# Out-of-Distribution and Anomaly Detection

IA Frameworks

Joseba Dalmau

# Learning Outcomes

→ Master the fundamental OOD / AD concepts

→ Understand the main algorithm families, their advantages and weaknesses

→ Familiarize with the most common evaluation metrics

# What is an Anomaly?

## Anomaly

"An observation that deviates considerably from some concept of normality"

# Why Care About Anomaly Detection?

→ Intrusion detection (cybersecurity)

→ Fraud detection (finance)

→ Industrial fault / damage detection

→ Medical diagnosis

→ [NEW!] Out-of-Distribution detection for ML and DL models

# Main Challenges

- Ill-defined problem
- Absence of abnormal data
- Deep Learning use-cases: data is very high-dimensional

# What is an Anomaly ?

"An observation that deviates considerably from some concept of normality"

A typical mathematical formalization :

$\mathcal{X} \subset \mathbb{R}^D$ data space

$\mathbb{P}^+$ ground-truth law of normal data on $\mathcal{X}$ with p.d.f. $p^+$

$$\mathcal{A}_\tau = \{ x \in \mathcal{X} \mid p^+(x) \leq \tau \}$$ ← set of anomalies

$\tau$ ↑ ?

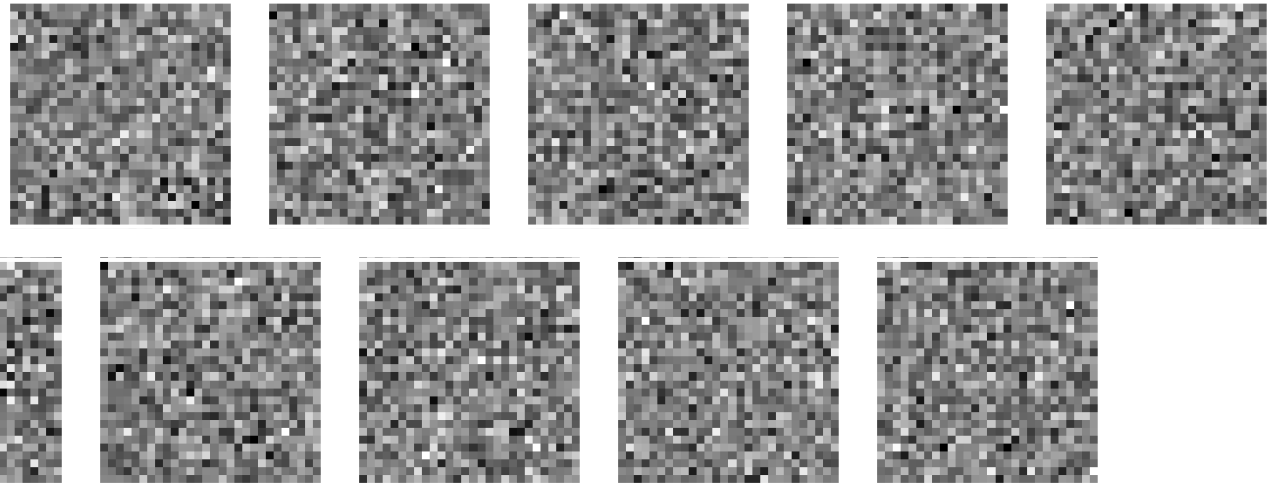
# CAREFUL!

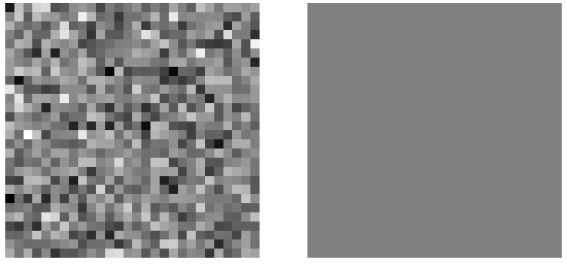
What is the concept of normality used in this formalization ?

# CAREFUL!

What is the concept of normality used in this formalization ?

10 NORMAL images →





← Question : which of these 2 images is normal ?

# CAREFUL!

Yet: pixels are i.i.d.

$\mathcal{N}(0,1)$ random variables :

$\Rightarrow$ Likelihood Ratio

$$\frac{f(\blacksquare)}{f(\blacksquare)} = \prod_{ij} \frac{1}{e^{-\frac{x_{ij}^2}{2}}}$$

$x_{ij}$: value of pixel $ij$ in $\blacksquare$

# Types of Anomalies

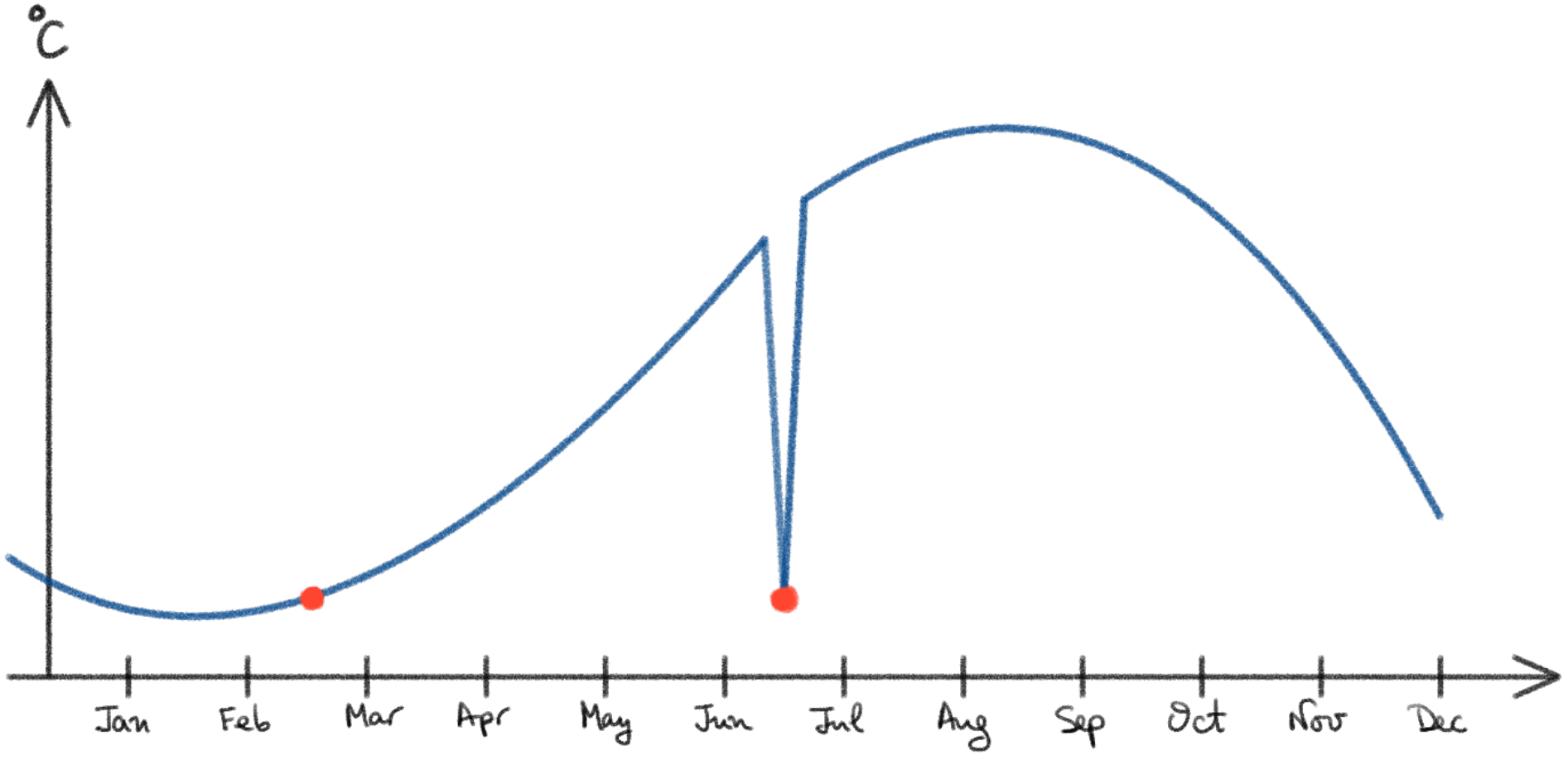
Traditionally:

