Anomaly detection with Neural Networks

Zaytsev Alexey, Lab head, Skoltech

12 May 2020

Anomaly detection problem statement

The problem is to find anomalous objects
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				

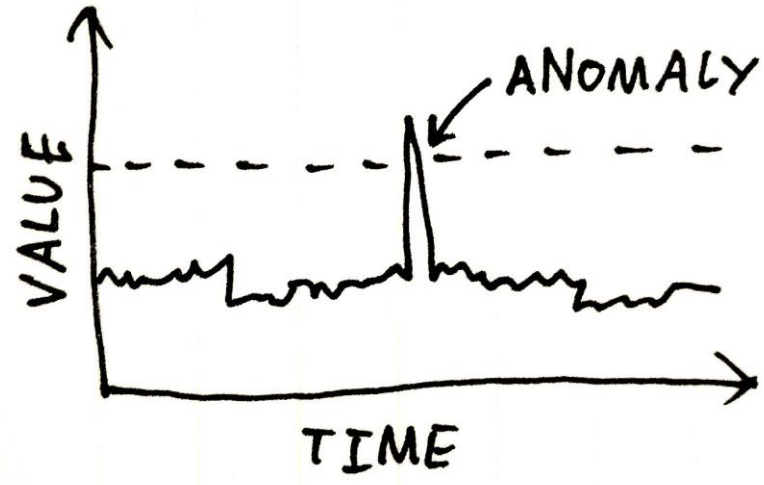
Anomaly

Racoon
1


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

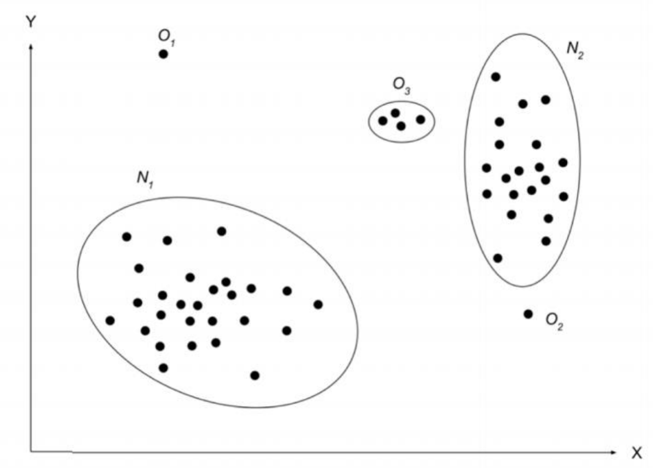
Problem examples

- Fraud detection
- Failure detection for an airplane
- Intrusion detection in cybersecurity
- 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



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

Taxonomy with respect to available data

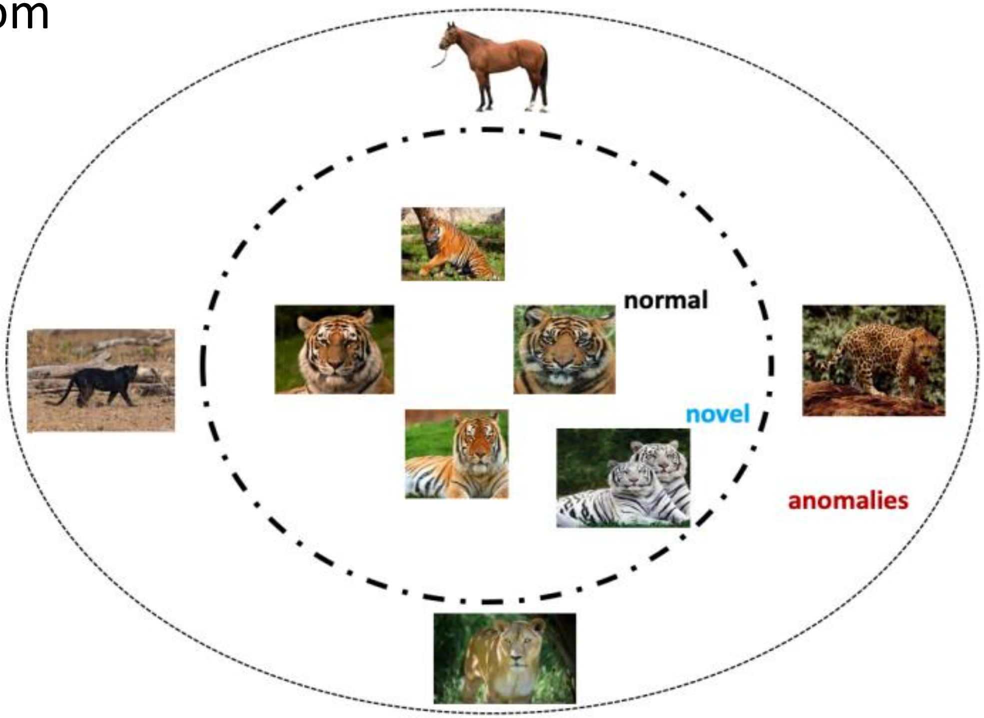
- Sequential data
- Non-sequential data

- Supervised
- Unsupervised
- Semi-supervised

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

Taxonomy with respect to problem statement

- Novelty detection – unseen objects from our distribution
- Anomaly detection – objects from another distribution



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

Taxonomy with respect to problem statement

- Supervised
- Unsupervised
- Semi-supervised

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

Anomaly type

- Point

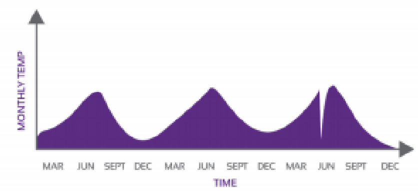
May-22	1:14 pm	FOOD	Monaco Café	$1,127.80
May-22	2:14 pm	WINE	Wine Bistro	$28.00
...				
Jun-14	2:14 pm	MISC	Mobil Mart	$75.00
Jun-14	2:05 pm	MISC	Mobil Mart	$75.00
Jun-15	2:06 pm	MISC	Mobil Mart	$75.00
Jun-15	11:49 pm	MISC	Mobil Mart	$75.00
May-28	6:14 pm	WINE	Acton shop	$31.00
May-29	8:39 pm	FOOD	Crossroads	$128.00
Jun-16	11:14 am	MISC	Mobil Mart	$75.00
Jun-16	11:49 am	MISC	Mobil Mart	$75.00

Point Anomaly

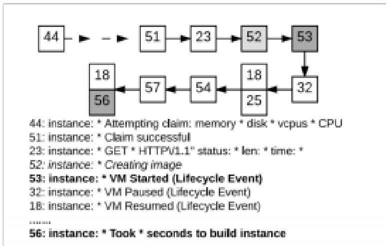
Collective Anomaly

- Collective or group

- Contextual or conditional



(a) Temperature data Hayes and Capretz [2015].



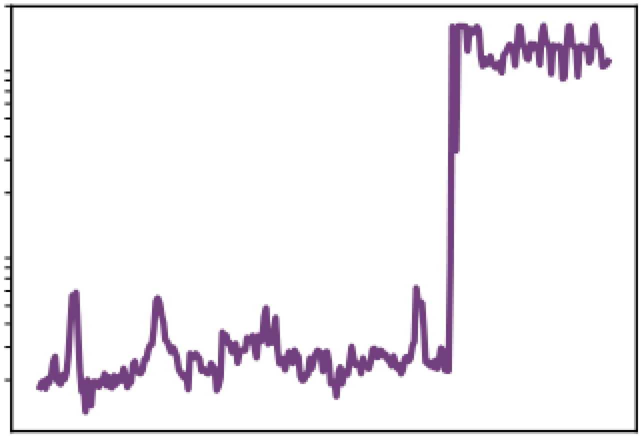
(b) System logs Du et al. [2017].

