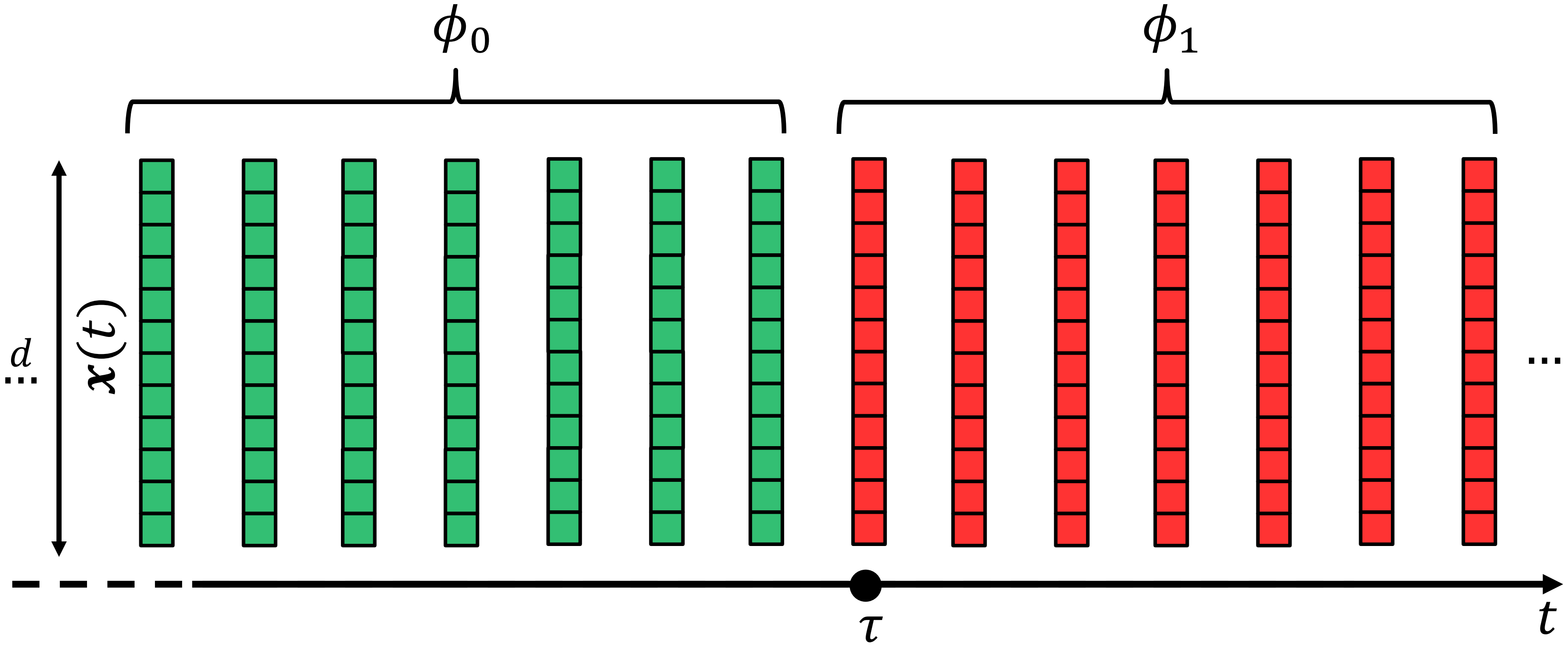
C. Alippi, G. Boracchi, D. Carrera, M. Roveri, "*Change Detection in Multivariate Datastreams: Likelihood and Detectability Loss*" IJCAI 2016, New York, USA, July 9 - 13

*Study how the **data dimension** d **influences** the **change detectability**, i.e., how difficult is to solve change/anomaly detection problems*



