Anomaly Detection
Skoltech, March 10, 2023

Alexey Zaytsev

Assistant professor, Lab Head
Skoltech

Slides by A. Zaytsev & E. Romanenkova

Lecture plan

- Intro to Anomaly detection
- Unsupervised approaches for Anomaly detection. General idea
- Autoencoders for Anomaly detection
- GAN-based Anomaly detection
- Anomaly detection for Time Series

Intro to Anomaly Detection

Problem statement

The problem is to find objects that anomalous given training data

Normal data

Image Type	No Cat	Cat not on approach	Cat on approach	Cat with prey
Count of Images	6,542	9,504	6,689	**260**
Example				

Anomaly data

https://www.theverge.com/tldr/2019/6/30/19102430/amazon-engineer-ai-powered-catflap-prey-ben-hamm

Problem statement

The problem is to find objects that anomalous given training data

Normal data

Image Type	No Cat	Cat not on approach	Cat on approach	Cat with prey
Count of Images	6,542	9,504	6,689	**260**
Example				

Anomaly data

https://www.theverge.com/tldr/2019/6/30/19102430/amazon-engineer-ai-powered-catflap-prey-ben-hamm

Problem examples

- Fraud detection 🕵️
- Failure detection for an airplane ✈️
- Intrusion detection 😱
- Earthquake prediction 💥

Problem examples

- Fraud detection
- Failure detection for an airplane
- Intrusion detection
- Earthquake prediction

Typical challenges:

- Requires problem-specific knowledge => new problem – new approach
- Hard to identify something we don’t see
- Bunch of various problem statements => how to define what is anomaly?

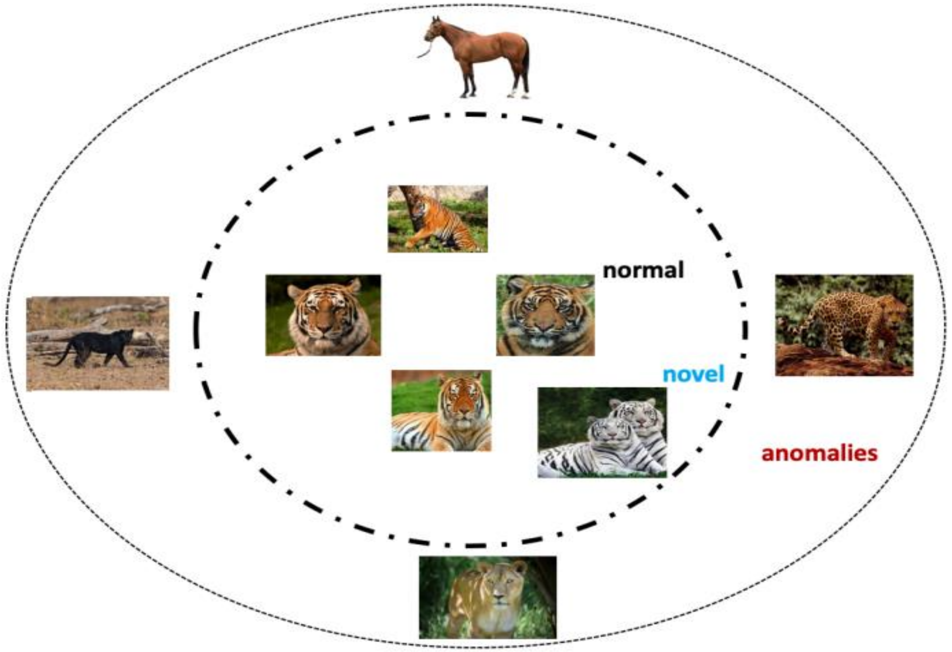
Different taxonomies

By method:
- Supervised
- Unsupervised
- Semi-supervised

By problem statement:
- Anomaly detection
- Novelty detection

Anomaly Detection

By data:
- Structured data
- Sequential data
- Unstructured data
- Semi-structured data



https://arxiv.org/pdf/1901.03407.pdf

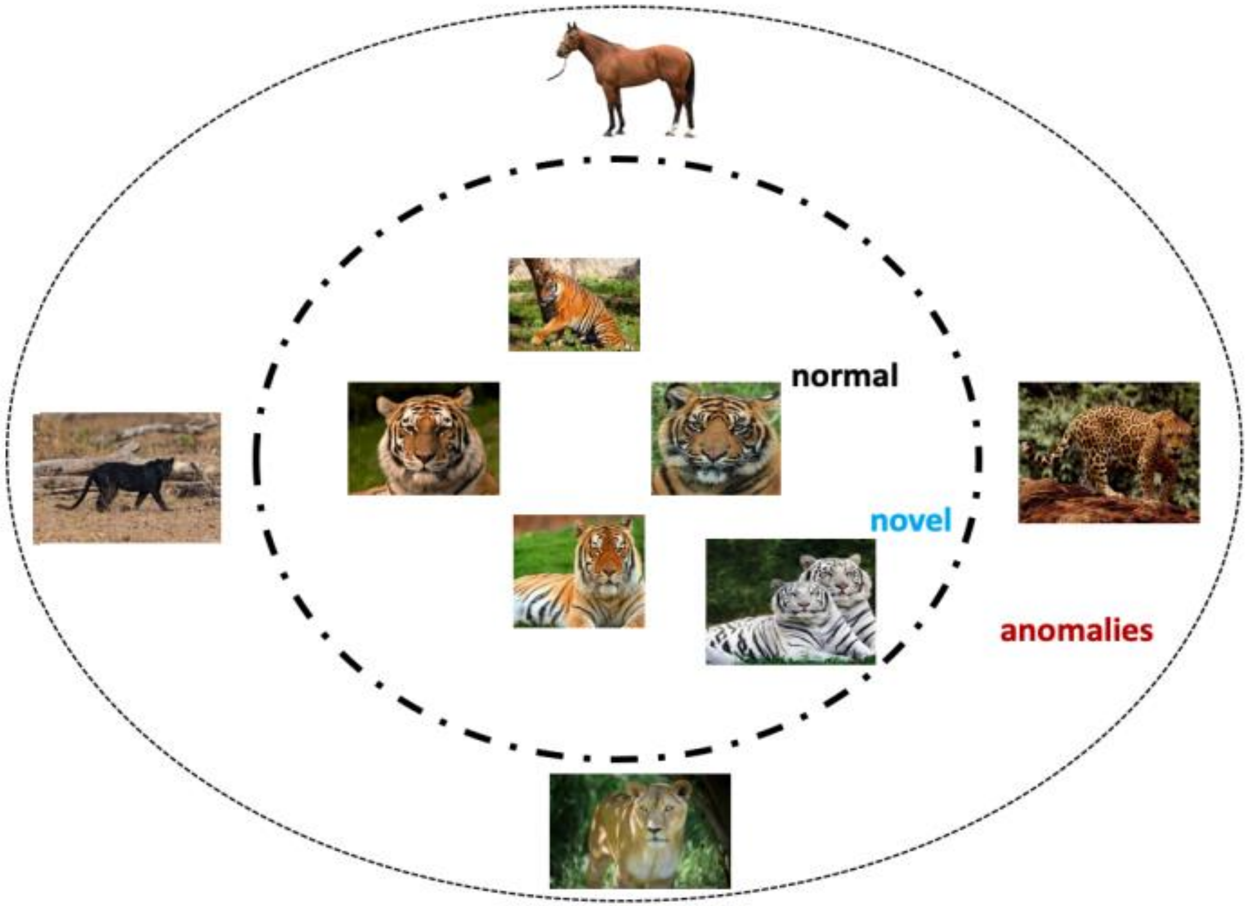
Taxonomy with respect to problem statement

Novelty detection

We have never seen the objects of this class

Anomaly detection

Outliers, can be already in the sample



https://arxiv.org/pdf/1901.03407.pdf

Anomaly type taxonomy

Normal data

Image Type	No Cat	Cat not on approach	Cat on approach	Cat with prey
Count of Images	6,542	9,504	6,689	**260**
Example				

Point Anomaly

Group Anomaly

Contextual Anomaly

In 2019 → He is mad! We should avoid him. (anomaly)

In 2020 → He takes care of himself and others. Well done! (normal)

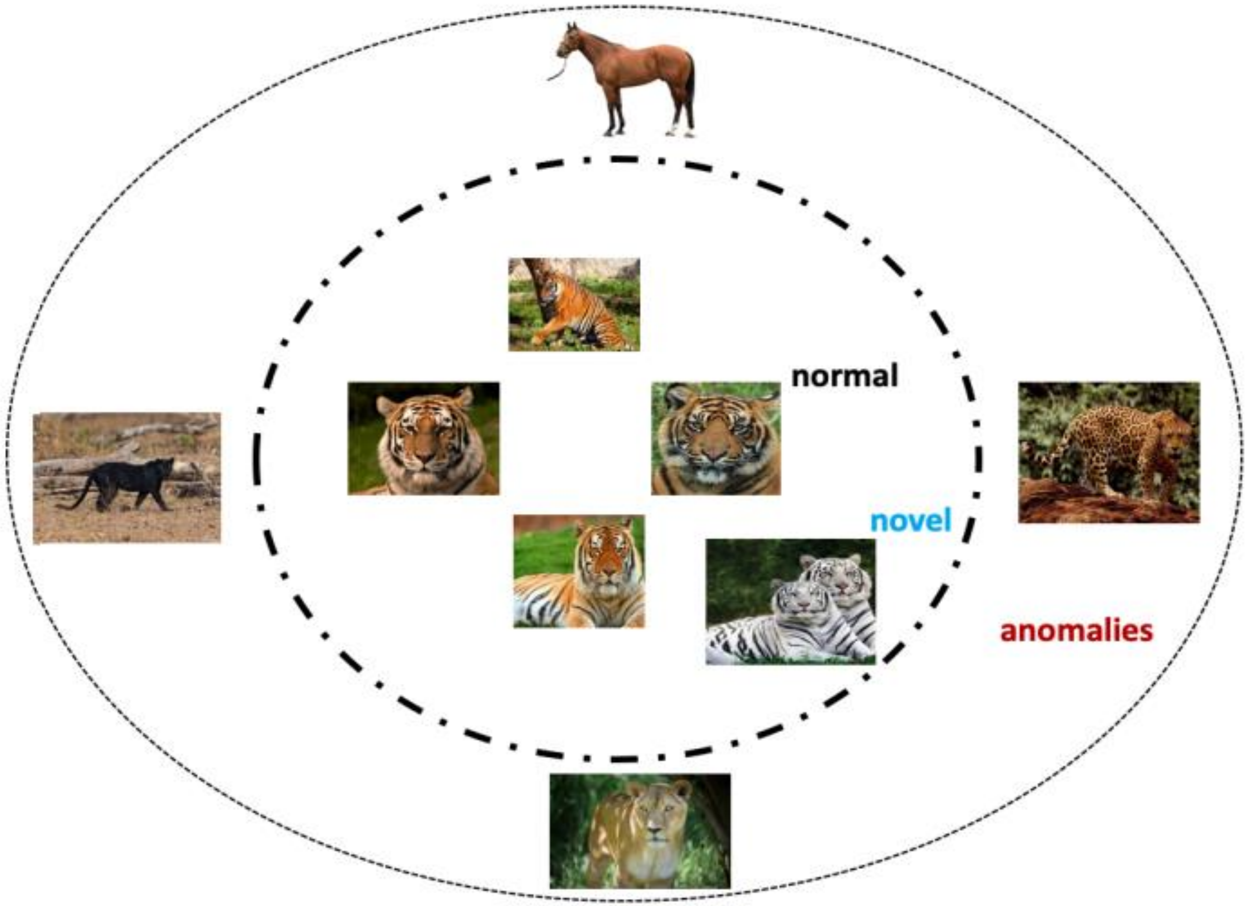
Taxonomy by type of anomalies

Novelty detection

We have never seen the objects of this class

Anomaly detection

Outliers, can be already in the sample



https://arxiv.org/pdf/1901.03407.pdf

Well-defined anomaly assumption

WDAD: the anomalies are drawn from a well-defined probability distribution *Example: repeated instances of known machine failures*

The WDAD assumption is often risky:

- adversarial situations (fraud, insider threats, cyber security)
- diverse set of potential causes (novel device failure modes)
- user’s notion of “anomaly” changes with time (e.g., anomaly == “interesting point”)

Supervised anomaly detection

Supervised ID is just imbalanced classification

Weights for classes

- *Proved not to be helpful in most cases*

Resampling methods

- *Undersampling*
- *Oversampling/data generation: SMOTE, etc.*

How to choose which method to use?
How to choose resampling parameter?