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

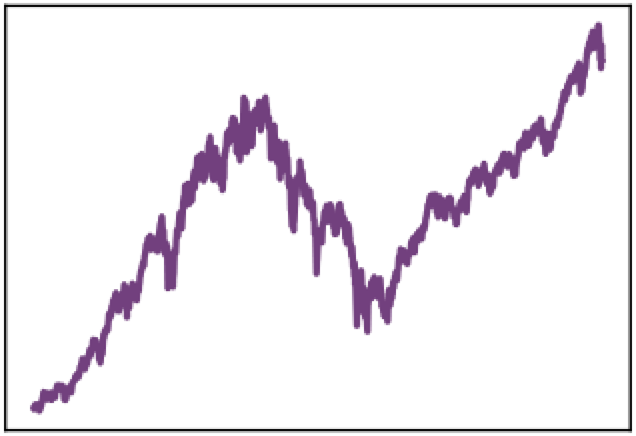
Skoltech
Skolkovo Institute of Science and Technology

Change point detection

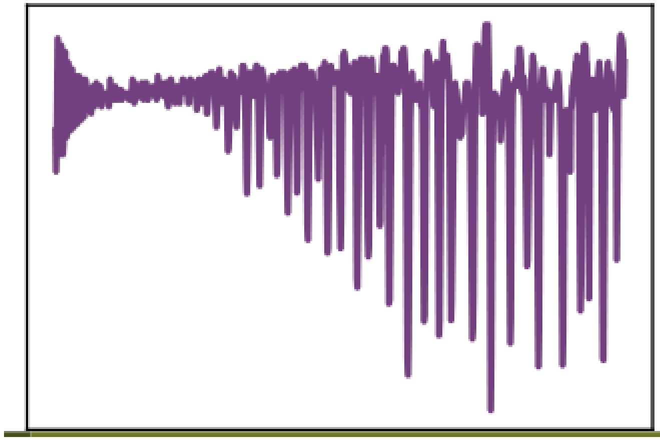
Discontinuity



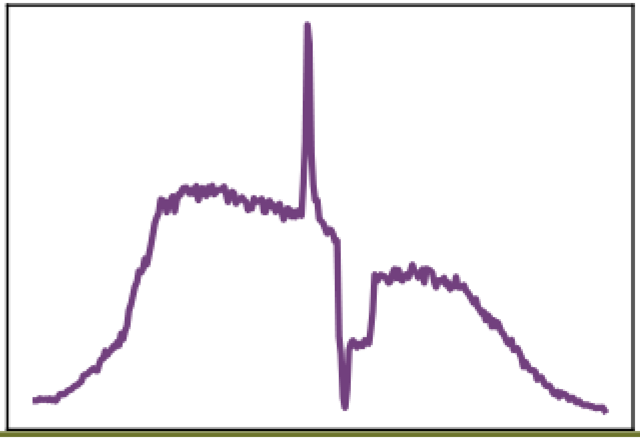
Kinks



Variance growth



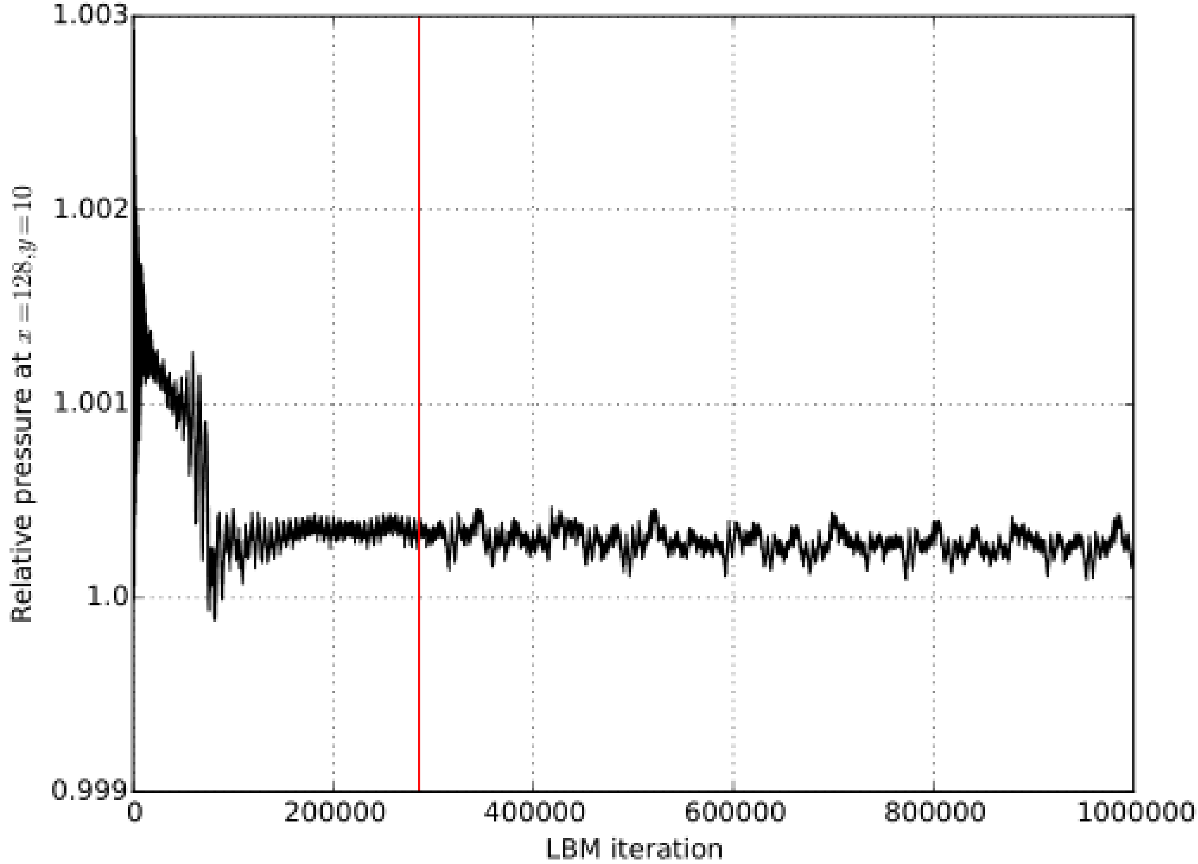
Violation of dynamic



Slide by A. Artemov, E. Burnaev

Real world change point: fluid pressure

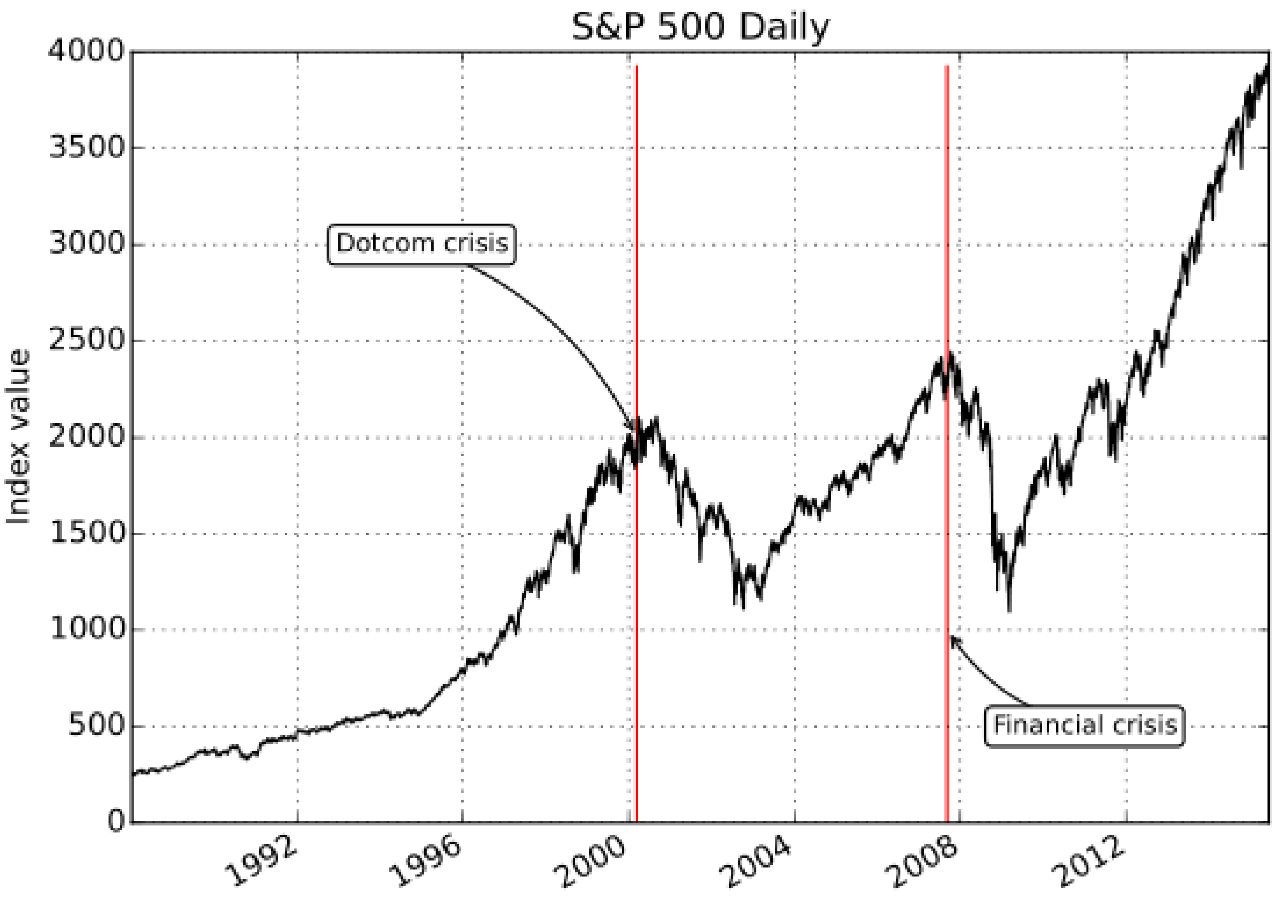
Fluid pressure in a hydrodynamic system for a numerical model based on Boltzmann method



