Outlier and anomaly detection

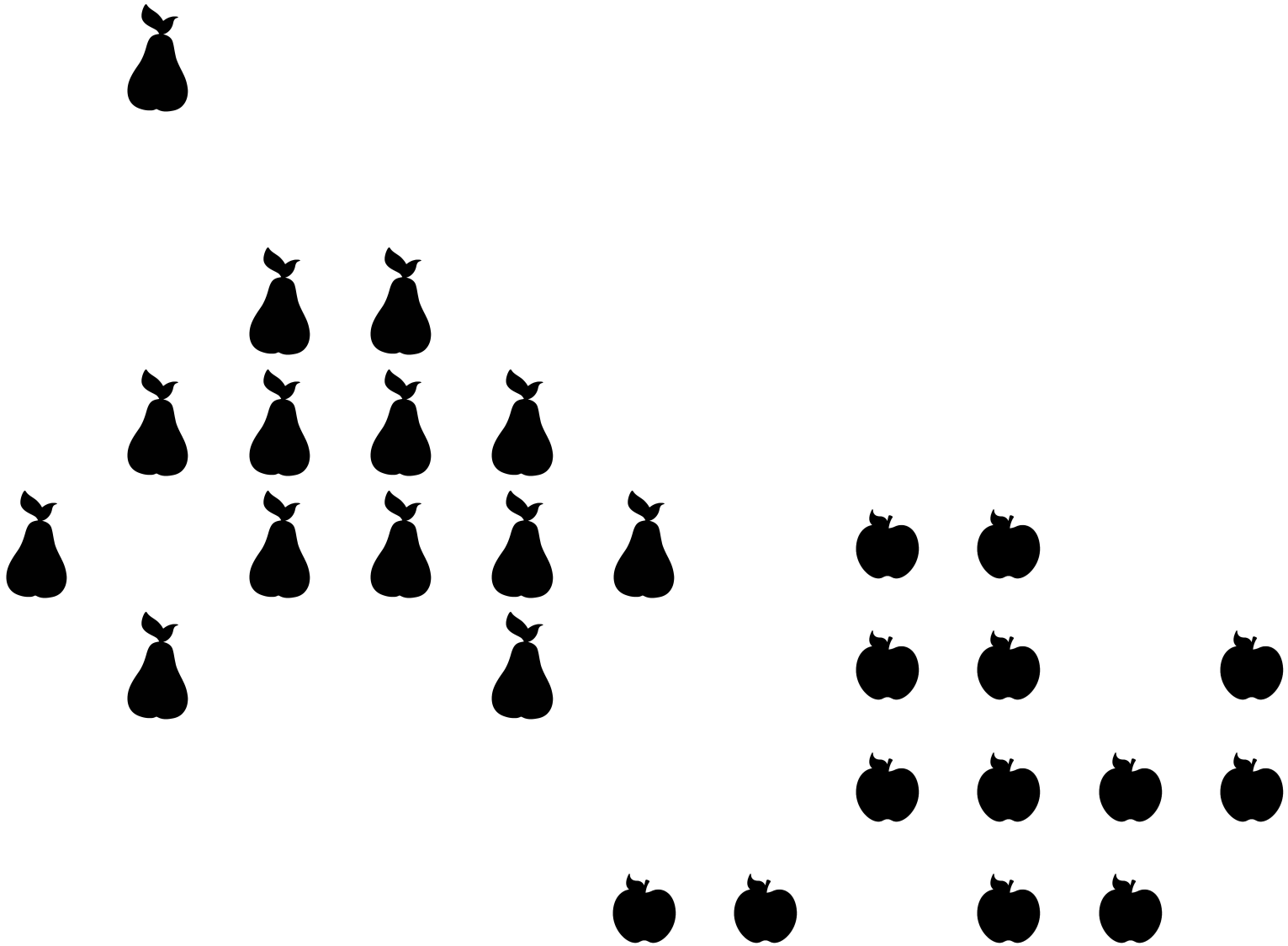
Tomáš Pevný

Department of Computers, Czech Technical University

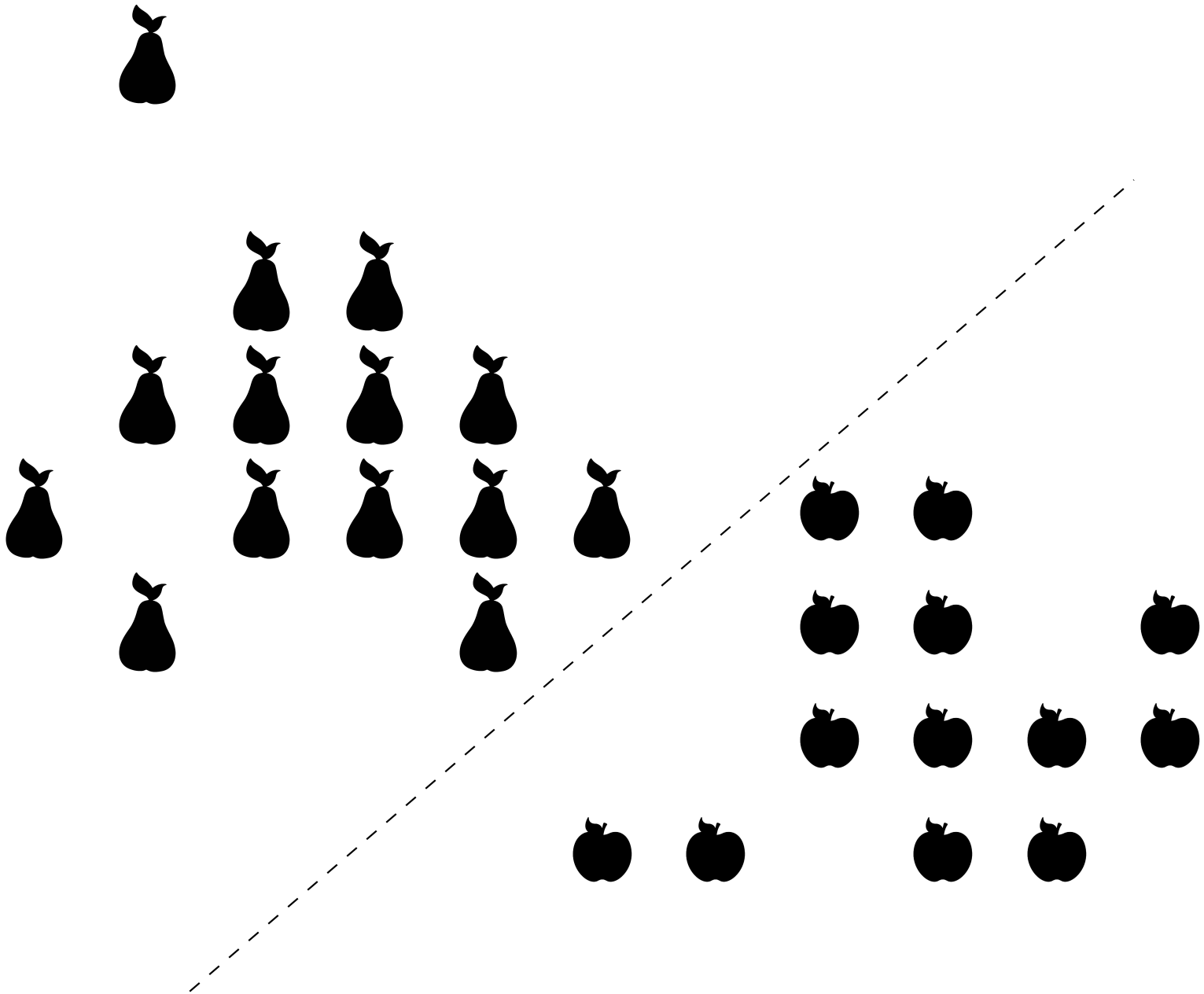
December 2, 2024

Plan

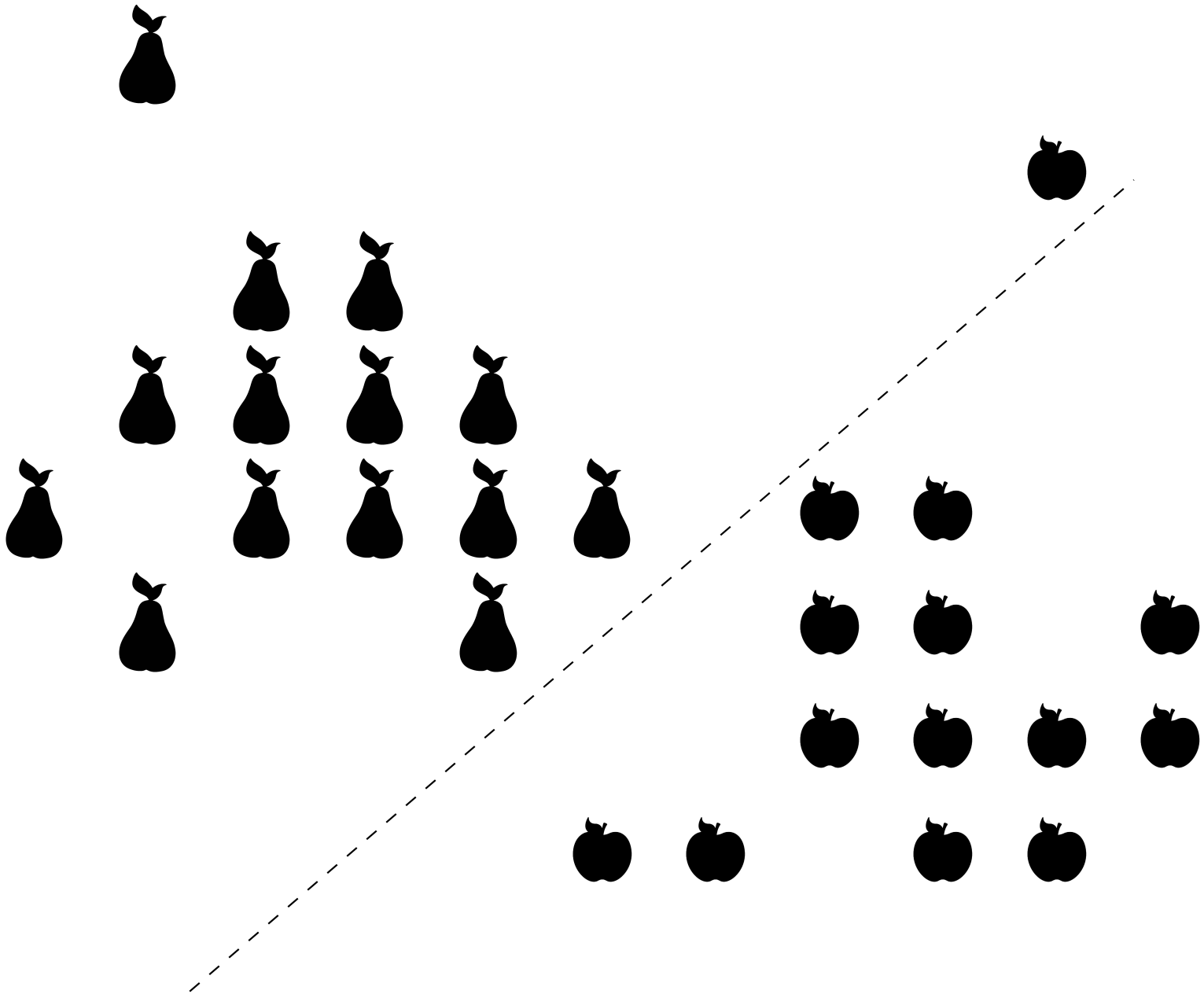
Supervised classification



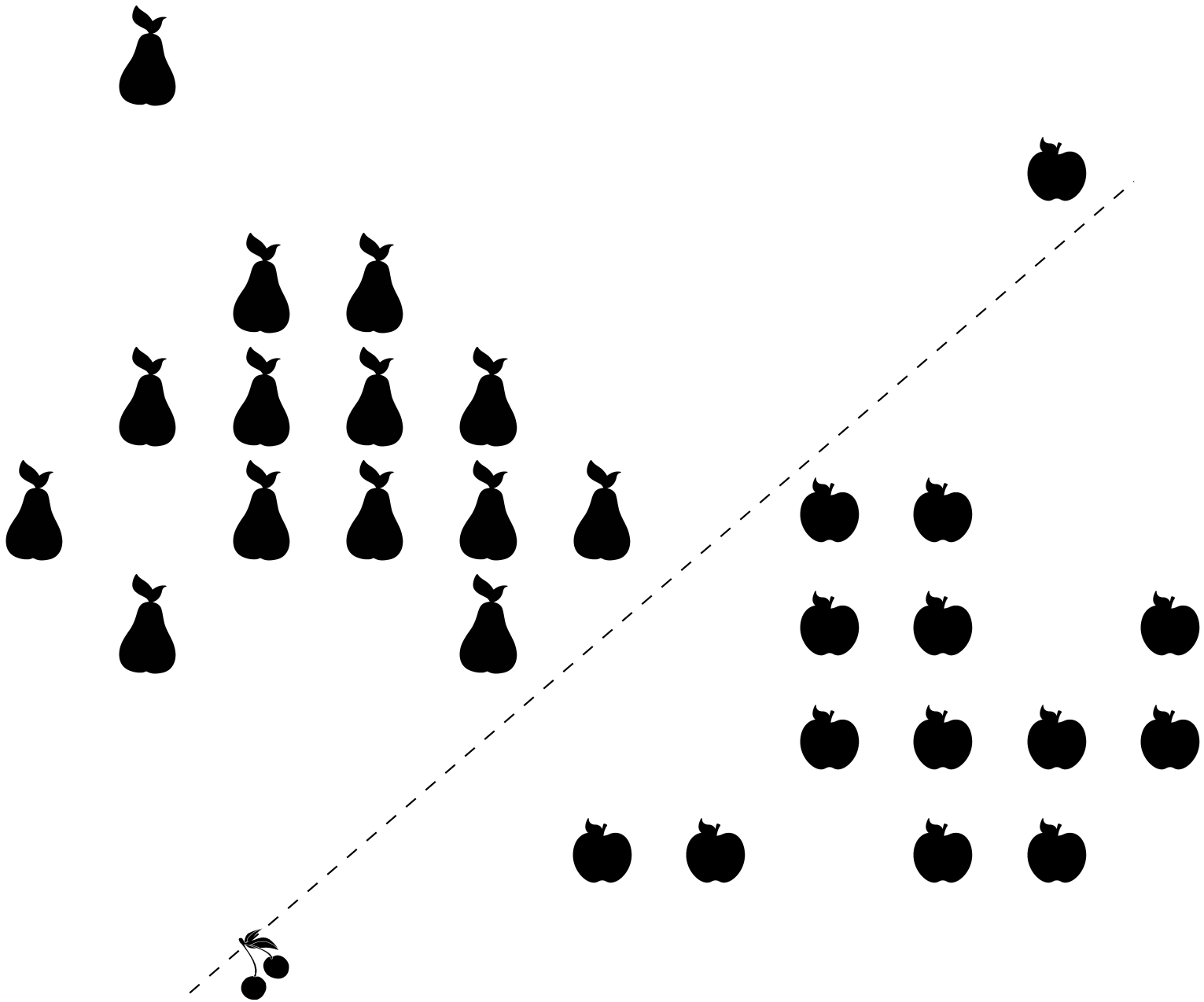