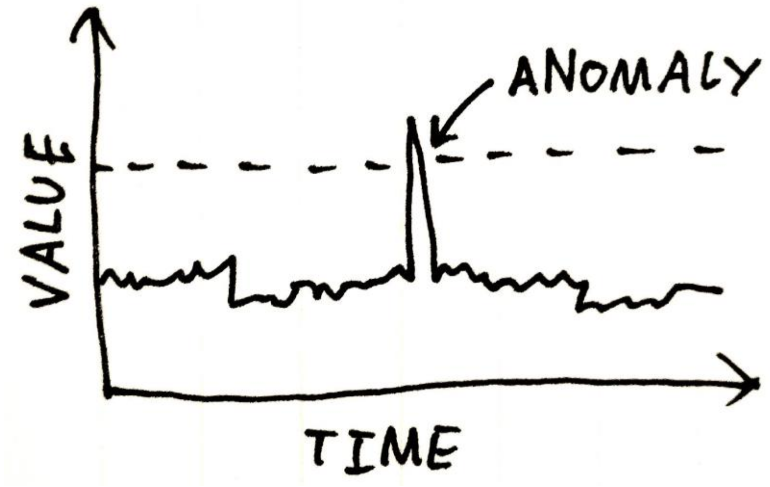
Unsupervised & semi-supervised approaches to Anomaly Detection

Classic approach to anomaly detection

1. Construct anomaly score *s(x)* using data

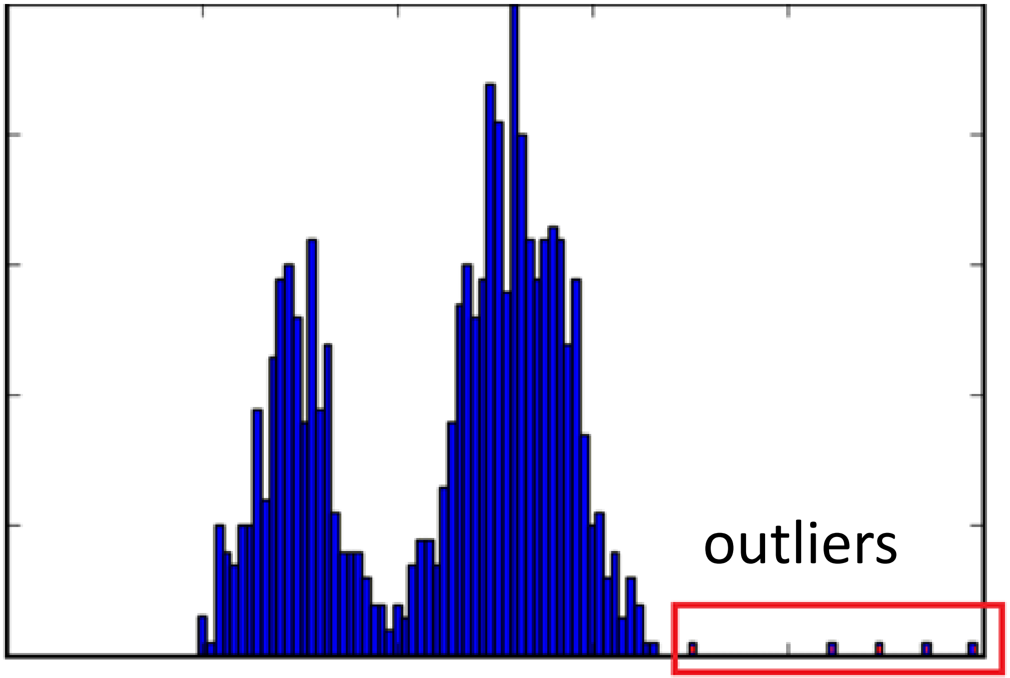
2. Signal about anomaly if anomaly score is greater than some threshold τ

Threshold selection τ is a separate problem, as we often have only positive examples