Slide by A. Artemov, E. Burnaev

Real world change point: S&P500 index

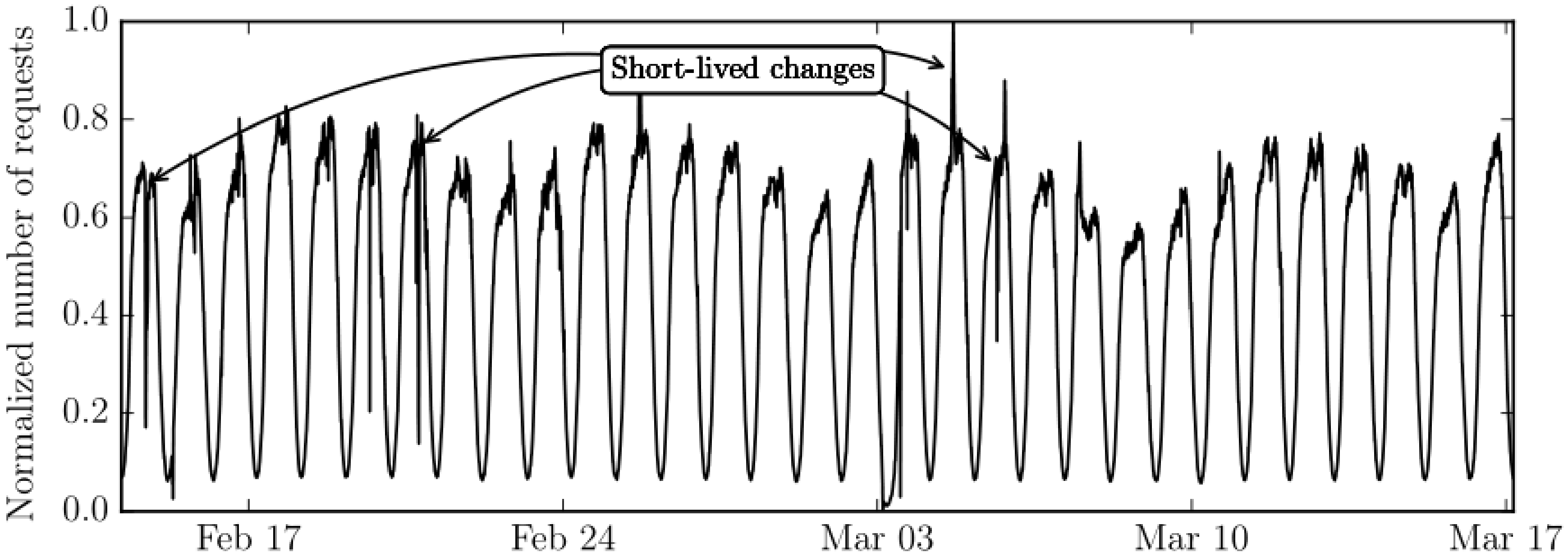
Dynamic of S&P500 stock exchange index for 7 years



Slide by A. Artemov, E. Burnaev

Real world change point: internet data

Web-site visits



Slide by A. Artemov, E. Burnaev

Real world change point: drilling data

Nuclear-magnetic response during the drilling of a well

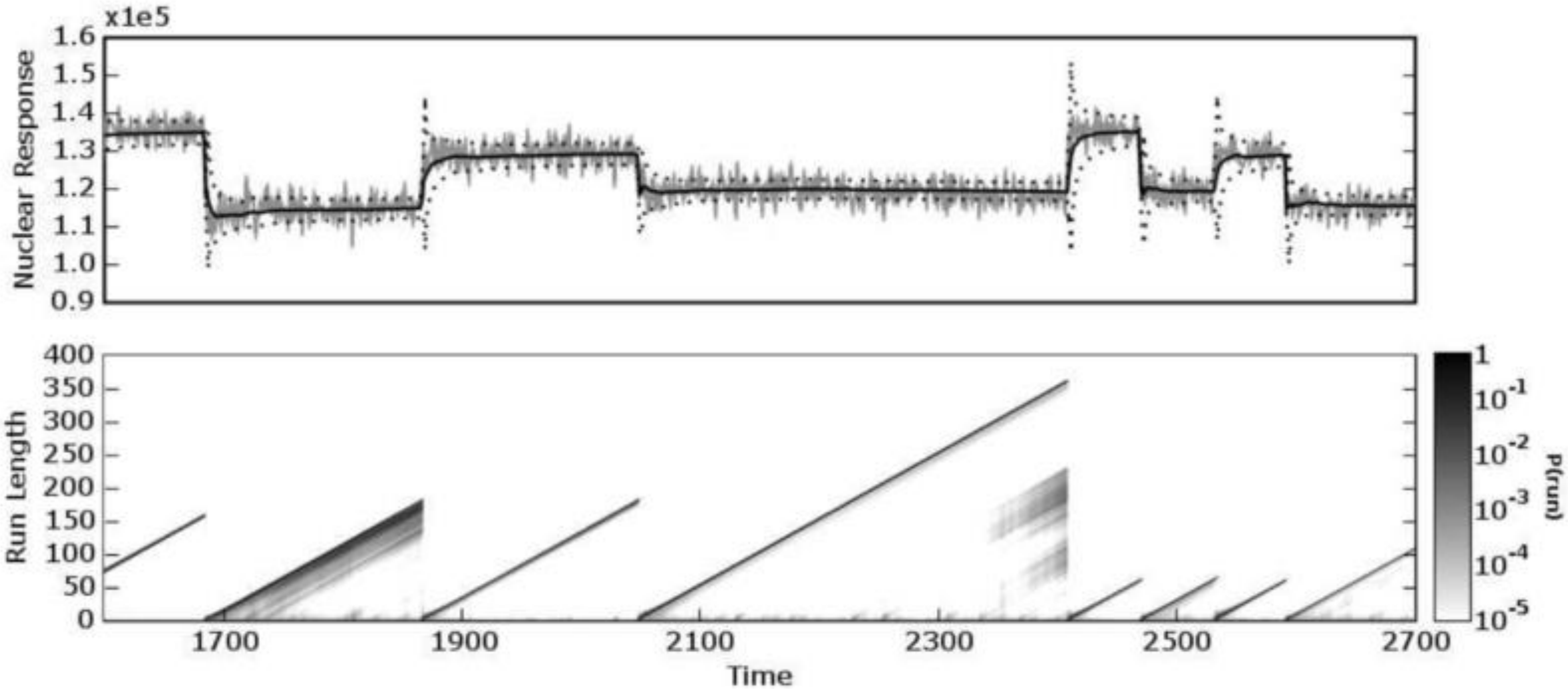


Figure 2: The top plot is a 1100-datum subset of nuclear magnetic response during the drilling of a well. The data are plotted in light gray, with the predictive mean (solid dark line) and predictive 1-σ error bars (dotted lines) overlaid. The bottom plot shows the posterior probability of the current run $P(r_t \,|\, \boldsymbol{x}_{1:t})$ at each time step, using a logarithmic color scale. Darker pixels indicate higher probability.