→ Point anomalies

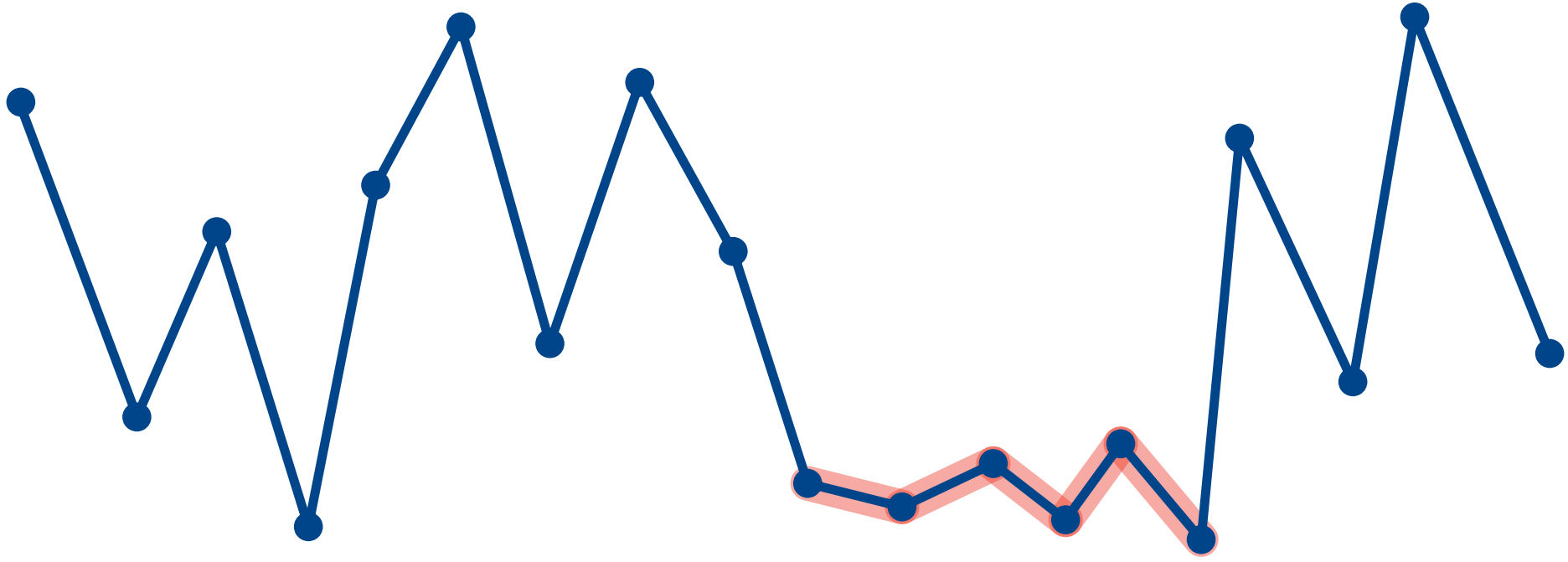
→ Conditional / Contextual anomalies

→ Group / Collective anomalies

# Contextual Anomalies



# Collective Anomalies



# Types of Anomalies

Since Deep Learning:

→ Low-level sensory anomalies

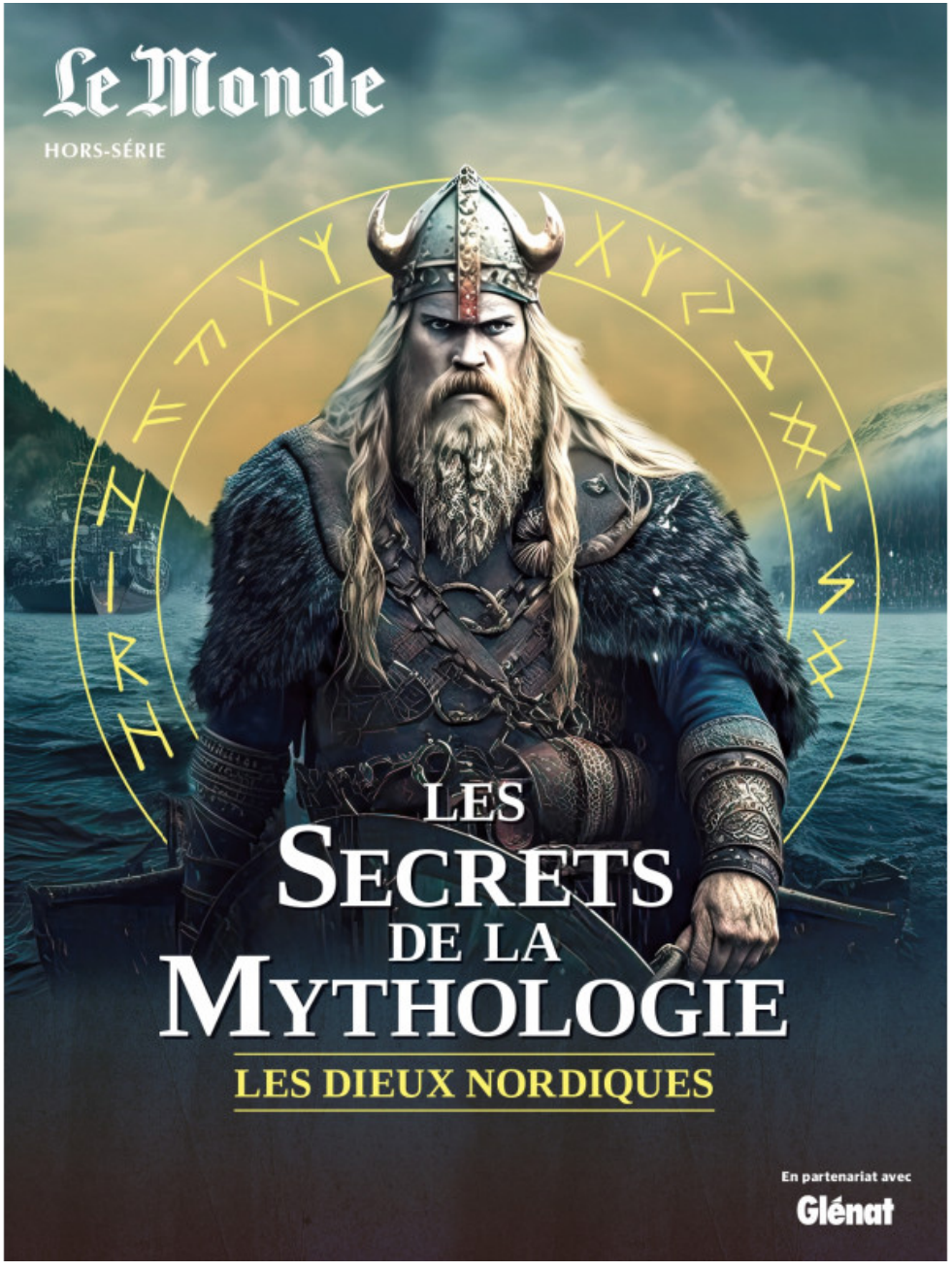
→ High-level semantic anomalies

# Types of Anomalies

Since Deep Learning:

- Low-level sensory anomalies
- High-level semantic anomalies

# Low Level, Sensory




The book is interesting and it conveys a meaningful messgae.

# High Level, Semantic

The girl was playing with a colorless green liquid.

# Types of Anomalies

Anomaly not drawn from $\mathbb{P}^+$

Outlier drawn from $\mathbb{P}^+$ w low proba

Novelty from an evolving $\mathbb{P}^+$

# Anomaly Detection Strategies

→ Model-agnostic

Given $X_1, \dots, X_n \sim \mathbb{P}^+$, $X_{test} \in A_c$ ?

