Anomaly Detection with Sparse Representations

Giacomo Boracchi

Dipartimento di Elettronica
Informazione e Bioingegneria,

Politecnico di Milano, Italy

IDSIA, Lugano, Dec. 11, 2014

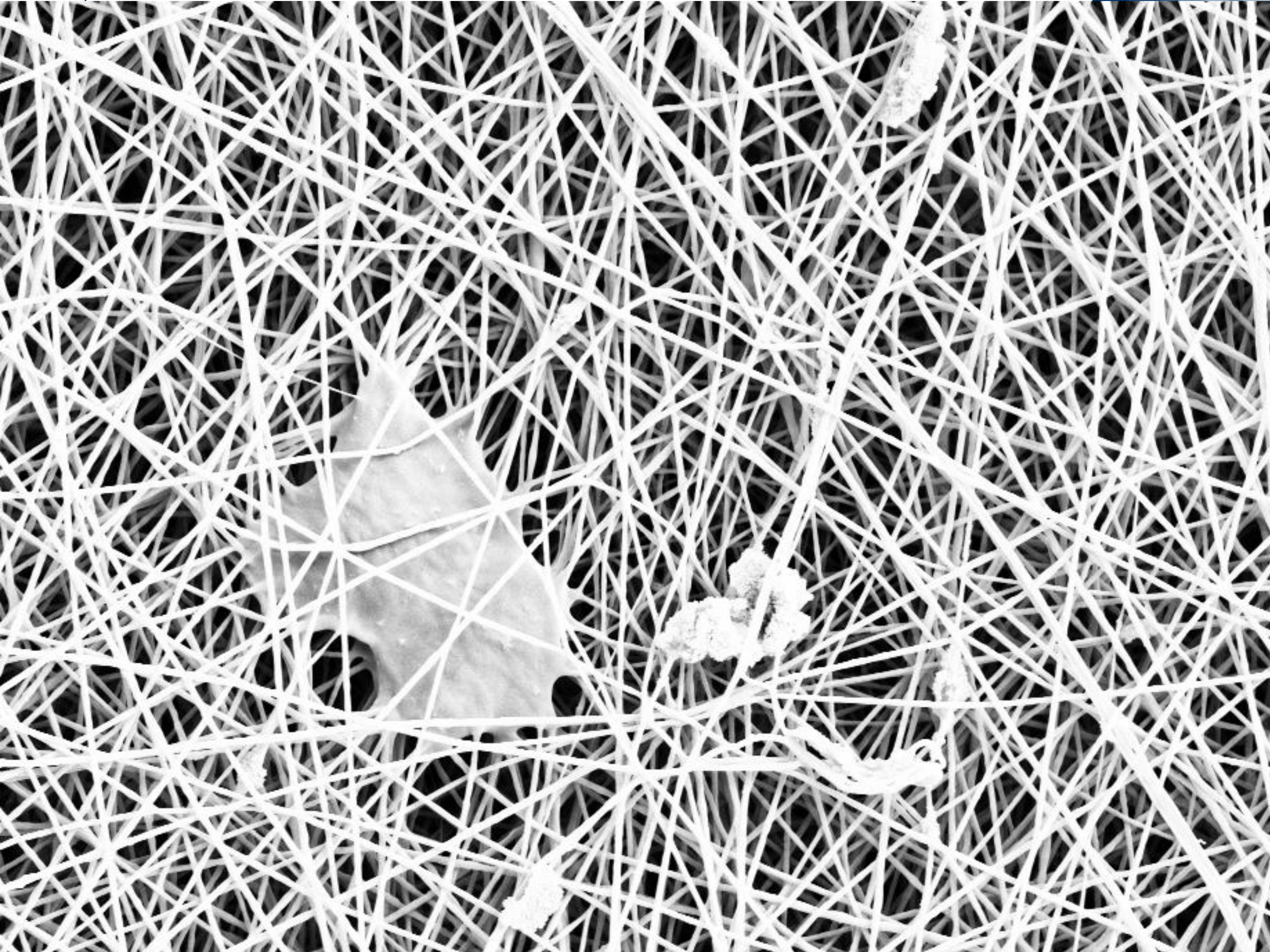
AN ONGOING WORK WITH

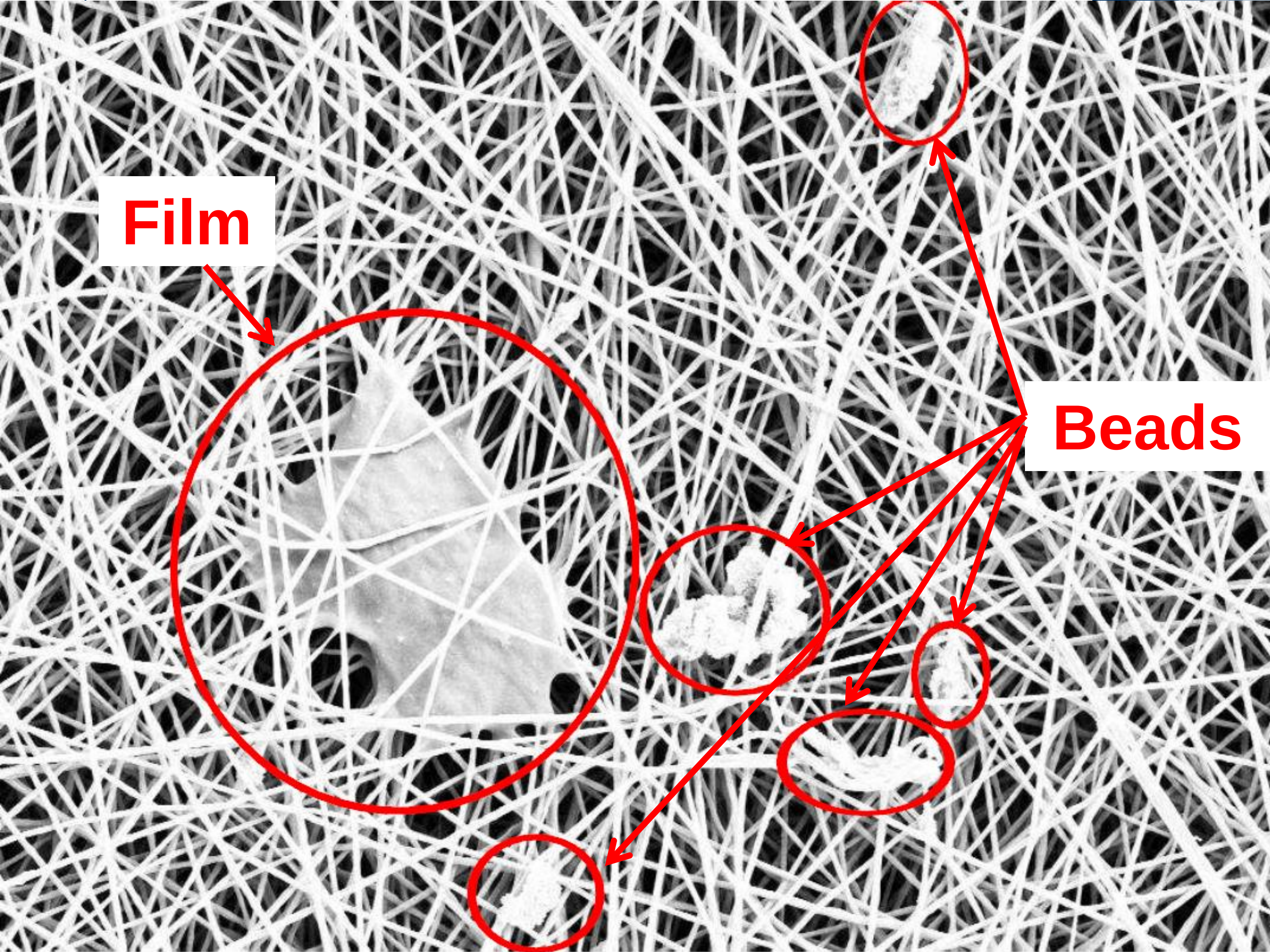
Cesare Alippi, Diego Carrera (Polimi)

Brendt Wohlberg (Los Alamos National Laboratory)

Anomaly (Novelty) Detection

- We consider **monitoring systems** acquiring and processing **images**, such as those employed in biomedical or industrial control applications.
- We assume that **images** acquired under **normal conditions** are characterized by **specific local structures**
- **Regions** that **do not conform** to these structures are considered **anomalies**
- We address the problem of **learning a model** for describing **normal structures** and **detect anomalies** as regions that cannot be properly described by the model
- As «running example» we consider scanning electron microscope (SEM) images for monitoring the production of nanofibers




Film
Beads

Outline

- Problem Formulation
- Sparse Representations for Anomaly Detection
- Anomaly indicators and Anomaly Detection
- Experiments on Anomaly Detection
  - Texture Images
  - SEM images for nanofiber production
- The Change-Detection Problem
- Experiments on Change Detection
  - Microacoustic bursts for rock-face monitoring

PROBLEM FORMULATION

Patch-Generating Process

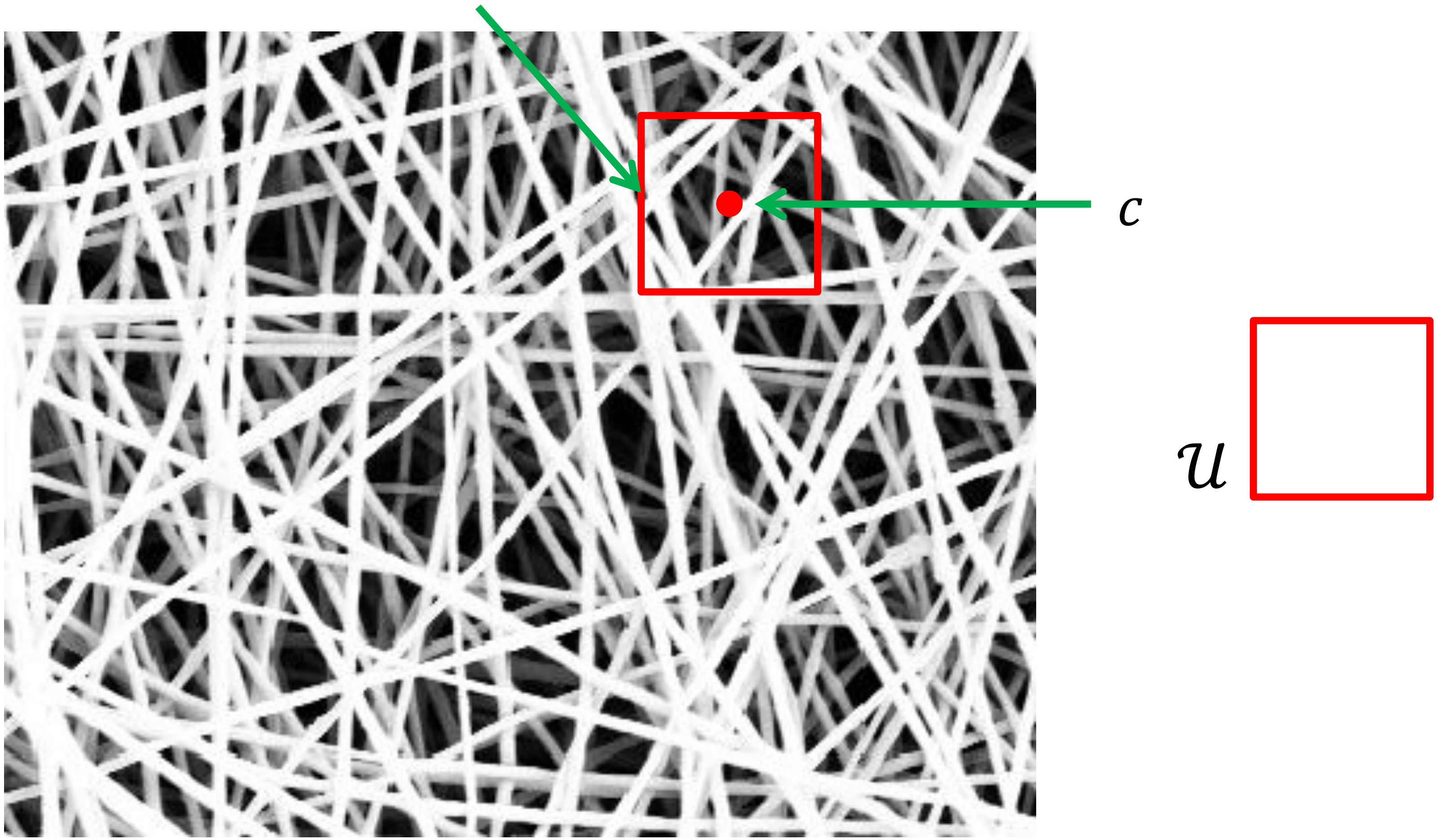
- Patches are small image regions of a predefined shape $\mathcal{U}$,

$$\mathbf{s}_c = \{s(c+u), u \in \mathcal{U}\}$$

Patch-Generating Process

- Patches are small image regions of a predefined shape $\mathcal{U}$,

$$\mathbf{s}_c = \{s(c+u), u \in \mathcal{U}\}$$



Patch-Generating Process

- Patches are small image regions of a predefined shape $\mathcal{U}$,

$$\mathbf{s}_c = \{s(c+u), u \in \mathcal{U}\}$$

- We assume that in **nominal** conditions, patches $\boldsymbol{s}_c \in \mathbb{R}^m$ are i.i.d. realizations from a stochastic process $\mathcal{P}_N$

$$\mathbf{s}_\text{c} \sim \mathcal{P}_N$$

- A training set of l normal patches $T \in \mathbb{R}^{m \times l}$ is given to learn a model $\widehat{D}$ approximating normal patches

The Anomaly-Detection Problem

- We assume that anomalous patches are generated by $\mathcal{P}_A$

$$\mathbf{s}_{\mathrm{c}} \sim \mathcal{P}_A$$

- The process generating anomalies $\mathcal{P}_A \neq \mathcal{P}_N$ is unknown
- Anomalies have to be detected as patches that **do not conform** the model learned to describe normal patches
  - We define **anomaly indicators** $f(\mathbf{s}_i)$ that measure the degree to which the learned model fits each patch $\mathbf{s}_i$
  - We detect anomalies as outliers in the anomaly indicators
- Peculiarity of the proposed approach is **to leverage models** $\widehat{D}$ **yielding sparse representation** of image patches