Slide by A. Artemov, E. Burnaev

Can you identify anomalies?

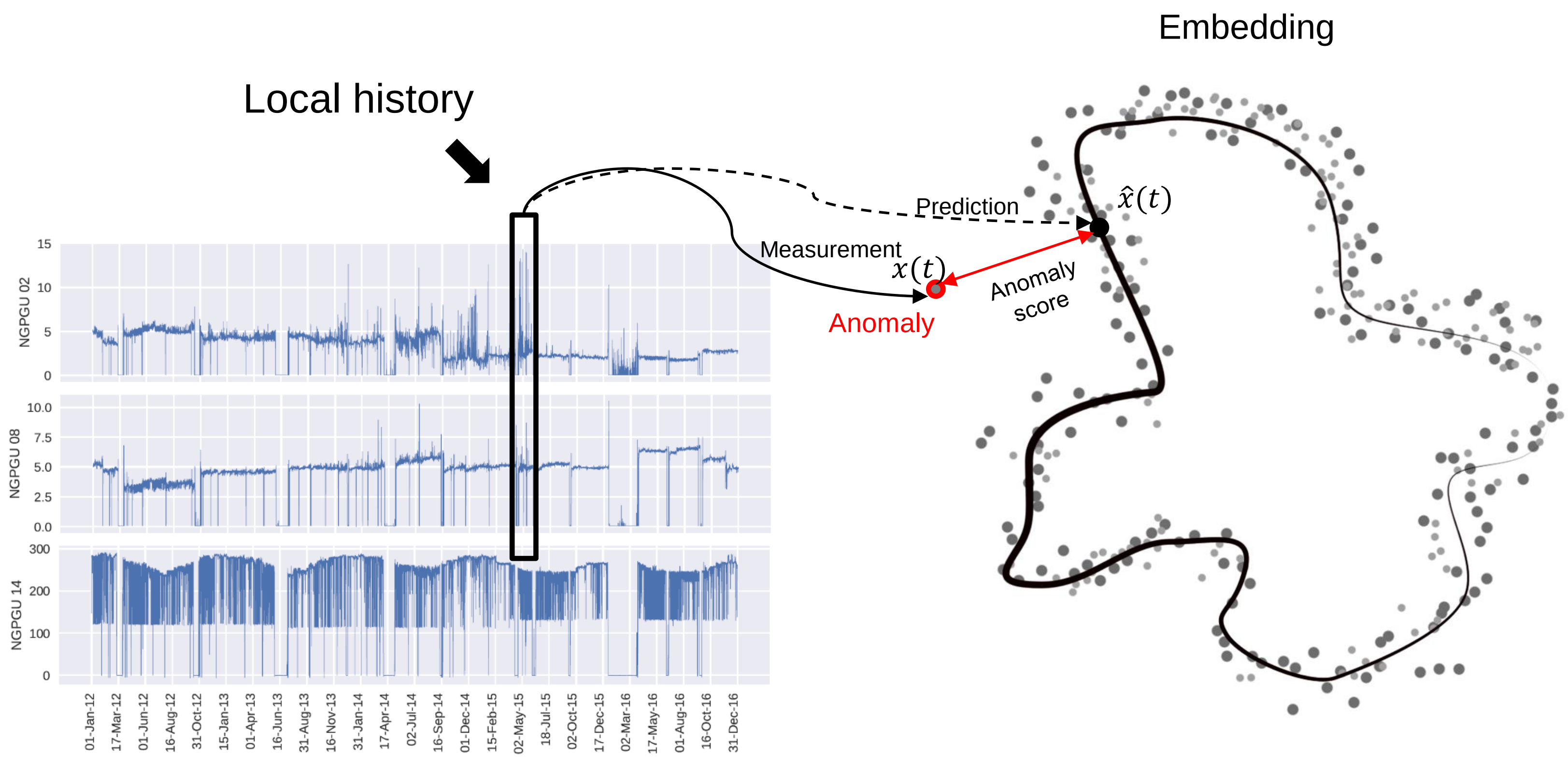
Let’s try!

https://speakerdeck.com/skibish
/a-note-about-anomalies-article

Focus today:
unsupervised anomaly detection

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				

General approach



General approach-1

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

$$
\begin{aligned}
\hat{x}_1 &= f_1(x_2, x_3, \dots, x_d), \\
&\dots \\
\hat{x}_d &= f_d(x_1, x_2, \dots, x_{d-1}).
\end{aligned}
$$

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

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

General approach-2 is more general

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

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

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

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

E.g. PCA, autoencoder

Autoencoder

Train such that features can be used to reconstruct original data

L2 Loss function:

$$\|x - \hat{x}\|^2$$

Don't need labels!

Reconstructed input data $\hat{x}$

Representation z

Input data x

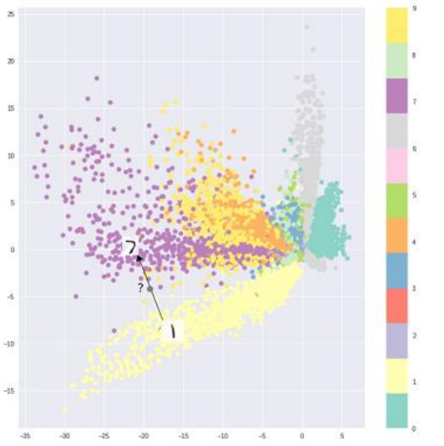
Reconstructed data

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

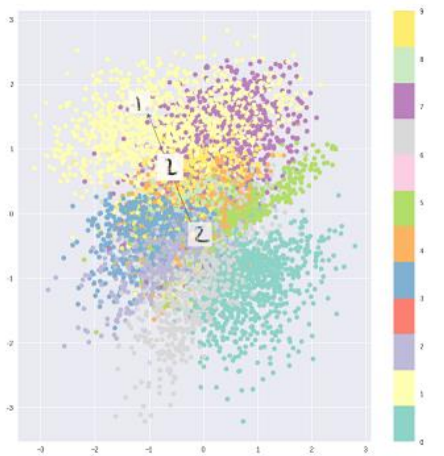
Input data

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

Taxonomy of autoencoders



AE



VAE

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

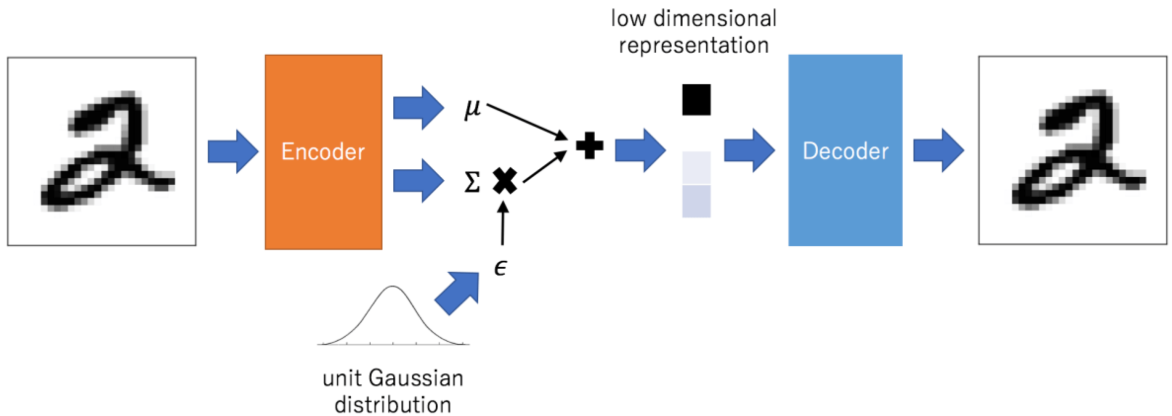
- **Variational autoencoder**
  If you want precise control over your latent
  representations and
  what you would like them to represent, then choose VAE.
  Sometimes, precise modeling can capture better
  representations