→ Model-based

Given $(X_1, Y_1), \dots, (X_n, Y_n) \sim \mathbb{P}^+ \times \mathbb{P}_Y$

and $h : \mathcal{X} \longrightarrow \mathcal{Y}$, $X_{test} \in A_c$ ?

# Anomaly Detection Algorithms

## AD algorithm

$s : \mathcal{X} \longrightarrow \mathbb{R}$ and $\tau \in \mathbb{R}$

$s$: score, $\tau$: threshold

such that

$$\begin{cases} \text{if} \quad s(x) \leq \tau \Rightarrow & x \text{ is declared "normal"} \\ \text{if} \quad s(x) > \tau \Rightarrow & x \text{ is declared "anomalous"} \end{cases}$$

# Anomaly Detection Algorithms

→ One-class classifier

→ Probability density estimation

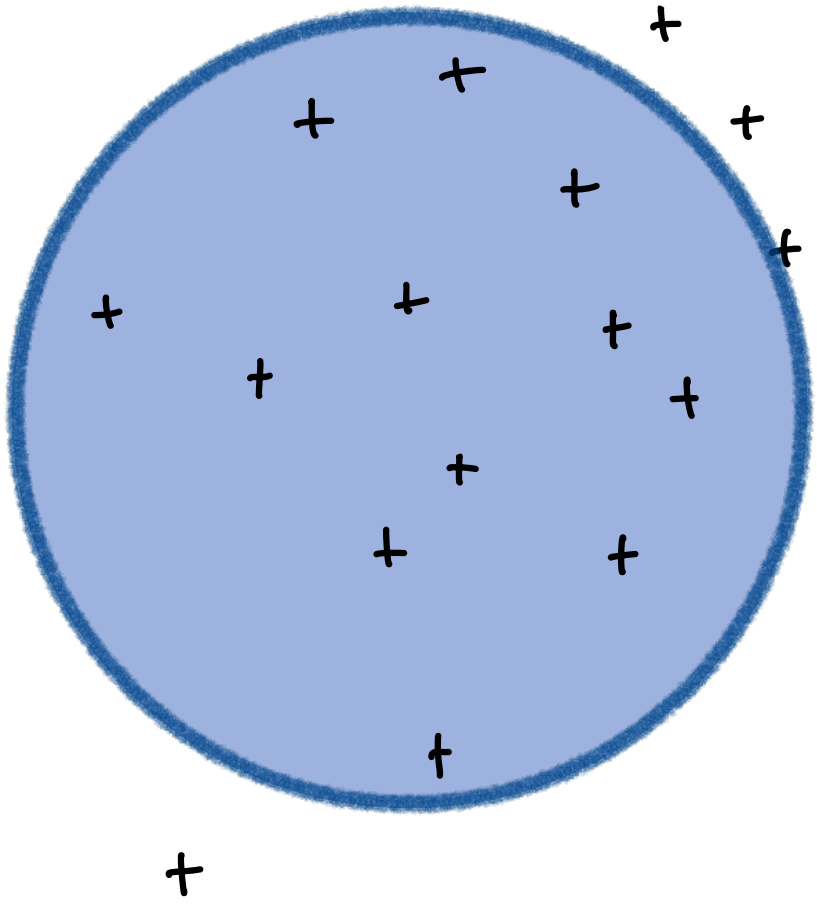
→ Reconstruction

→ Distance

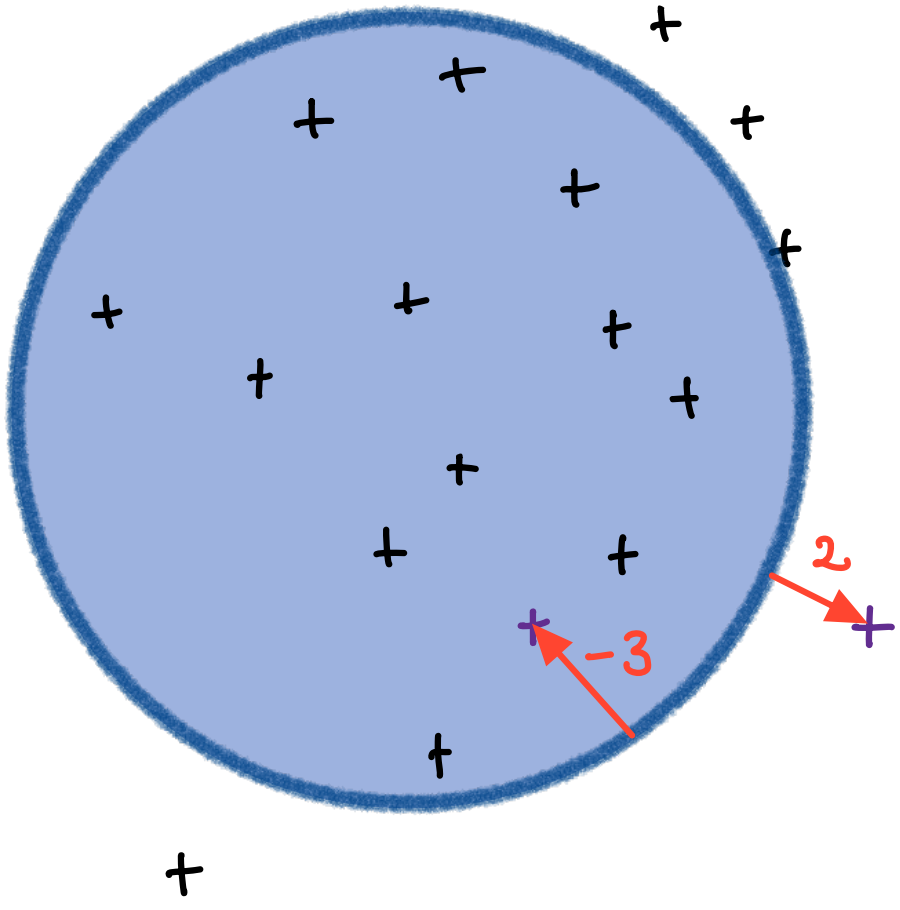
+ predictor $h : \mathcal{X} \to \mathcal{Y}$

(if available)

# One-Class SVM



# ONE-CLASS SVM




$s(x)$ = signed distance from $x$ to ○

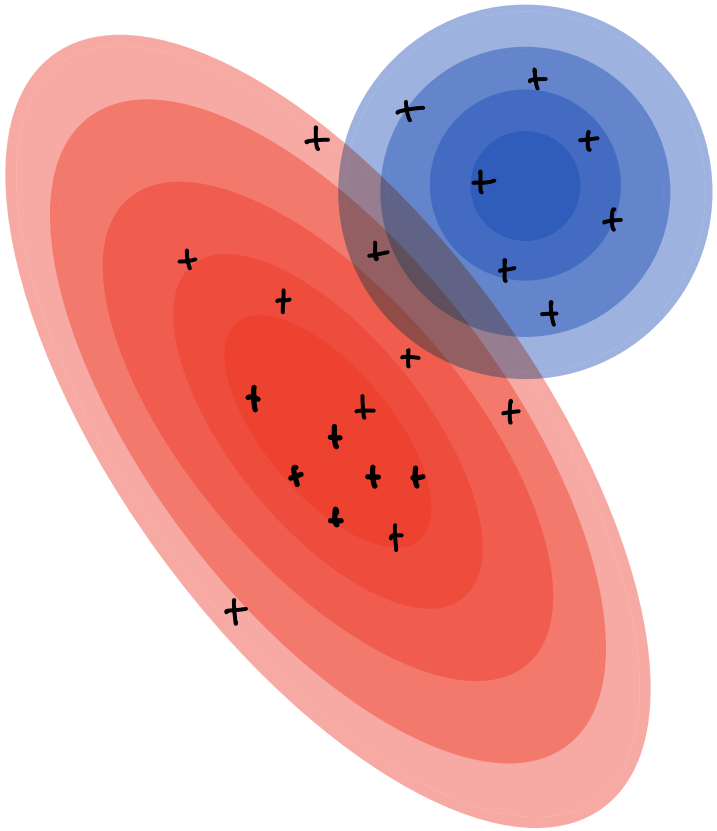
$$\tau = 0$$

$$\Rightarrow \begin{cases} x \in \bigcirc \Rightarrow \text{"normal"} \\ x \notin \bigcirc \Rightarrow \text{"anomaly"} \end{cases}$$