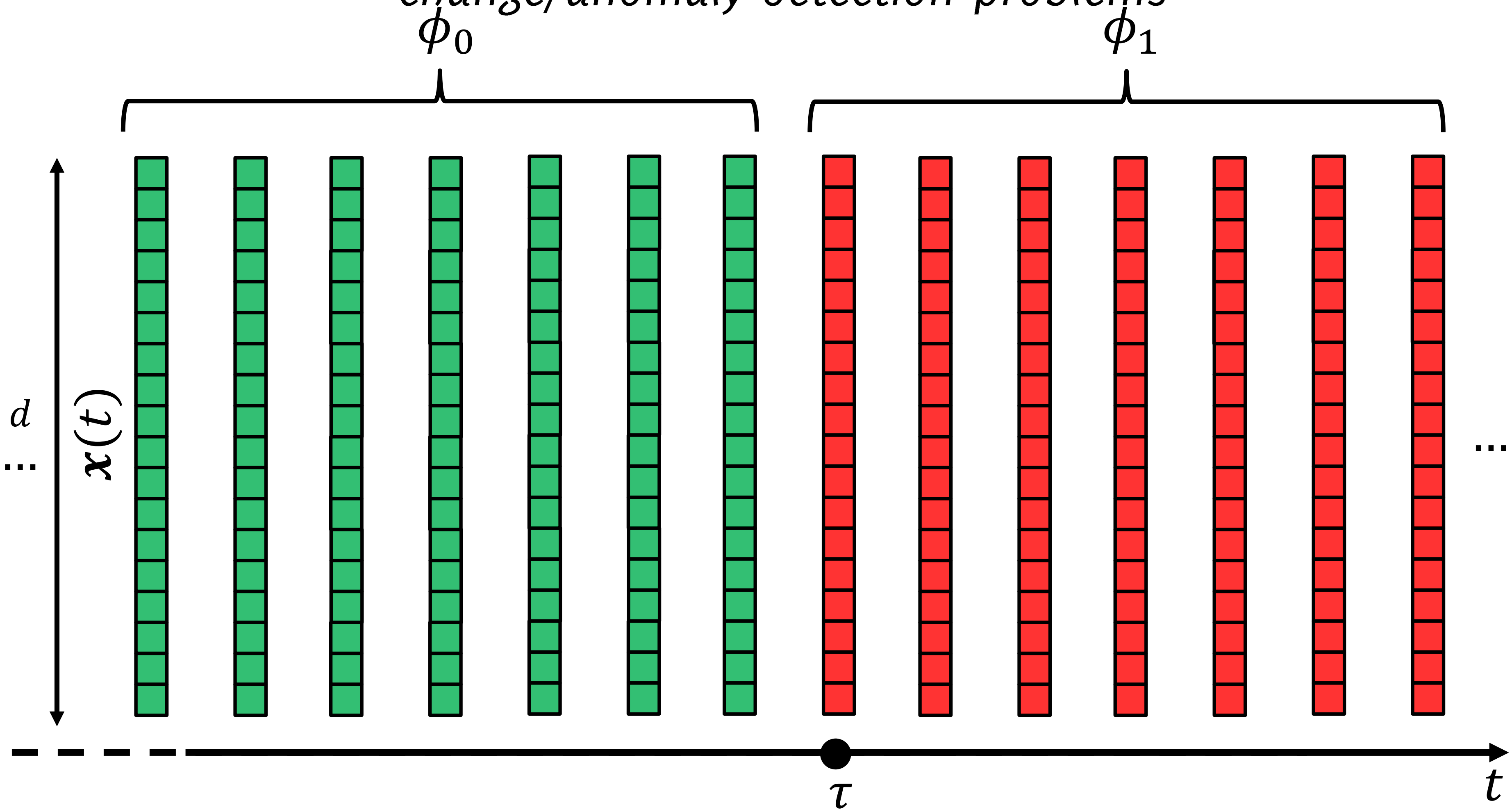
OUR GOAL

Study how the ***data dimension*** d ***influences*** *the* ***change detectability****, i.e., how difficult is to solve change/anomaly detection problems*



OUR GOAL

Study how the ***data dimension*** d ***influences*** *the* ***change detectability****, i.e., how difficult is to solve change/anomaly detection problems*



OUR APPROACH

To study the impact of the **sole data dimension** d in **change-detection problems** we need to:

1. Consider a **change-detection approach**
2. Define a measure of **change detectability** that well correlates with traditional performance measures
3. Define a measure of **change magnitude** that refers only to differences between ϕ_0 and ϕ_1

OUR APPROACH

To study the impact of the **sole data dimension** d in **change-detection problems** we need to:

1. Consider a **change-detection approach**
2. Define a measure of **change detectability** that well correlates with traditional performance measures
3. Define a measure of **change magnitude** that refers only to differences between ϕ_0 and ϕ_1

Our goal (reformulated):
Studying how the **change detectability varies** in **change-detection problems** that have

- **different data dimensions** d
- **constant change magnitude**

OUR RESULT

We show there is a **detectability loss** problem, i.e. that change **detectability** steadily **decreases** when d increases.

Detectability loss is shown by:

- Analytical derivations: when ϕ_0 and ϕ_1 are **Gaussians**
- Empirical analysis: measuring the **power of hypothesis tests** in change-detection problems on real data

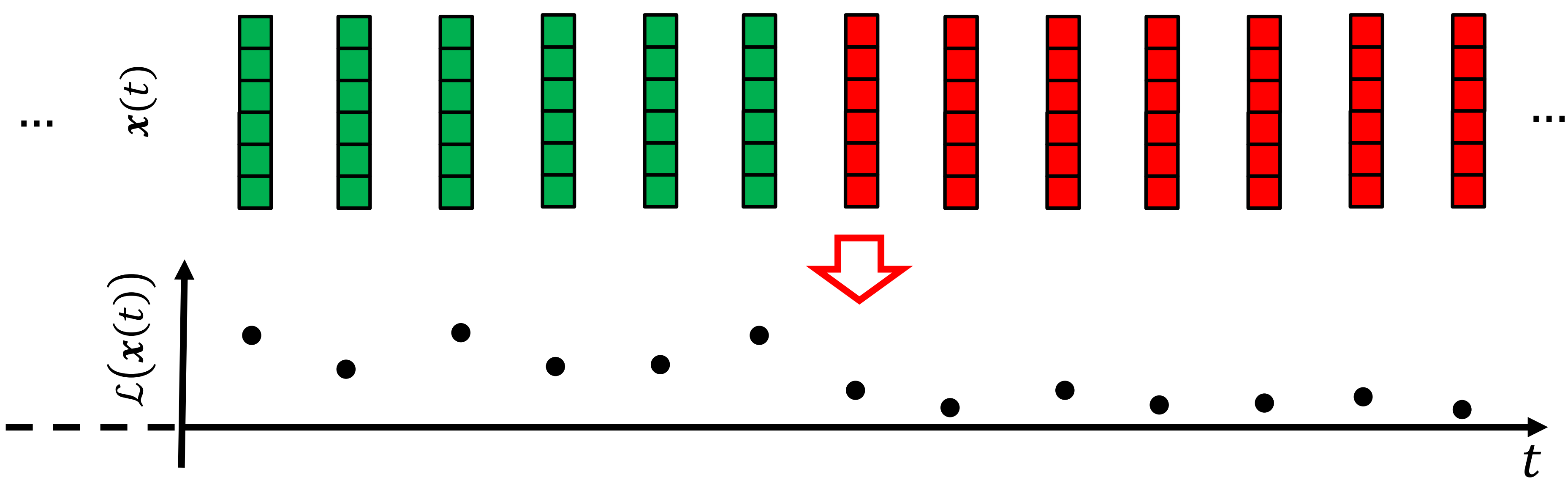
ROADMAP TO DETECTABILITY LOSS

- Preliminaries:
  - The change-detection approach
  - The measure of change detectability
  - The change magnitude
- The *detectability loss*
  - Analytical results
  - Empirical analysis

HOW? MONITORING THE LOG-LIKELIHOOD

A typical approach to monitor the log-likelihood

1. During training, estimate $\hat{\phi}_0$ from TR
2. During testing, compute
$$\mathcal{L}\big(\boldsymbol{x}(t)\big) = \log(\hat{\phi}_0(\boldsymbol{x}(t)))$$
3. Monitor $\{\mathcal{L}\big(\boldsymbol{x}(t)\big),\ t = 1, \dots\}$