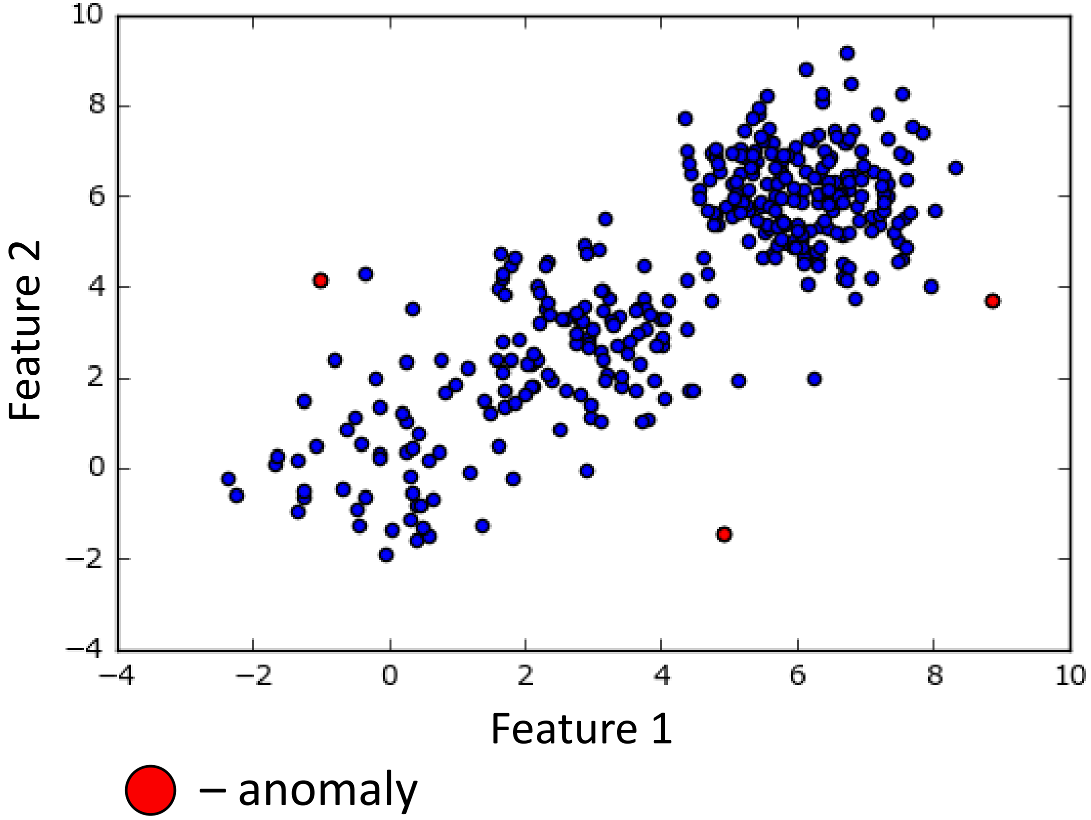


Analytics for selected features

One feature anomaly



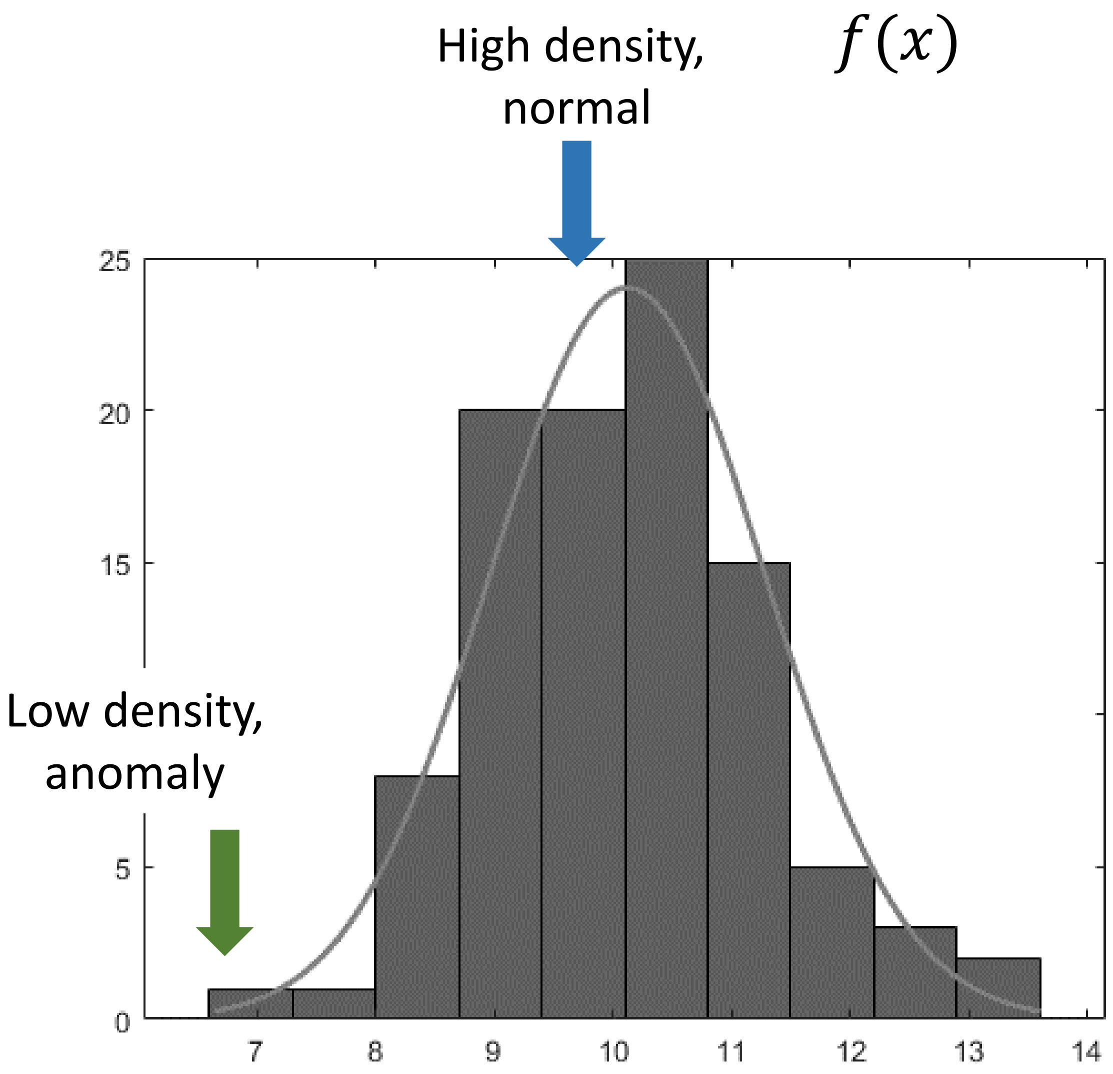
Pair of features anomaly



Anomaly based on one feature

Histogram: real values from density

Curve – recovered density for data $f(x)$

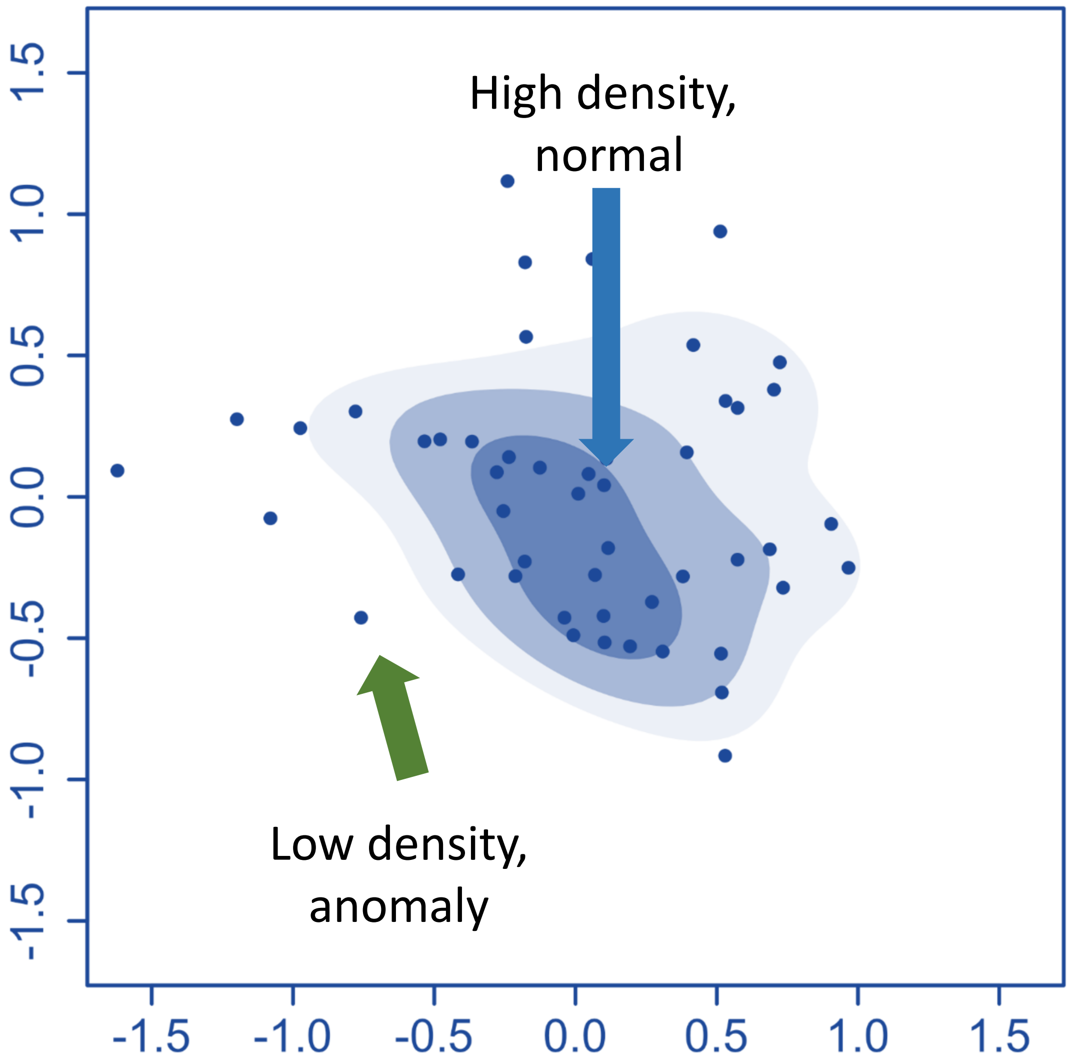


Skoltech

Anomaly based on a pair of features

Points are real data

Shading is a density estimation



LODA anomaly detector

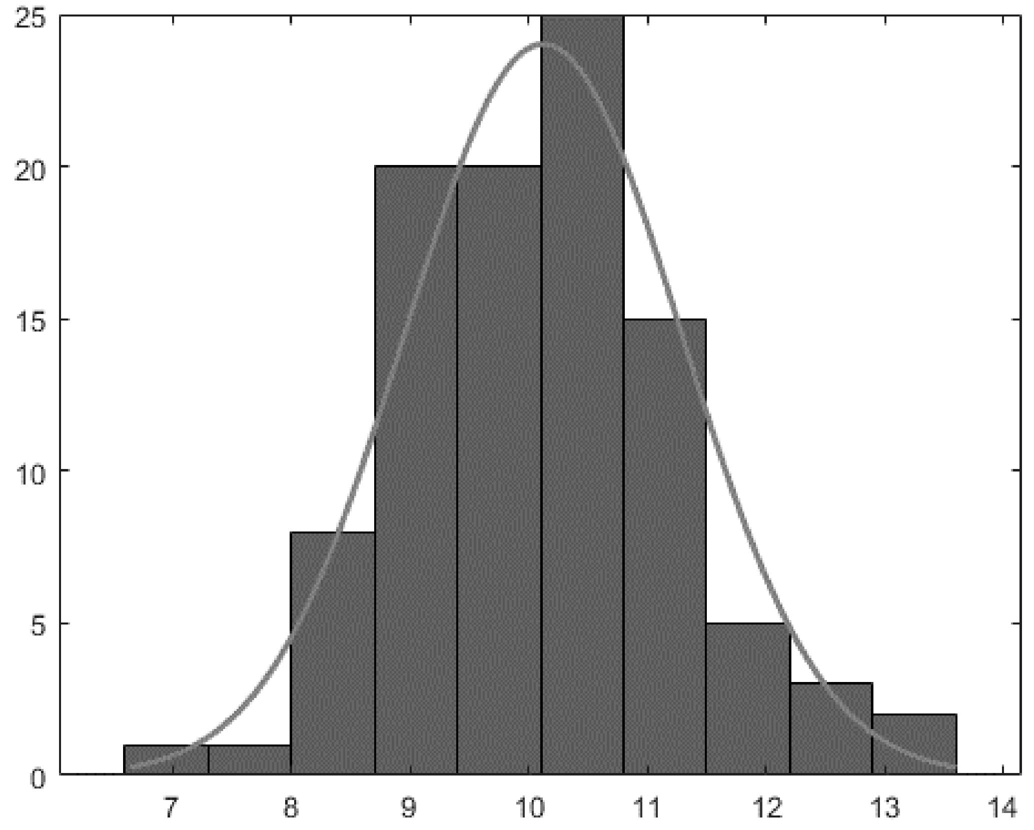
1. Build *M* random projections

2. Estimate density for each projection $f_i(x)$

3. Mean density for all projections

$$\mathrm{S}(x) = -\frac{1}{M}\sum_{i=1}\log f_i(x)$$

Get mean «surprise» $\mathrm{E}\,\mathrm{S}(x)$



Details on methods for anomaly detection

- Isolation forest **iForest**
- Isolation Nearest Neighbours Ensembles **INNE**

https://federation.edu.au/__data/assets/pdf_file/0011/443666/ICDM2018-Tutorial-Final.pdf

Top recommendations from the three studies

- Isolation based methods: iForest and iNNE [8]
- All known weaknesses of iForest are overcome by iNNE, with higher time cost; but still has significantly lower time cost than kNN.
- Nearest neighbour-based methods: aNNE and kNN
- Clustered anomalies: ABOD & LOF
- iNNE [5] can do well in detecting clustered anomalies.
- Kernel Mahalanobis [6]: The key weakness is the time cost

- Among the compared methods, iForest and iNNE have the highest detection accuracy and also have the lowest time cost.
- Simple methods should be used as the baseline to justify any more complicated methods.

[6] Hoffmann, H. (2007). Kernel PCA for Novelty Detection, Pattern Recognition, 40(3), 863–874.
[8] Bandaragoda, T. R., Ting, K. M., Albrecht, D., Liu F. T., Wells, J. R. (2018). Isolation-based Anomaly Detection using Nearest Neighbour Ensembles. Computational Intelligence. Doi:10.1111/coin.12156.

https://federation.edu.au/__data/assets/pdf_file/0011/443666/ICDM2018-Tutorial-Final.pdf

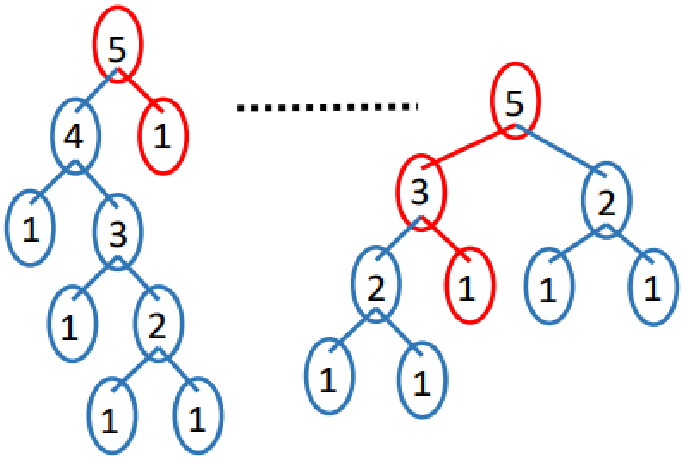
Isolation forest

- A collection of isolation trees (iTrees)

- Each iTree isolates every instance from the rest of the instances in a given sample

- Anomalies are ‘few and different’
  - More susceptible to isolation
  - Shorter average path

$$Score(\boldsymbol{x}) = \frac{1}{t}\sum_{i=1}^{t} \ell_i(\boldsymbol{x})$$

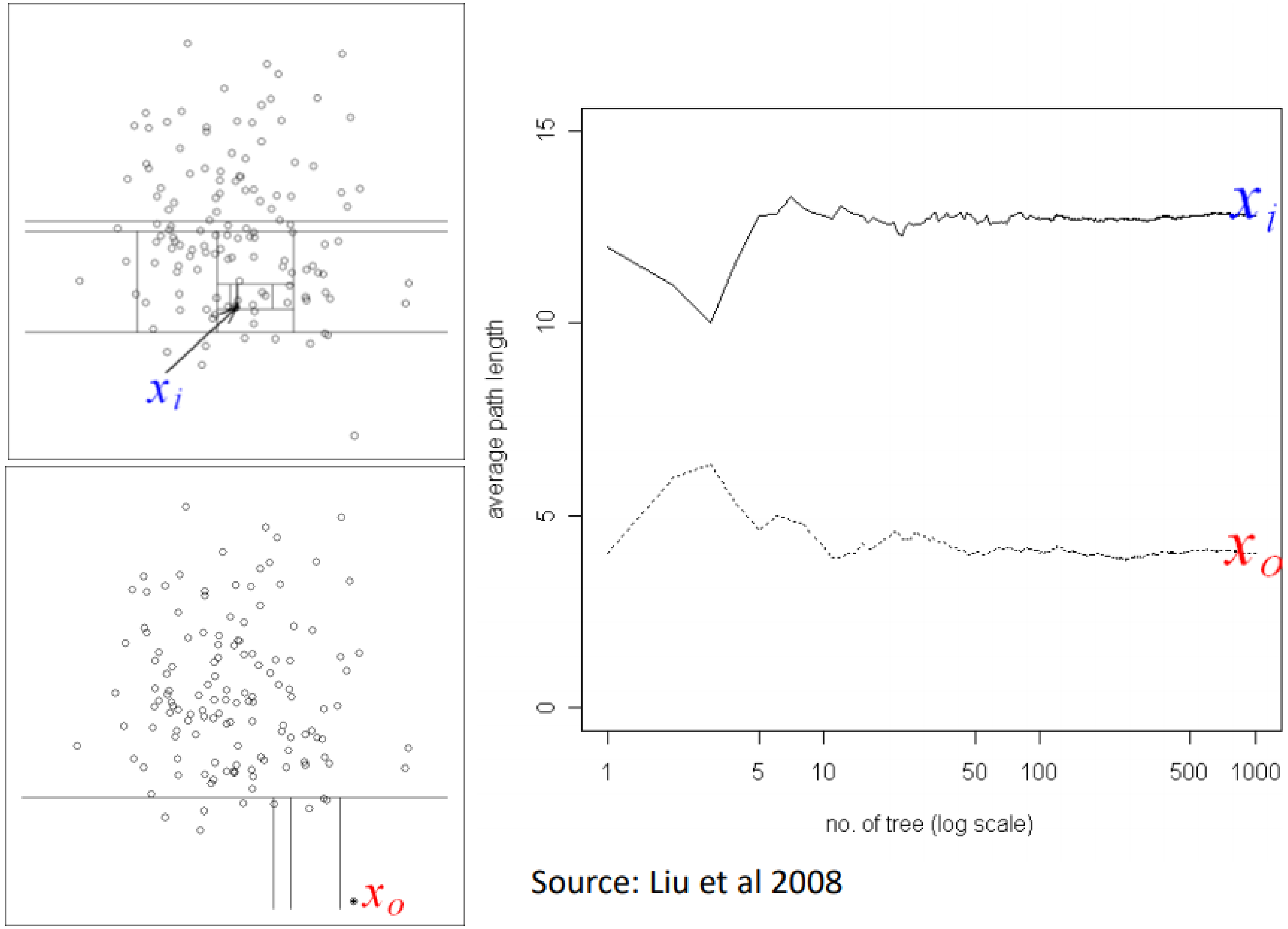


where $\ell_i(\boldsymbol{x})$ is the path length of $\boldsymbol{x}$ traversed in tree i

[Liu et al ICDM 2008]

https://federation.edu.au/__data/assets/pdf_file/0011/443666/ICDM2018-Tutorial-Final.pdf

Isolation example



Source: Liu et al 2008

[Liu et al ICDM 2008]

https://federation.edu.au/__data/assets/pdf_file/0011/443666/ICDM2018-Tutorial-Final.pdf

INNE definitions

Let $D \subset \Re^d$ be a given data set, and let $\|a - b\|$ denote the Euclidean distance between a and b, where $a, b \in \Re^d$.

Let $\mathcal{S} \subset D$ be a subsample of size ψ selected randomly without replacement from a dataset $D \subset \Re^d$; and η_x be the nearest neighbour of x.

Definition 1: A hypersphere $B(c)$ centred at c with radius $\tau(c) = \|c - \eta_c\|$, is defined to be $\{x : \|x - c\| < \tau(c)\}$, where $x \in \Re^d$ and $c, \eta_c \in \mathcal{S}$.

Definition 2: Isolation score for $x \in \Re^d$ based on $\mathcal{S}$ is defined as follows:

$$I(x) = \begin{cases} 1 - \frac{\tau(\eta_{cnn(x)})}{\tau(cnn(x))}, & \text{if } x \in \bigcup_{c \in \mathcal{S}} B(c) \\ 1, & \text{otherwise} \end{cases}$$

where $cnn(x) = \arg\min_{c \in \mathcal{S}} \{\tau(c) : x \in B(c)\}$.