SPARSE REPRESENTATIONS

for anomaly detection

Sparse Representations

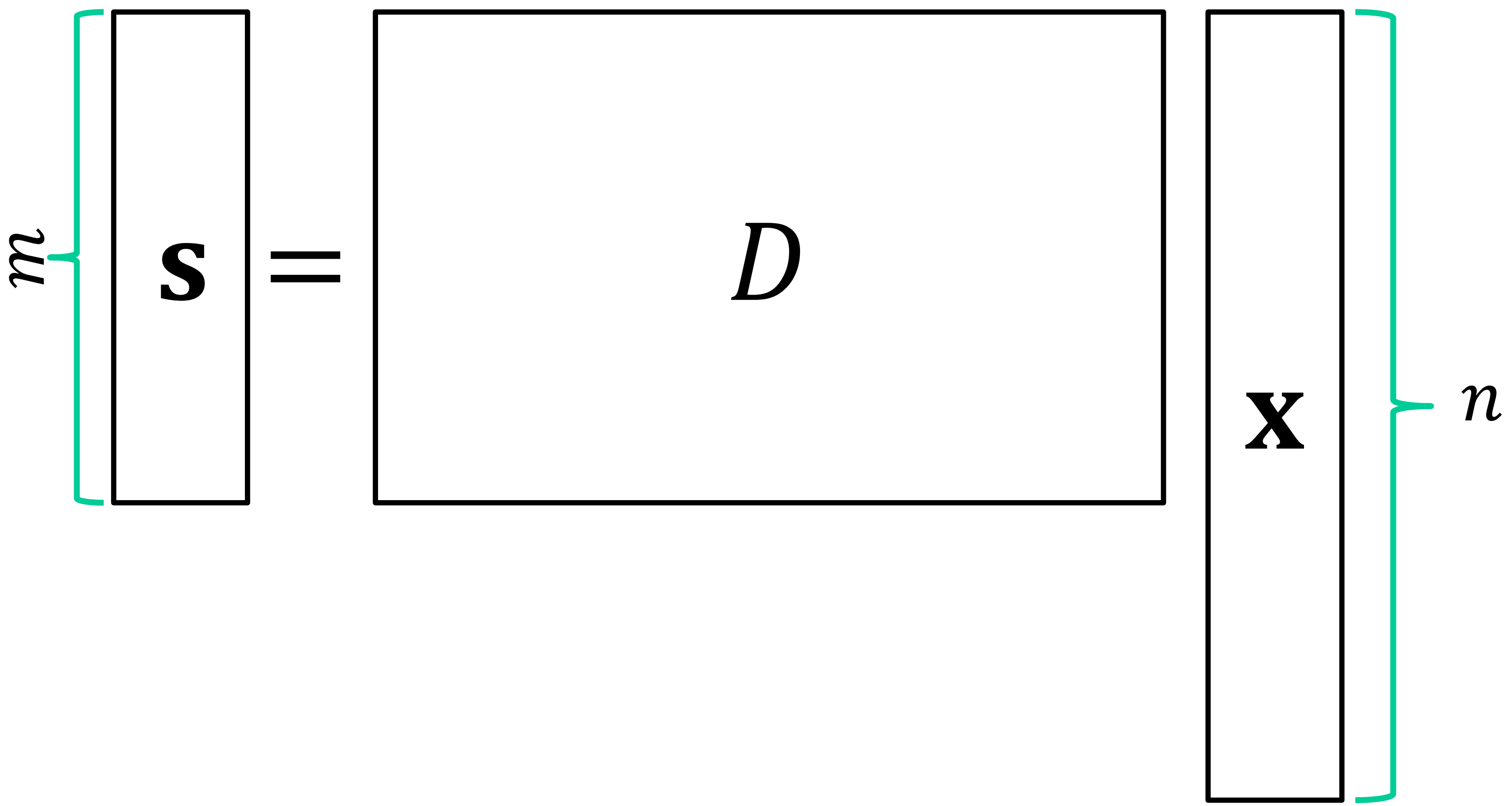
- **Sparse representations** have shown to be a very useful method for **constructing signal models**
- The underlying assumption is that

$$\mathbf{s} \approx D\mathbf{x} \quad \text{i.e,} \qquad \|\mathbf{s} - D\mathbf{x}\|^2 \approx 0$$

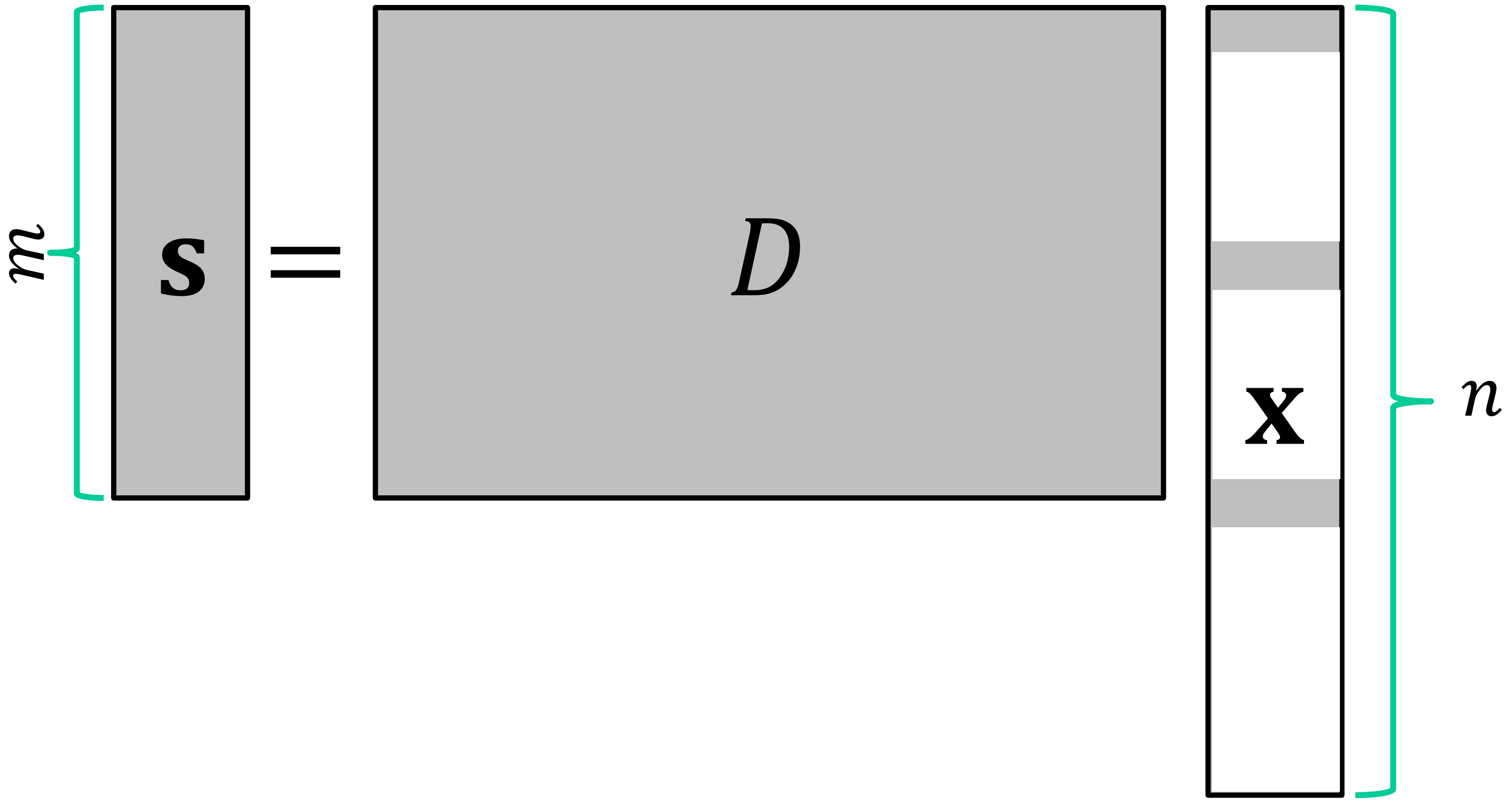
and $\mathbf{x} \in \mathbb{R}^n$ where:

  - $D \in \mathbb{R}^{m \times n}$ is the **dictionary**, columns are called **atoms**
  - the coefficient vector $\mathbf{x}$ is sparse ($\|\mathbf{x}\|_0 = L \ll n$)

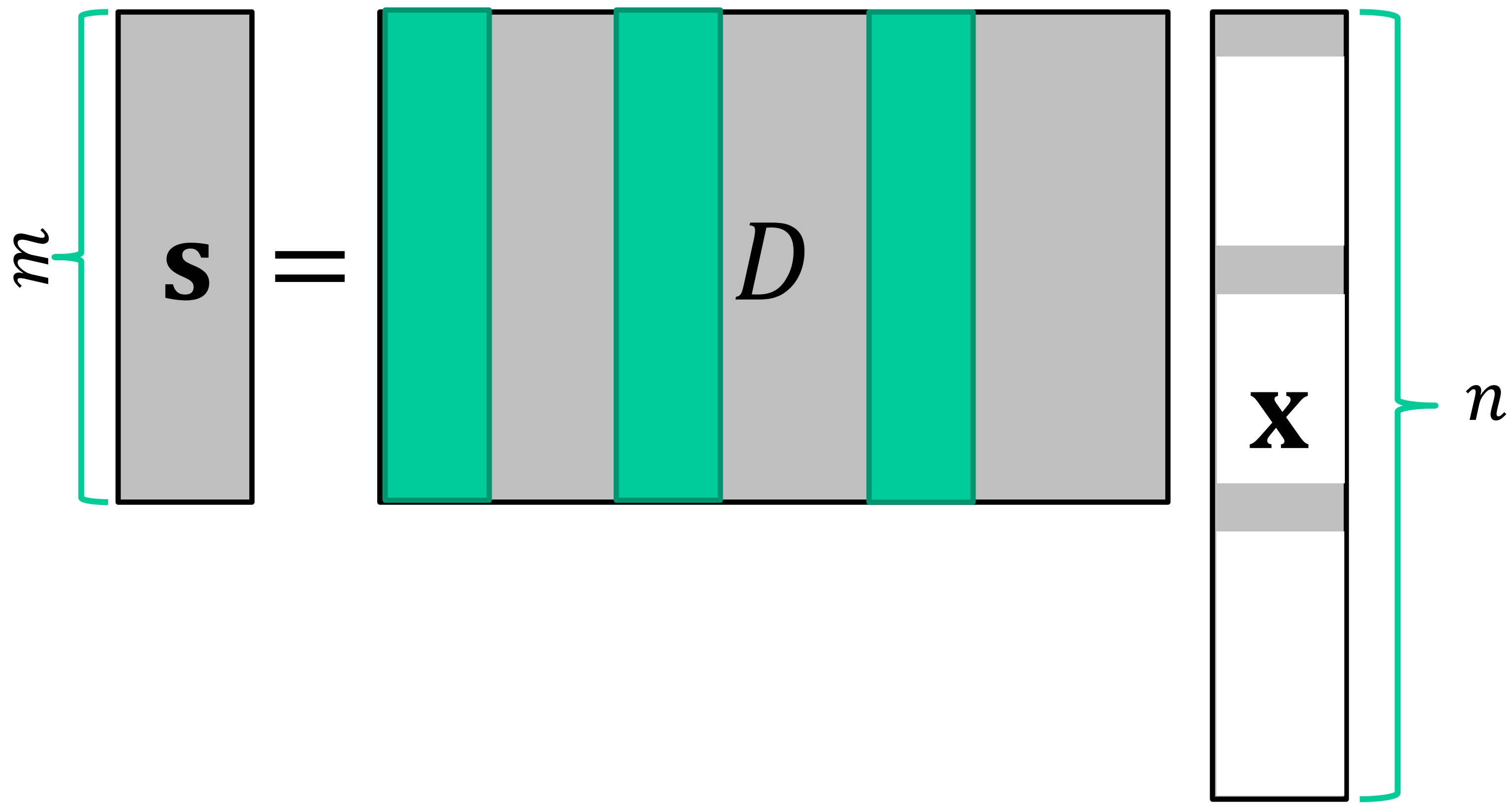
Sparse Representations



Sparse Representations



Sparse Representations



Sparse Representations

- **Sparse representations** have shown to be a very useful method for **constructing signal models**
- The underlying assumption is that

$$\mathbf{s} \approx D\mathbf{x} \quad \text{i.e,} \qquad ||\mathbf{s} - D\mathbf{x}||^2 \approx 0$$

  and $\mathbf{x} \in \mathbb{R}^n$ where:
  - $D \in \mathbb{R}^{m \times n}$ is the **dictionary**, columns are called **atoms**
  - the coefficient vector $\mathbf{x}$ is sparse ($||\mathbf{x}||_0 = L \ll n$)
- Sparse signals **live in a union of low-dimensional subspaces** of $\mathbb{R}^m$, each having maximum dimension L, defined by dictionary atoms $\{\mathbf{d}_i\}$ (columns of D).

$$\exists\, \mathbf{x} \in \mathbb{R}^n \text{ s.t. } \mathbf{s} = \sum_{i=1}^{n} x_i \mathbf{d}_i$$

Learning a Dictionary for Modeling Stationarity

- **Learning** $\widehat{D}$ corresponds to learning the **union of subpaces** where patches in T – the **normal** ones- live.
- Dictionary learning is a joint optimization over the dictionary and coefficients of a sparse representation of T

$$\widehat{D} = \underset{D \in \mathbb{R}^{m \times n}, X \in \mathbb{R}^{n \times l}}{\operatorname{argmin}} \|DX - T\|_F$$

such that $\|\mathbf{x}_k\|_0 \leq L, \forall k$

- We consider here the KSVD algorithm [Aharon 06]

[Aharon 06] M. Aharon, M. Elad, and A. M. Bruckstein, "K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation," Transactions on Signal Processing vol. 54, no. 11, November 2006, pp. 4311–4322.

Sparse Coding

- The dictionary $\widehat{D}$ can be used for **computing the sparse representation** of any **patch to be tested**
- There are efficient tools for computing $\mathbf{x}$, the sparse approximation of a patch $\boldsymbol{s}$ w.r.t. a given dictionary $\widehat{D}$

$$\widehat{D}\mathbf{x} \approx \boldsymbol{s}$$

- This operation is referred to as the **sparse coding**

Sparse Coding - ℓ^0 norm problem

- Sparse coding solving the constrained problem

$$\text{P0:} \quad \hat{\mathbf{x}}_{\mathbf{0}} = \underset{\mathbf{x}\in\mathbb{R}^n}{\operatorname{argmin}} \left\| \widehat{D}\mathbf{x} - \mathbf{s} \right\|_{\mathbf{2}} \ \text{s.t.} \ \|\mathbf{x}\|_0 \leq L$$

- The sparsity of the solution is constrained to be at most L
- Typically solved by means of Greedy Algoritms, such as the Orthogonal Matching Pursuit (OMP).
- Solving this problem actually corresponds to projecting the observed data into the union of subspaces (determined by at most L atoms).