ROADMAP TO DETECTABILITY LOSS

- Preliminaries:
  - The change-detection approach
  - The measure of change detectability
  - The change magnitude
- The *detectability loss*
  - Analytical results
  - Empirical analysis

THE CHANGE DETECTABILITY

The *Signal to Noise Ratio of the change*

$$\mathrm{SNR}(\phi_0 \to \phi_1) = \frac{\left(\underset{\boldsymbol{x}\sim\phi_0}{\mathrm{E}}[\mathcal{L}(\boldsymbol{x})] - \underset{\boldsymbol{x}\sim\phi_1}{\mathrm{E}}[\mathcal{L}(\boldsymbol{x})]\right)^2}{\underset{\boldsymbol{x}\sim\phi_0}{\mathrm{var}}[\mathcal{L}(\boldsymbol{x})] + \underset{\boldsymbol{x}\sim\phi_1}{\mathrm{var}}[\mathcal{L}(\boldsymbol{x})]}$$

measures the extent to which $\phi_0 \to \phi_1$ is **detectable by statistical tools** designed to **detect changes** in $\mathrm{E}[\mathcal{L}(\boldsymbol{x})]$

ROADMAP TO DETECTABILITY LOSS

- Preliminaries:
  - The change-detection approach
  - The measure of change detectability
  - The change magnitude
- The *detectability loss*
  - Analytical results
  - Empirical analysis

THE CHANGE MAGNITUDE

We measure the **magnitude of a change** $\phi_0 \rightarrow \phi_1$ by the *symmetric Kullback-Leibler divergence*

$$\mathrm{sKL}(\phi_0, \phi_1) = \mathrm{KL}(\phi_0, \phi_1) + \mathrm{KL}(\phi_1, \phi_0) =$$
$$= \int \log\left(\frac{\phi_0(\boldsymbol{x})}{\phi_1(\boldsymbol{x})}\right) \phi_0(\boldsymbol{x}) d\boldsymbol{x} + \int \log\left(\frac{\phi_1(\boldsymbol{x})}{\phi_0(\boldsymbol{x})}\right) \phi_1(\boldsymbol{x}) d\boldsymbol{x}$$

In practice, **large values** of $\mathrm{sKL}(\phi_0, \phi_1)$ correspond to **changes** $\phi_0 \rightarrow \phi_1$that are very apparent, since $\mathrm{sKL}(\phi_0, \phi_1)$ identifies an upperbound of the power of hypothesis tests designed to detect either $\phi_0 \rightarrow \phi_1$ or $\phi_1 \rightarrow \phi_0$

T. Dasu, K. Shankar, S. Venkatasubramanian, K. Yi, "*An information-theoretic approach to detecting changes in multi-dimensional data streams*" In Proc. Symp. on the Interface of Statistics, Computing Science, and Applications, 2006

ROADMAP TO DETECTABILITY LOSS

- Preliminaries:
  - The change-detection approach
  - The measure of change detectability
  - The change magnitude
- The *detectability loss*
  - Analytical results
  - Empirical analysis

THE DETECTABILITY LOSS

Theorem

Let $\phi_0 = \mathcal{N}(\mu_0, \Sigma_0)$ *and let* $\phi_1(\boldsymbol{x}) = \phi_0(Q\boldsymbol{x} + \boldsymbol{v})$ *where* $Q \in \mathbb{R}^{d \times d}$ *and orthogonal* , $\boldsymbol{v} \in \mathbb{R}^d$, *then*

$$\mathrm{SNR}(\phi_0 \to \phi_1) < \frac{C}{d}$$

Where C *is a constant that depends only on* $\mathrm{sKL}(\phi_0, \phi_1)$

C. Alippi, G. Boracchi, D. Carrera, M. Roveri, "*Change Detection in Multivariate Datastreams: Likelihood and Detectability Loss*" IJCAI 2016, New York, USA, July 9 - 13

THE DETECTABILITY LOSS: REMARKS

Theorem

Let $\phi_0 = \mathcal{N}(\mu_0, \Sigma_0)$ *and let* $\phi_1(\boldsymbol{x}) = \phi_0(Q\boldsymbol{x} + \boldsymbol{v})$ *where* $Q \in \mathbb{R}^{d \times d}$ *and orthogonal* , $\boldsymbol{v} \in \mathbb{R}^d$*, then*

$$\mathrm{SNR}(\phi_0 \to \phi_1) < \frac{C}{d}$$

Where C *is a constant that depends only on* $\mathrm{sKL}(\phi_0, \phi_1)$

Remarks:

- Changes of a given magnitude, $\mathrm{sKL}(\phi_0, \phi_1)$, become more difficult to detect when d increases
- DL does not depend on how ϕ_0 changes
- DL does not depend on the specific detection rule
- DL does not depend on estimation errors on $\hat{\phi}_0$

THE DETECTABILITY LOSS: THE CHANGE MODEL

Theorem

Let $\phi_0 = \mathcal{N}(\mu_0, \Sigma_0)$ *and let* $\phi_1(\boldsymbol{x}) = \phi_0(Q\boldsymbol{x} + \boldsymbol{v})$ *where* $Q \in \mathbb{R}^{d\times d}$ *and orthogonal ,* $\boldsymbol{v} \in \mathbb{R}^d$*, then*