Definition 3: iNNE has a set of t sets of hyperspheres, generated from t subsamples S_i, defined as follows:

$$\left\{ \left\{ B(c) : c \in \mathcal{S}_i \right\} : i = 1, \ldots, t \right\}$$

[Bandaragoda et al, 2018]

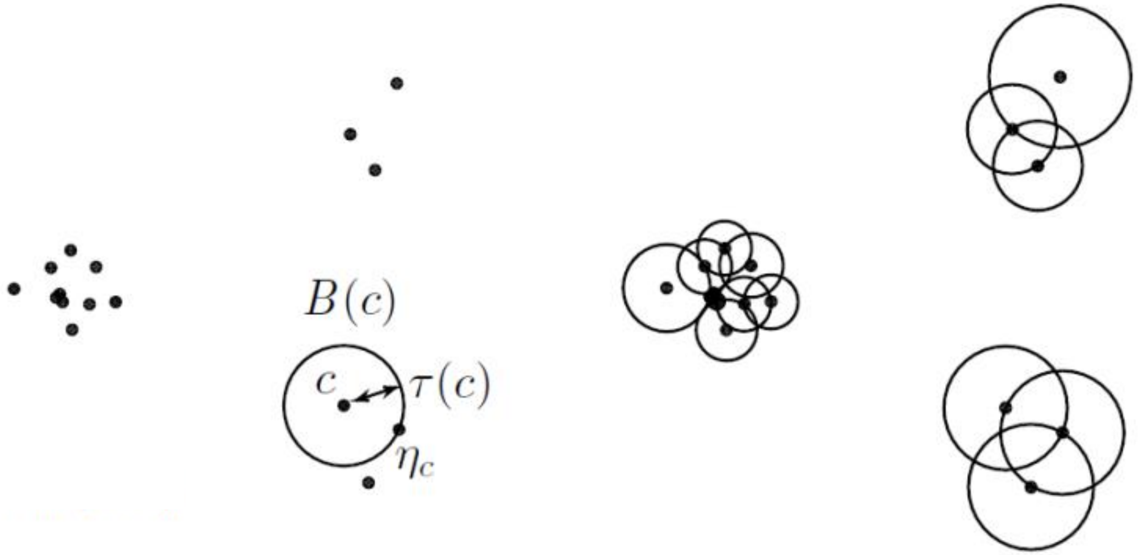
https://federation.edu.au/__data/assets/pdf_file/0011/443666/ICDM2018-Tutorial-Final.pdf

Skoltech
Skolkovo Institute of Science and Technology

INNE idea

- Sample S is selected randomly from the given dataset
- Ball $B(c)$ is created centering each $c \in S$
- Radius $\tau(c) = ||c - \eta_c||$

  η_c is the nearest neighbour of c where $c, \eta_c \in S$

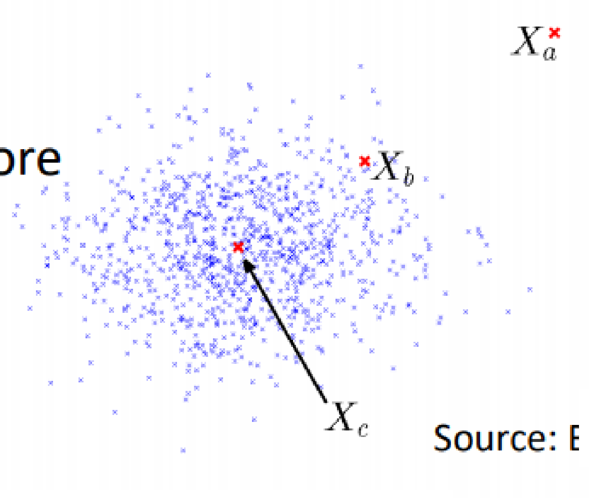


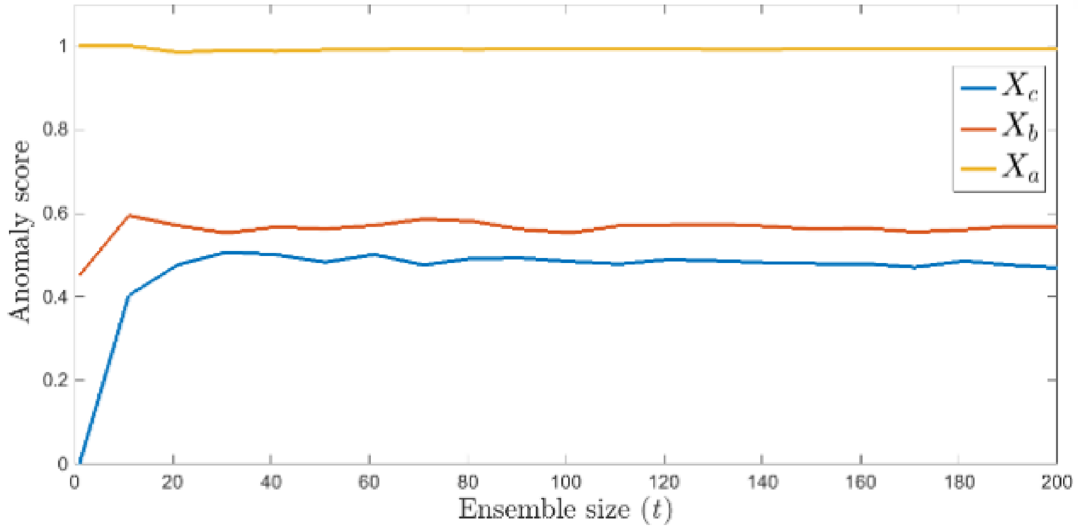
[Bandaragoda et al, 2018]

https://federation.edu.au/__data/assets/pdf_file/0011/443666/ICDM2018-Tutorial-Final.pdf

Example

- X_a has the maximum anomaly score
- X_b has a lower anomaly score
- X_c has the lowest anomaly score





[Bandaragoda et al, 2018]

https://federation.edu.au/__data/assets/pdf_file/0011/443666/ICDM2018-Tutorial-Final.pdf

Classic approach revisited

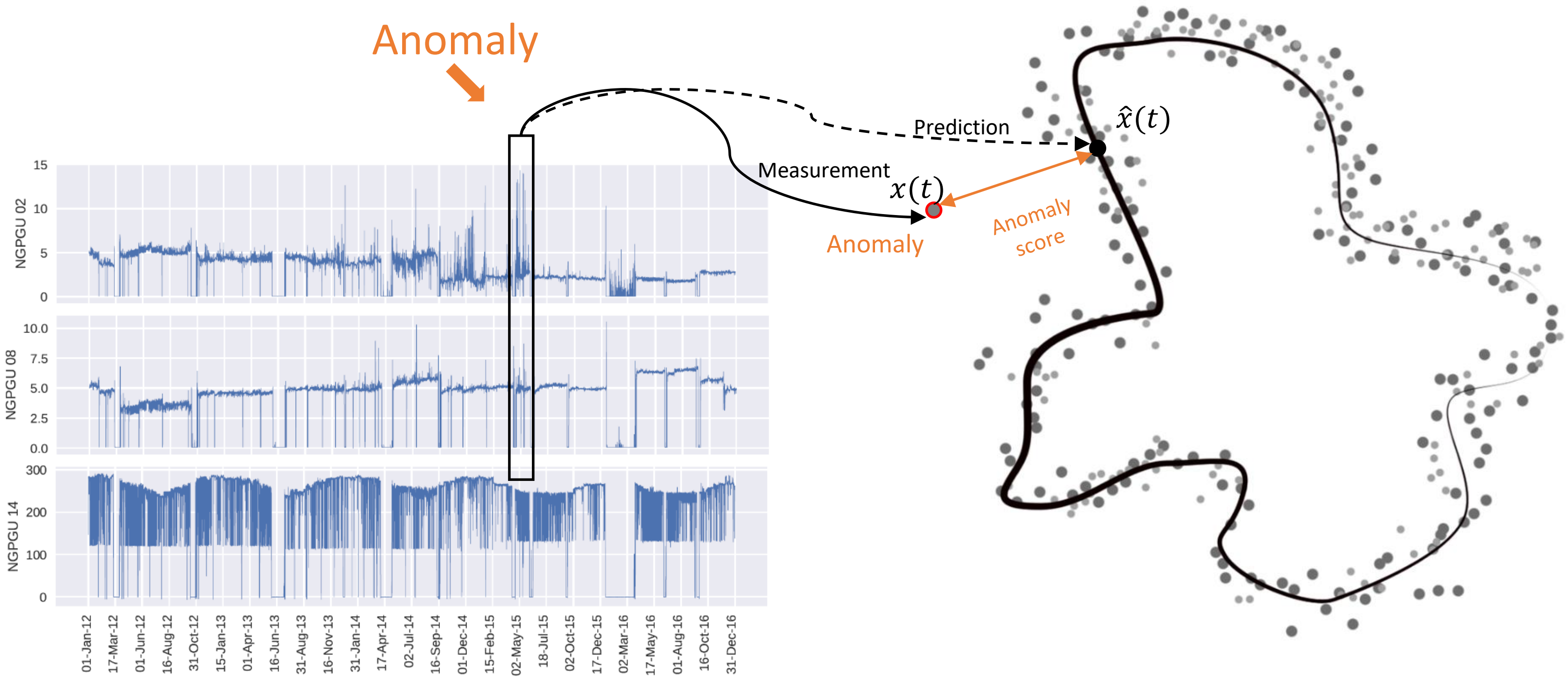
- A sample $D = \{\mathbf{x}_i\}_{i=1}^{n}$ is given, each $\mathbf{x} \in \mathbb{R}^d$.
- Construct models

$$\hat{x}_1 = f_1(x_2, x_3, \ldots, x_d),$$
$$\ldots$$
$$\hat{x}_d = f_d(x_1, x_2, \ldots, x_{d-1}).$$

- We have d anomaly scores for $\mathbf{x} = \{x_1, x_2, \ldots, x_d\}$:

$$s_i(\mathbf{x}) = |\hat{x}_i - x_i|, i = \overline{1, d}.$$

Unsupervised anomaly detection. General approach



Local history

Embedding

Structured data anomaly detection

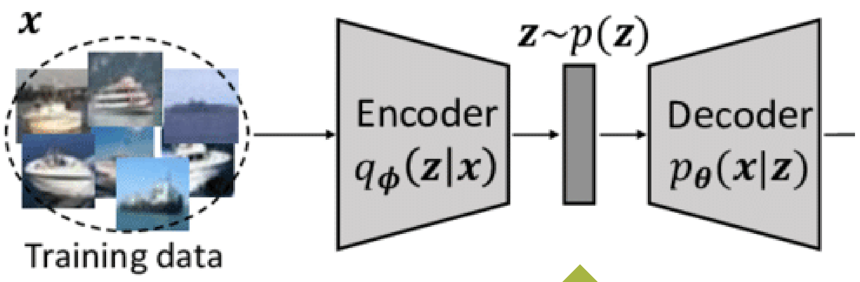
Unsupervised anomaly detection. General approach

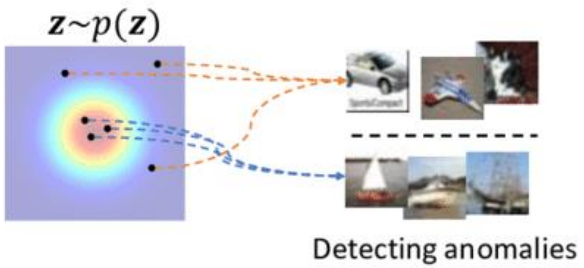
- A sample $D = \{\mathbf{x}_i\}_{i=1}^n$ is given, each $\mathbf{x} \in \mathbb{R}^d$.
- Construct encoder and decoder model

$$\begin{aligned} \mathbf{z}_i &= e(\mathbf{x}_i), \\ \mathbf{x}_i \approx \hat{\mathbf{x}}_i &= d(\mathbf{z}_i) = d(e(\mathbf{x}_i)). \end{aligned}$$

- We have an anomaly score $s(\mathbf{x})$ for any $\mathbf{x}$:

$$s(\mathbf{x}) = \|\hat{\mathbf{x}}_i - \mathbf{x}_i\|.$$

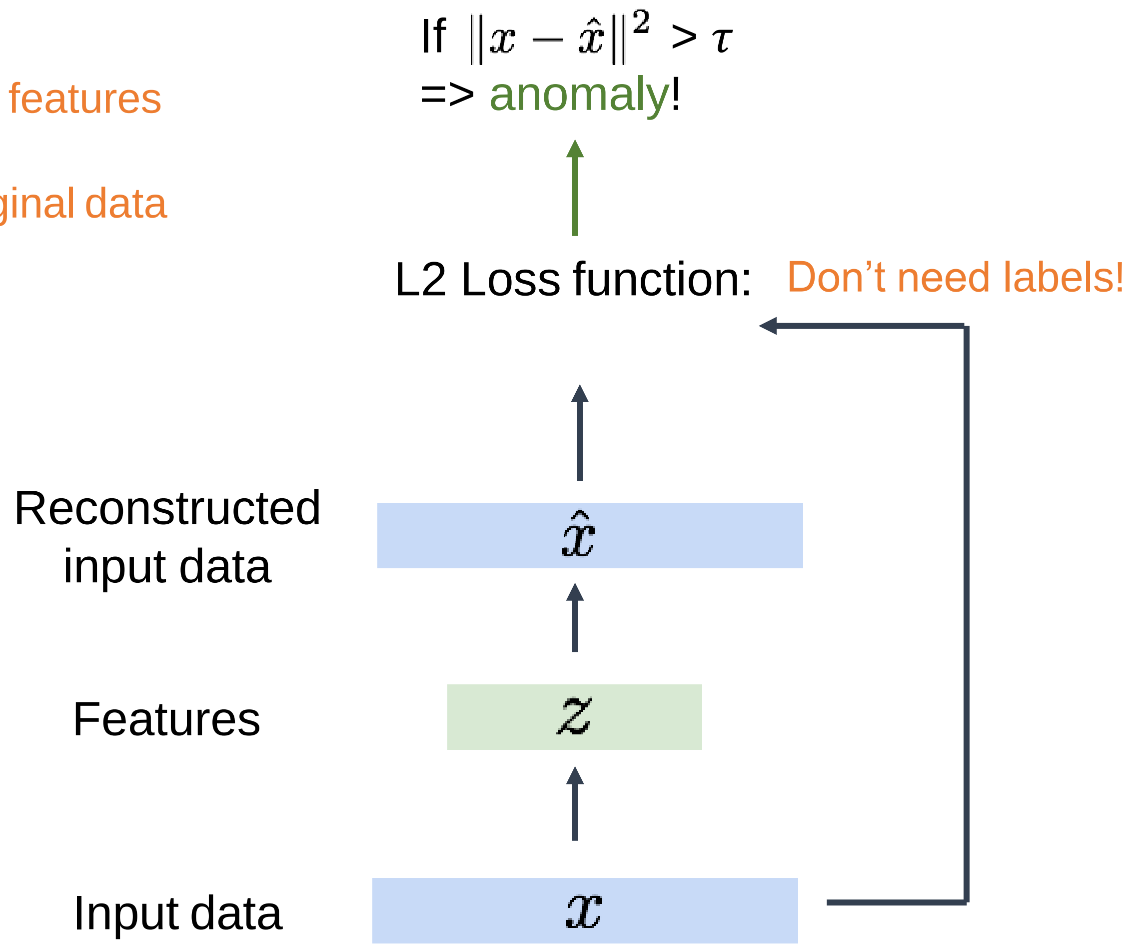




Encoder, decoder examples: PCA, Autoencoder

Autoencoder. General idea

Train such that features can be used to reconstruct original data



Reconstructed data

Encoder: 4-layer conv
Decoder: 4-layer upconv

Input data

Slides were adapted from lecture by Fei-Fei Li & Justin Johnson & Serena Yeung

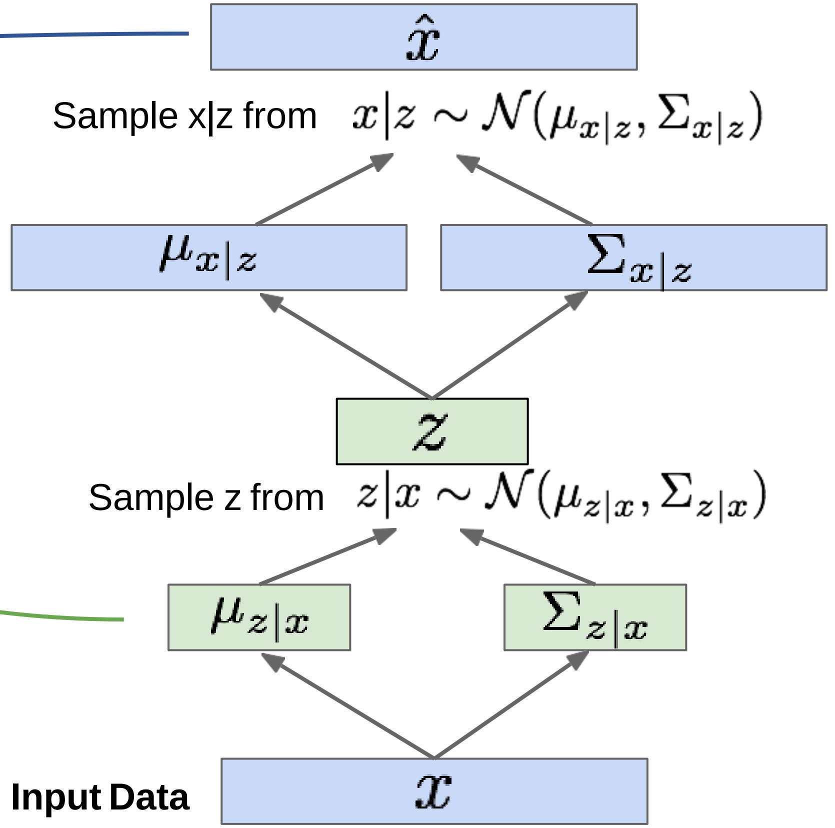
Variational Autoencoder

We maximize the likelihood lower bound

Maximize likelihood of original input being reconstructed

$$\underbrace{\mathbf{E}_z\left[\log p_\theta(x^{(i)} \mid z)\right] - D_{KL}(q_\phi(z \mid x^{(i)}) \,||\, p_\theta(z))}_{\mathcal{L}(x^{(i)}, \theta, \phi)}$$

Make approximate posterior distribution close to prior



Decoder network $p_\theta(x|z)$

Encoder network $q_\phi(z|x)$

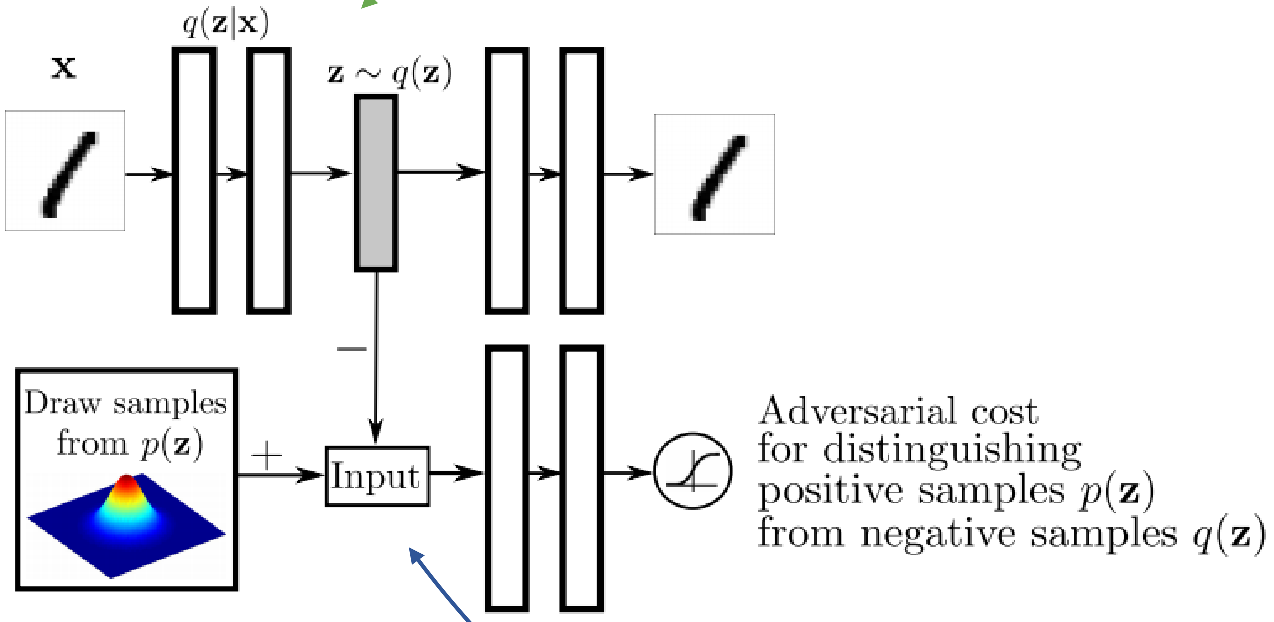
https://arxiv.org/pdf/1606.05908.pdf

Adversarial Autoencoder

Reconstruction Loss

Autencoder / Generator

Discriminator



Adversarial loss (cross-entropy for true VS fake detection)

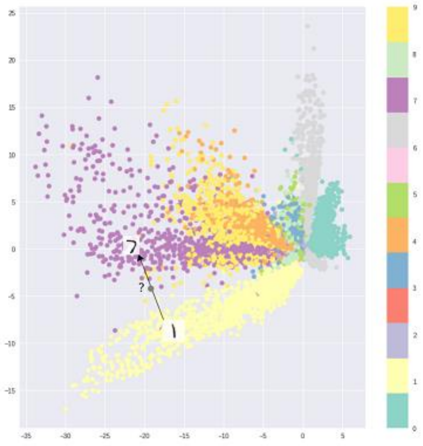
https://arxiv.org/pdf/1511.05644.pdf

Taxonomy of autoencoders

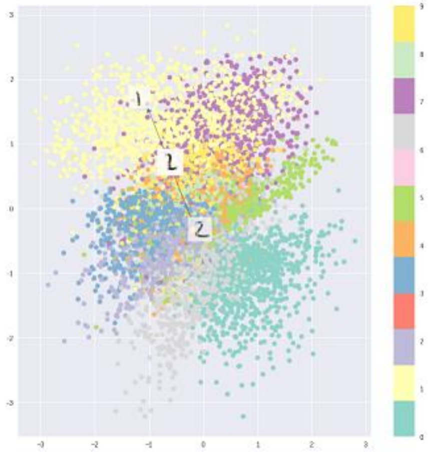
- **Just autoencoder**
  Sometimes we need just nonlinear PCA
  The latent space may not be continuous
  or allow easy interpolation.

- **Variational autoencoder**
  If you want precise control over your latent representations and
  what you would like them to represent, then choose VAE.
  Sometimes, precise modeling can capture better representations

- **Adversarial autoencoder**



AE



VAE

https://towardsdatascience.com/intuitively-understanding-variational-autoencoders-1bfe67eb5daf

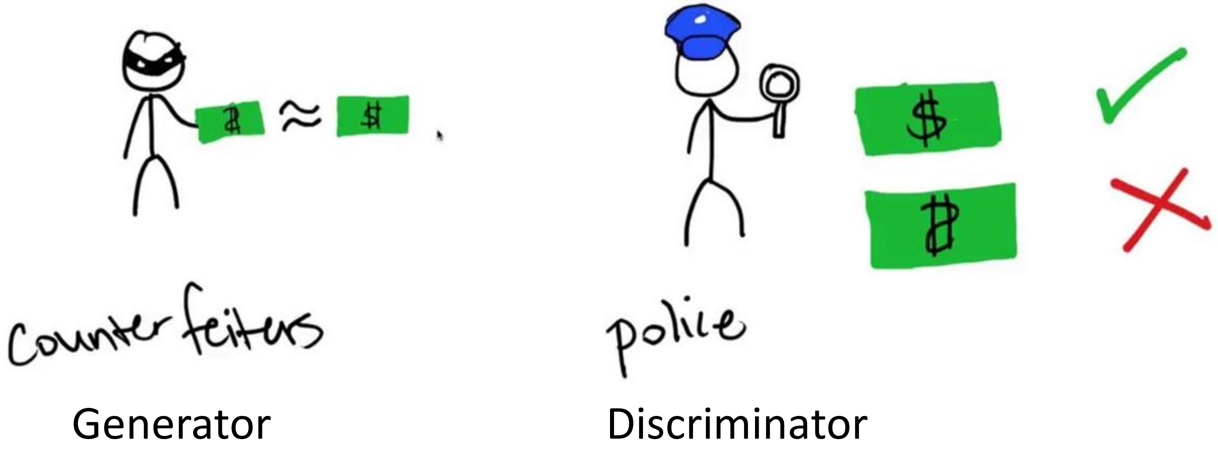
GANs for Anomaly Detection

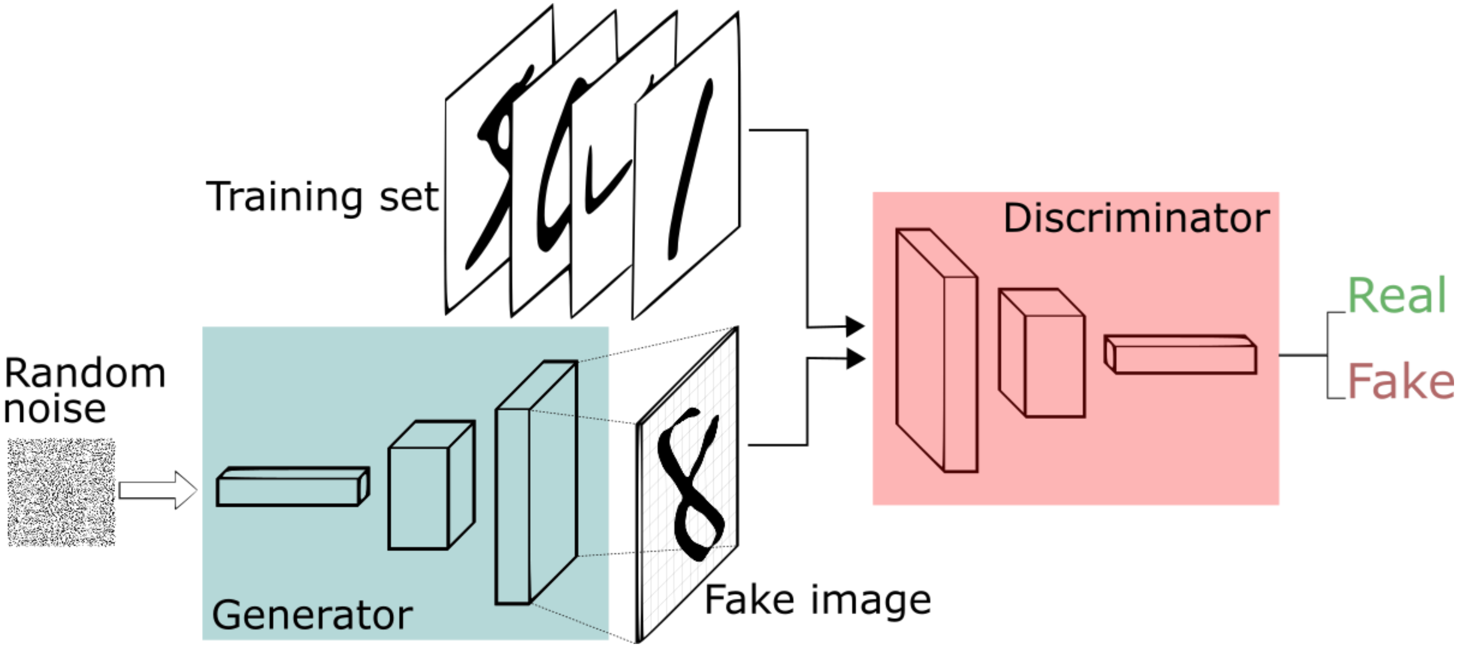
Generative Adversarial Network

Generator network: try to fool the discriminator by generating real-looking images