Sparse Coding - ℓ^1 norm problem

- Sparse coding solving the unconstrained problem

$$\text{P1:} \quad \hat{\mathbf{x}}_{\mathbf{1}} = \underset{\mathbf{x}\in\mathbb{R}^n}{\operatorname{argmin}} \; J_\lambda\left(\mathbf{x}, \widehat{D}, \mathbf{s}\right)$$

where the functional is

$$J_\lambda\left(\mathbf{x}, \widehat{D}, \mathbf{s}\right) = \left\|\widehat{D}\mathbf{x} - \mathbf{s}\right\|_{\mathbf{2}}^{\mathbf{2}} + \lambda\|\mathbf{x}\|_1$$

- The sparsity requirement is relaxed by a penalization term on the ℓ_1- norm of the coefficients
- Under some conditions the solution of P0 and P1 do coincide
- This is a Basis Pursuit Denoising (BPDN) problem: there are several optimization methods in the literature.
- We adopt Alternating Direction Method of Multipliers (ADMM)

ANOMALY INDICATORS

Tools to quantitatively assess «patch normality»

Anomaly Indicators

- Given a dictionary $\widehat{D}$ learned to describe the training set T
- We measure the extent to which a given patch $\mathbf{s}$ is **consistent with the nominal conditions,** by computing the **sparse coding** of $\mathbf{s}$ w.r.t. $\widehat{\boldsymbol{D}}$

$$\mathbf{s} \to \hat{\mathbf{s}}, \quad \text{where } \hat{\mathbf{s}} = \widehat{D}\hat{\mathbf{x}} \text{ and } \hat{\mathbf{s}} \approx \mathbf{s}$$

- When solving the $\mathrm{P0}$ problem, $\hat{\mathbf{s}}$ is the projection of $\mathbf{s}$ on the best subspace of at most L atoms of $\widehat{D}$.
- We need suitable **anomaly-indicators** that **quantitatively assess** how close $\mathbf{s}$ is to nominal patches.
  - **anomaly indicators** have to take into account both **accuracy** and **sparsity** of the representation

Anomaly Indicators

- The **following anomaly indicators** have been considered:

  - When solving $\mathrm{P0}$ the reconstruction error

    $e(\mathbf{s}) = \left\|\mathbf{s} - \widehat{D}\hat{\mathbf{x}}_{\mathbf{0}}\right\|_{\mathbf{2}}$, being $\hat{\mathbf{x}}_{\mathbf{0}}$ the solution of $\mathrm{P0}$

  - When solving $\mathrm{P1}$, the value of the functional

    $f(\mathbf{s}) = \left\|\mathbf{s} - \widehat{D}\hat{\mathbf{x}}_{\mathbf{1}}\right\|_{\mathbf{2}} + \lambda\|\hat{\mathbf{x}}_{\mathbf{1}}\|_{\mathbf{1}}$, being $\hat{\mathbf{x}}_{\mathbf{1}}$ the solution of $\mathrm{P1}$

  - When solving $\mathrm{P1}$, jointly the sparsity and the error

    $g(\mathbf{s}) = [\left\|\mathbf{s} - \widehat{D}\hat{\mathbf{x}}_{\mathbf{1}}\right\|_{\mathbf{2}};\ \lambda\|\hat{\mathbf{x}}_{\mathbf{1}}\|_{\mathbf{1}}]$, being $\hat{\mathbf{x}}_{\mathbf{1}}$ the solution of $\mathrm{P1}$

ANOMALY DETECTION

on the anomaly indicators

Anomaly Detection

- The anomaly indicators captures the degree to which the structure of $\boldsymbol{s}$ is similar to that of normal patches
- Patches are processed independently
- We treat the anomaly indicators as realization from an unknown random variable: thus
- **Detecting** patches having **anomalous structures** becomes **detecting outliers** in **anomaly indicators**
  - Several statistical techniques have been developed ranging from graphical, confidence intervals-based, density-based
  - Outliers are detected as point in low-denisty regions
  - We perform outlier detection using confidence intervals which behaves quite well for unimodal distribution

Anomaly Detection from 1D Anomaly Indicators

- We treat **anomaly indicators** computed from i.i.d. stationary data as **random variables**.
- We define **high-density regions** for the empirical distribution of anomaly indicators from T
- In case of 1D-anomaly indicators, such a region is

$$\mathcal{J}_{\alpha}^{e} = [q_{\frac{\alpha}{2}}, q_{1-\frac{\alpha}{2}}]$$

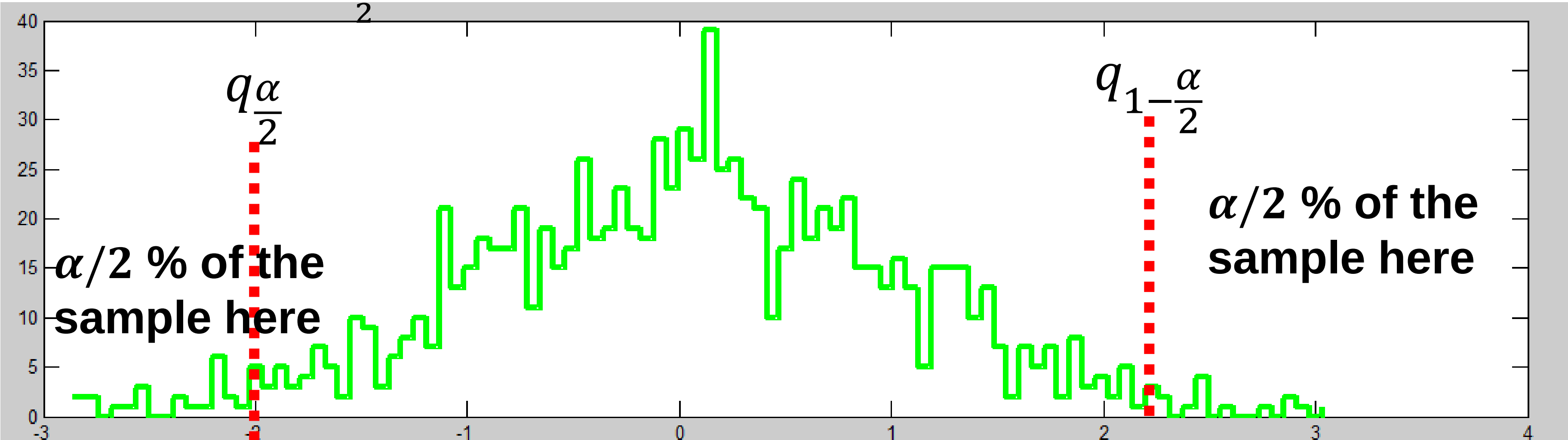
where $q_{\frac{\alpha}{2}}$ is the $\alpha/2$ quantile of the empirical distribution

Anomaly Detection from 1D Anomaly Indicators

- We treat **anomaly indicators** computed from i.i.d. stationary data as **random variables**.
- We define **high-density regions** for the empirical distribution of anomaly indicators from T
- In case of 1D-anomaly indicators, such a region is

$$\mathcal{J}_\alpha^e = [q_{\frac{\alpha}{2}}, q_{1-\frac{\alpha}{2}}]$$

where $q_{\frac{\alpha}{2}}$ is the $\alpha/2$ quantile of the empirical distribution



Anomaly Detection from 1D Anomaly Indicators

- We treat **anomaly indicators** computed from i.i.d. stationary data as **random variables**.
- We define **high-density regions** for the empirical distribution of anomaly indicators from T
- In case of 1D-anomaly indicators, such a region is

$$\mathcal{J}_{\alpha}^{e} = [q_{\frac{\alpha}{2}}, q_{1-\frac{\alpha}{2}}]$$

where $q_{\frac{\alpha}{2}}$ is the $\alpha/2$ quantile of the empirical distribution

- We detect anomalies as data yielding anomaly indicators, out of high-density regions (outliers)

$$e(\mathbf{s}) \notin \mathcal{J}_{\alpha}^{e}$$

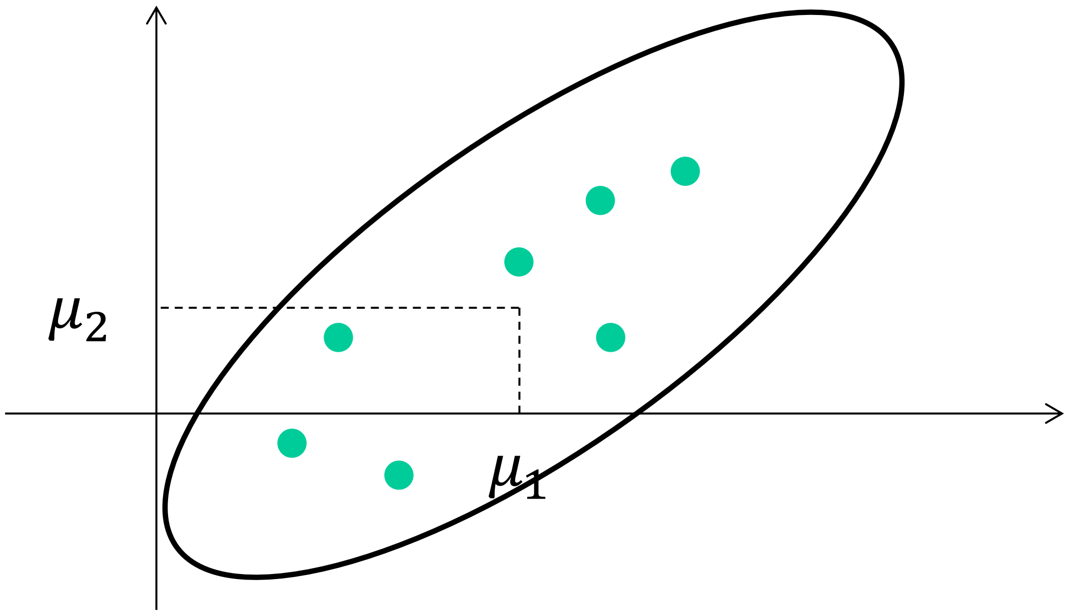
- The same for anomaly indicator $f(\cdot)$

Anomaly Detection from 2D Anomaly Indicators

- For the bivariate indicator $g(\cdot)$ we build a confidence region

$$R_\gamma = \left\{\xi \in \mathbb{R}^2, \text{s.t.} \sqrt{(\xi-\mu)'\Sigma^{-1}(\xi-\mu)} \le \gamma\right\}$$

where μ and Σ are the sample mean and sample covariance of the anomaly indicators from T.



Anomaly Detection from 2D Anomaly Indicators

- For the bivariate indicator $g(\cdot)$ we build a confidence region

$$R_\gamma = \left\{\xi \in \mathbb{R}^2, \mathrm{s.t.} \sqrt{(\xi - \mu)' \Sigma^{-1} (\xi - \mu)} \leq \gamma\right\}$$

  where μ and Σ are the sample mean and sample covariance of the anomaly indicators from T.

- The Chebyshev's inequality ensures that a normal patch falls outside R_γ with probability $\leq 2/\gamma^2$
- Anomalies are detected as

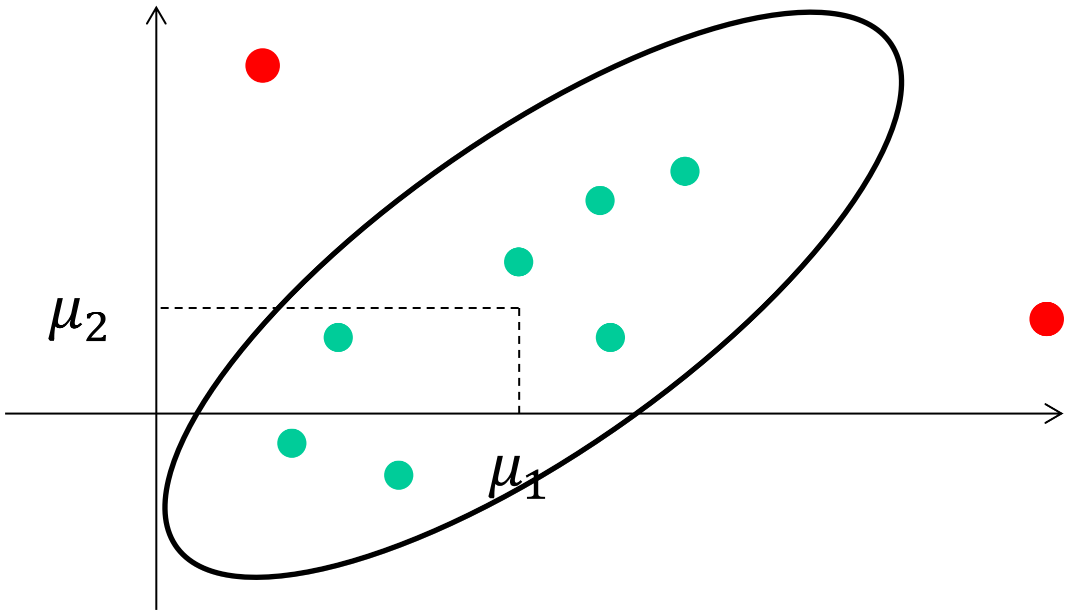
$$\mathbf{s} \quad \mathrm{s.t.} \sqrt{(\boldsymbol{g}(\mathbf{s}) - \mu)' \Sigma^{-1} (\boldsymbol{g}(\mathbf{s}) - \mu)} > \gamma$$

Anomaly Detection from 2D Anomaly Indicators

- For the bivariate indicator $g(\cdot)$ we build a confidence region

$$R_\gamma = \left\{\xi \in \mathbb{R}^2, \mathrm{s.t.} \sqrt{(\xi - \mu)'\Sigma^{-1}(\xi - \mu)} \leq \gamma\right\}$$

where μ and Σ are the sample mean and sample covariance of the anomaly indicators from T.



EXPERIMENTS

On Texture and SEM images

Anomaly detection in images

- Data are 15×15 patches extracted from textured images characterized by a specific structure