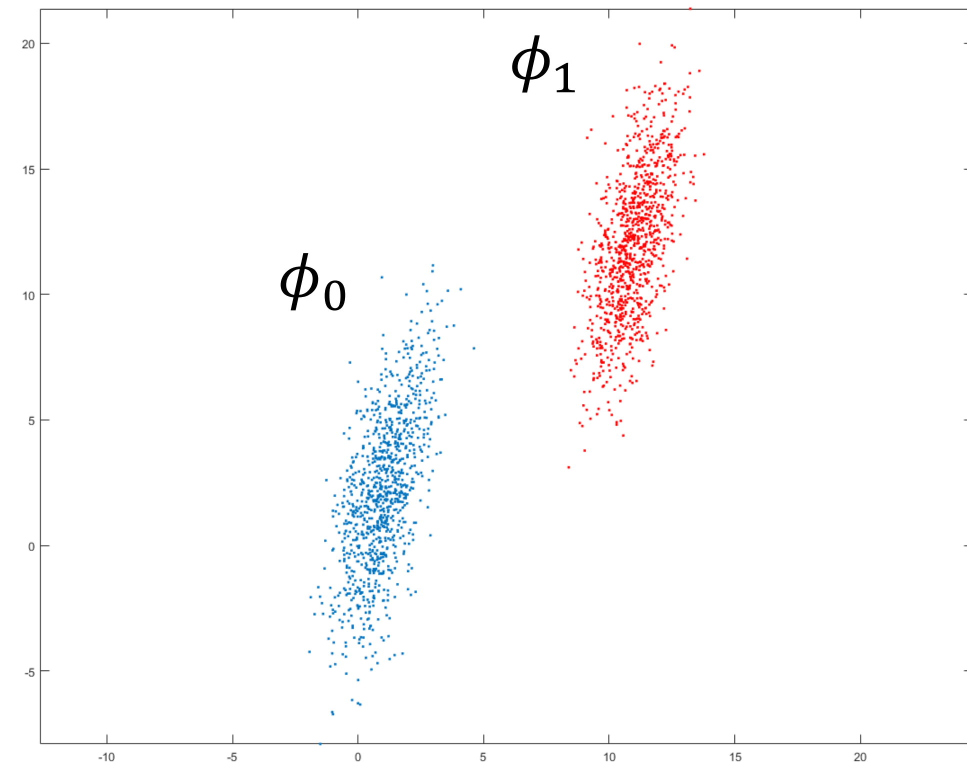
$$\mathrm{SNR}(\phi_0 \to \phi_1) < \frac{C}{d}$$

Where C *is a constant that depends only on* $\mathrm{sKL}(\phi_0, \phi_1)$

THE DETECTABILITY LOSS: THE CHANGE MODEL

The change model $\phi_1\,(\boldsymbol{x}) = \phi_0(Q\boldsymbol{x}\, + \boldsymbol{v})$ includes:

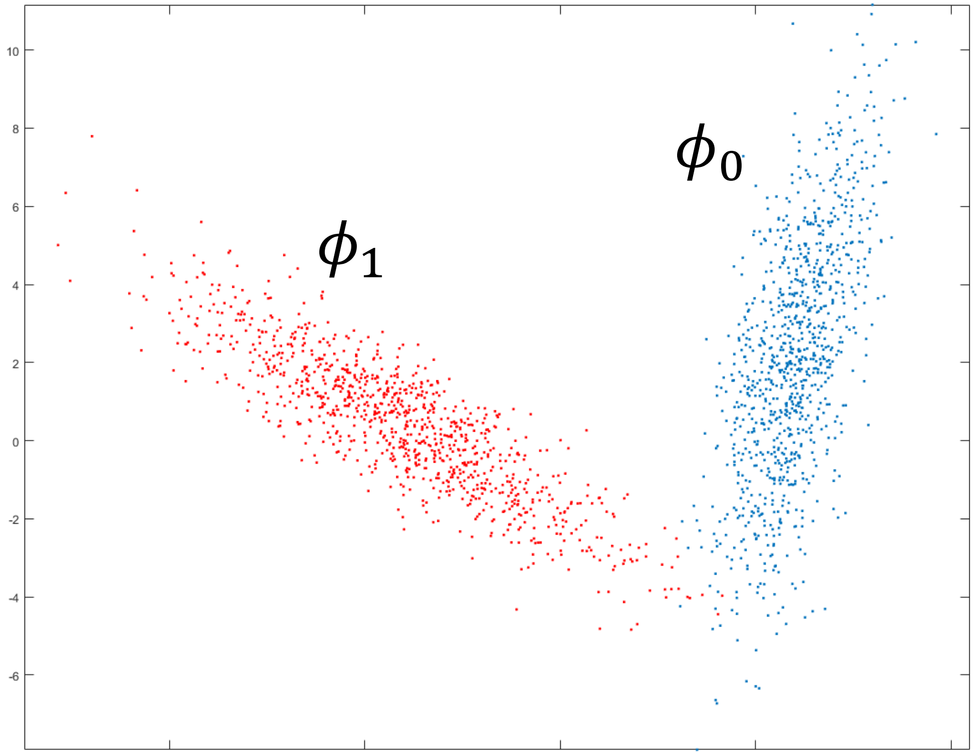
- Changes in the location of ϕ_0 (i.e, $+\boldsymbol{v}$)



THE DETECTABILITY LOSS: THE CHANGE MODEL

The change model $\phi_1(\boldsymbol{x}) = \phi_0(Q\boldsymbol{x} + \boldsymbol{v})$ includes:

- Changes in the location of ϕ_0 (i.e, $+\boldsymbol{v}$)
- Changes in the correlation of $\boldsymbol{x}$ (i.e, $Q\boldsymbol{x}$)