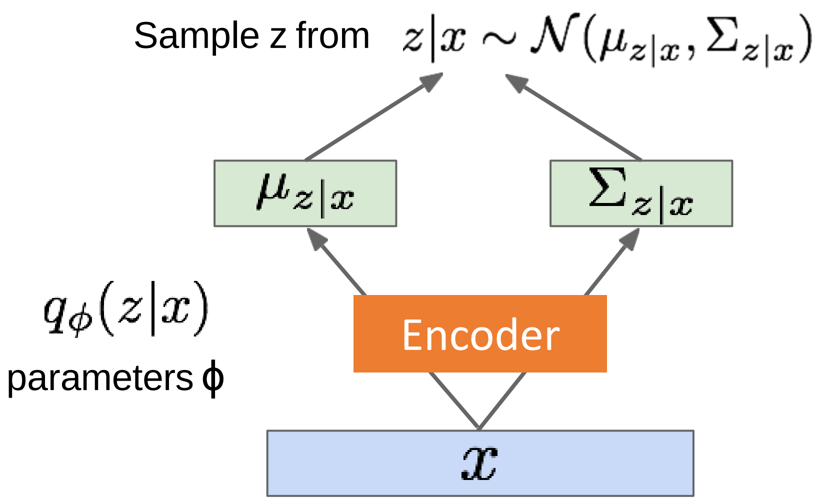
- **Adversarial autoencoder**

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

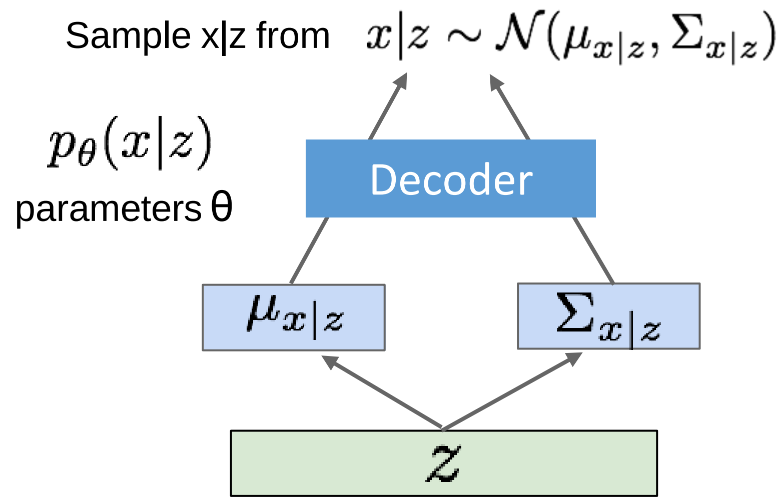
https://arxiv.org/abs/1506.02216

Variational autoencoder





Encoder network



Decoder network

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

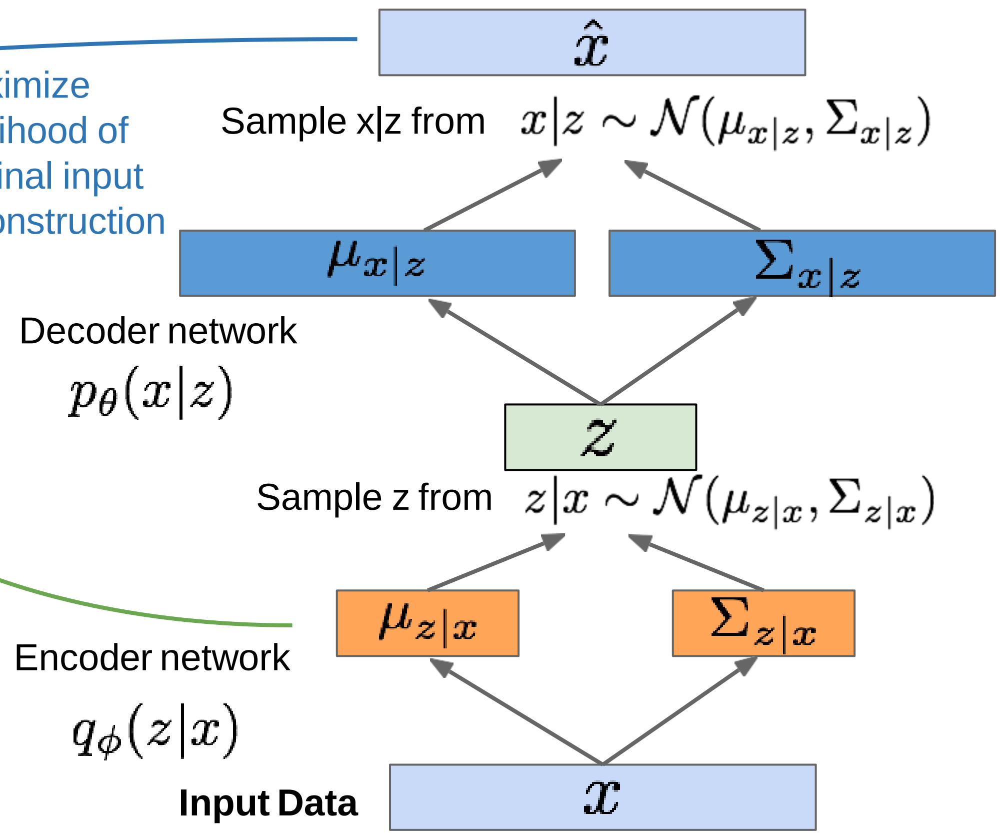
Variational autoencoder: formulas

We maximize the likelihood lower bound

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

Maximize likelihood of original input reconstruction

Approximate posterior should be close to prior

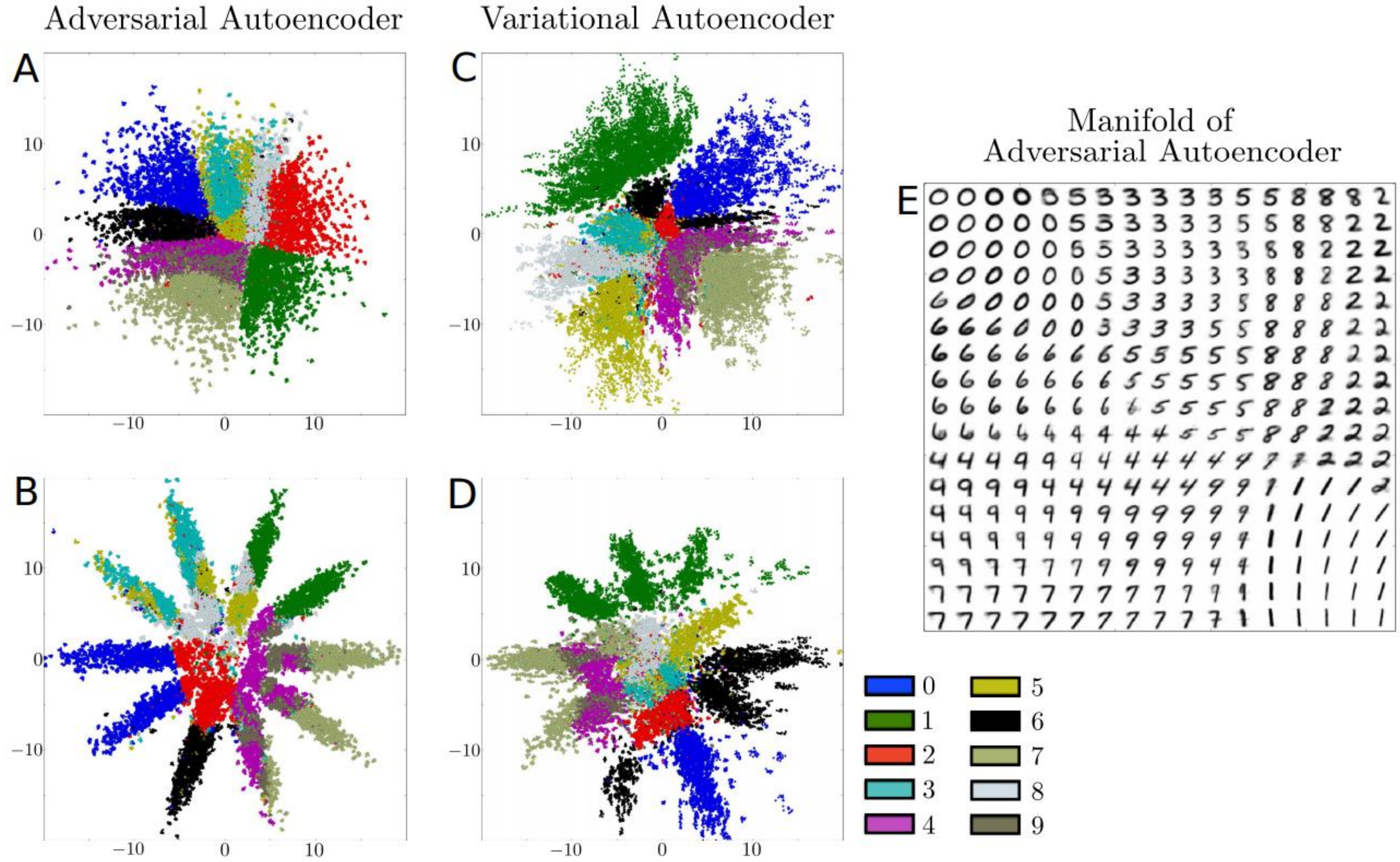


https://arxiv.org/pdf/1606.05908.pdf

We can do better with Adversarial autoencoder

2-D Gaussian

10-D Gaussian

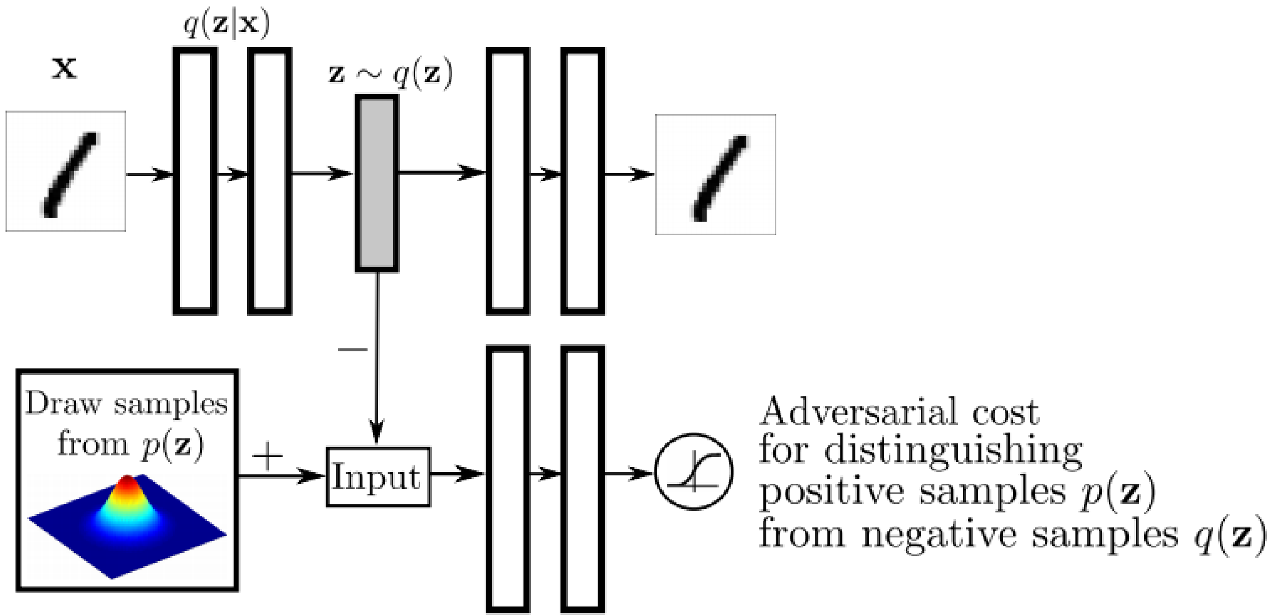


https://arxiv.org/pdf/1511.05644.pdf

Adversarial autoencoder

Autencoder / Generator

Discriminator



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