Test on Synthetic Images

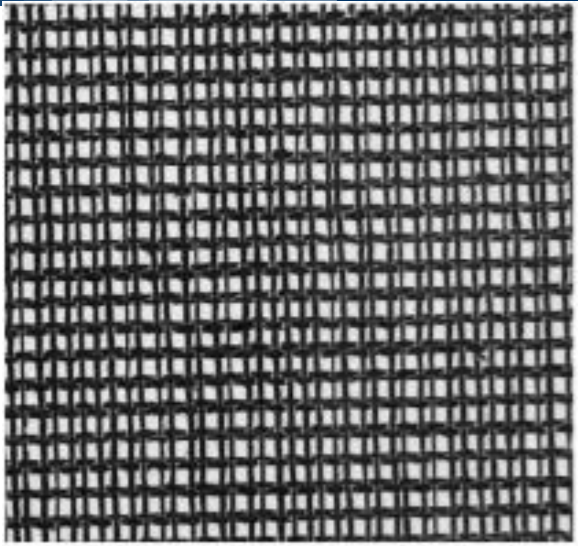


Image 1

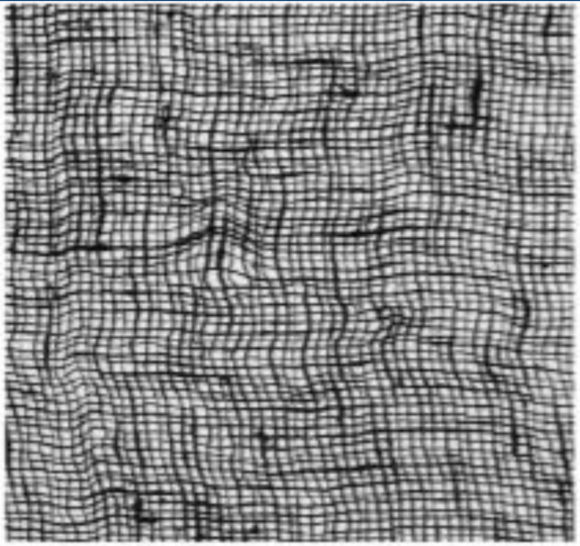


Image 2

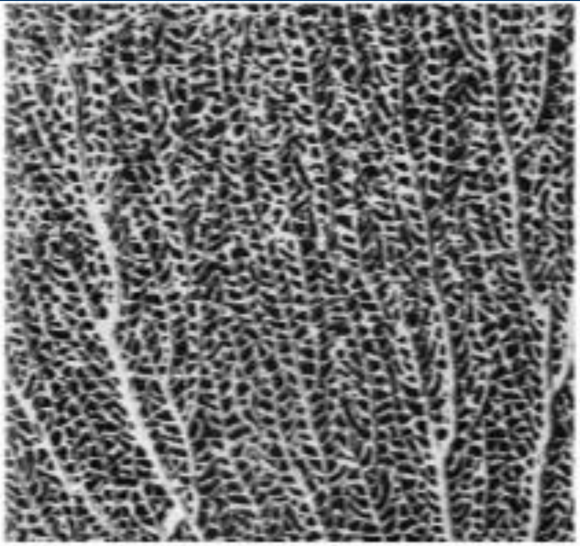


Image 3

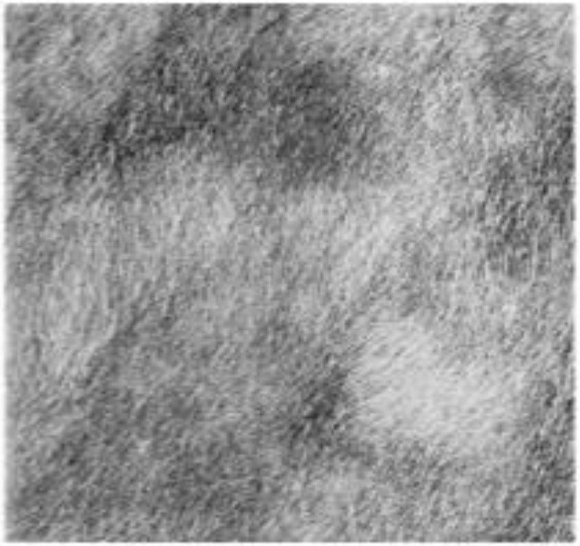


Image 4

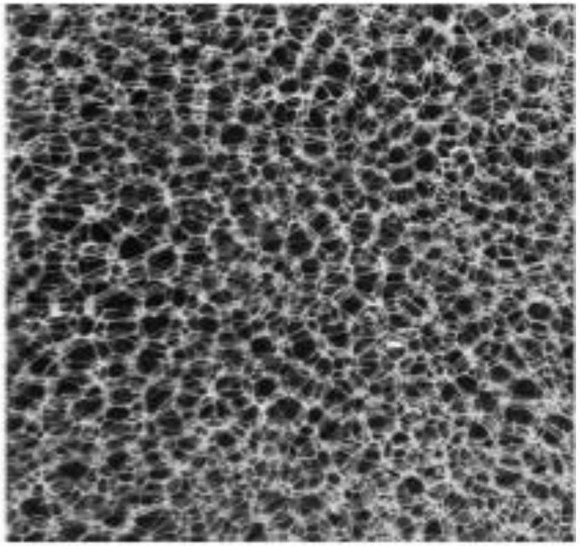


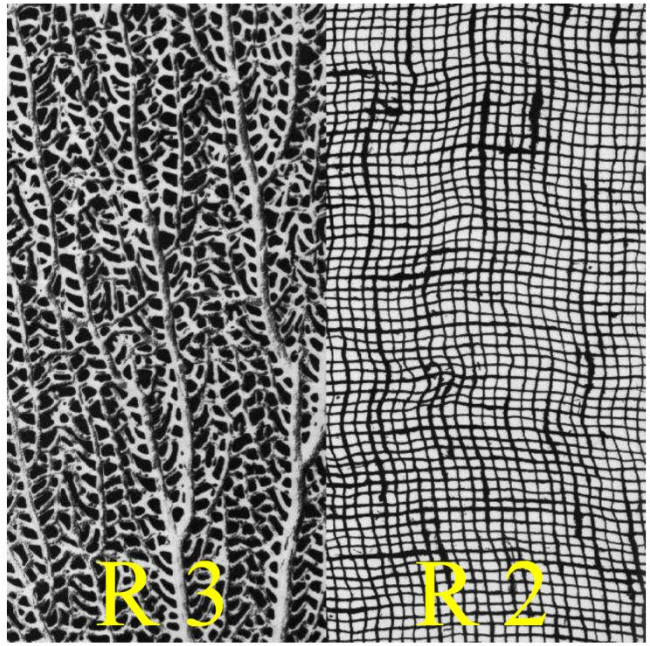
Image 5

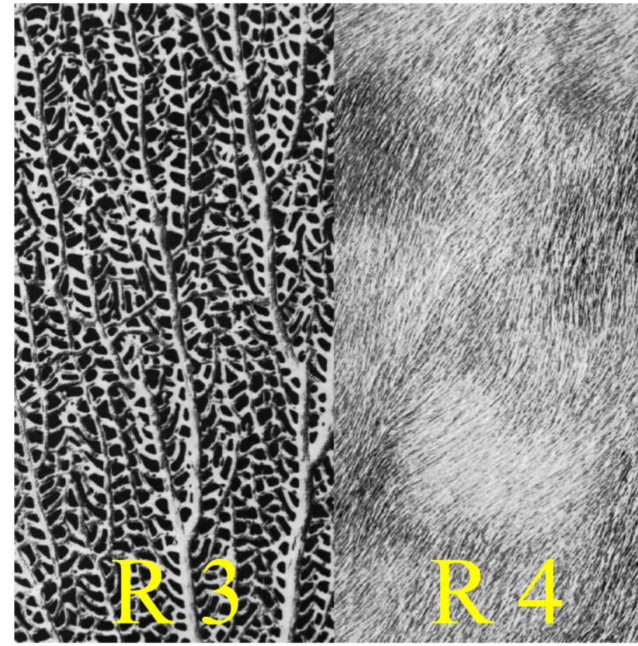
Anomaly detection in images

- We extract 15×15 patches from textured images, each characterized by a specific structure
- Anomaly detection problems are simulated by assembling test images that contains patches from different texture
  - The left half of each image is used to learn $\widehat{D}$
  - The right half is used for testing and juxtaposed with other half images

Test Images

Test images





We learn a dictionary from L3

Anomaly detection in images

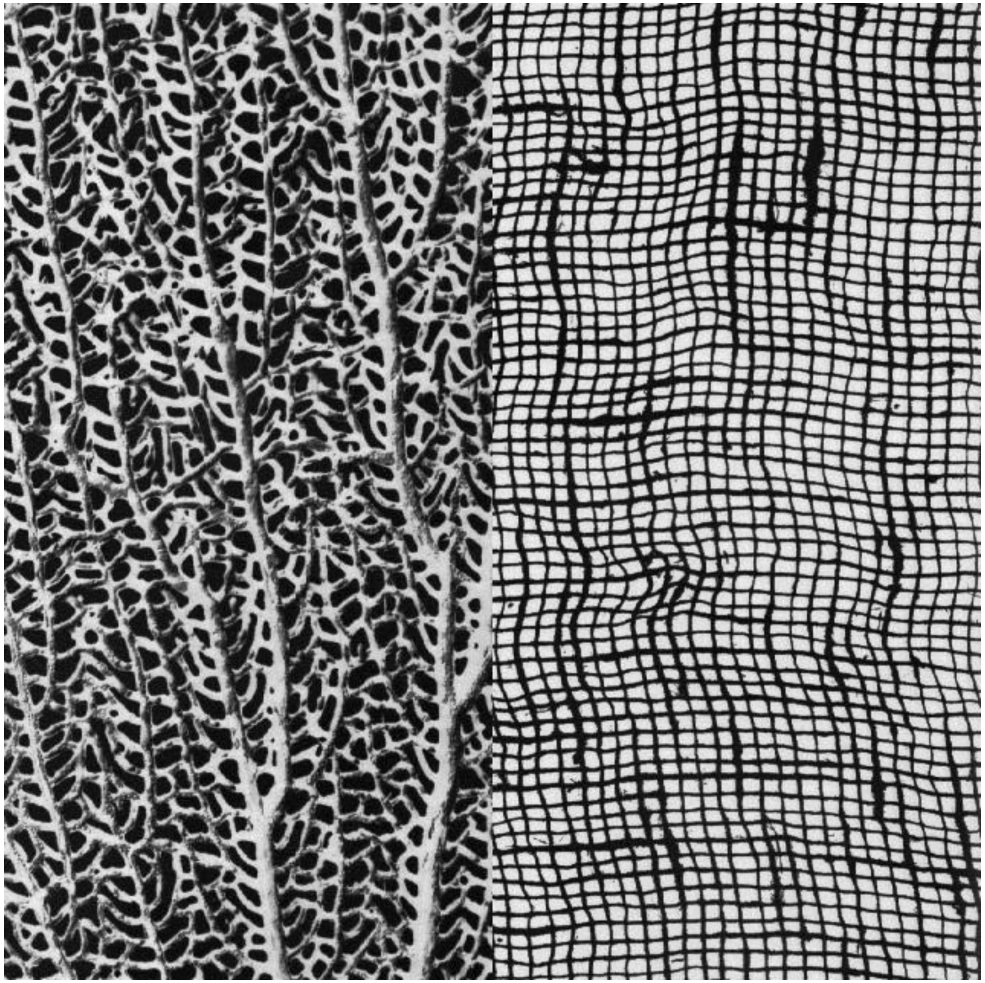
- Data are 15×15 patches extracted from textured images characterized by a specific structure
- Anomaly detection problems are simulated by syntetically creating test images gathering patches from different texture
- Each patch is **pre-processed** by subtracting its mean
- **No post-processing** to aggregate decision spatially is performed
- For further details, please refer to [Boracchi 2014]

[Boracchi 2014] Giacomo Boracchi, Diego Carrera, Brendt Wohlberg «Anomaly Detection in Images By Sparse Representations» SSCI 2014

Figures of Merit

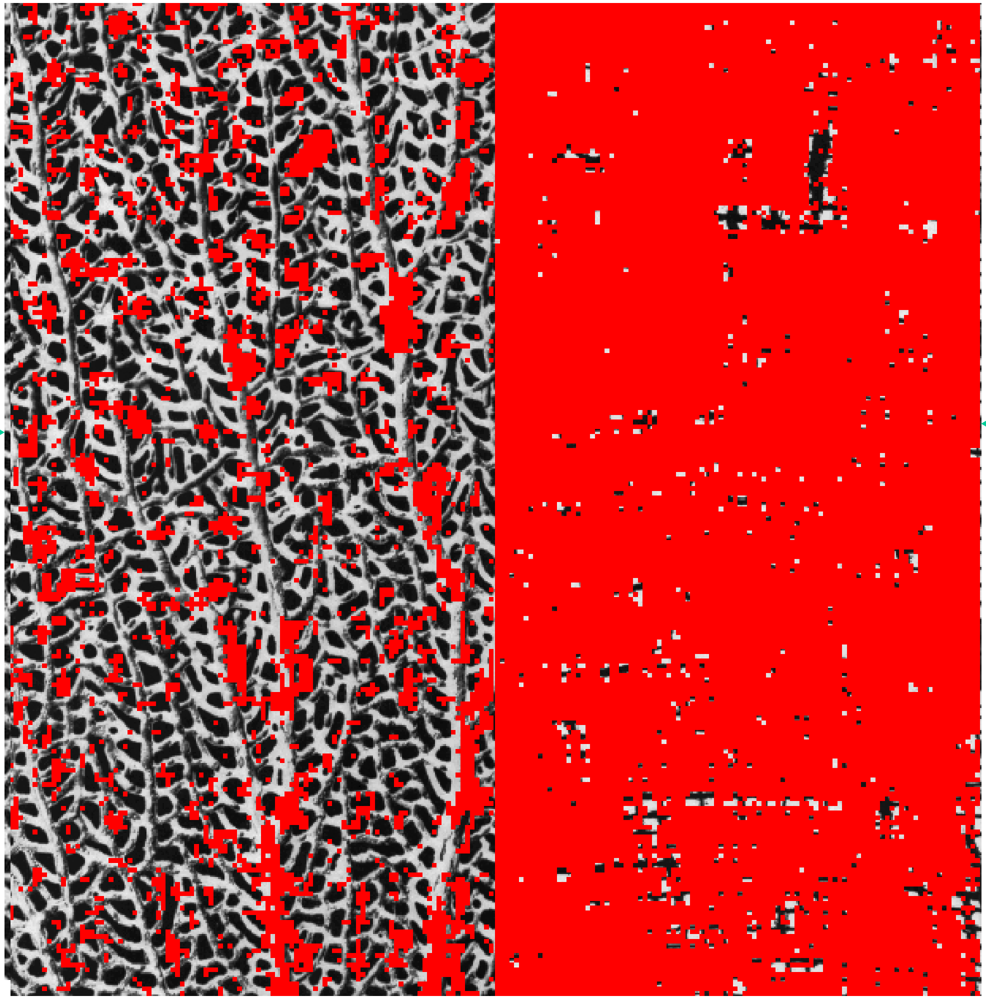
- FPR: the false positive rate, i.e. the percentage of normal patches labelled as anomalous
- TPR: the true positive rate, i.e., the percentage of anomalies correctly detected

Figures of Merit



Figures of Merit

False Positives



True Positives

Alternative Solution

- In [Adler 2013] the anomaly detection is performed during the sparse coding. The following model is consider

$$\mathbf{s} = D\mathbf{x} + \mathbf{a} + \boldsymbol{\nu} \quad \text{where } \boldsymbol{\nu} \text{ is a noise term}$$

  and $\mathbf{a}$ collects all the components of $\mathbf{s}$ that cannot be sparsely approximated.

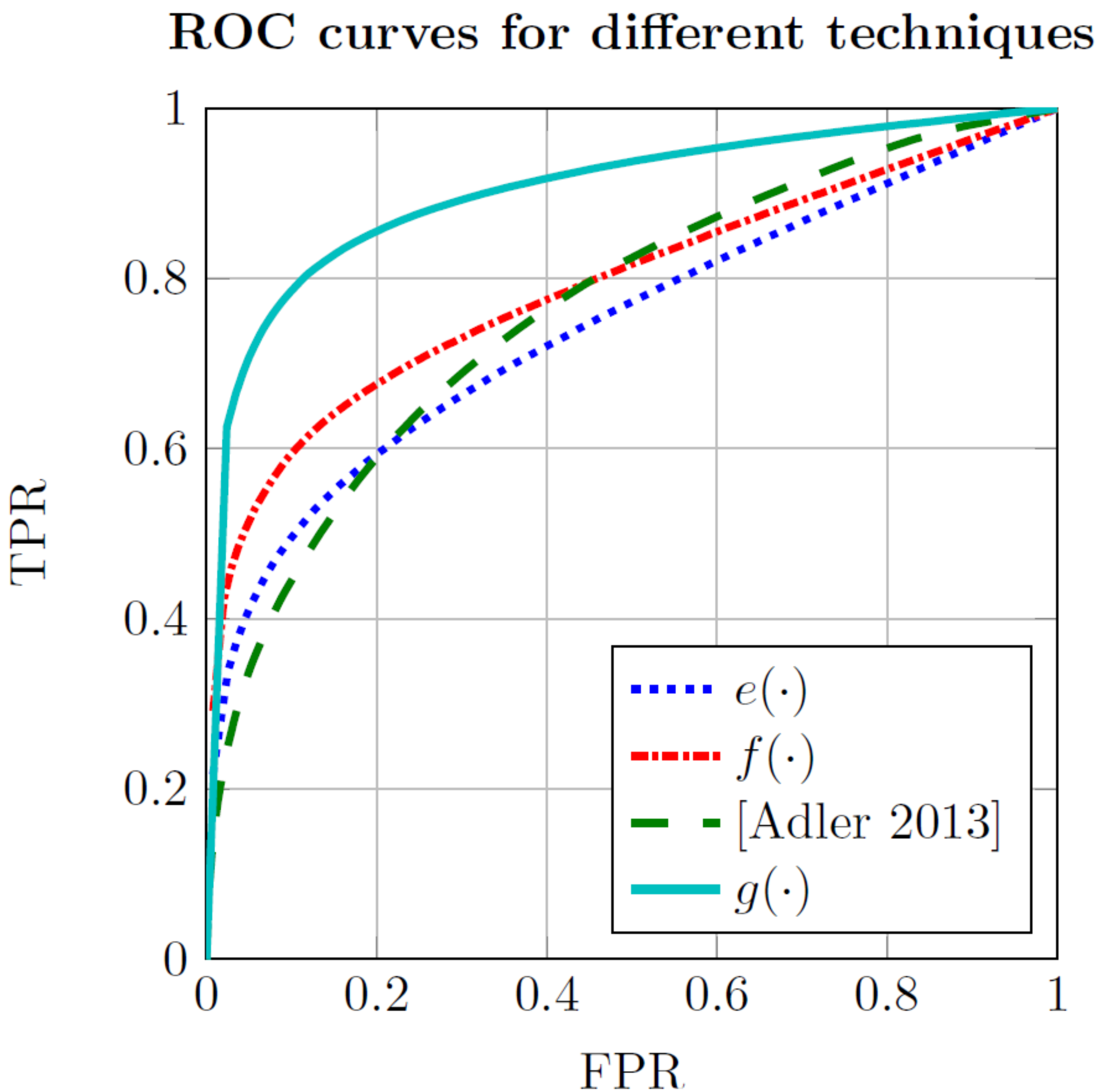
- Sparse coding is performed solving the following problem

$$\hat{\mathbf{x}} = \underset{\mathbf{x}\in\mathbb{R}^n}{\operatorname{argmin}} \frac{1}{2}\|\boldsymbol{s} - D\boldsymbol{x} - \boldsymbol{a}\|_2 + \|\boldsymbol{x}\|_1 + \|\boldsymbol{a}\|_2$$