# Probability Density Estimation

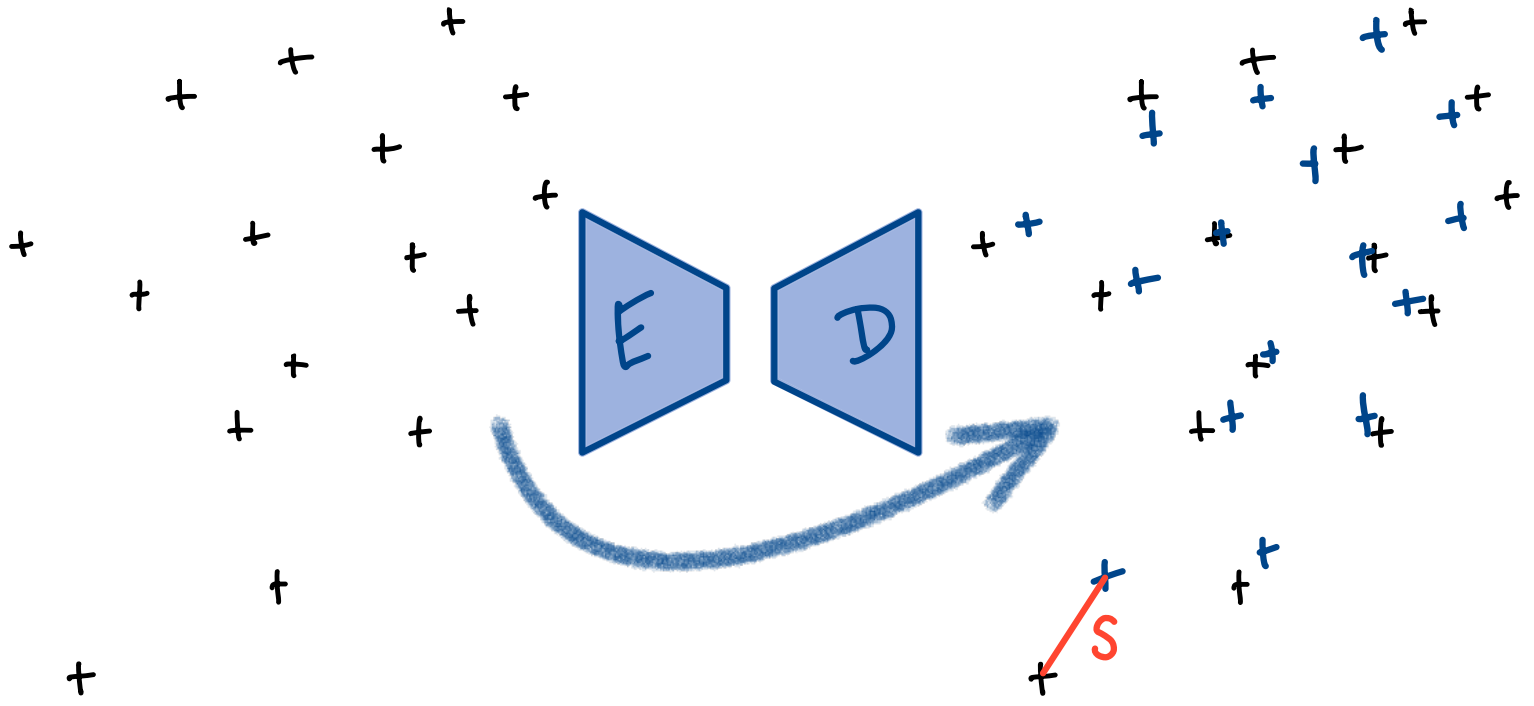
e.g. with a Gaussian Mixture Model (GMM)

i.e. $\mathbb{P}^+ \simeq \underbrace{\sum_{i=1}^{K} w_i \mathcal{N}(\mu_i, \Sigma_i)}_{\text{GMM fitted on } x_1, \dots, x_n}$



$S(x) = -$ log-likelihood of $x$ under the GMM

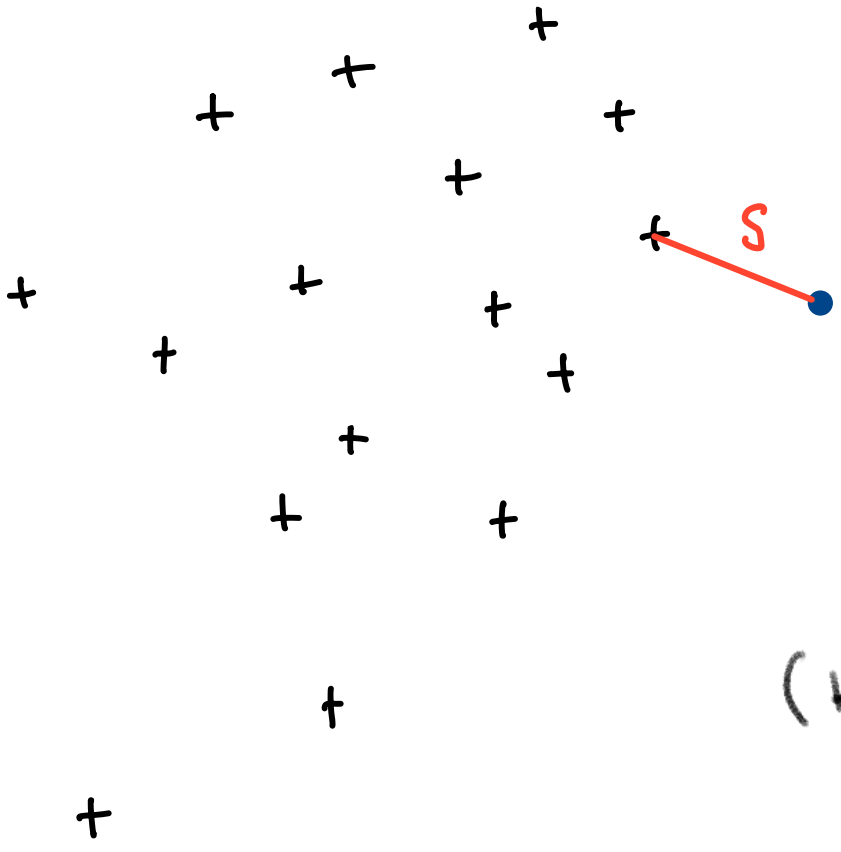
# Reconstruction w. Autoencoders




$$S(x) = d(x, D(E(x)))$$

# Distance w. K-nearest neighbors

$+ \sim \mathbb{P}^+$




$S(x)$ = distance to (K-th) nearest neighbor

# Quiz

| | sensitive to outliers? | scales to large dim? | ε easy to pick? |
|---|---|---|---|
| SVM | | | |
| GMM | | | |
| AE | | | |
| K-NN | | | |

# Evaluation Metrics

Normal = −  Anomaly = +

| | | Predicted | |
|---|---|---|---|
| | | − | + |
| Ground Truth | − | True Negative | False Positive |
| | + | False Negative | True Positive |

Confusion Matrix

# Evaluation Metrics

For a fixed threshold $\tau$:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

$$\text{Precision} = \frac{TP}{TP + FP}$$

| | | Predicted | |
|---|---|---|---|
| | | − | + |
| Ground Truth | − | True Negative | False Positive |
| | + | False Negative | True Positive |

$$\text{Recall / True Positive Rate} = \frac{TP}{TP + FN}$$

# Evaluation Metrics

For a fixed threshold $\tau$:

$$F_1 \text{ score} = 2 \cdot \frac{\text{Recall} \cdot \text{Precision}}{\text{Recall} + \text{Precision}}$$

$$\text{False Positive Rate} = \frac{FP}{TN + FP}$$

# Quiz

1. Is accuracy an adequate metric for AD? Why?

2. Give an example of a confusion matrix where precision and recall are both high.

3. Express the $F_1$ score in terms of the elements of the confusion matrix

4. The $F_\beta$ score is $F_\beta = (1+\beta^2)\frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \beta^2 \text{Recall}}$

What term do we give more importance to

(i) if $\beta < 1$ ? (ii) if $\beta > 1$ ?