Adversarial autoencoder: formulas

$$\mathcal{L} = \mathbb{E}_x \left[\underbrace{\mathbb{E}_{q(z|x)}[-\log p(x|z)]}_{\text{Reconstruction Error}} \right] + \mathbb{E}_x \left[\underbrace{\mathrm{KL}(q(z|x) || p(z))}_{\text{KL Regularizer}} \right]$$

Replaced by adversarial loss in AAE

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

GANs

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**

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

Minimax objective function:

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

Discriminator output for real data x

Discriminator output for generated fake data G(z)

- Discriminator (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 (θ_g) wants to **minimize objective** such that D(G(z)) is close to 1 (discriminator is fooled into thinking generated G(z) is real)

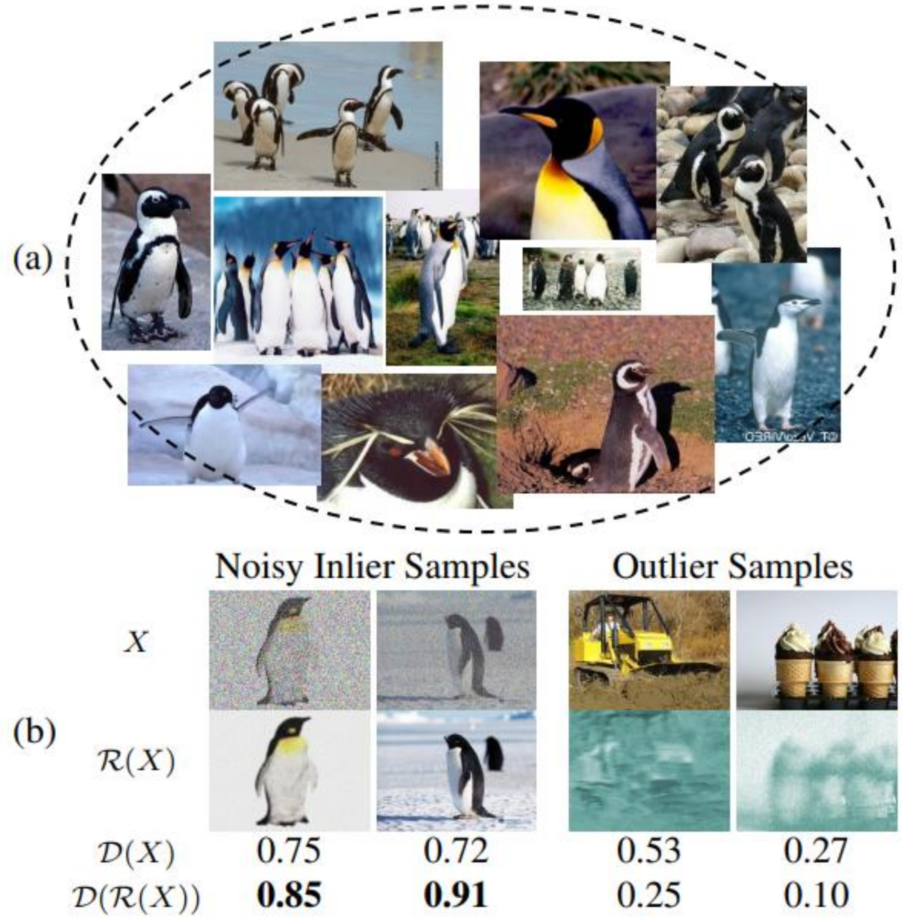
More general approach to anomaly detection