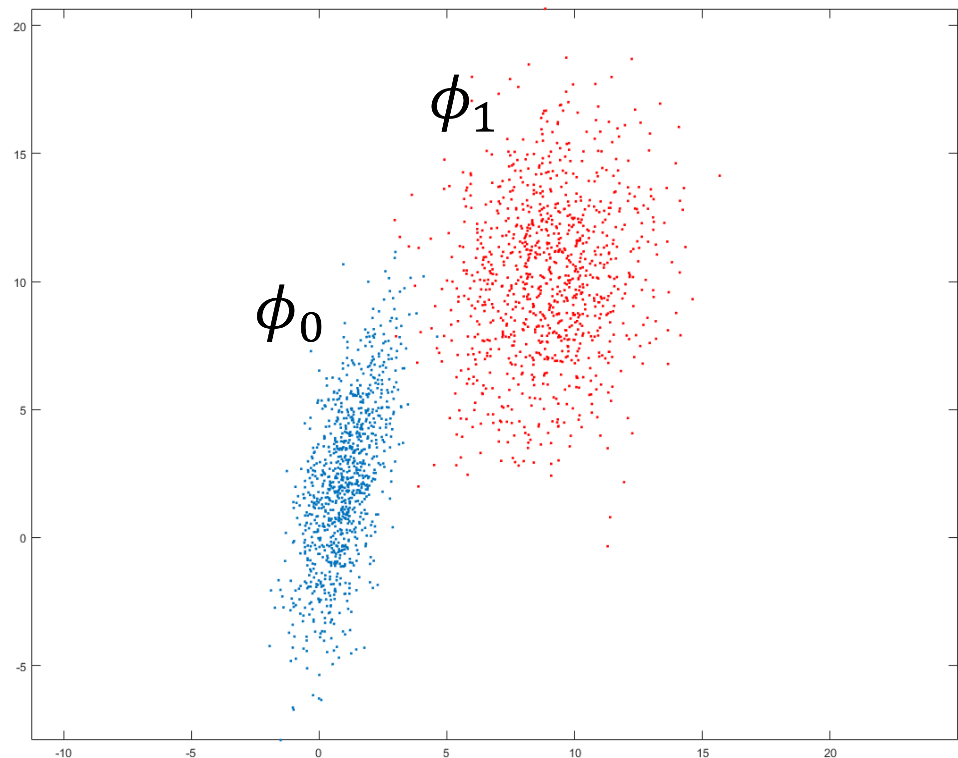


THE DETECTABILITY LOSS: THE CHANGE MODEL

The change model $\phi_1(\boldsymbol{x}) = \phi_0(Q\boldsymbol{x} + \boldsymbol{v})$ includes:

- Changes in the location of ϕ_0 (i.e, $+\boldsymbol{v}$)
- Changes in the correlation of $\boldsymbol{x}$ (i.e, $Q\boldsymbol{x}$)

It does not include changes in the scale of ϕ_0 that can be however detected monitoring $||\boldsymbol{x}||$



THE DETECTABILITY LOSS: THE GAUSSIAN ASSUMPTION

Theorem

Let $\phi_0 = \mathcal{N}(\mu_0, \Sigma_0)$ *and let* $\phi_1(\boldsymbol{x}) = \phi_0(Q\boldsymbol{x} + \boldsymbol{v})$ *where* $Q \in \mathbb{R}^{d \times d}$ *and orthogonal ,* $\boldsymbol{v} \in \mathbb{R}^d$*, then*

$$\mathrm{SNR}(\phi_0 \to \phi_1) < \frac{C}{d}$$

Where C *is a constant that depends only on* $\mathrm{sKL}(\phi_0, \phi_1)$

THE DETECTABILITY LOSS: THE GAUSSIAN ASSUMPTION

Assuming $\phi_0 = \mathcal{N}(\mu_0, \Sigma_0)$ looks like a severe limitation.

- Other distributions are not easy to handle analytically
- We can prove that DL occurs also in random variables having independent components
- The result can be empirically extended to approximations of $\mathcal{L}(\cdot)$ typically used for Gaussian mixtures

ROADMAP TO DETECTABILITY LOSS

- Preliminaries:
  - The change-detection approach
  - The measure of change detectability
  - The change magnitude
- The *detectability loss*
  - Analytical results
  - Empirical analysis

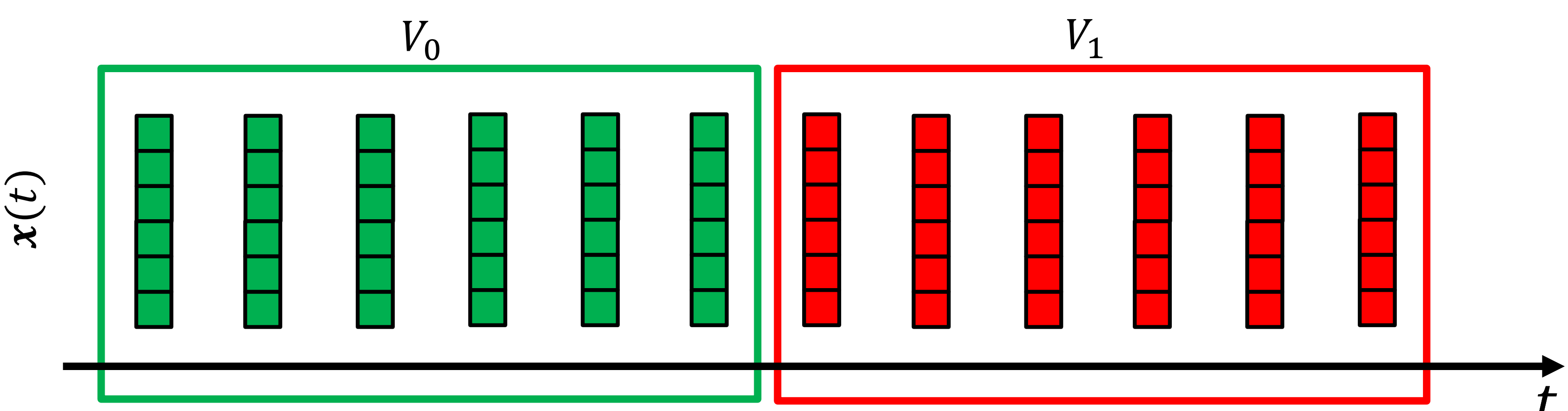
THE DETECTABILITY LOSS: EMPIRICAL ANALYSIS

The data

- Synthetically generate streams having different dimension d
- Estimate $\hat{\phi}_0$ by GM from a **stationary training set**
- In each stream we introduce $\phi_0 \rightarrow \phi_1$ such that

$$\phi_1\,(\boldsymbol{x}) = \phi_0(Q\boldsymbol{x}\,+\boldsymbol{v}) \text{ and } \mathrm{sKL}(\phi_0,\phi_1) = 1$$

- **Test data: two windows** V_0 and V_1 (500 samples each) selected before and after the change.