# Solutions

$$F_\beta = (1+\beta^2)\frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \beta^2 \text{Recall}}$$

# Evaluation Metrics

Threshold independent: AUC-ROC / AUROC

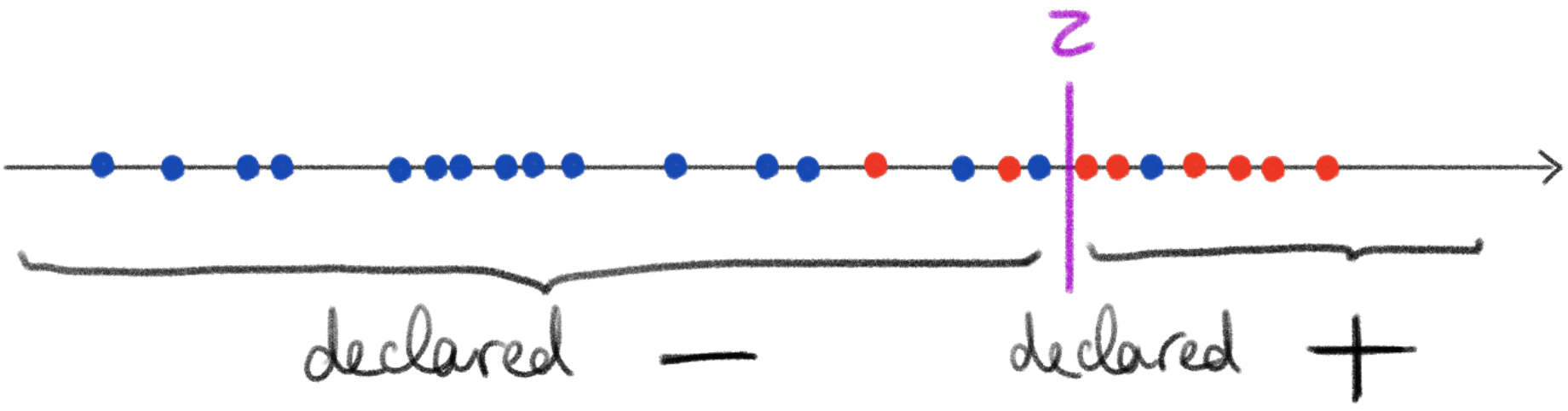
ROC = "Receiver Operating Characteristic"

Given: $s : \mathcal{X} \longrightarrow \mathbb{R}$

● = normal (ground-truth)

● = anomaly (ground-truth)

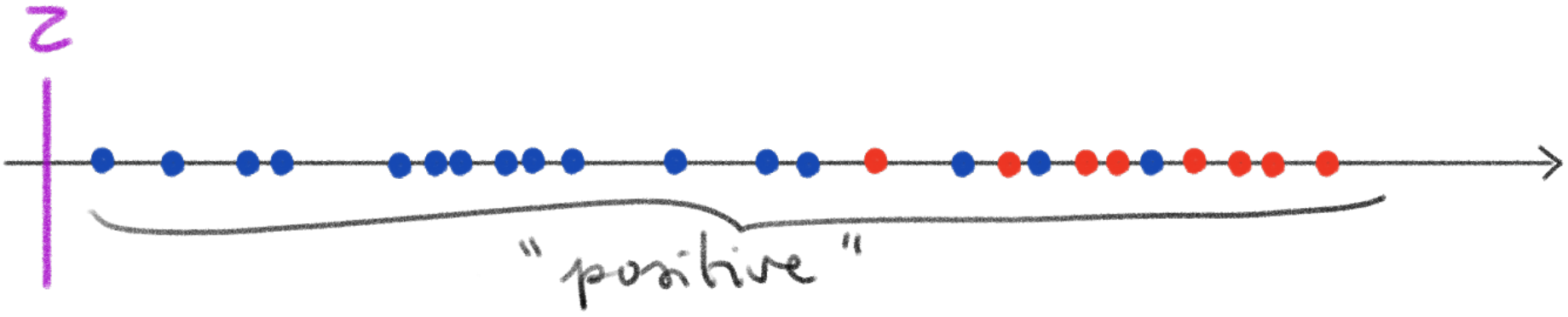
# Evaluation Metrics




$$TPR(z) = \frac{\#\text{ of } \bullet \text{ and } +}{\#\text{ of } \bullet}$$

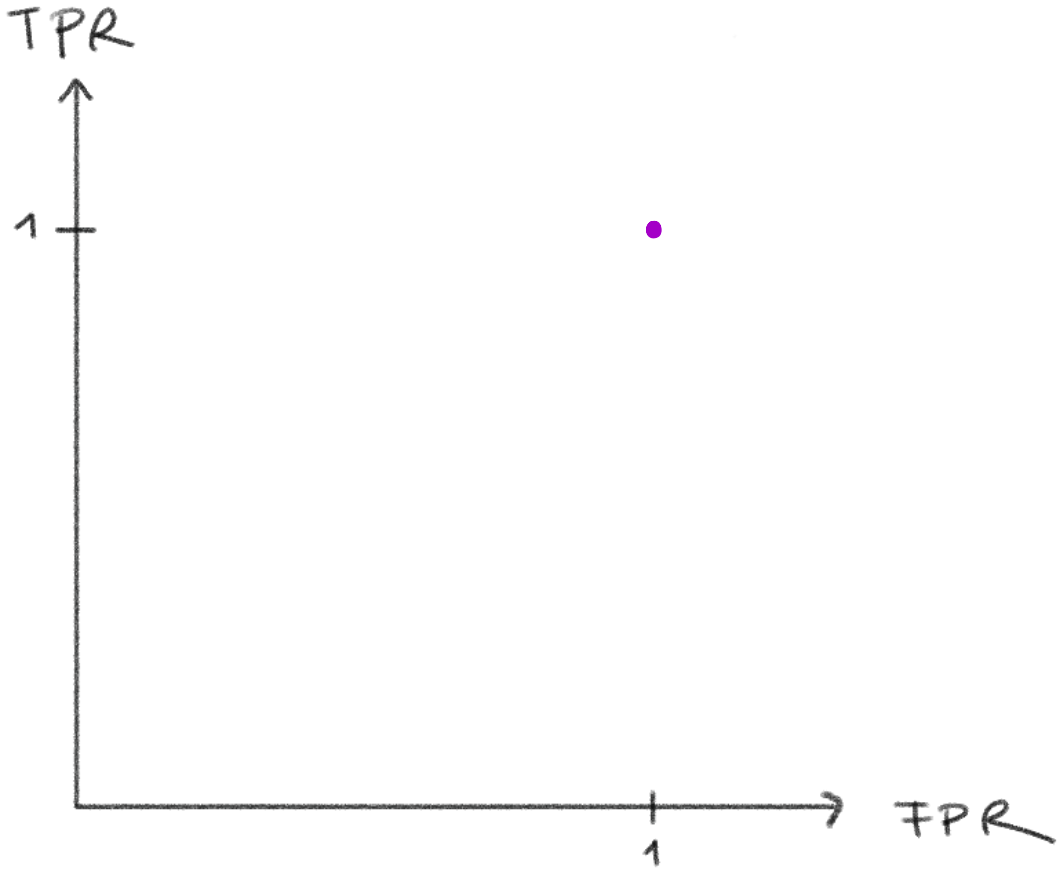
$$FPR(z) = \frac{\#\text{ of } \bullet \text{ and } -}{\#\text{ of } \bullet}$$