- Normal patches: $\|\boldsymbol{a}\|_2$ is negligible, anomalous patches: $\|\boldsymbol{a}\|_2$ is large.
- Anomalies detected comparing $\|\boldsymbol{a}\|_2$ against a threshold

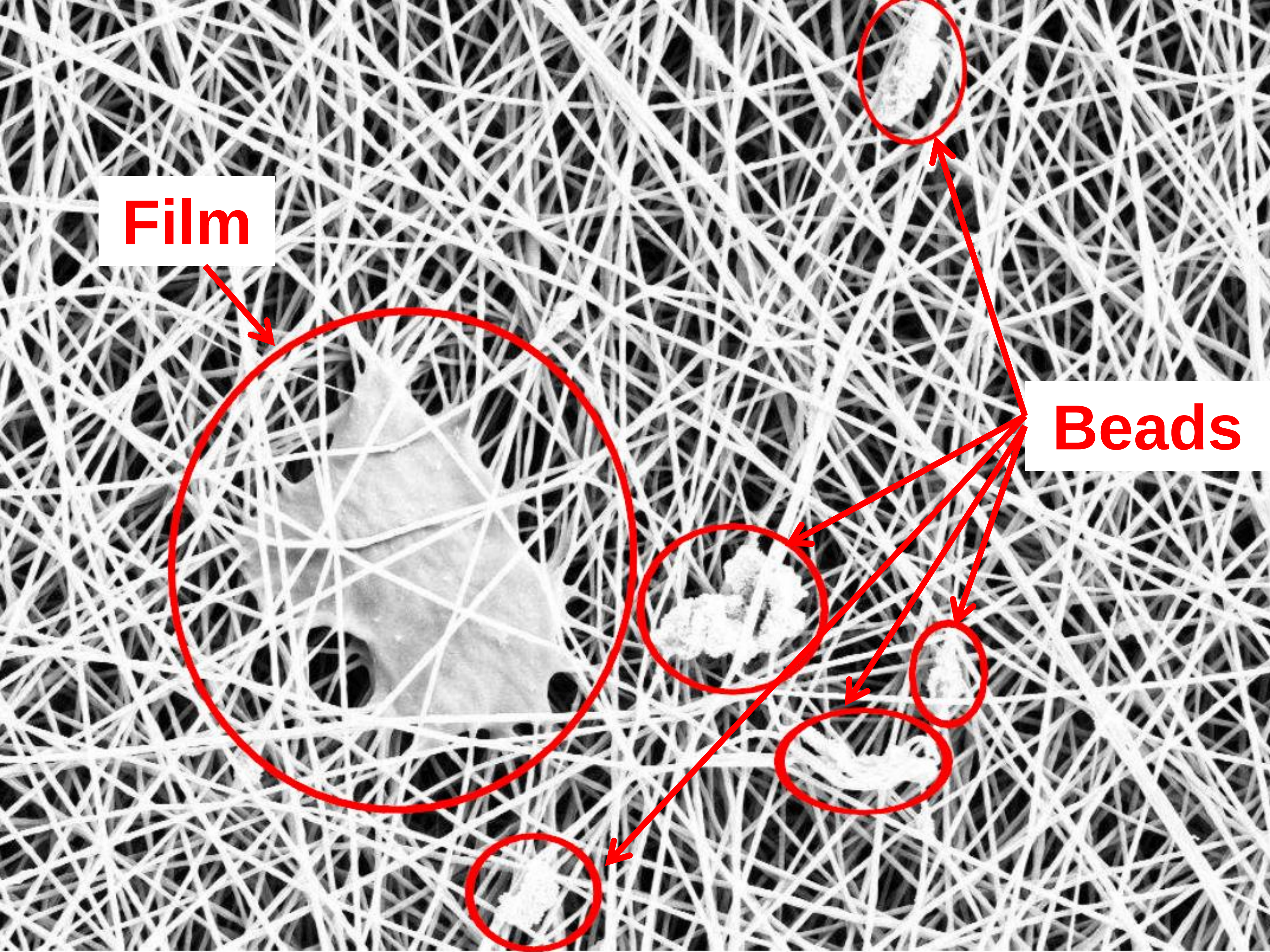
[Adler 2013] A. Adler, M. Elad, Y. Hel-Or, and E. Rivlin, “Sparse coding with anomaly detection,” in Proc. of IEEE MLSP, September 2013,

ROC curves when varying the threshold



Anomaly detection in SEM images

- Problem Description: we consider the production of nanofibrous materials by an electrospinning process
- An scanning electron microscope (SEM) is used to monitor the production process and detect the presence of
  - Beads
  - Films
- Detecting anomalies and assessing how large they are is very important for supervising the monitoring process

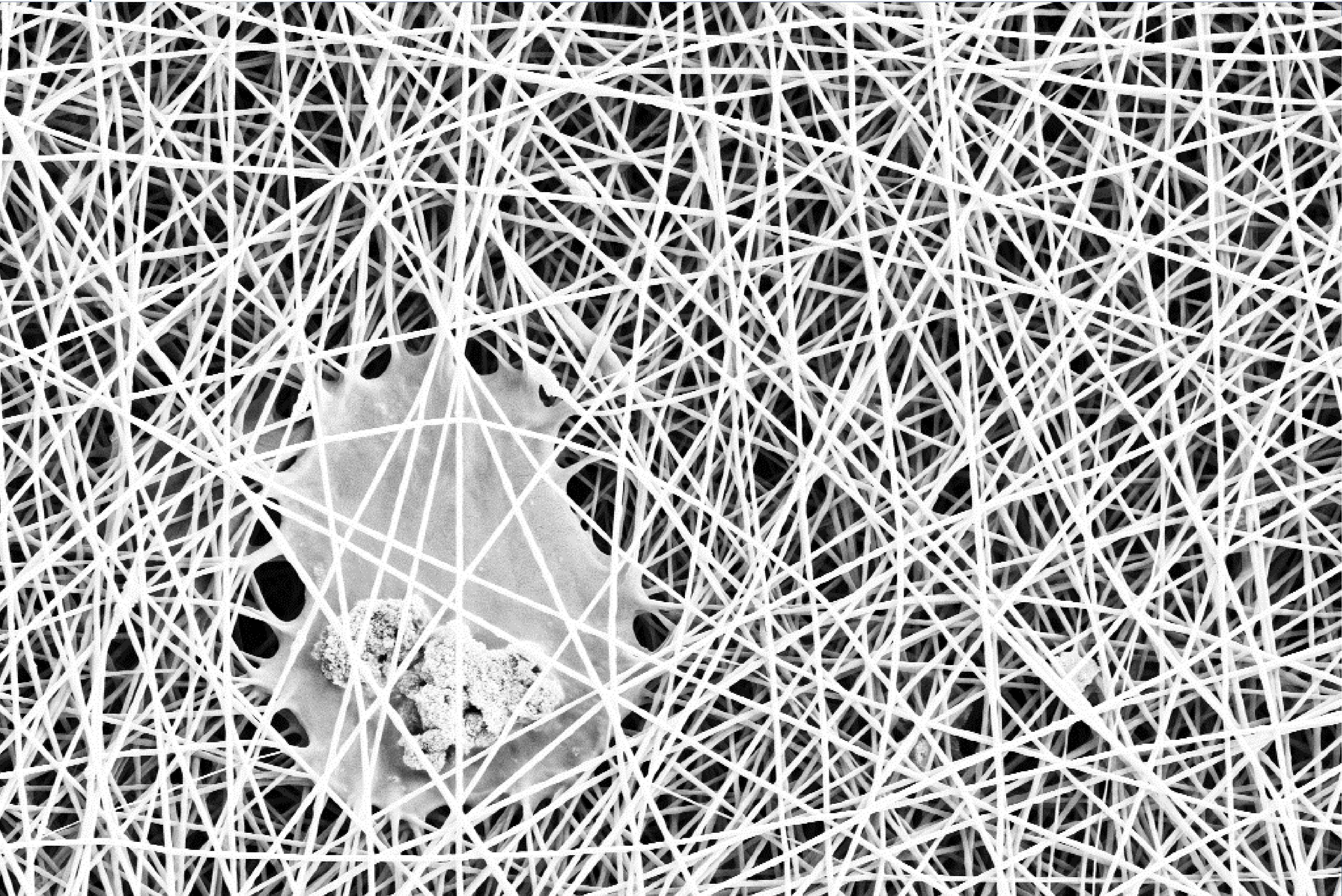

Film
Beads

Anomaly detection in SEM images

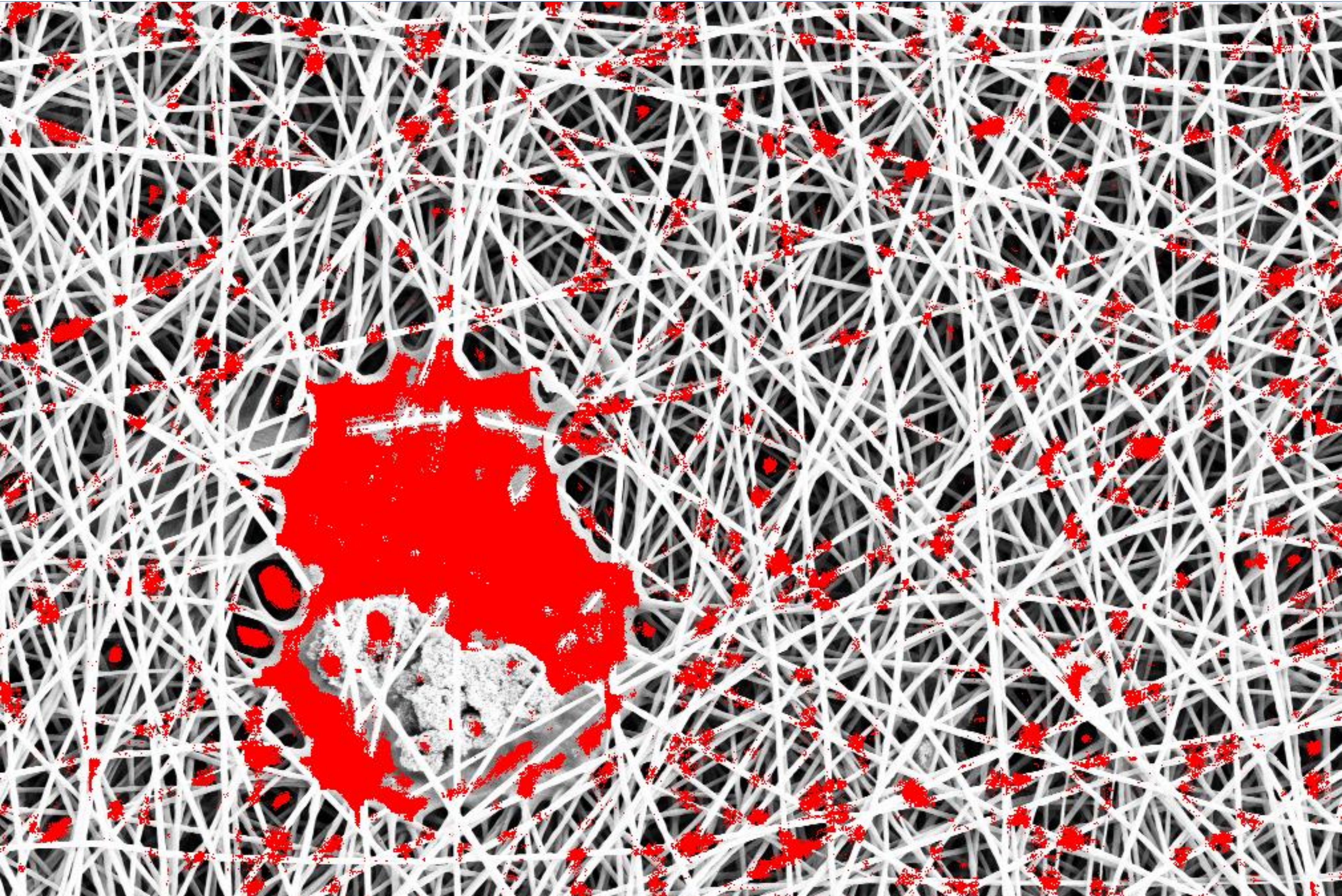
- Problem Description: we consider the production of nanofibrous materials by an electrospinning process
- An scanning electron microscope (SEM) is used to monitor the production process and detect the presence of
  - Beads
  - Films
- Detecting anomalies and assessing how large they are is very important for supervising the monitoring process
- Each anomaly detection method has been manually tuned to operate at its best performance
- Further details can be found in [Boracchi 2014]

[Boracchi 2014] Giacomo Boracchi, Diego Carrera, Brendt Wohlberg «Anomaly Detection in Images By Sparse Representations» SSCI 2014