THE DETECTABILITY LOSS: EMPIRICAL ANALYSIS

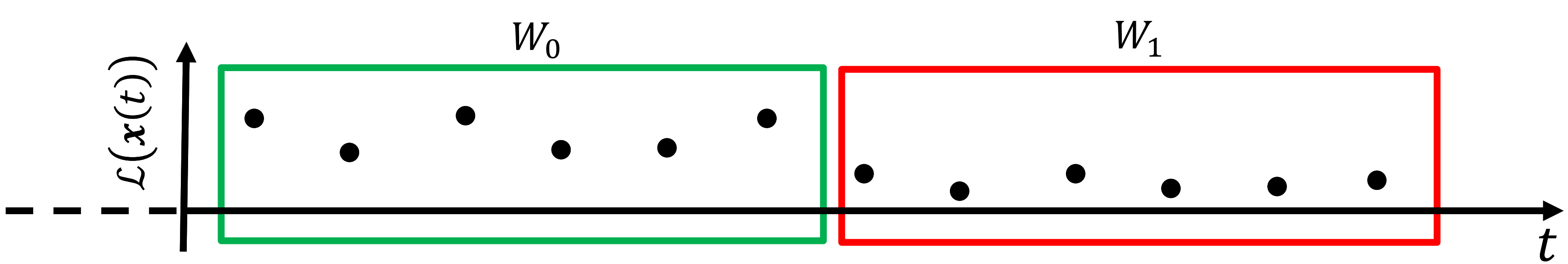
The change-detectability measure:

- Compute $\mathcal{L}\left(\hat{\phi}_0(\boldsymbol{x})\right)$ from V_0 and V_1, obtaining W_0 and W_1
- Compute a test statistic $\mathcal{T}(W_0, W_1)$ to compare the two
- Detect a change by an hypothesis test