# Evaluation Metrics

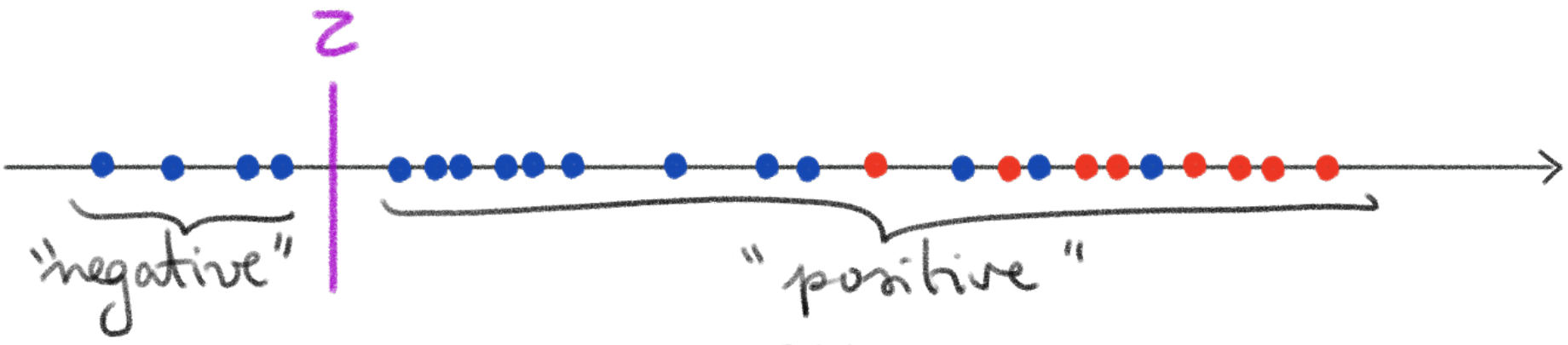



$FPR(z) = 1$

$TPR(z) = 1$

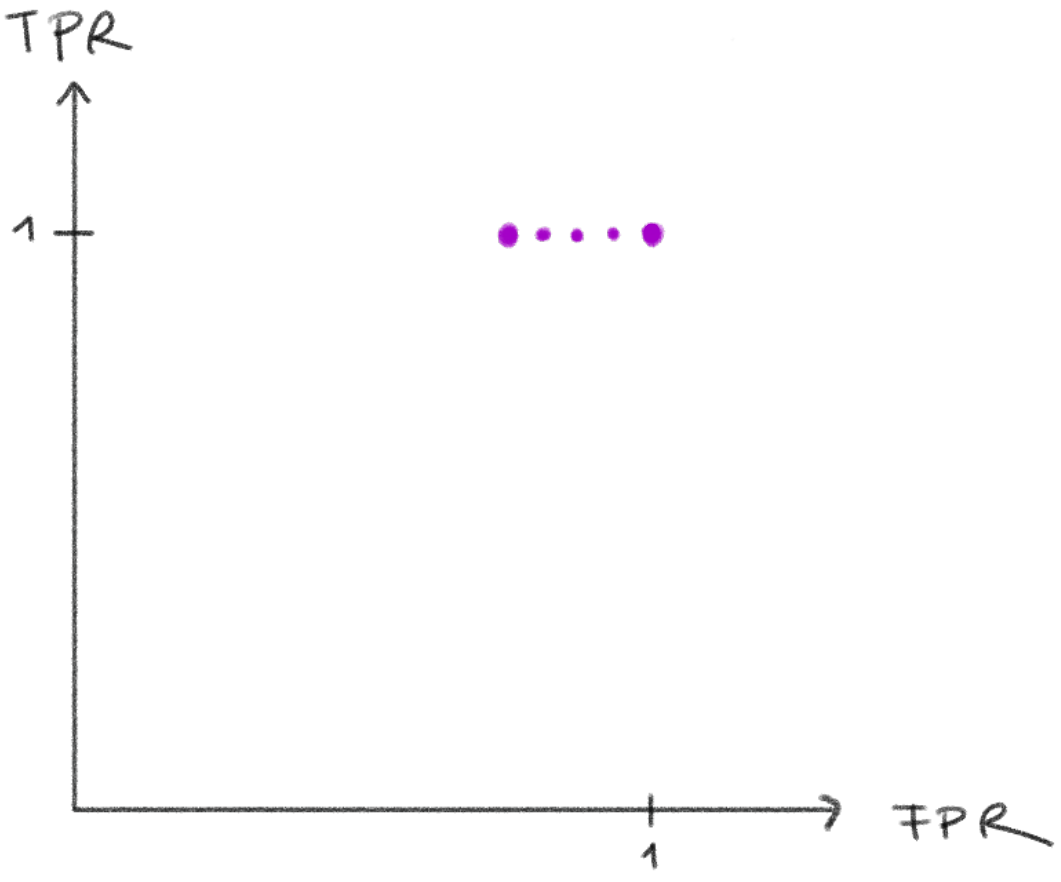


# Evaluation Metrics

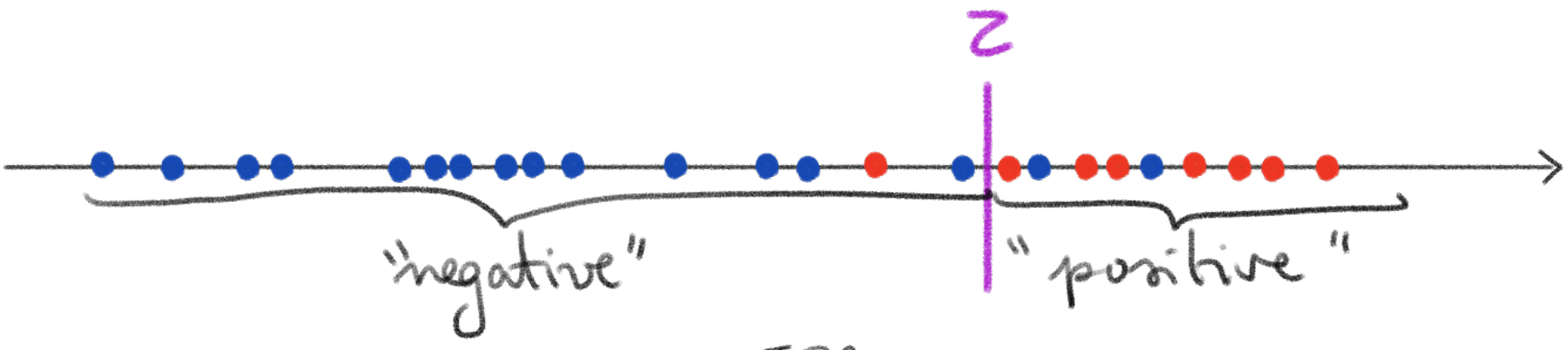



$$FPR(z) = \frac{12}{16} = \frac{3}{4}$$

$$TPR(z) = 1$$

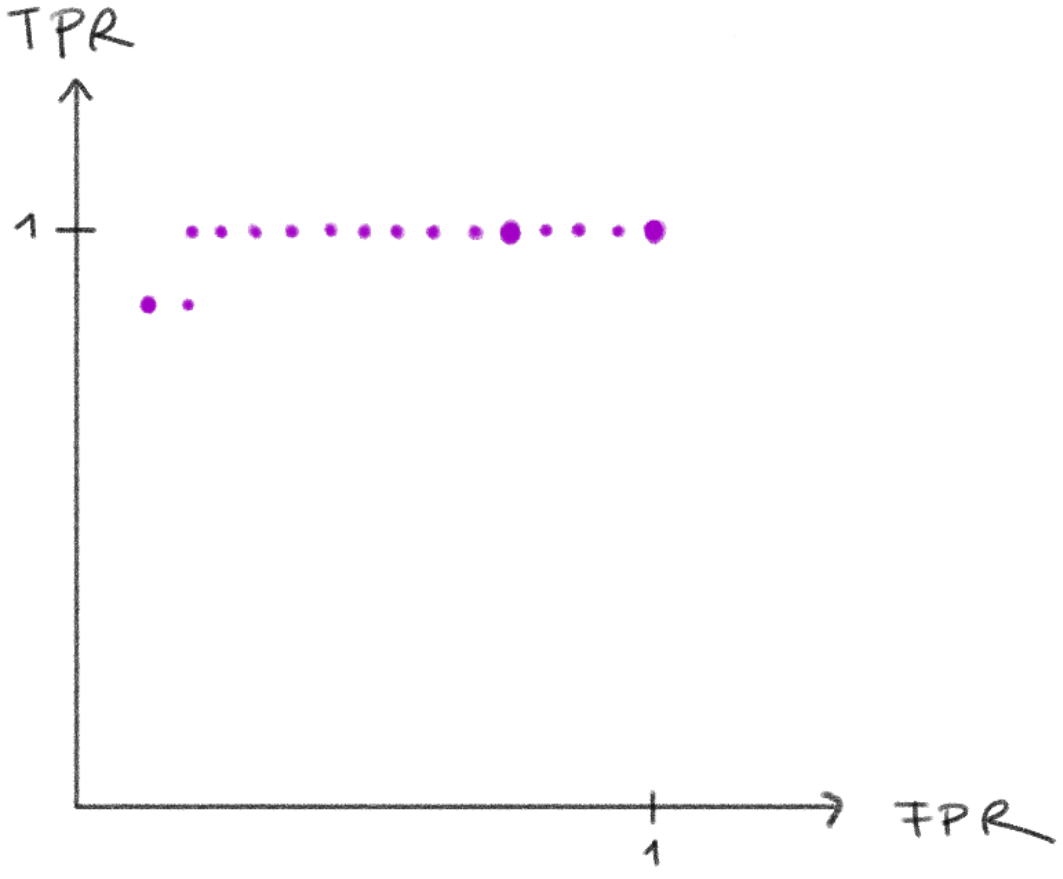


# Evaluation Metrics

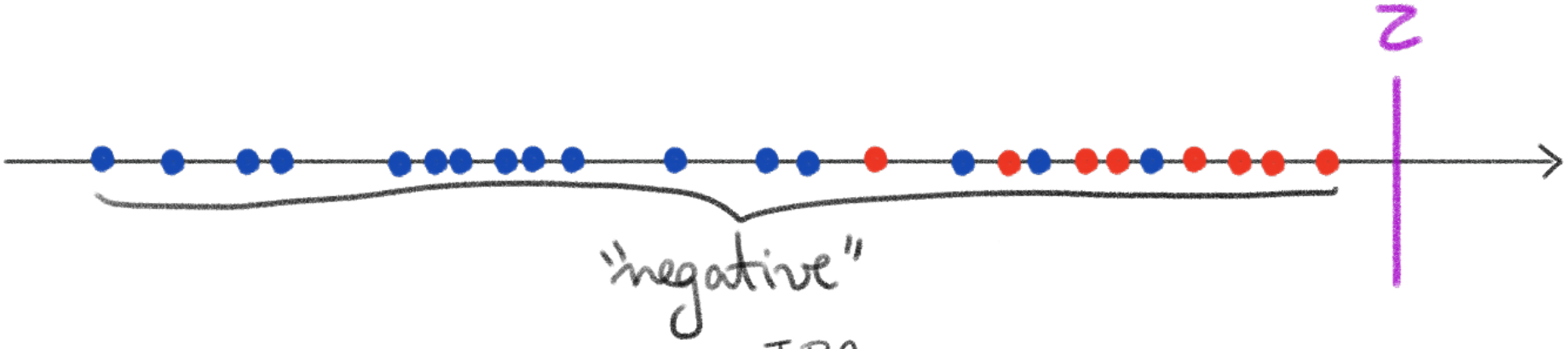



$$FPR(z) = \frac{1}{8}$$

$$TPR(z) = \frac{7}{8}$$

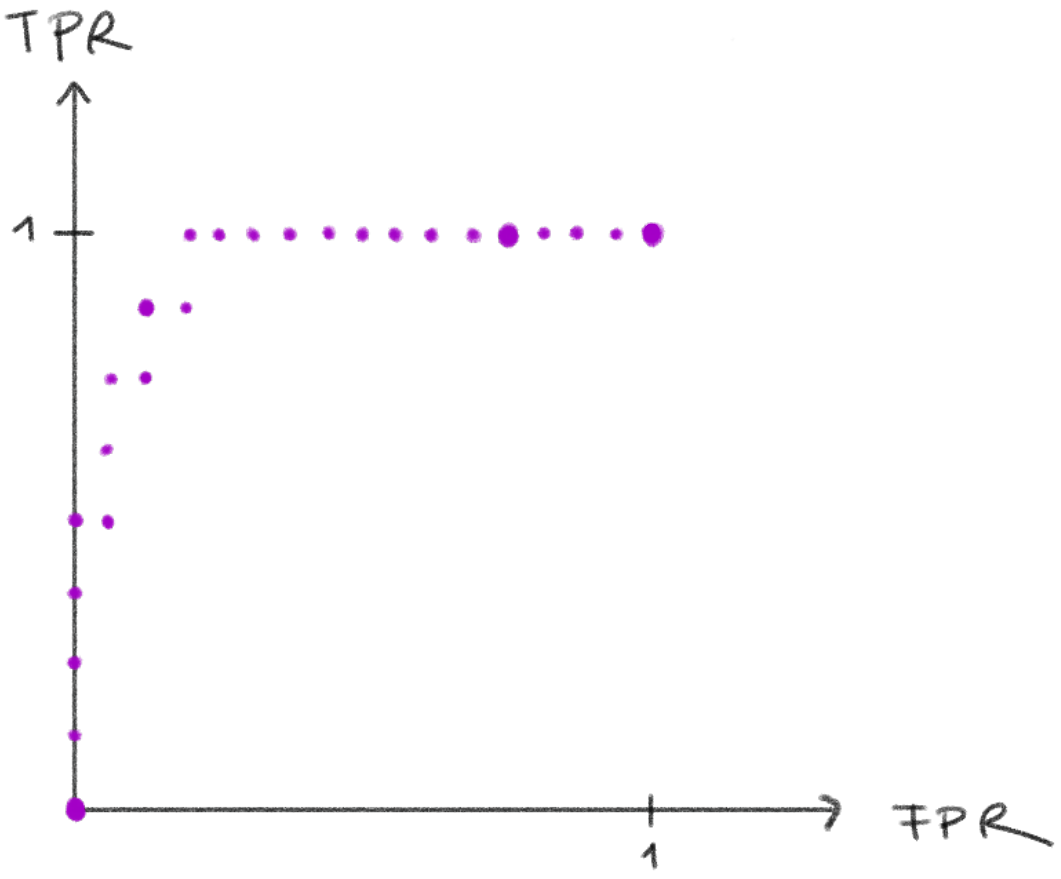


# Evaluation Metrics

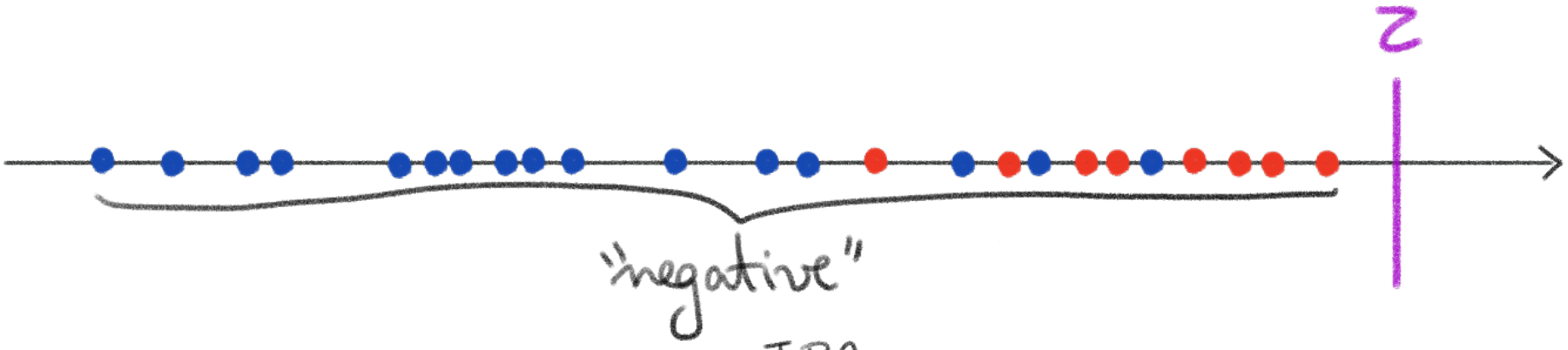



FPR(z) = 0

TPR(z) = 0

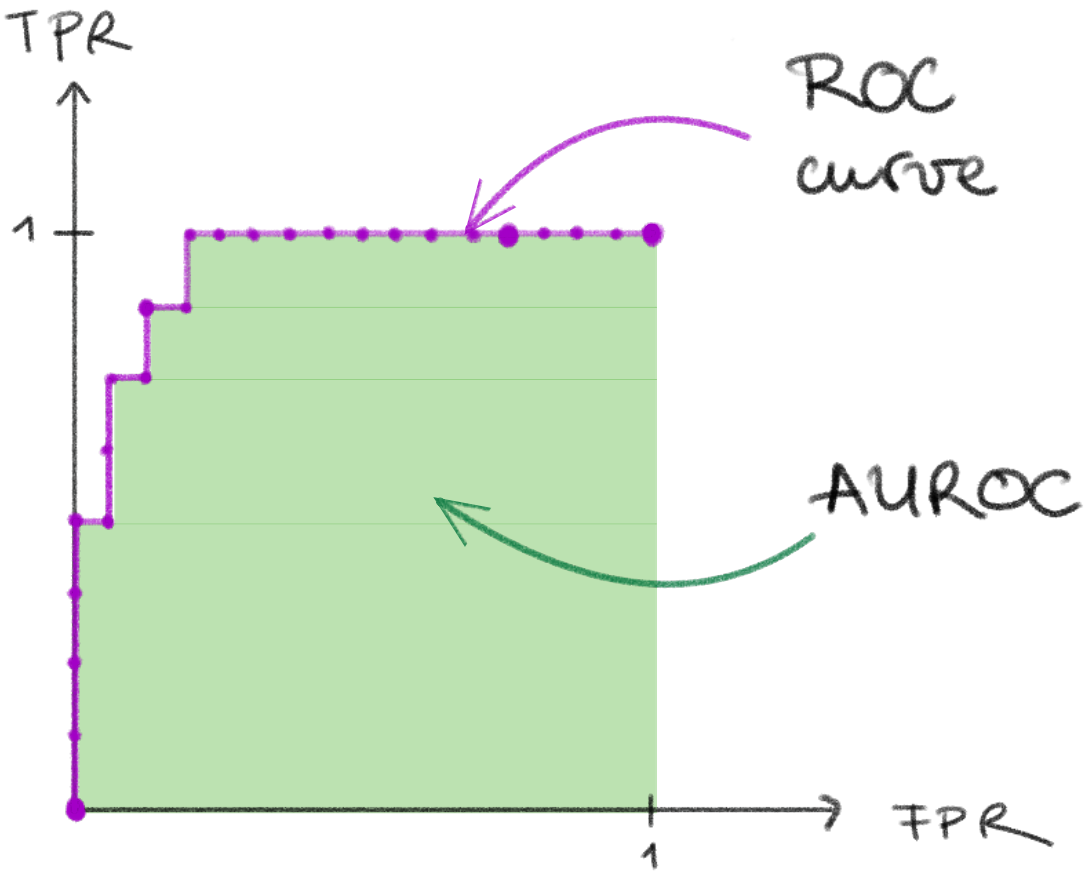


# Evaluation Metrics




$FPR(z) = 0$

$TPR(z) = 0$



# Quiz

1. What would a "perfect" score look like? And the ROC curve / AUROC associated?

2. Same question for a "random" score.

3. Same question for a "bad" score.

4. Can we generalize the ROC curve / AUROC to the case of probability distributions?

# Solutions

# BENCHMARKING

Lack of datasets for high-dimensional data (images, text)

see for example MVTecAD

Solution 1.

normal

anomalies

# BENCHMARKING

Lack of datasets for high-dimensional data (images, text)

see for example MVTecAD

Solution 2 .

MNIST → normal

CIFAR10 → anomaly

# Out-of-Distribution Detection


NN $\hat{f}_\theta$ trained on Cifar10

"deer"

# OoD = AD ?

→ They accept similar mathematical formulations, but:

1. In OoD detection the space of anomalous images can be much larger

2. In OoD detection we have access to feature representations via $\hat{f}_\theta$

# OoD Scores

For a classifier $\hat{f}_\theta$ :

$$\hat{f}_\theta : \mathcal{X} \xrightarrow{\hat{g}_\theta} \mathbb{R}^D \xrightarrow{\hat{h}_\theta} \mathbb{R}^C$$

feature space (usually the penultimate layer)

logit space

$\longrightarrow$ logit-based scores: $S(x) = \tilde{S}(\hat{f}_\theta(x))$

$\longrightarrow$ feature-based scores : $S(x) = \tilde{S}(\hat{g}_\theta(x))$

# Logit-Based Scores

→ Maximum Logit Score MLS

$$\tilde{S}(z) = -\max_{1 \leq c \leq C} z_c$$

→ Maximum Softmax Probability

$$\tilde{S}(z) = -\max_{1 \leq c \leq C} P_c(z)$$

# Logit-Based Scores

→ Energy

tunable hyperparameter

$$\tilde{S}(z) = -T \log \sum_{c=1}^{C} e^{z_c/T}$$

→ Entropy

$$\tilde{S}(z) = -\sum_{c=1}^{C} p_c(z) \log p_c(z)$$

# Feature-Based Scores

→ Deep K-nearest neighbor (DKNN)

Given $z_1, \dots, z_N$ and $z^*$ :

$z_1, \dots, z_N$: fitting dataset

$z^*$: test point

(i) Compute $d_1 = d(z_1, z^*), \dots, d_N = d(z_N, z^*)$

(ii) Rank $d_{(1)} \leq d_{(2)} \leq \dots \leq d_{(N)}$

(iii) DKNN score $\tilde{s}(z^*) = d_{(k)}$ ($k$: hyper-parameter)

# Feature-Based Scores

→ Mahalanobis

Given $x_1, \dots, x_N$ and $x^*$ :

$x_1, \dots, x_N$: fitting dataset