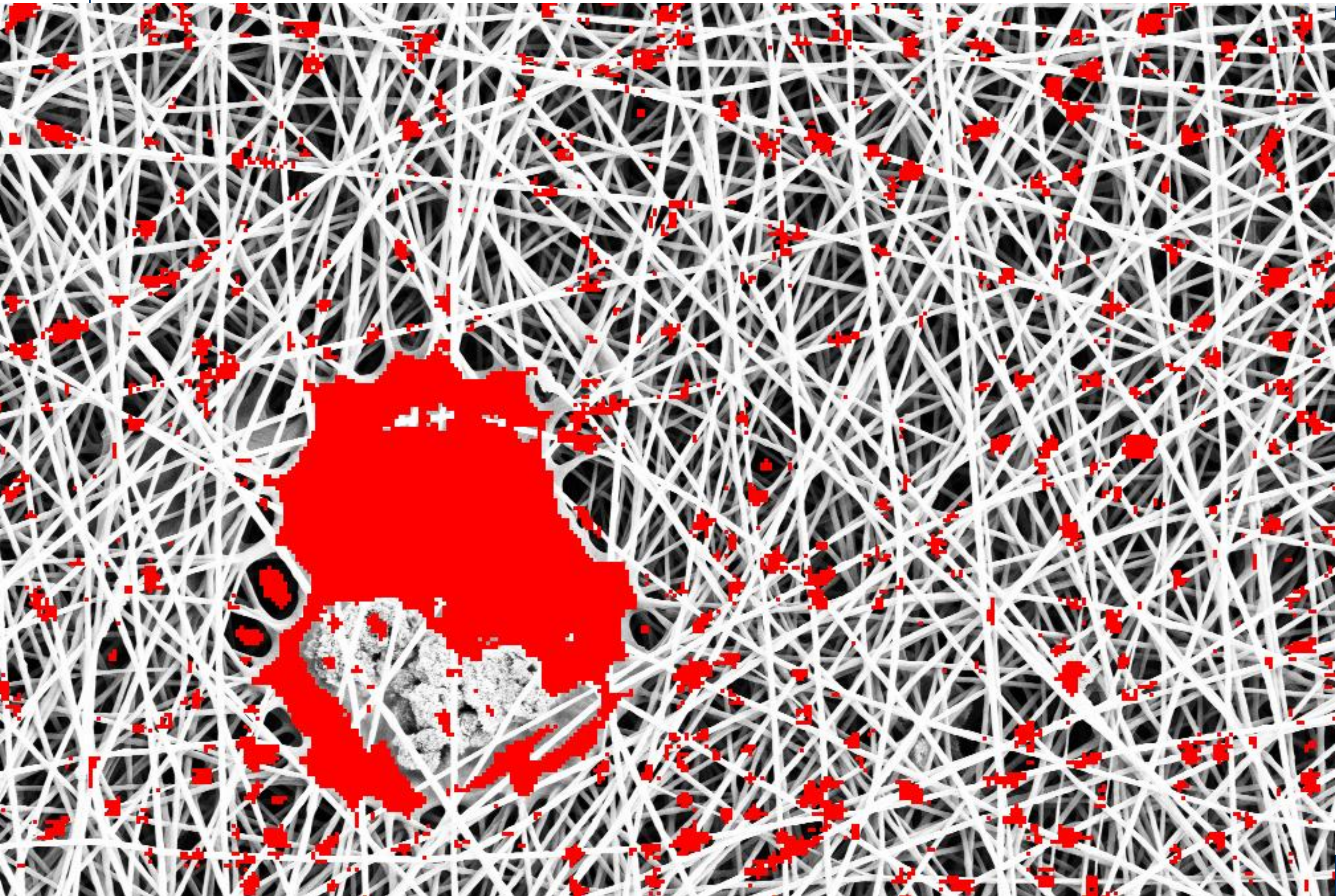
Original Image



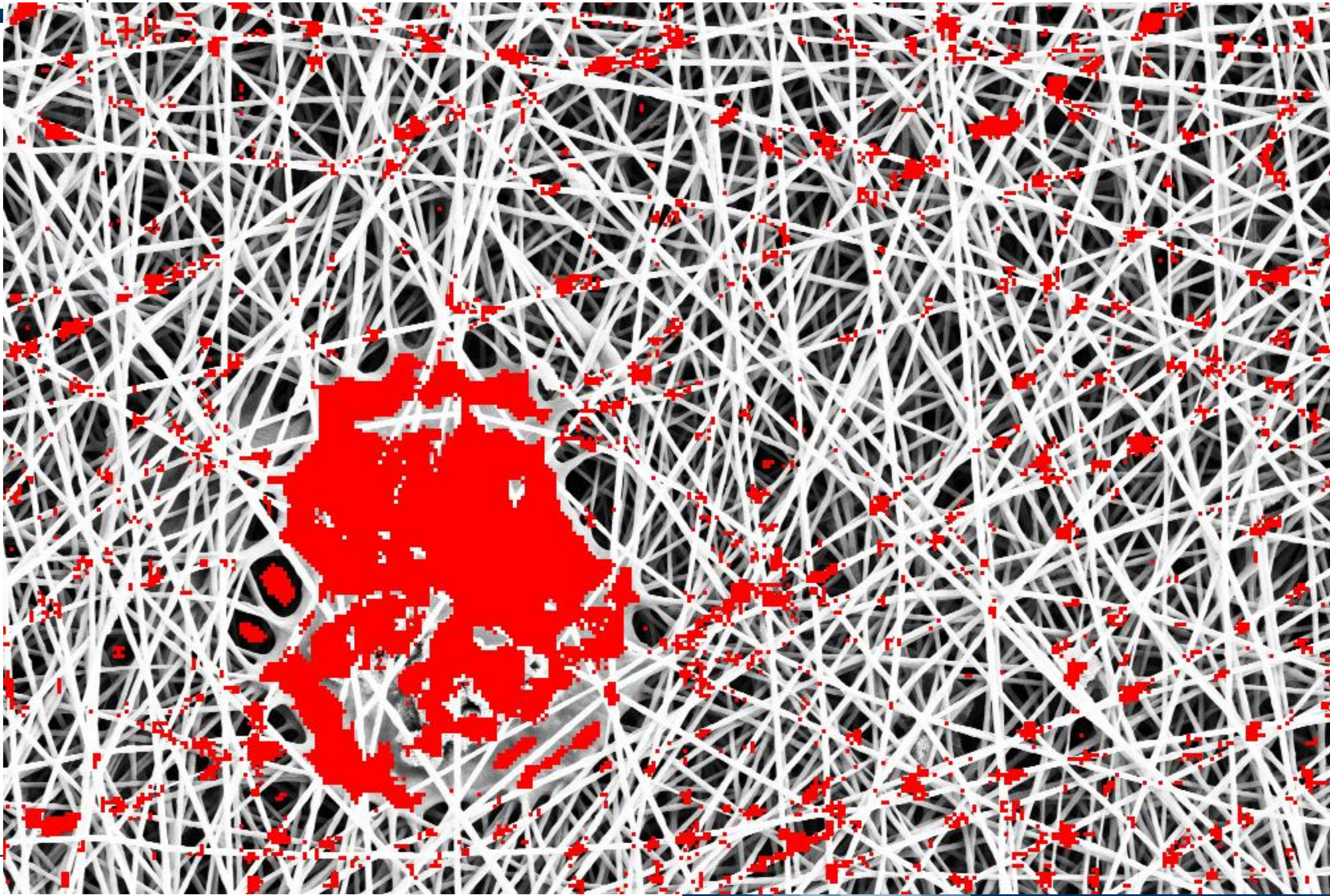
Anomaly detection by means of $e(\cdot)$



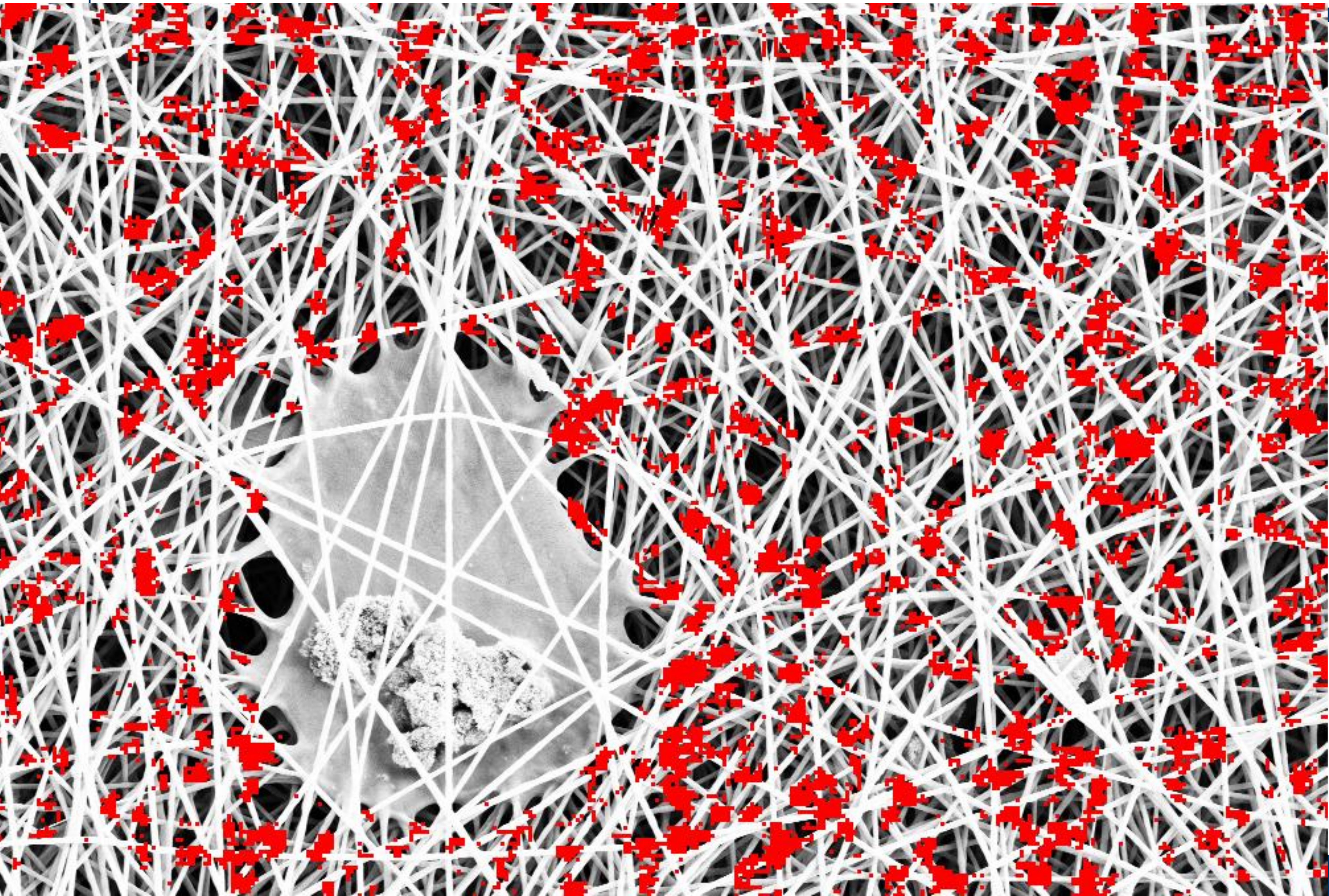
Anomaly detection by means of $f(\cdot)$



Anomaly detection by means of $g(\cdot)$



Anomaly detection by means of Adler



SOME REMARKS

From a more general perspective…

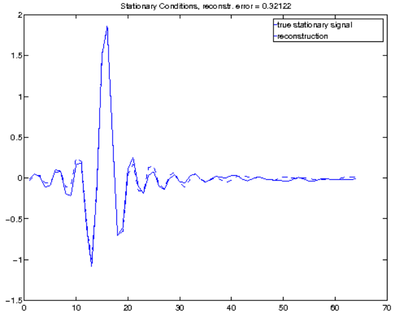
- This approach can be applied to any data-generating process as far as:
  - Observations are **signals** whos **structure** characterizes the stationarity
  - It is possible to **learn a dictionary** to describe these signals
  - **Anomalies** exhibit **different structures** (or different noise levels)

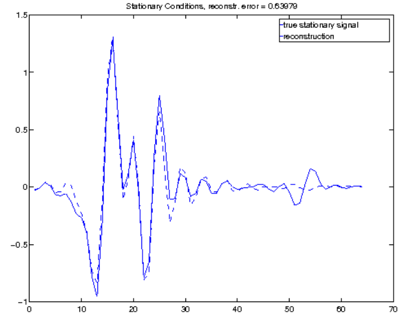
Data-Generating Process

- We assume that in **normal (stationary)** conditions, we observe data $\boldsymbol{s} \in \mathbb{R}^m$ drawn from a stochastic process $\mathcal{P}_N$

$$\mathbf{s} \sim \mathcal{P}_N$$

- We do not know the process, we only assume that data are i.i.d. realizations from $\mathcal{P}_N$.





From a more general perspective…

- This approach can be also applied to **sequential monitoring** applications, where we are interested in **detecting persistent changes** in the **data-generating process**
- Permanent shifts of the process could be due to
  - Faults
  - Unforeseen evolution of the environment

CHANGE DETECTION ON STREAMS OF SIGNALS

A very related problem

The change-detection problem

- The **change-detection problem** consists in monitoring a sequence of data (datastream), vectors of $\mathbb{R}^m$

$$\{\mathbf{s}_t\}_{t=1,\dots}$$

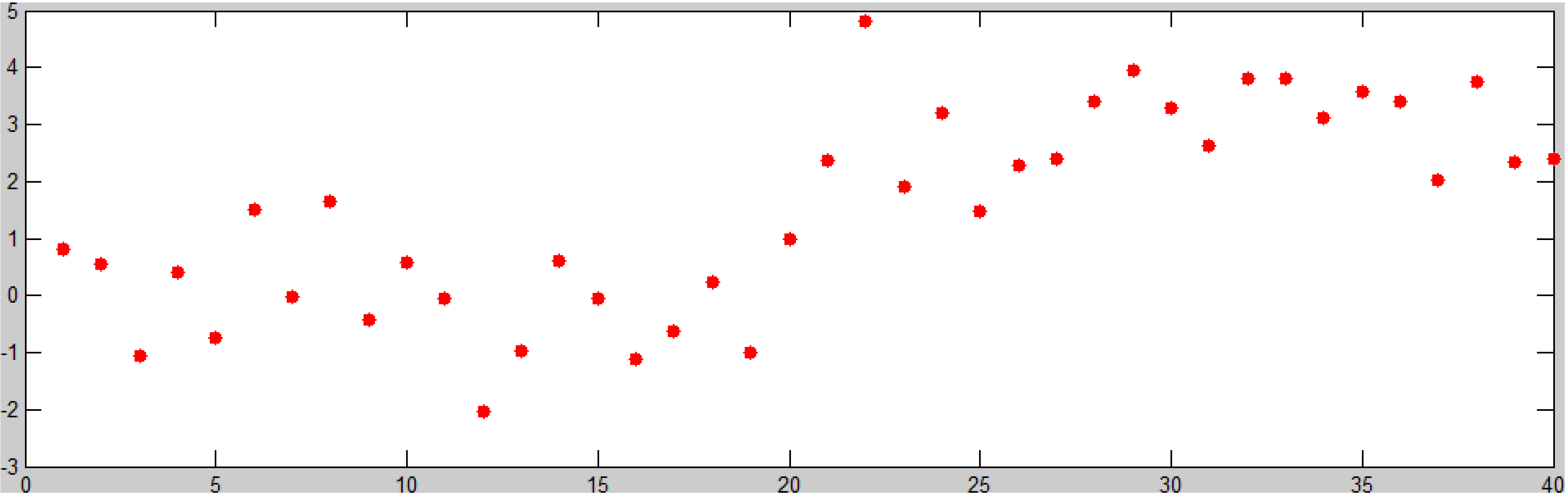
and determining when the data-generating process changes.

$$\mathbf{s}_t = \begin{cases} \mathbf{s}_t \sim \mathcal{P}_N & t < T^* \\ \mathbf{s}_t \sim \mathcal{P}_A & t \geq T^* \end{cases}$$

- **Unpredictability** of the change, $\mathcal{P}_A$ is unknown and sometimes also $\mathcal{P}_N$ is unknown.
- T^* is denoted the **change point**