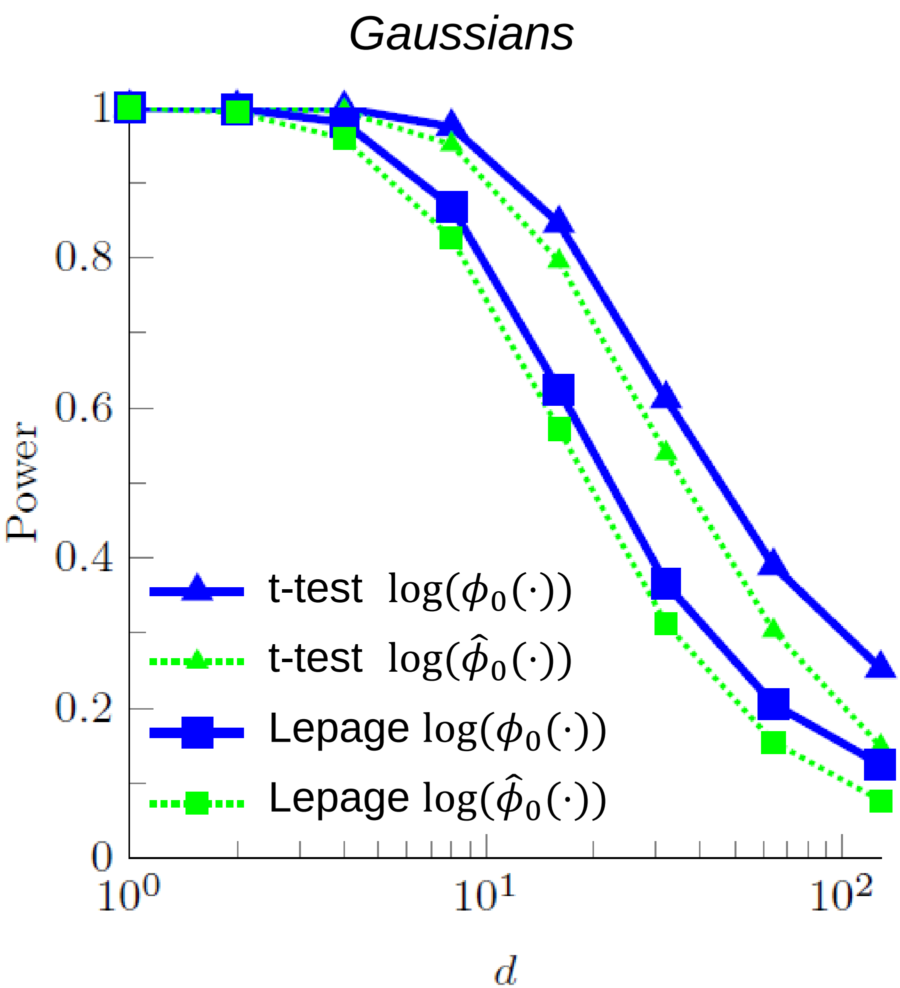
$$\mathcal{T}(W_0, W_1) \lesseqgtr h$$

where h controls the amount of false positives

- Use the **power** of this **test** to assess change detectability



DL: THE POWER OF HTS ON GAUSSIAN STREAMS



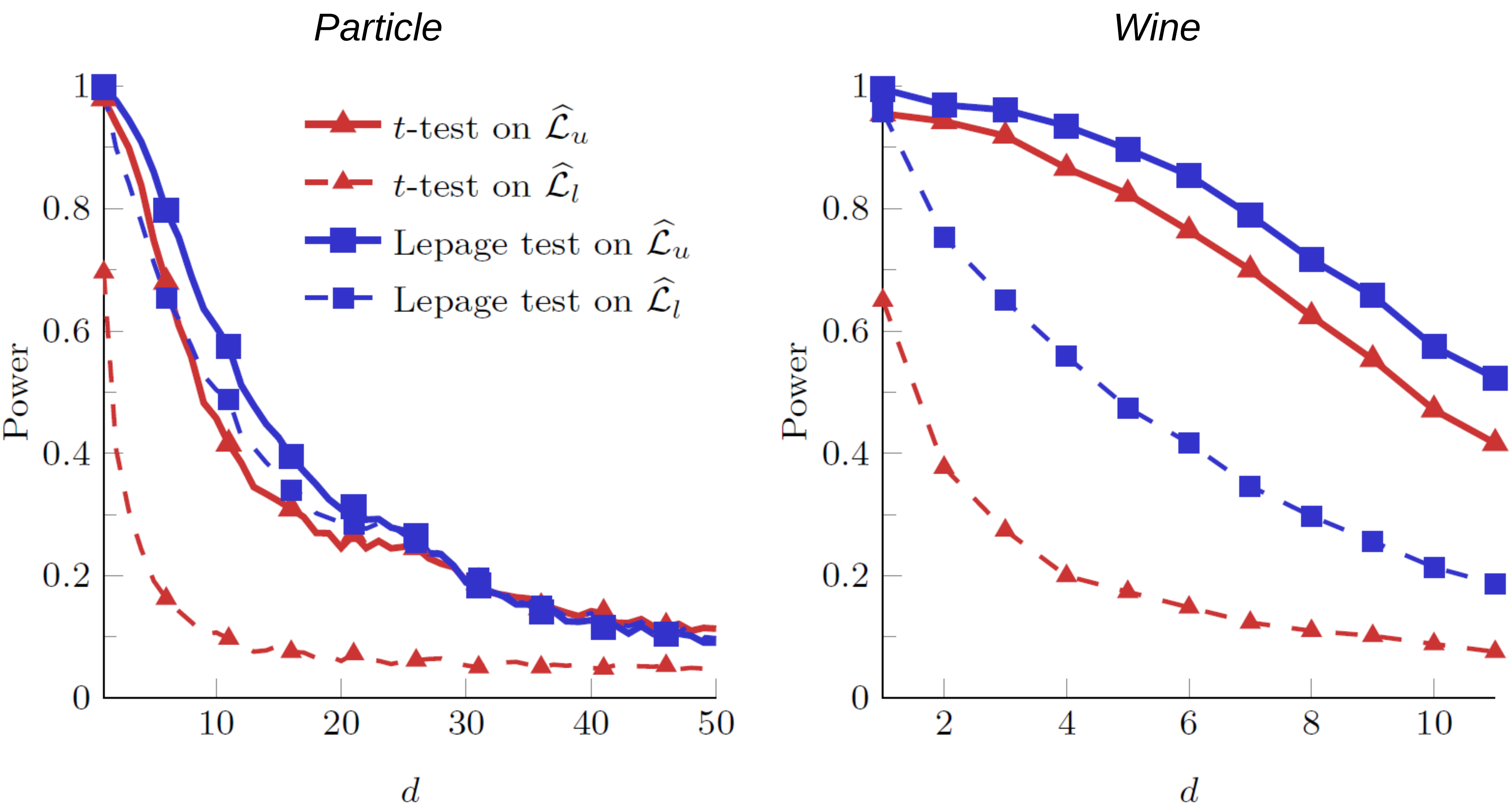
Remarks:

- ϕ_1 is defined analytically
- The t-test detects changes in expectation
- The Lepage test detects changes in the location and scale

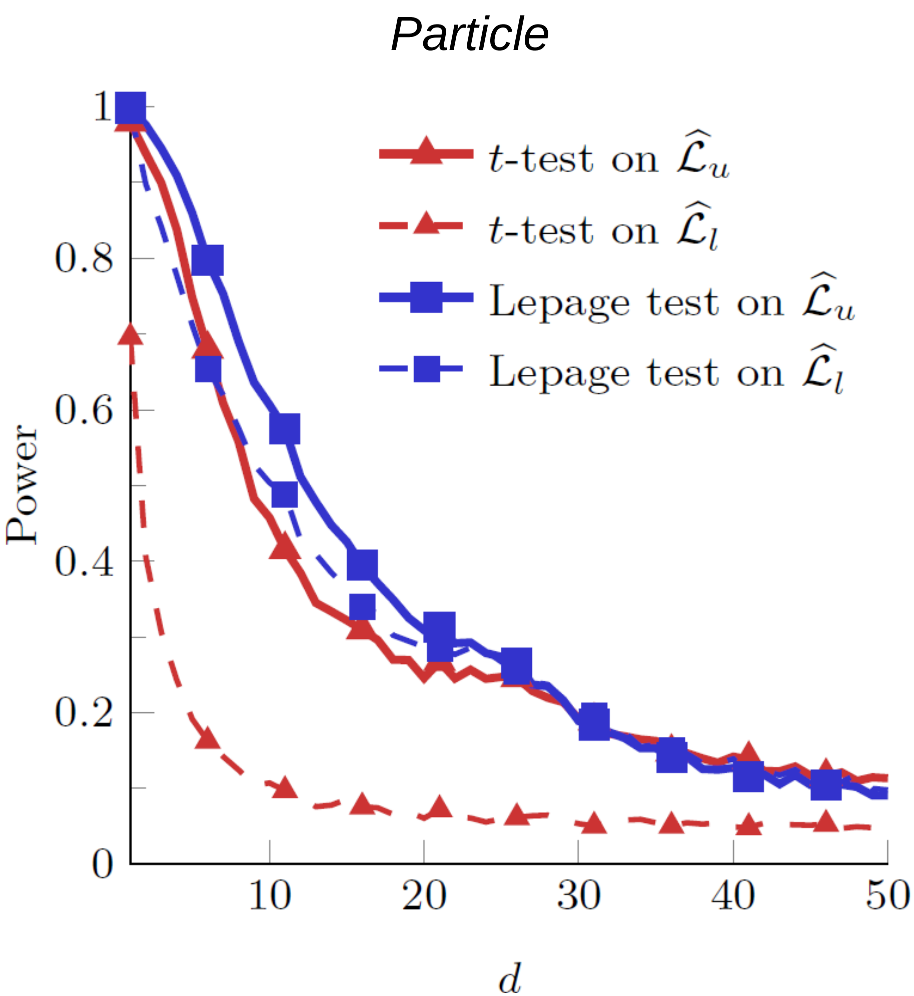
Results

- The HT power decays with d: DL does not only concern the upperbound of SNR.
- DL is not due to estimation errors, but these make things worst.
- The power of the Lepage HT also decreases, which indicates that the change is more difficult to detect also monitoring the variance