- Construct anomaly score a(x) using data
- Signal about anomaly if anomaly score is greater than some threshold t
- Threshold selection is a separate problem, as we have only positive examples

Adversarial autoencoders help

https://arxiv.org/abs/1802.09088

- The model is trained using images of penguins



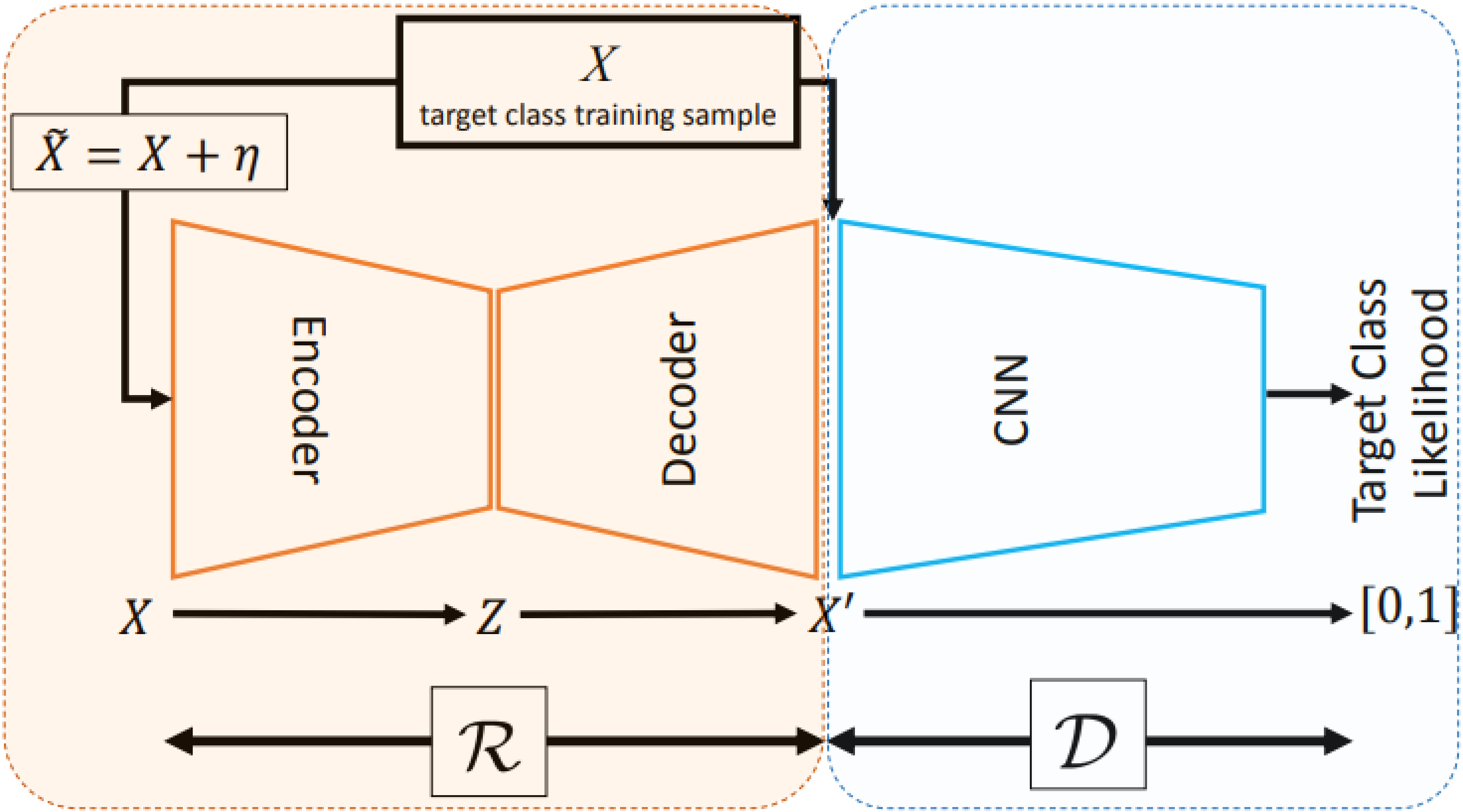
D is discriminator score

- If we use noisy inputs and pass them to our network **R**, we get enhanced images as the output
- If we use outlier input instead, the output of **R** is distorted

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

Skoltech
Skolkovo Institute of Science and Technology

Overall architecture has a generator and a discriminator

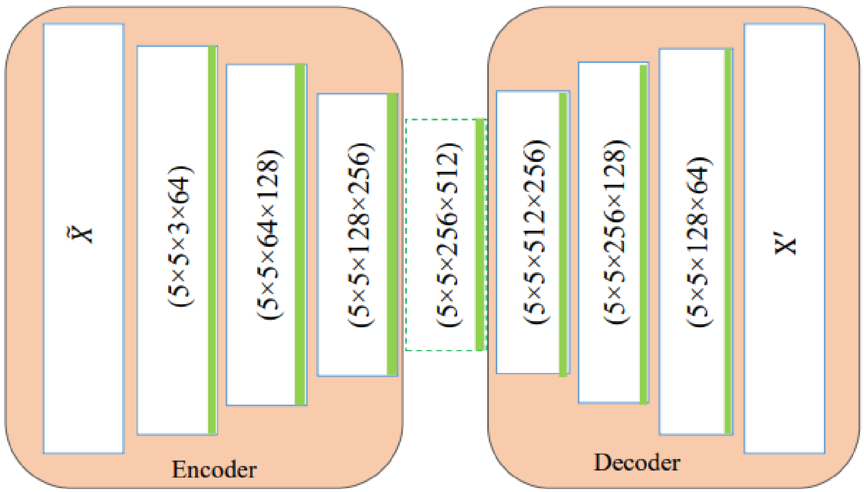


Autoencoder generator

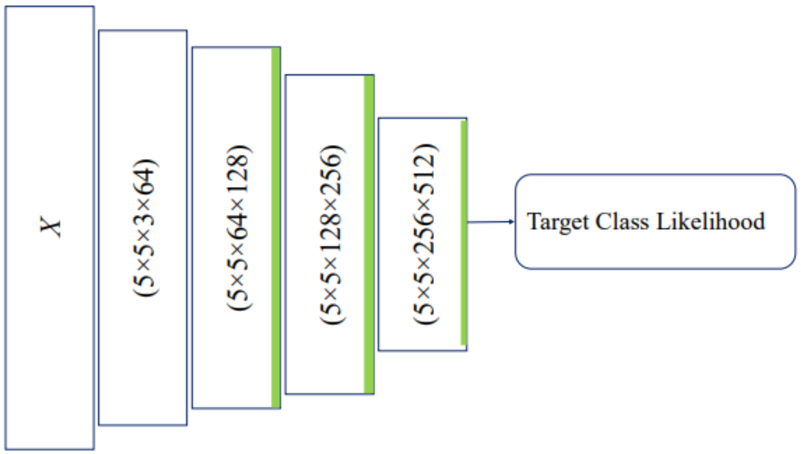
Discriminator gives anomaly score

https://arxiv.org/abs/1802.09088

Internal architectures



Autoencoder generator



Discriminator gives anomaly score

https://arxiv.org/abs/1802.09088

Loss function

Full loss function:

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

Loss function for difference between the initial object and reconstruction (can be log loss instead)