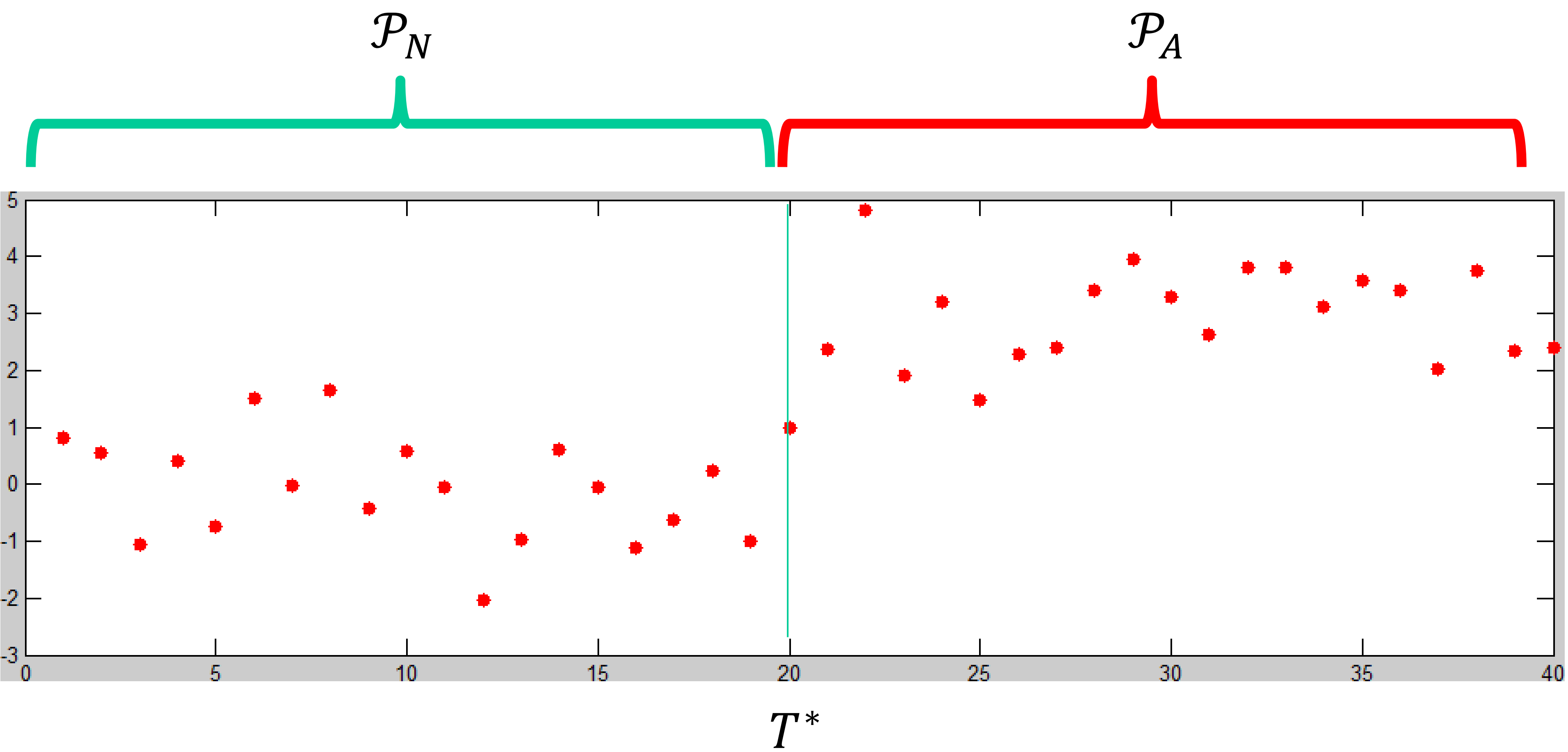
The change-detection problem

- There is a **temporal dimension** and we want do detect permanent shifts of the process



The change-detection problem

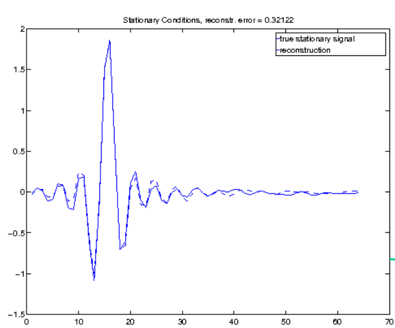
- There is a **temporal dimension** and we want do detect permanent shifts of the process



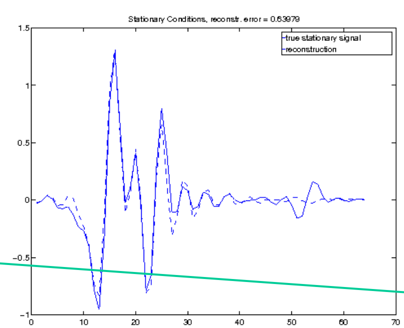
Sequential Monitoring

- We assueme a **training set** T of signals generated in stationary conditions are given
- We use these data **to learn a dictionary** $\widehat{D}$
- During the operational life, **signals arrives steadily**
- We perform **sparse coding** of each incoming signal $\mathbf{s}_i$ w.r.t. $\widehat{D}$ and comptute the change indicator $e(\mathbf{s}_i)$
- Use a sequential decision tool to determine, at each time t if the sequence $\{e(\mathbf{s}_i), i < t\}$ contains stationary data

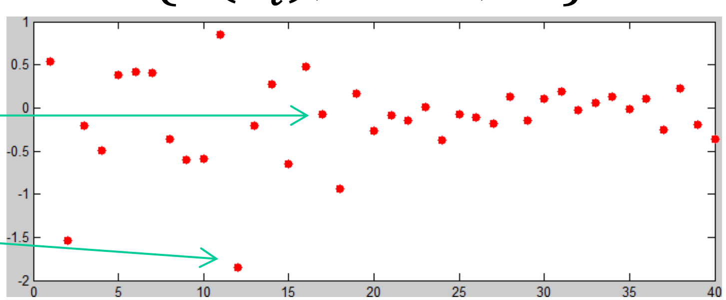
$$\{\mathbf{s}_i, i = 1, \dots\}$$



...



$$\{e(\mathbf{s}_i), i = 1, \dots\}$$



Sequential Monitoring

- Sequential Change-Detection Tests (CDTs) can be used for detecting changes in a stream of anomaly indicators [Basseville 93]
  - Data are analyzed incrementally
  - Decisions are taken online considering in principle the whole past sequence
- We adopt the **Change-Point Method** in [Ross 2011] based on the Lepage Test Statistic
- The **Lepage** test Statistic detects **changes in the scale and location** of an unknown random variable

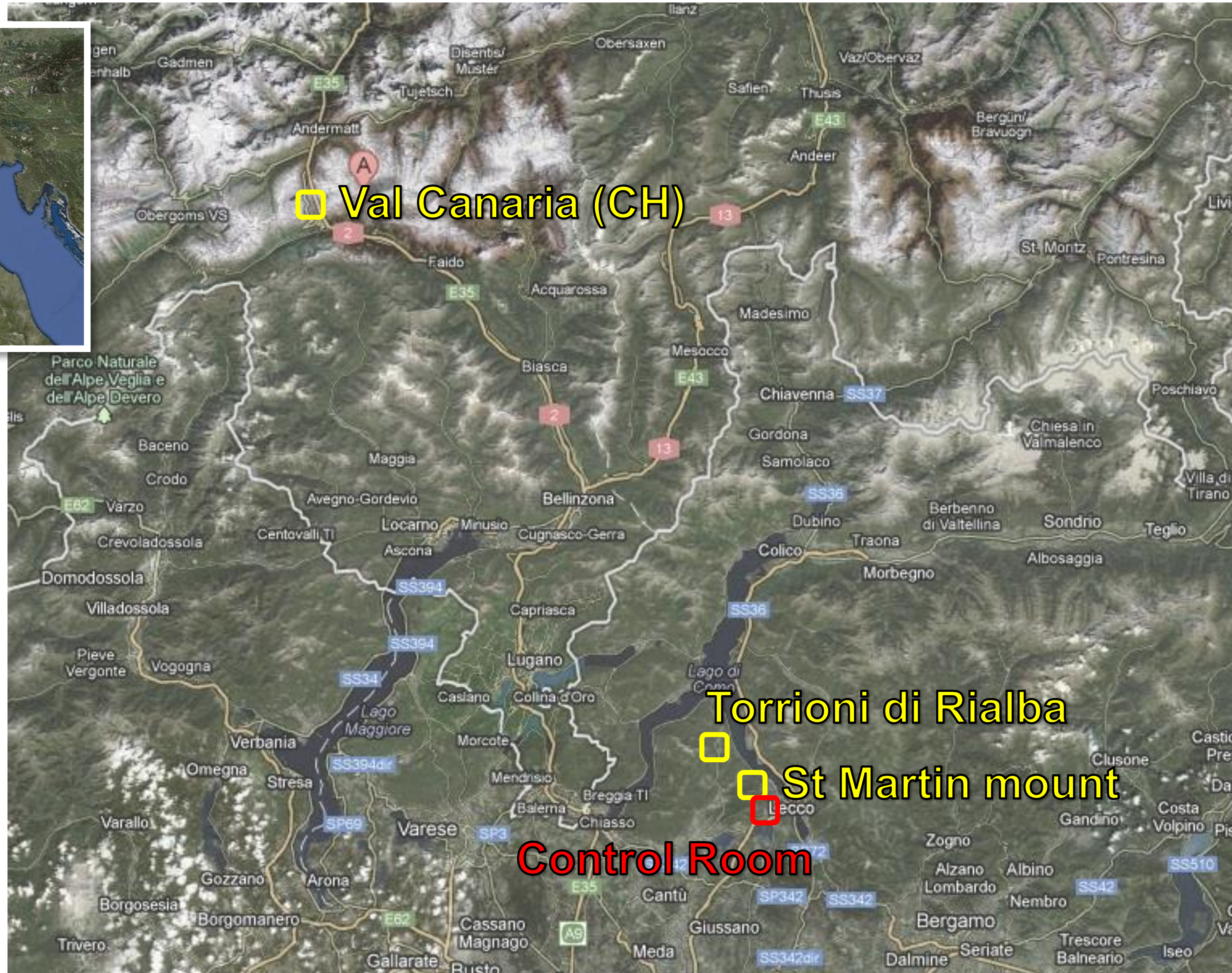
[Basseville 93] M. Basseville and I. V. Nikiforov, Detection of abrupt changes: theory and application. Upper Saddle River, NJ, USA:Prentice-Hall, Inc., 1993.

[Ross 2011] G. J. Ross, D. K. Tasoulis, and N. M. Adams, "Nonparametric monitoring of data streams for changes in location and scale," Technometrics, vol. 53, no. 4, pp. 379–389, 2011.