RESULTS: THE POWER OF THE HYPOTHESIS TESTS



RESULTS: THE POWER OF THE HYPOTHESIS TESTS



Remarks:

- ϕ_1 is defined through a numerical procedure to yield $\text{sKL}(\phi_0, \phi_1) \approx 1$
- $\hat{\phi}_0$ is a Gaussian Mixture where k is selected by cross-validation
- Approximated expression of $\mathcal{L}(\cdot)$ to prevent numerical approximations

Results:

- DL occurs also in non-Gaussian data approximated by GM
- DL is clearly visible at quite a low dimensions

C. Alippi, G. Boracchi, D. Carrera *"CCM: Controlling the Change Magnitude in High Dimensional Data"*, INNS Conference on Big Data, 10 pages, 2016 https://home.deib.polimi.it/carrerad/projects.html

CONCLUDING REMARKS

FEW CONCLUDING REMARKS

Change/Anomaly detection problems are very popular nowadays in engineering applications.

Most of the algorithms in the literature refer to the presented framework and often boil down to applying statistics and decision rules to a stream of random variables.

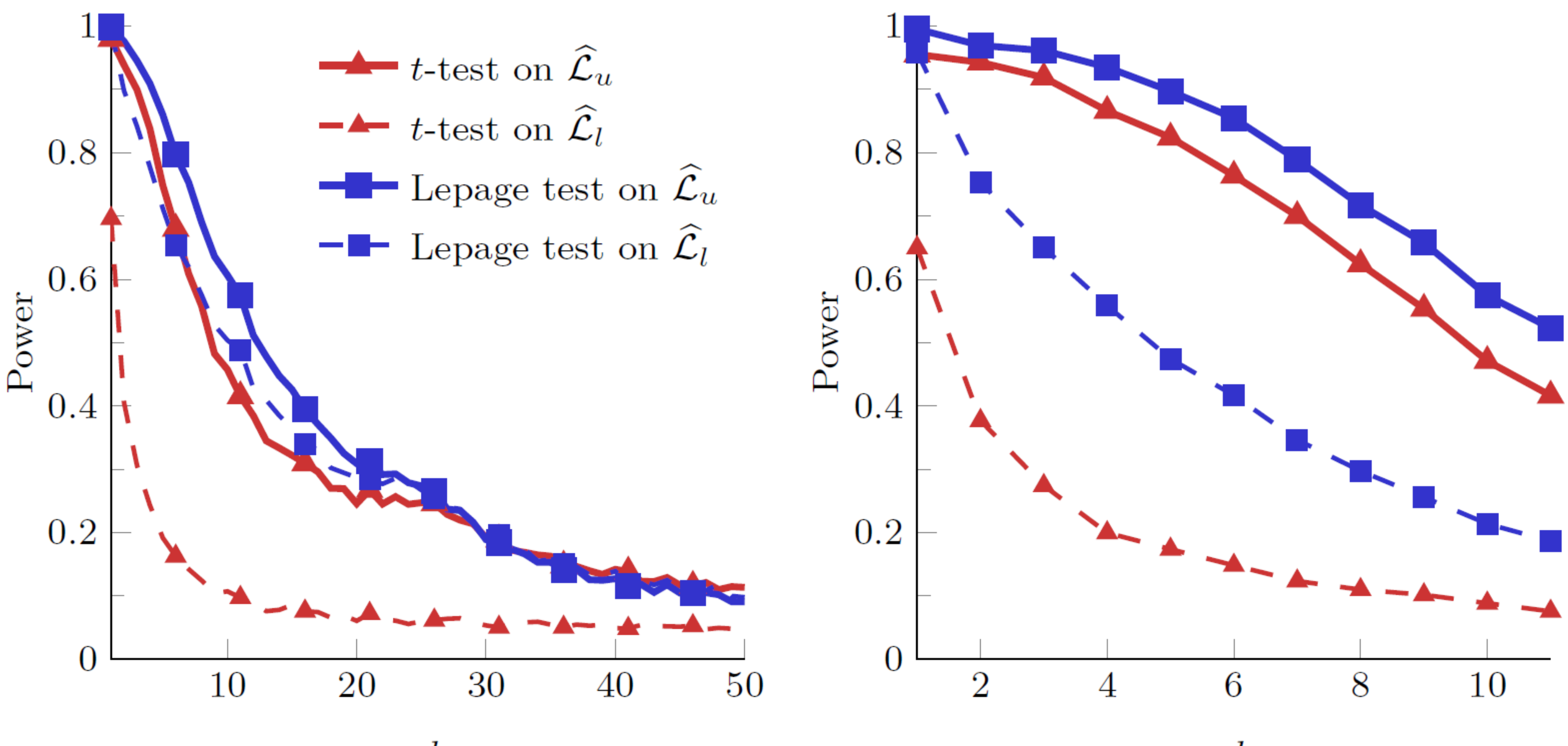
When designing/learning features, one should consider the detectability loss: irrelevant components are harmful!

To rigorously investigate change-detection problems when d increases it is necessary to control the change magnitude.

Interesting research direction are:

- Designing statistics / feature extraction methods specifically designed for anomaly/change detection
- Rigorously combining change and anomaly detection

THANKS, QUESTIONS?



C. Alippi, G. Boracchi, D. Carrera, M. Roveri, "*Change Detection in Multivariate Datastreams: Likelihood and Detectability Loss*" IJCAI 2016, New York, USA, July 9 - 13

D. Carrera, G. Boracchi, A. Foi and B. Wohlberg "*Detecting Anomalous Structures by Convolutional Sparse Models*" IJCNN 2015 Killarney, Ireland, July 12

D. Carrera, F. Manganini, G. Boracchi, E. Lanzarone *"Defect Detection in Nanostructures"*, IEEE Transactions on Industrial Informatics -- Submitted, 11 pages.