$x^*$: test point

(i) Fit $C$ Gaussians: $\mathcal{N}(\mu_c, \Sigma)$, $c = 1, \dots, C$

$$\hat{\mu}_c := \frac{1}{N_c} \sum_{i : y_i = c}^{N} x_i, \quad \hat{\Sigma} = \frac{1}{N} \sum_{c=1}^{C} \sum_{i : y_i = c} (x_i - \hat{\mu}_c)(x_i - \hat{\mu}_c)^T$$

# Feature-Based Scores

→ Mahalanobis

(ii) Mahalanobis score:

$$\tilde{s}(z) = \max_{1 \le c \le C} (z - \hat{\mu}_c)^T \hat{\Sigma}^{-1} (z - \hat{\mu}_c)$$

Exercise: If $C = 1$, $\hat{\mu}_c = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$ and $\hat{\Sigma} = \begin{pmatrix} 1 & 0 \\ 0 & 4 \end{pmatrix}$, draw $A \subseteq \mathbb{R}^2$ given by $A = \{z : \tilde{s}(z) \le 1\}$

# Resources

## Surveys

1. A unifying review of deep and shallow anomaly detection, Ruff et al. 2021

2. Generalized out-of-distribution detection, Yang et al. 2021

# RESOURCES

## Github repositories

1. OpenOOD

2. ADBench

# CHALLENGES

→ Representation learning for AD

→ Characterization of anomalies

→ Interpretability of AD

→ Creation of a unified methodology for AD research:

benchmarks, metrics, ...

# Quiz

1. Why is AD important for ML/DL?
2. Name 2 classes of AD algorithms.
3. Write down the formula for 1 metric.
4. Write down 1 sentence stating 1 thing you understood well.
5. Write down 1 question with something you would like to know more about.