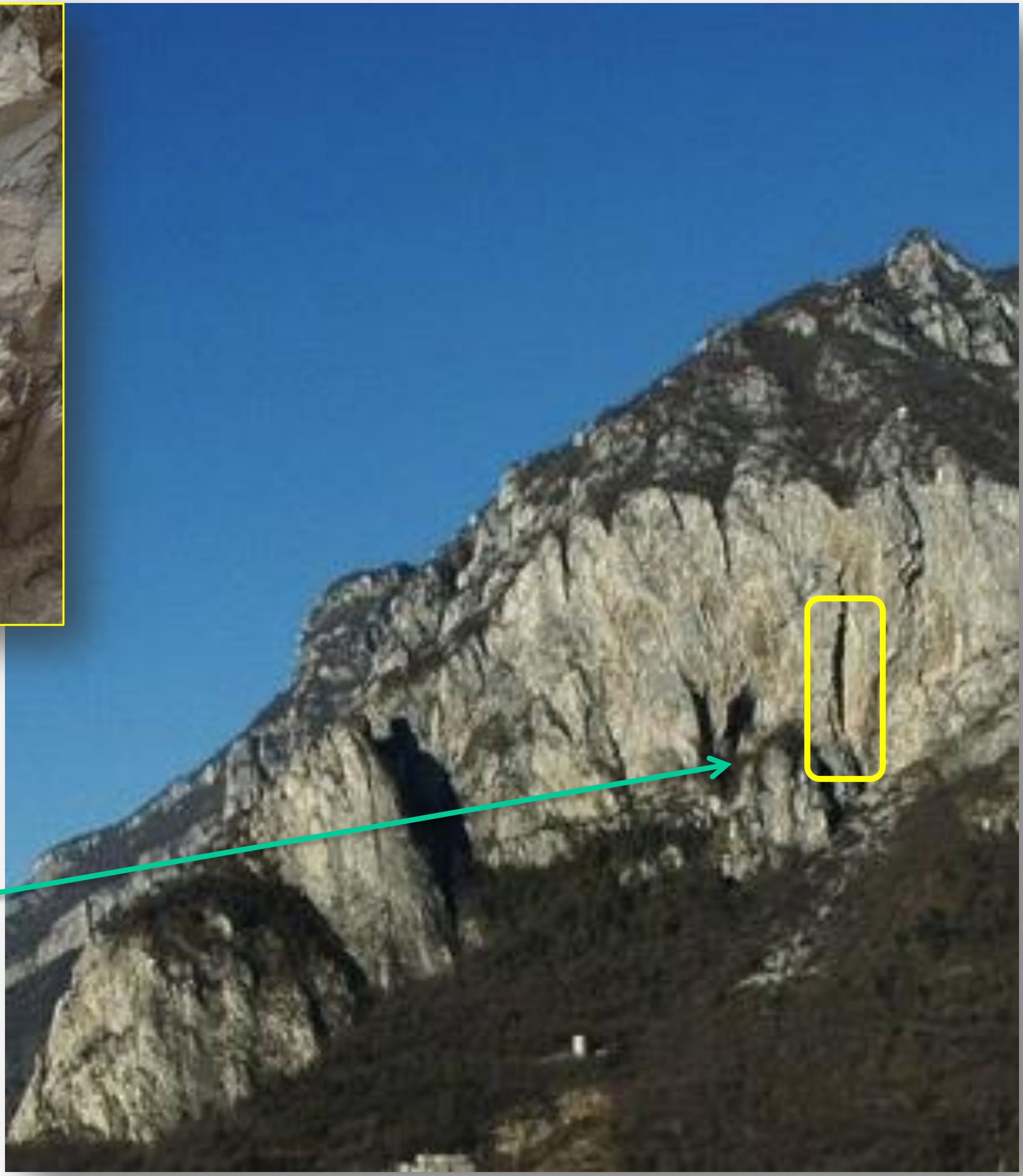
A CHANGE-DETECTION EXPERIMENT

on environmental monitoring application

Current deployments

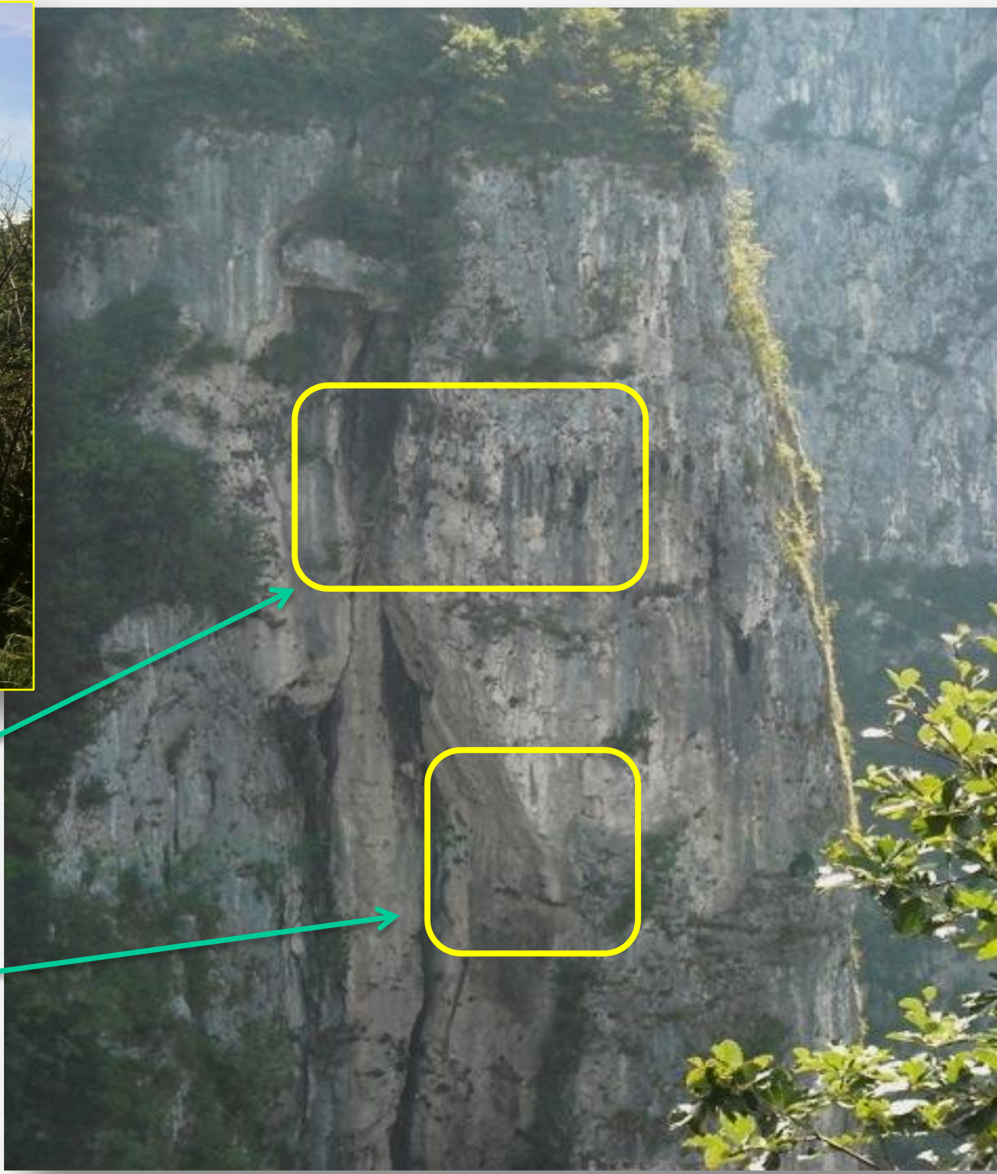


St Martin mount – LC, Italy



Hybrid monitoring system

Torrioni di Rialba - LC, Italy



Wireless monitoring system

Hybrid monitoring system

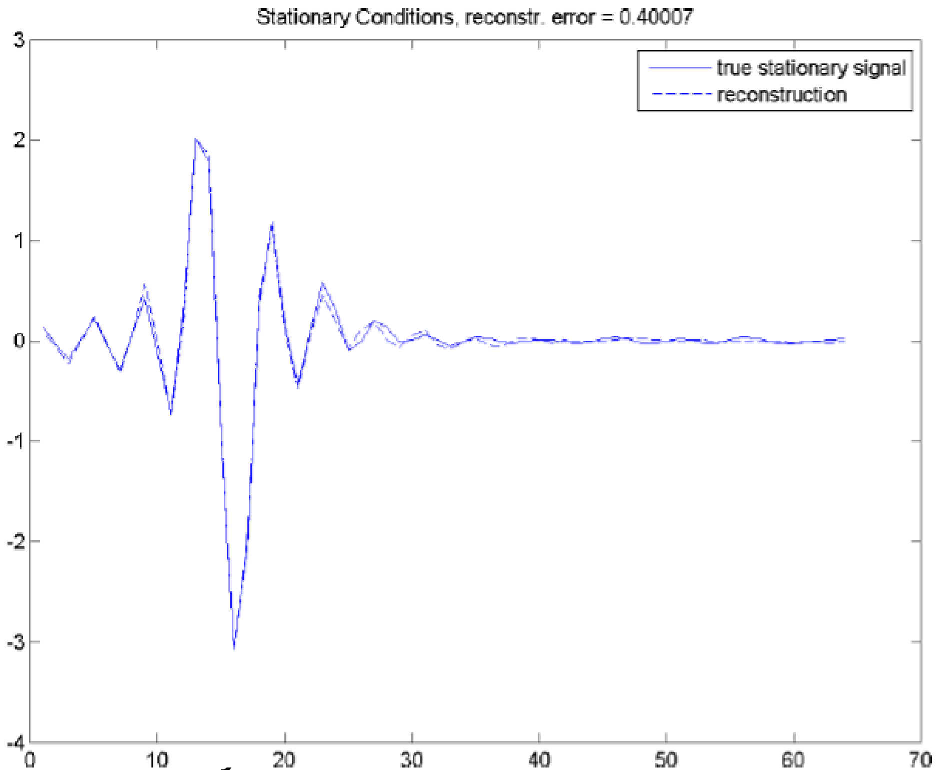
Environmental Monitoring

- We consideres acoustic emissions acquired by an wired/wireless sensor networks meant to monitor a rock faces
- 64 samples signals acquired at 2 KHz by a MEMS.
- Anomalies have been synthetically modified by randomly adding a DB4 wavelet basis atom

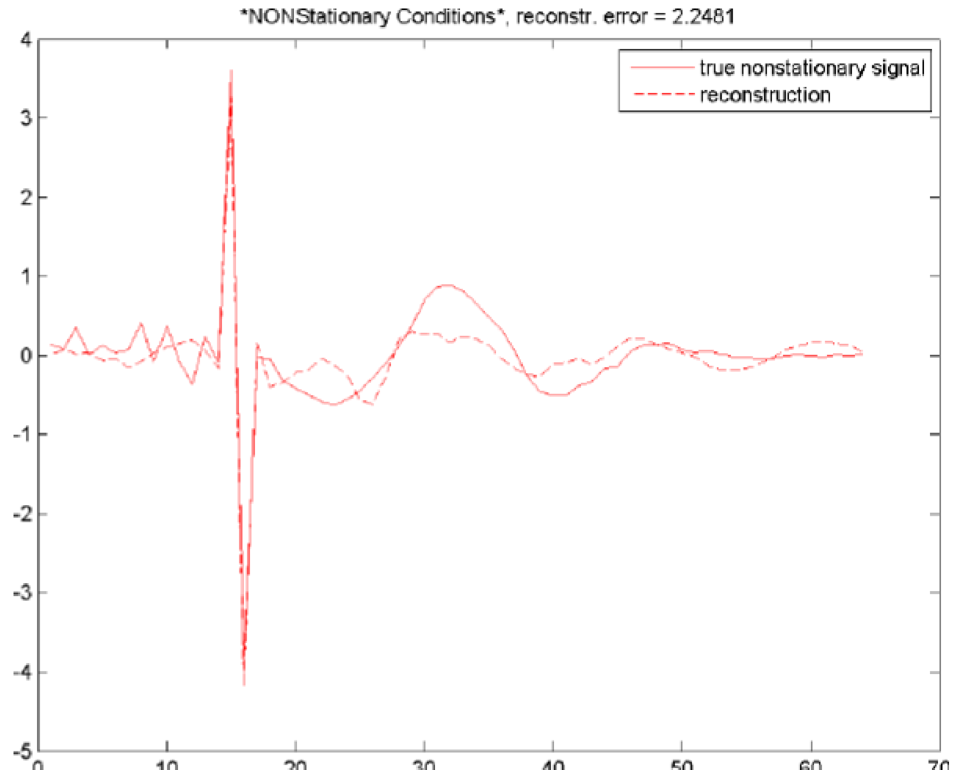
Environmental Monitoring

- We consideres acoustic emissions acquired by an wired/wireless sensor networks meant to monitor a rock faces
- 64 samples signals acquired at 2 KHz by a MEMS.

Example of Original Bursts



Example of Bursts Modified adding atoms from D_1



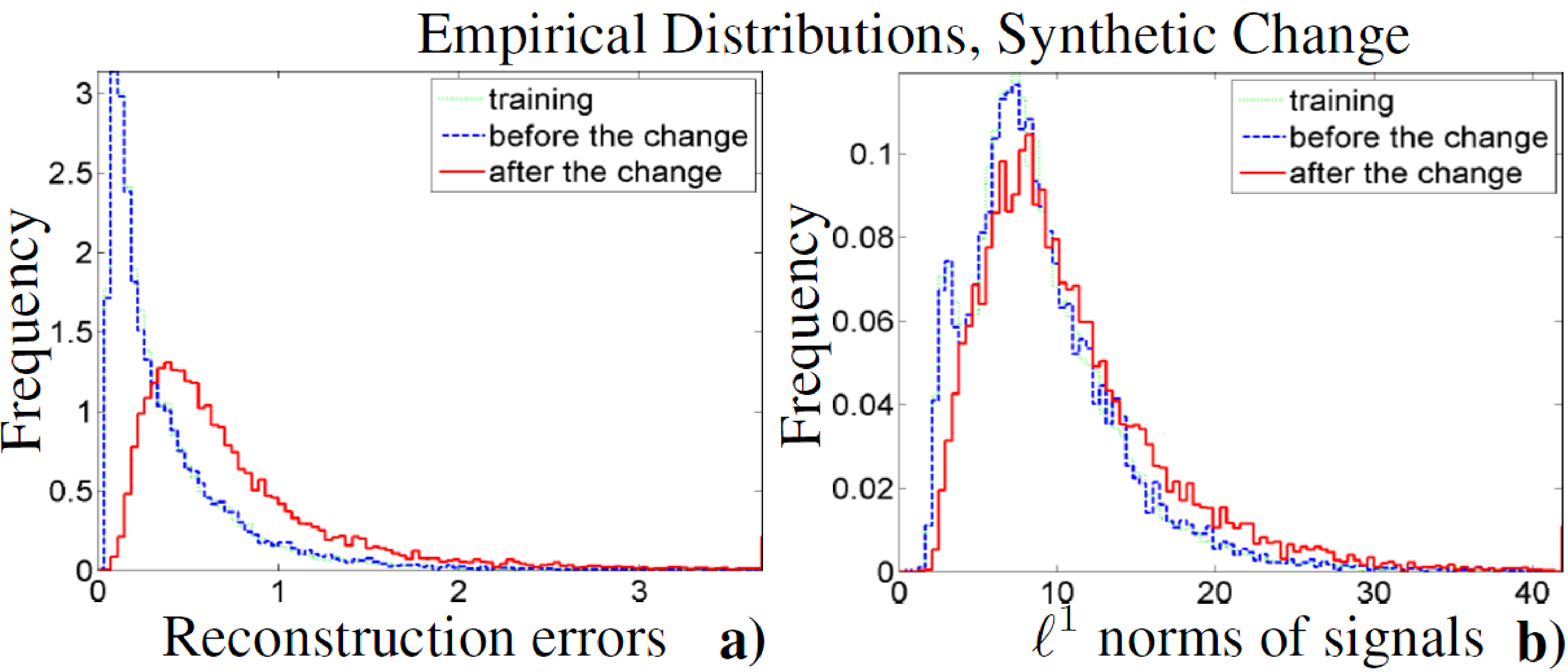
Environmental Monitoring

- We consideres acoustic emissions acquired by an wired/wireless sensor networks meant to monitor a rock faces
- 64 samples signals acquired at 2 KHz by a MEMS.
- Anomalies have been synthetically modified by randomly adding a DB4 wavelet basis atom
- We perform change detection by means of the Lepage CPM using the $e(\cdot)$ change indicator.
- We synthetically generate sequences containing 500 signals before and after the change
- Further details are provided in [Alippi 2014]

[Alippi 2014] C. Alippi, G. Boracchi, and B. Wohlberg, "Change detection in streams of signals with sparse representations," ICASSP 2014 , pp. 5252 – 5256.

Distribution of the change indicators

- To show the detectablity of the change we plot the empirical distribution of change indicator before and after the change.
- And compare it with the distirbution of $\|\mathbf{x}_i\|_2$ and $\|\mathbf{x}_i\|_1$



Distribution of the change indicators

- To show the detectablity of the change we plot the empirical distribution of change indicator before and after the change.
- And compare it with the distirbution of $||\mathbf{x}_i||_2$ and $||\mathbf{x}_i||_1$
- Change-detection performance using CPM are in line with the detectability of the change
  - Using $e(\cdot)$ all the changes are detected with no false positive with an average detection delay of 25 samples
  - Using $||\mathbf{x_i}||_1$ delay increased at 124, with 33% of FN
  - Using $||\mathbf{x_i}||_2$ no detections

CONCLUDING REMARKS

Conclusions

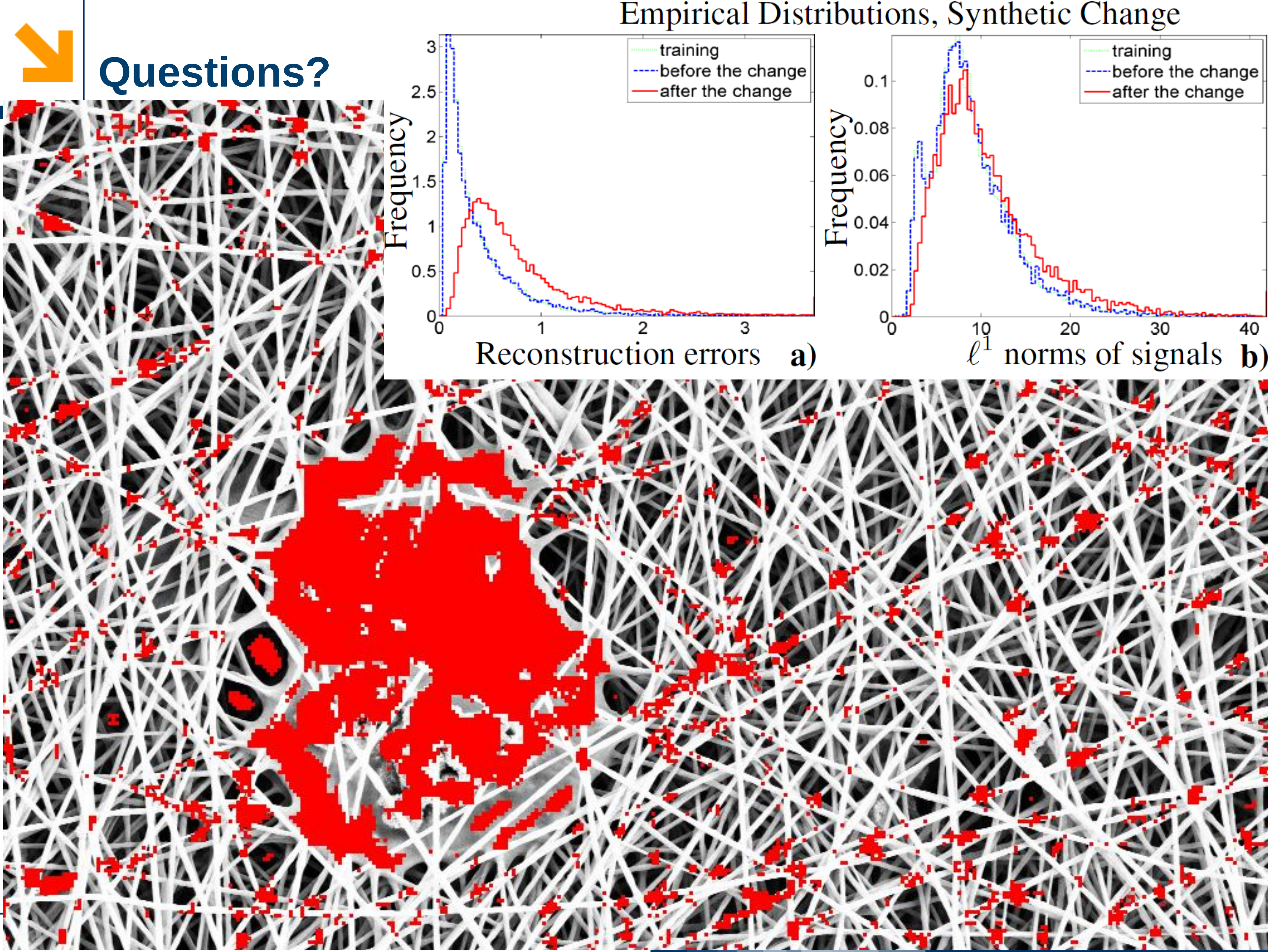
- Our experiments show that **sparse representation** allows to build **effective models** for **detecting**
  - anomalies
  - process changes

  affecting data **structures**
- Sparse models describe data that in stationary conditions are **heterogenous**: e.g., atoms of $\widehat{D}$ might be from different *classes.*

Ongoing works

- Ongoing works include:
  - the study of customized dictionary learning metods for performing change/anomaly detection
  - the application of the proposed system to other application domains such as EGC analysis to detect arrhythmia
  - For the specific case of SEM images we are performing a wider experimental campaign, also comparing with more starightforward techniques

Questions?



Empirical Distributions, Synthetic Change
training
before the change
after the change
Frequency
Reconstruction errors
a)
ℓ^1 norms of signals
b)