Discriminator network: try to distinguish between real and fake images

Train jointly in **minimax game!**





Generative Adversarial Network: formulas

Minimax objective function:

Discriminator outputs likelihood in (0,1) of real image

$$\min_{\theta_g} \max_{\theta_d} \left[\mathbb{E}_{x \sim p_{data}} \log D_{\theta_d}(x) + \mathbb{E}_{z \sim p(z)} \log(1 - D_{\theta_d}(G_{\theta_g}(z))) \right]$$

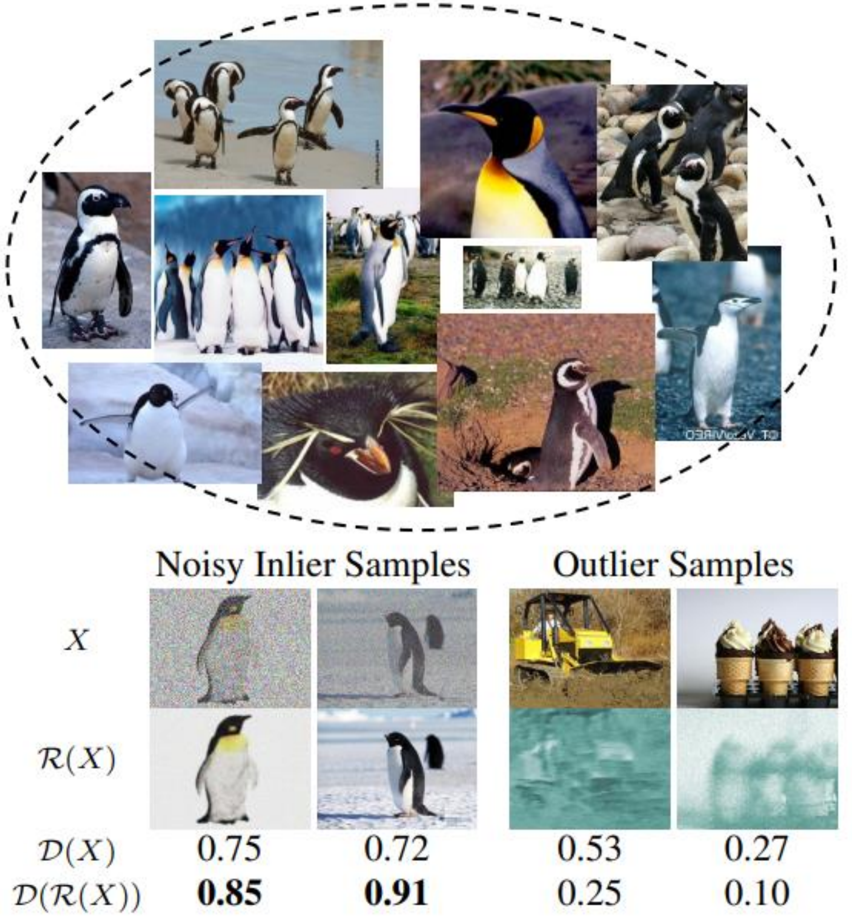
Discriminator output for real data x

Discriminator output for generated fake data G(z)

- *Discriminator* with parameters θ_d wants to **maximize objective** such that D(x) is close to 1 (real) and D(G(z)) is close to 0 (fake)
- *Generator* with parameters θ_g wants to **minimize objective** such that D(G(z)) is close to 1: the discriminator is fooled into thinking generated G(z) is real

GAN: example for novelty detection

The model is trained using images of penguins



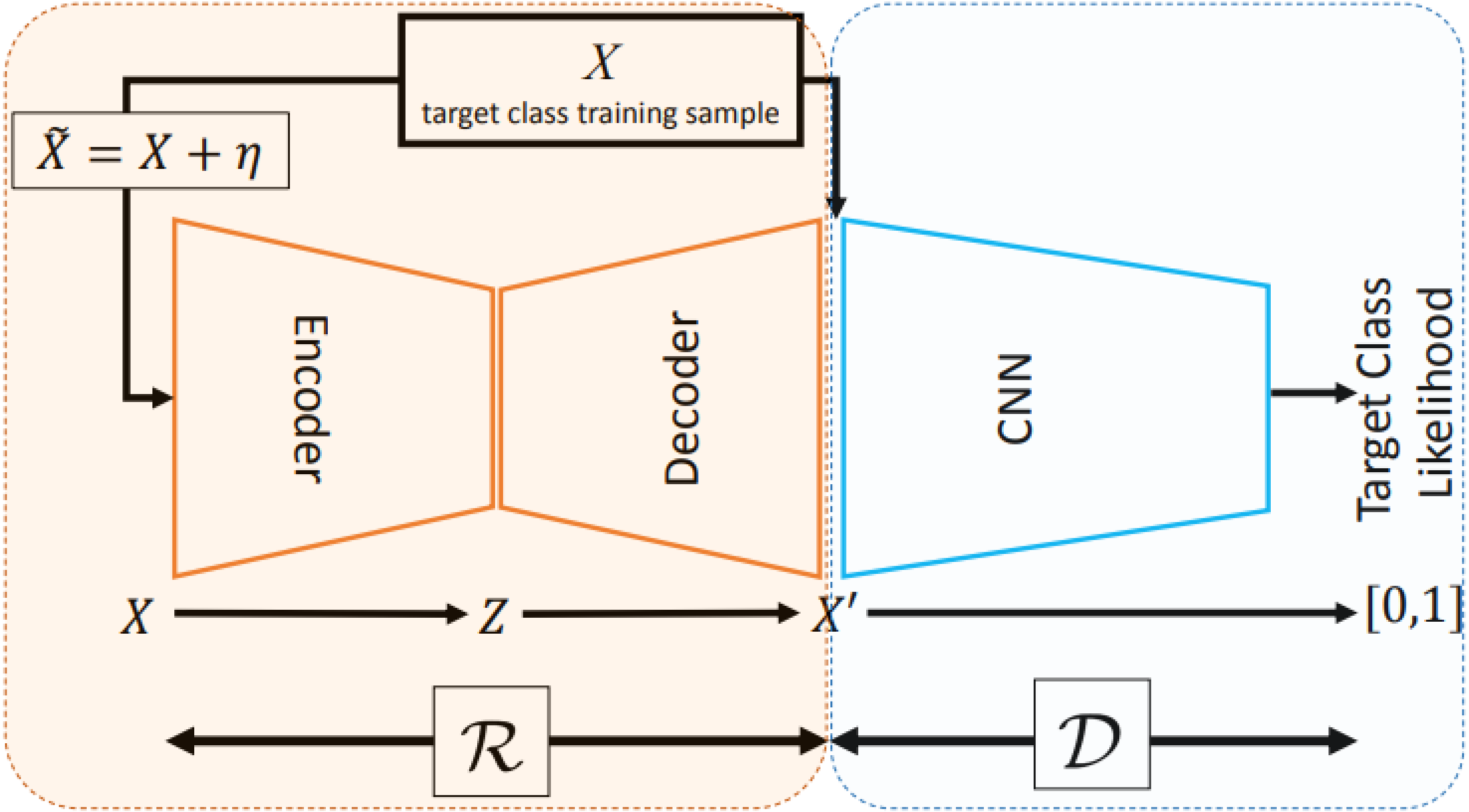
- If we use noisy inliers and pass them to "autoencoder" ***R*** NN, we get enhanced images as the output

- If we use outlier sample instead, the output of ***R*** is distorted

M. Sabokrou et al. *Adversarially Learned One-Class Classifier for Novelty Detection*, CVPR, 2018

GAN: example for novelty detection

Architecture



Autoencoder generator

Discriminator gives anomaly score

Adversarial training

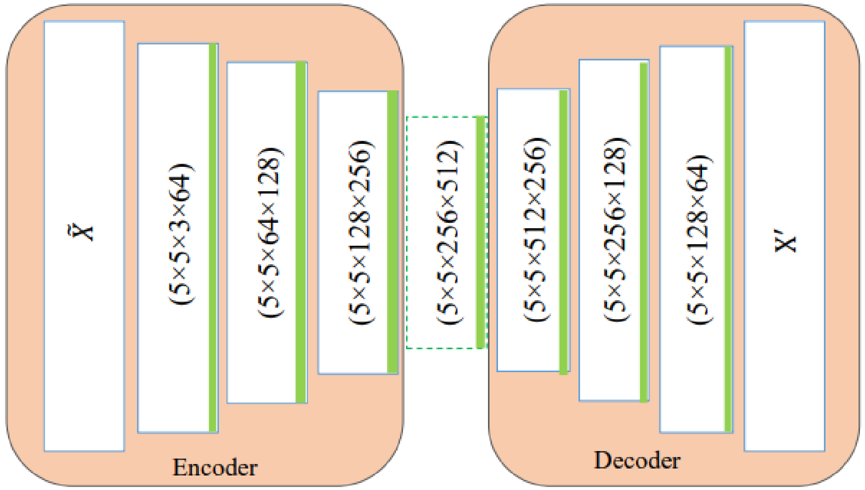
$$\mathcal{L} = \mathcal{L}_{\mathcal{R}+\mathcal{D}} + \lambda \mathcal{L}_{\mathcal{R}}$$

GAN loss

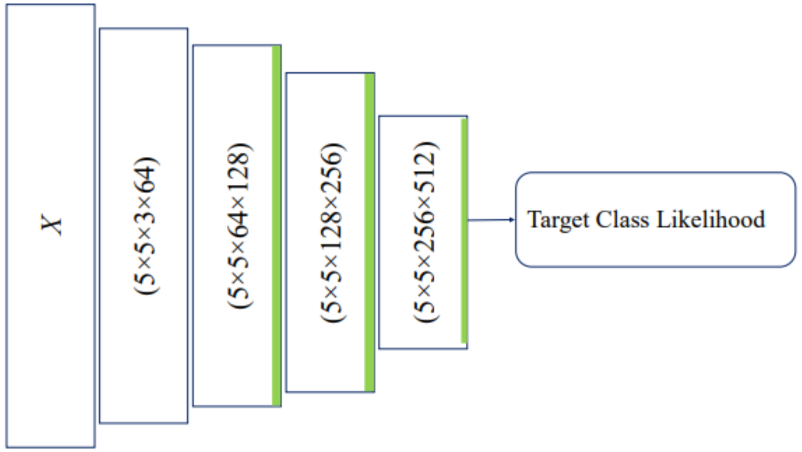
$$\mathcal{L}_{\mathcal{R}} = \|X - X'\|^2$$

https://arxiv.org/abs/1802.09088

Internal architectures



Autoencoder generator



Discriminator gives anomaly score

https://arxiv.org/abs/1802.09088

Skoltech
Skolkovo Institute of Science and Technology

GAN: example for novelty detection

Anomaly score with state-of-the-art performance:

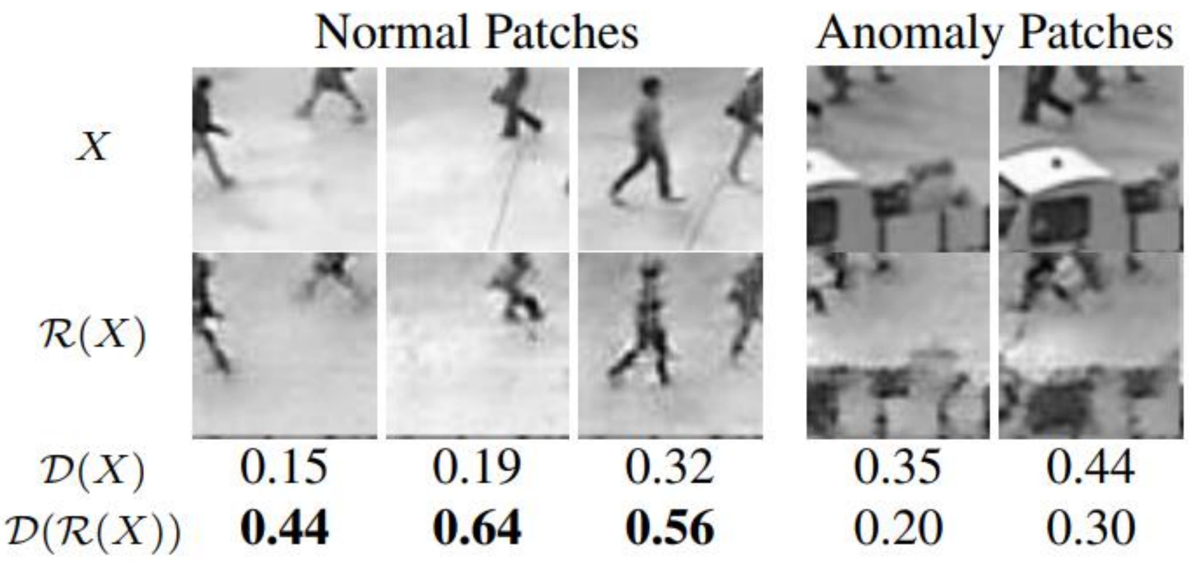
$$\mathrm{OCC}_1(X) = \begin{cases} \text{Target Class} & \text{if } \mathcal{D}(X) > \tau, \\ \text{Novelty (Outlier)} & \text{otherwise,} \end{cases}$$

PCA, VAE, AAE

Anomaly score that utilizes encoder-decoder

$$\mathrm{OCC}_2(X) = \begin{cases} \text{Target Class} & \text{if } \mathcal{D}(\mathcal{R}(X)) > \tau, \\ \text{Novelty (Outlier)} & \text{otherwise.} \end{cases}$$

Model quality

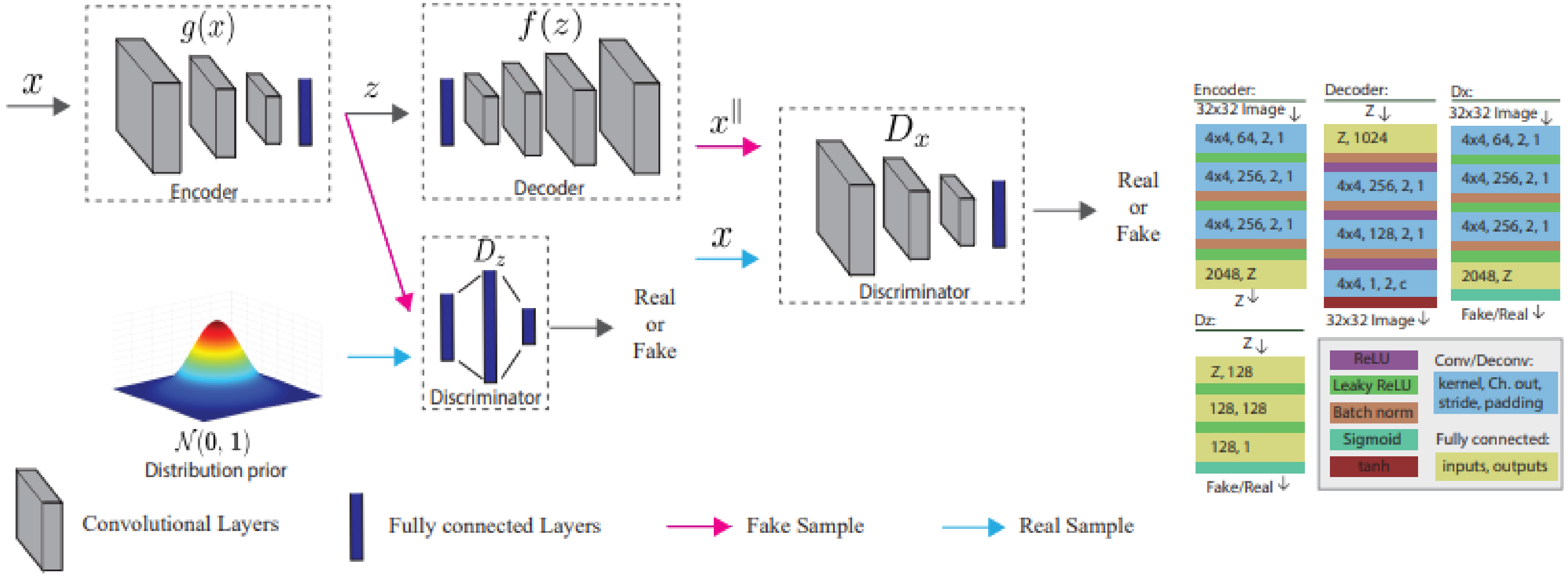


	CoP [32]	REAPER [22]	OutlierPursuit [50]	LRR [24]	DPCP [45]	R-graph [52]	Ours $\mathcal{D}(X)$	Ours $\mathcal{D}(\mathcal{R}(X))$
AUC	0.905	0.816	0.837	0.907	0.783	**0.948**	0.932	*0.942*
F_1	0.880	0.808	0.823	0.893	0.785	0.914	*0.916*	**0.928**
AUC	0.676	0.796	0.788	0.479	0.798	0.929	*0.930*	**0.938**
F_1	0.718	0.784	0.779	0.671	0.777	0.880	*0.902*	**0.913**
AUC	0.487	0.657	0.629	0.337	0.676	*0.913*	*0.913*	**0.923**
F_1	0.672	0.716	0.711	0.667	0.715	0.858	*0.890*	**0.905**

Skoltech
Skolkovo Institute of Science and Technology

Adversarial autoencoders help

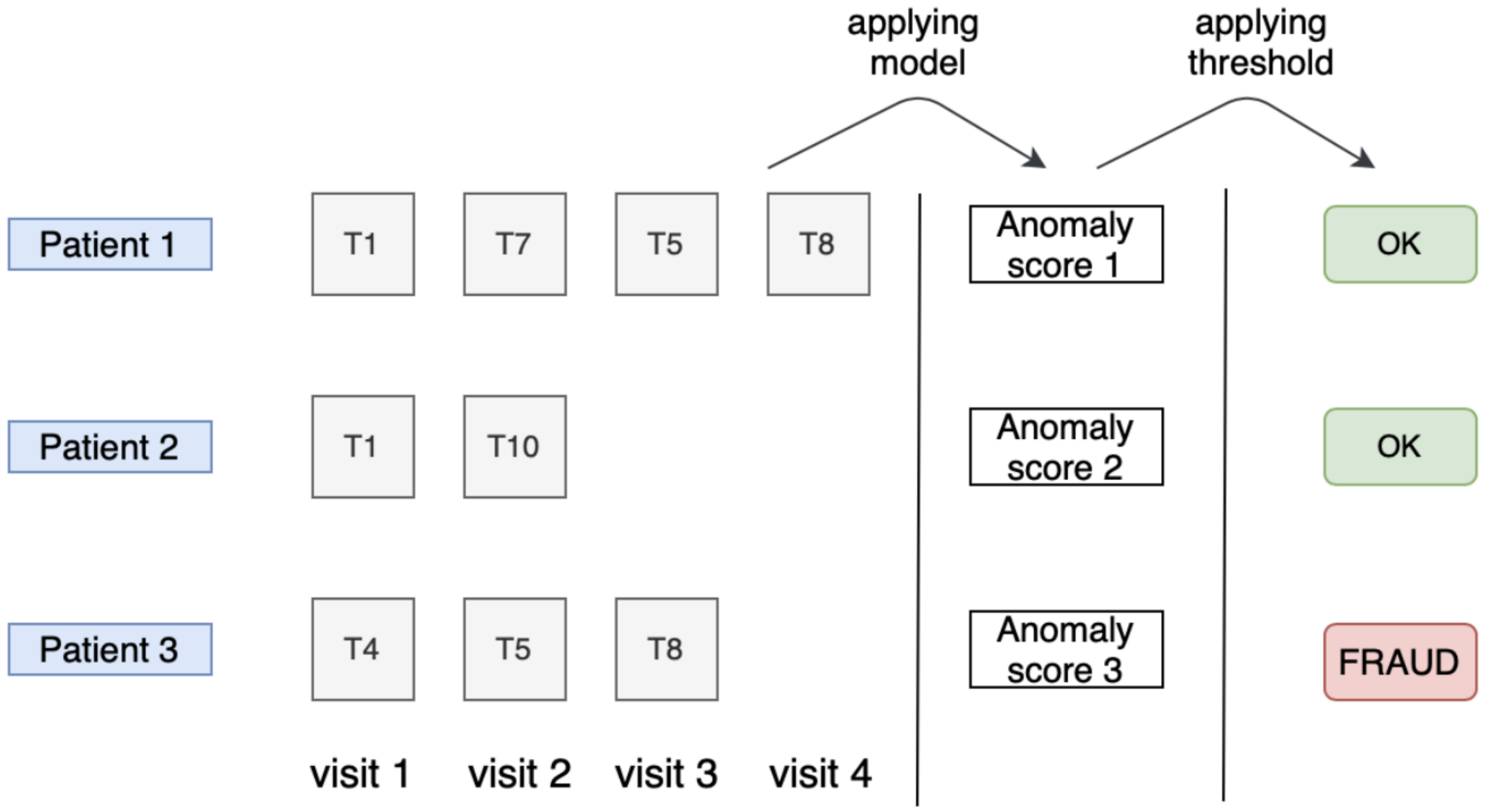
- Construct anomaly score s(x) using data
- Signal about anomaly if anomaly score is greater than some threshold t



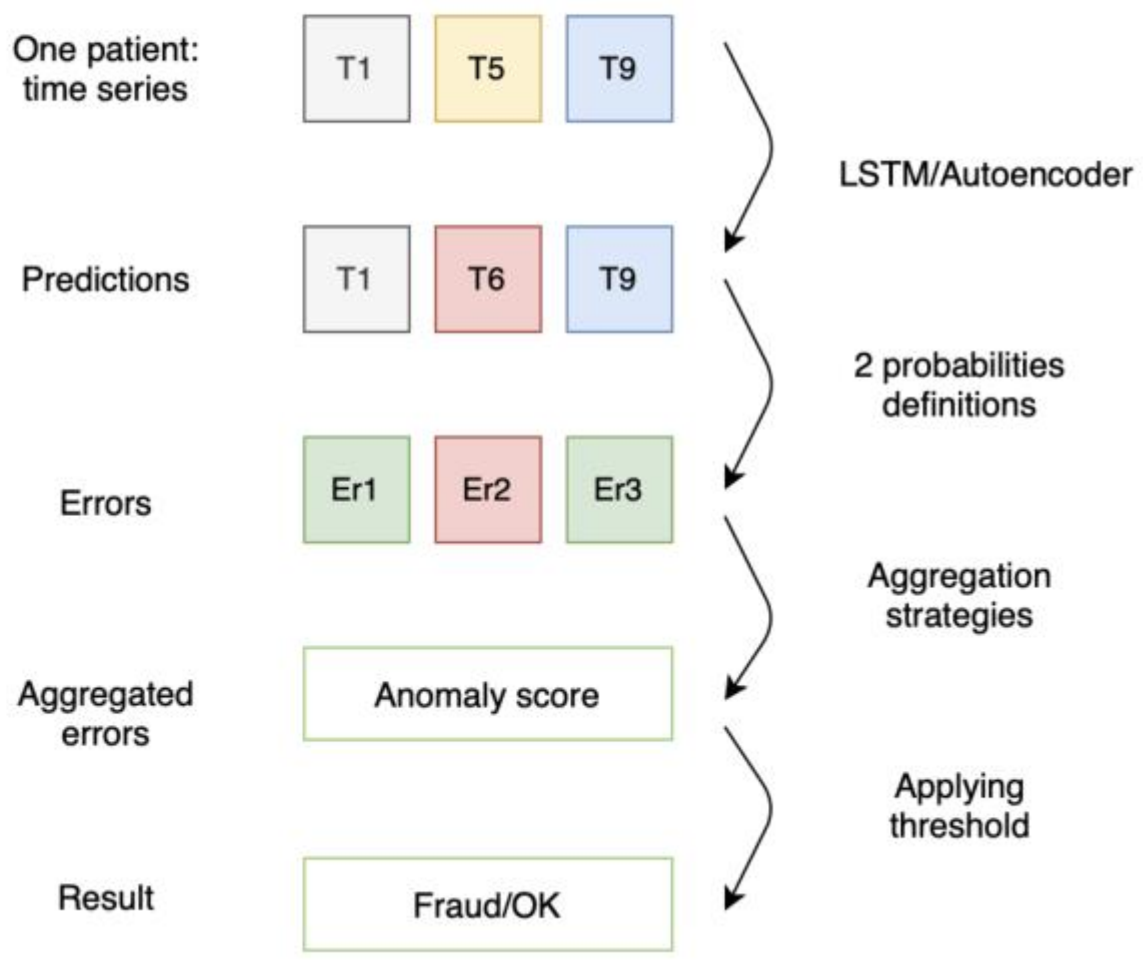
https://papers.nips.cc/paper/7915-generative-probabilistic-novelty-detection-with-adversarial-autoencoders.pdf