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

Loss function similar to VAE & GAN

$$\min_{\mathcal{R}} \max_{\mathcal{D}} \Big(\mathbb{E}_{X \sim p_t}[\log(\mathcal{D}(X))] \\ + \mathbb{E}_{\tilde{X} \sim p_t + \mathcal{N}_\sigma}[\log(1 - \mathcal{D}(\mathcal{R}(\tilde{X})))] \Big)$$

 https://arxiv.org/abs/1802.09088

Usage of the model

https://arxiv.org/abs/1802.09088

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

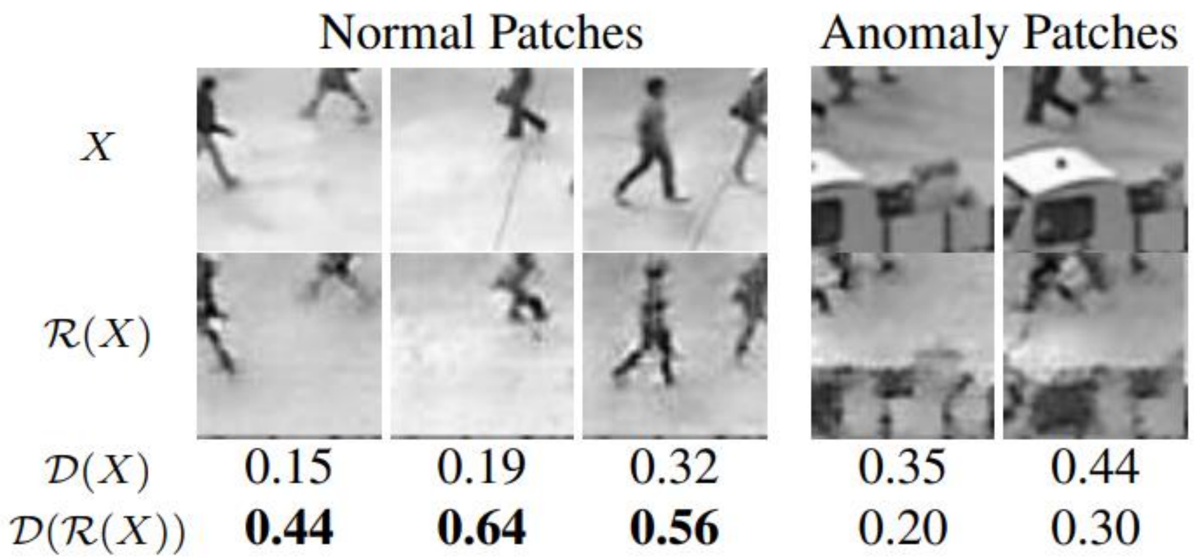
$$\mathrm{OCC}_1(X) = \begin{cases} \text{Target Class} & \text{if } \mathcal{D}(X) > \tau, \\ \text{Novelty (Outlier)} & \text{otherwise,} \end{cases}$$

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

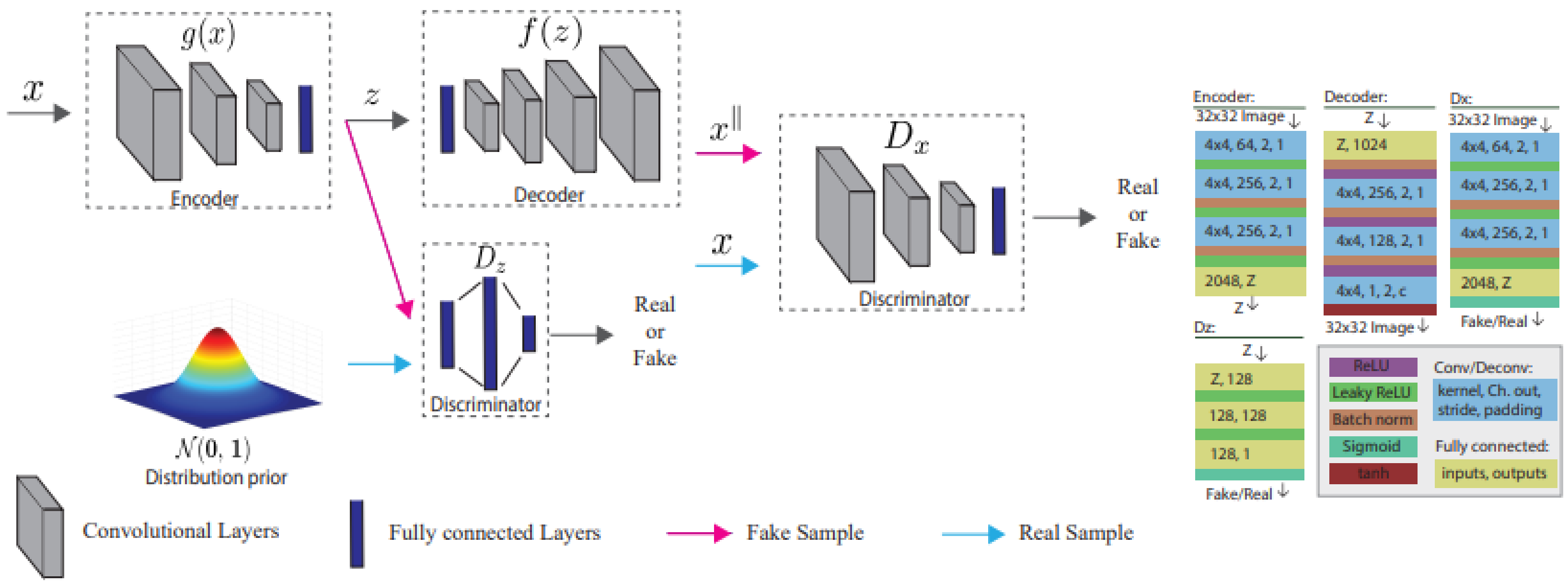
https://arxiv.org/abs/1802.09088



	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**

Adversarial autoencoders help

- Construct anomaly score a(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

Many Discriminators help

Figure 1. Limitations of in-class representation based novelty detection. Top: Input images; Middle: Output of an auto-encoder network trained on digit 8. Bottom: Output produced by OCGAN, the proposed method. Even though auto-encoder network is trained only on digits of 8, it provides good reconstruction for digits from classes 1,5,6 and 9. In contrast, OCGAN forces the latent representation of any example to reconstruct a digit 8. As a result, all out-of-class examples produce high Mean Squared Error (MSE). The intensity of red color in the bottom two rows is proportional to the MSE.

Table 1. Mean One-class novelty detection using Protocol 1.

	MNIST	COIL	fMNIST
ALOCC DR [22]	0.88	0.809	0.753
ALOCC D [22]	0.82	0.686	0.601
DCAE [23]	0.899	0.949	0.908
GPND [18]	0.932	0.968	0.901
OCGAN	**0.977**	**0.995**	**0.924**

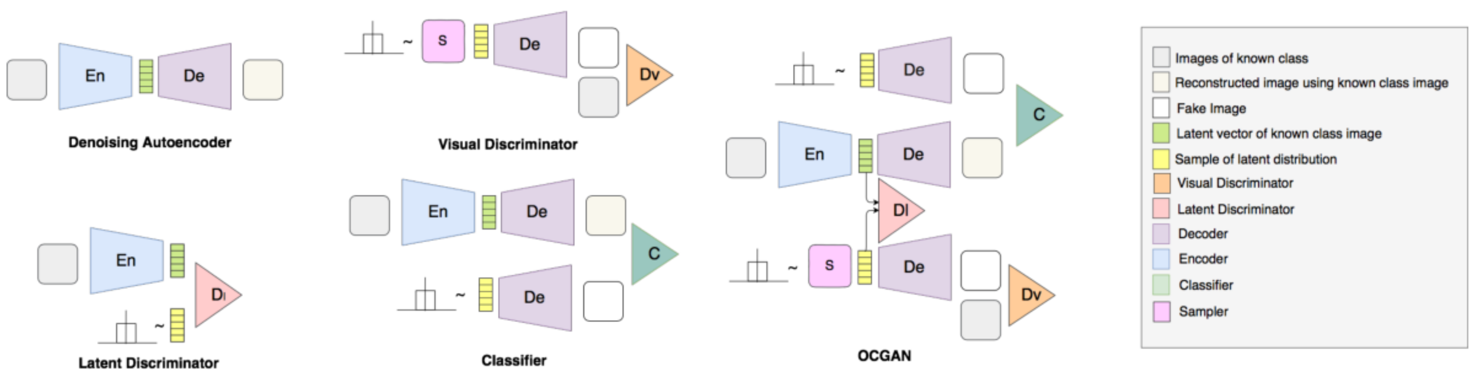


Figure 5. Illustration of OCGAN architecture: the network consists of four sub-networks : an auto-encoder, two discriminators and a classifier.

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

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

- Understand the nature of anomalies and the best match algorithm

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 close to the change detection problem