Anomaly Detection for Time Series

Anomaly detection for sequential data: healthcare insurance



Anomaly detection for sequential data: healthcare insurance



Anomaly and novelty for Time Series

Problems:

1. SotA techniques of anomaly detection are rarely used for classic time series

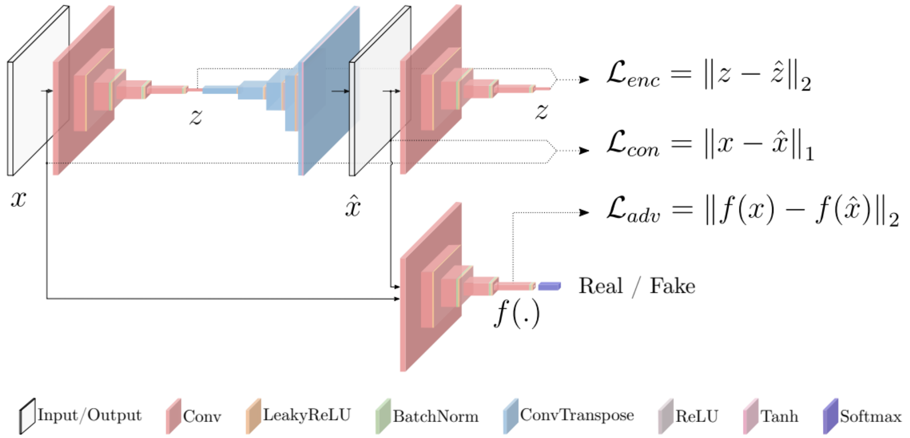
2. Available solutions don't take into account the statistical nature of Time Series

Proposed solution:

1. To develop a loss function for GAN-based anomaly detection for time series

2. To take into account requirements of statistical change point detection models: low number of false alarms and small detection delay

3. To develop new resampling techniques by learning data distribution



Li D. et al. MAD-GAN: Multivariate anomaly detection for time series data with generative adversarial networks // arXiv:1901.04997. – 2019.

Skoltech
Skolkovo Institute of Science and Technology

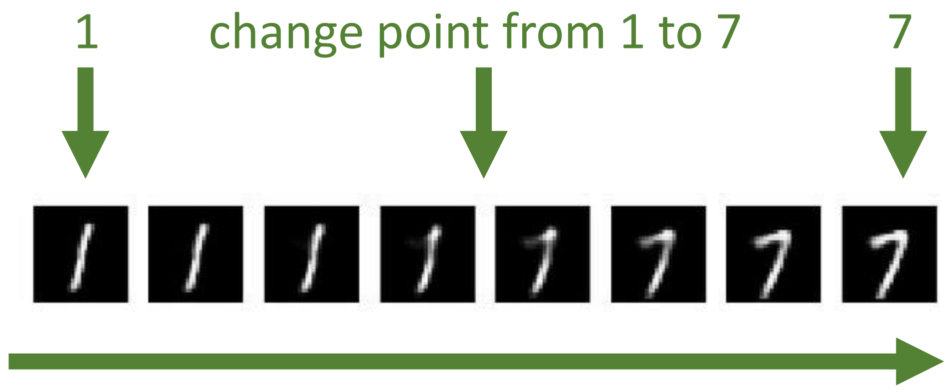
Change detection in semi-structured data

The change-point detection (CPD) model signals about time of change in the data distribution

Semi-structured data – sequences of semi-structured data (images, texts)

Goal: minimize Detection Delay & minimize number of False Alarms

Problems: Can’t apply classic method for semi-structured data



A sequence of MNIST images is an example of semi-structured data

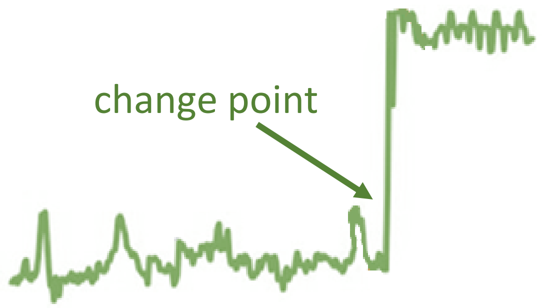
Change detection in semi-structured data

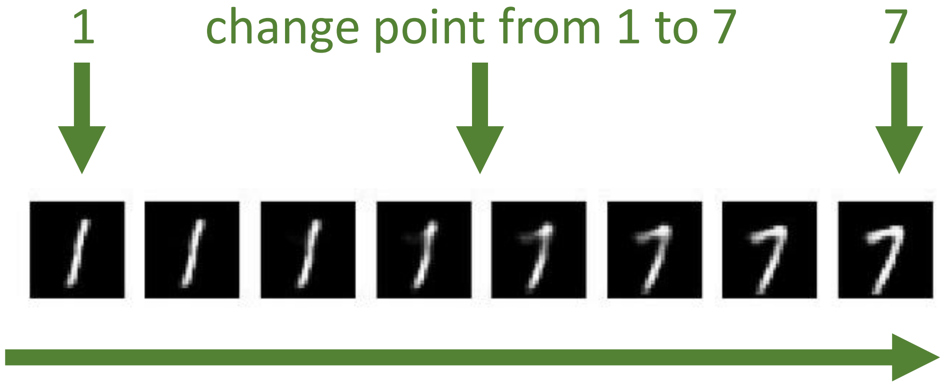
Proposed solution:

1. Develop a data embedding procedure

2. End2end methods based on statistical tests or detection outliers in embedded space – unsupervised anomaly detection

3. Develop new loss function for direct minimization of the problem specific metrics

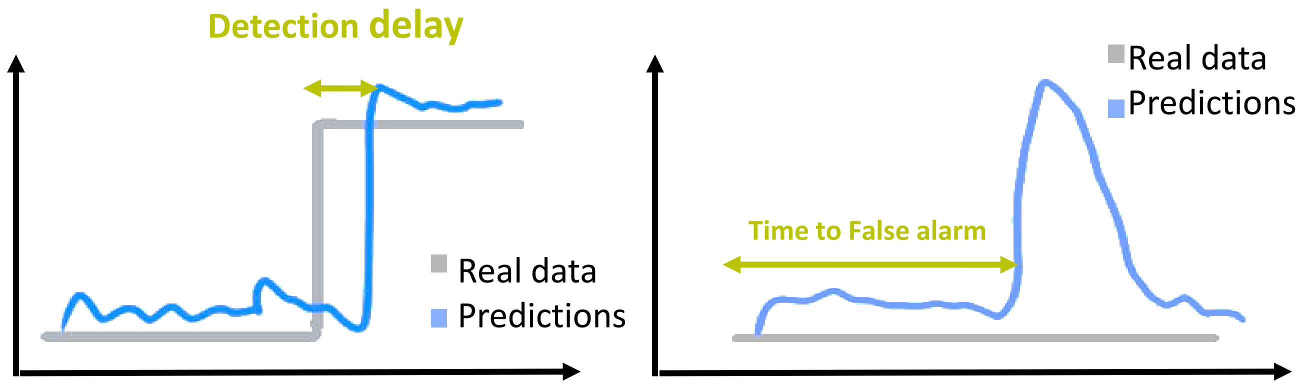




A sequence of MNIST images is an example of semi-structured data

Our end2end approach

We concentrate on typical quality metrics for change-point detection: delay detection and mean time to False alarm.



We optimize lower bounds for these metrics:

- p_k is the model's change point probability at moment k,
- T – hyperparameter that restricts the length of the considered sequence.

$$Loss_{detection\_delay} = \sum_{t=\theta}^{T}(t - \theta)\, p_t \prod_{k=\theta}^{t-1}(1 - p_k) + (T + 1) \prod_{k=\theta}^{T}(1 - p_k),$$

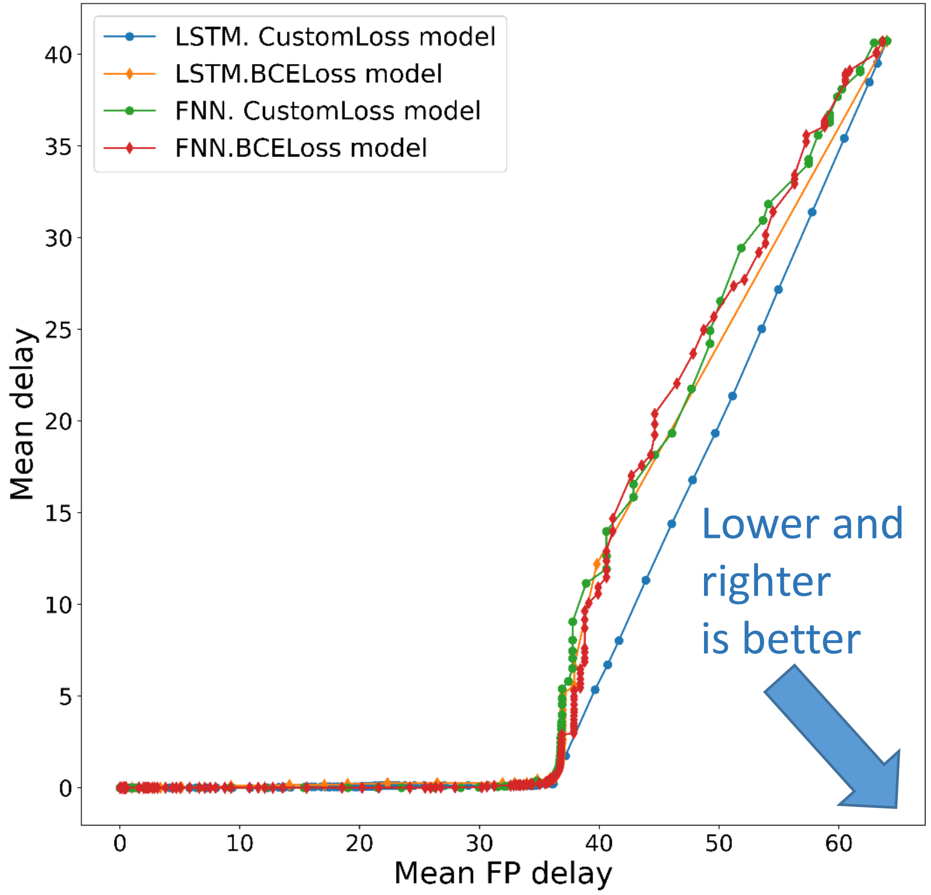
$$Loss_{FP\_delay} = 1 - \sum_{t=0}^{\theta}(t - \theta)p_t \prod_{k=0}^{\theta}(1 - p_k)$$

Results

Dataset: sequences with images from MNIST. There are sequences with (e.g. from 1 to 7) and without (e.g. from 1 to 1) change point.

We compare LSTM and fully connected neural network (FNN) architectures, as well as binary cross entropy loss (BCELoss) and our proposed loss.

LSTM with proposed loss function has a better Pareto frontier with respect to the mean detection delay and the mean false positive delay.



Skoltech
Skolkovo Institute of Science and Technology

Future work

- Try proposed approach for other datasets of semi-structured data
- Consider different neural network architectures for processing of semi-structured sequential data
- Combine representation learning and statistical change point detection procedures

53

References

Change point detection (CPD). Basic knowledge and main statistical approaches:

- Shiryaev A. N. Stochastic disorder problems. – Springer International Publishing, 2019.
- Romanenkova E. et al. Real-Time Data-Driven Detection of the Rock-Type Alteration During a Directional Drilling //IEEE Geoscience and Remote Sensing Letters. – 2019.

Supervised CPD:

- Malhotra P. et al. Long short term memory networks for anomaly detection in time series //ESANN proceedings, 2015.
- Hundman K. et al. Detecting spacecraft anomalies using lstms and nonparametric dynamic thresholding //Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. – 2018

Factors to consider when choosing an Anomaly Detector

- Few parameters
  - parameter-free the best
  - easy to tune; not too sensitive to parameter setting

- Fast runtime: can scale up to large datasets and high dimensional datasets

- Low space complexity

- Known behaviours under different data properties

- Can deal with different types of anomalies

- Its ability to deal with high dimensional problems

- <u>Understand the nature of anomalies and the best match algorithm</u>

https://federation.edu.au/__data/assets/pdf_file/0011/443666/ICDM2018-Tutorial-Final.pdf

Take-home messages

- Anomaly detection is a challenging problem
- Often problem-specific knowledge helps

- Common approaches are Autoencoder-based and Isolation forest
- There are some time-series specific approaches: the problem is similar to the change point detection problem

More references?

See

- Overview of anomaly detection for tabular data
  https://www.youtube.com/watch?v=12Xq9OLdQwQ

- A collection of *awesome* anomaly detection papers
  https://awesomeopensource.com/project/hoya012/awesome-anomaly-detection

- A collection of *awesome* anomaly detection resources
  https://github.com/yzhao062/anomaly-detection-resources

- Link-based list of anomaly detection methods
  https://github.com/zhuyiche/awesome-anomaly-detection