Supervised classification



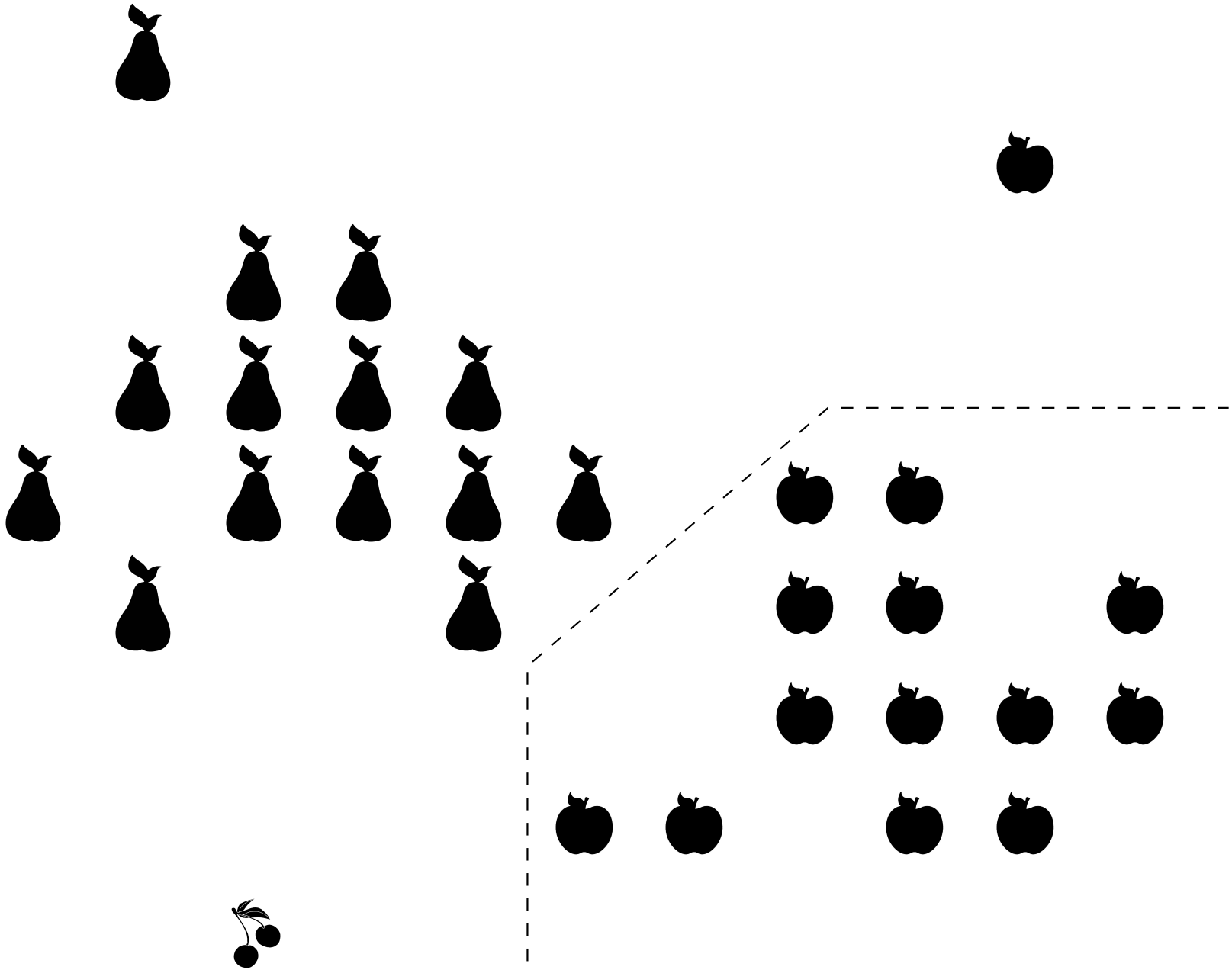
Supervised classification



Supervised classification



Anomaly detection



Where is anomaly detection used

- Security
  - Fraud detection
  - Intrusion detection
  - Airport security
- Safety
  - Monitoring of industrial processes
  - Detecting disease outbreaks
  - Detecting environmental hazards

Definition of anomaly

An outlier is an observation which deviates so much from the other observations as to arouse suspicions that it was generated by a different mechanism[1].

Anomaly detection refers to the problem of finding patterns in data that do not conform to expected behavior[2].

---

[1] *D. M. Hawkins, Identification of Outliers, 1980*

[2] *V. Chandola, A. Banerjee, and V. Kumar, Anomaly detection: a survey, 2009*

Pros and cons

Anomaly detection

- Does not need labelings (sort of).
- Need mostly clean datasest.
- Can identify unseen samples.
- Not all anomalies are interesting (harmful).

Supervised classification

- Needs a lot of labels.
- Very precise
- Provides labels of samples from known classes.
- Arbitrary results on unseen samples.

Plan

Problem definition

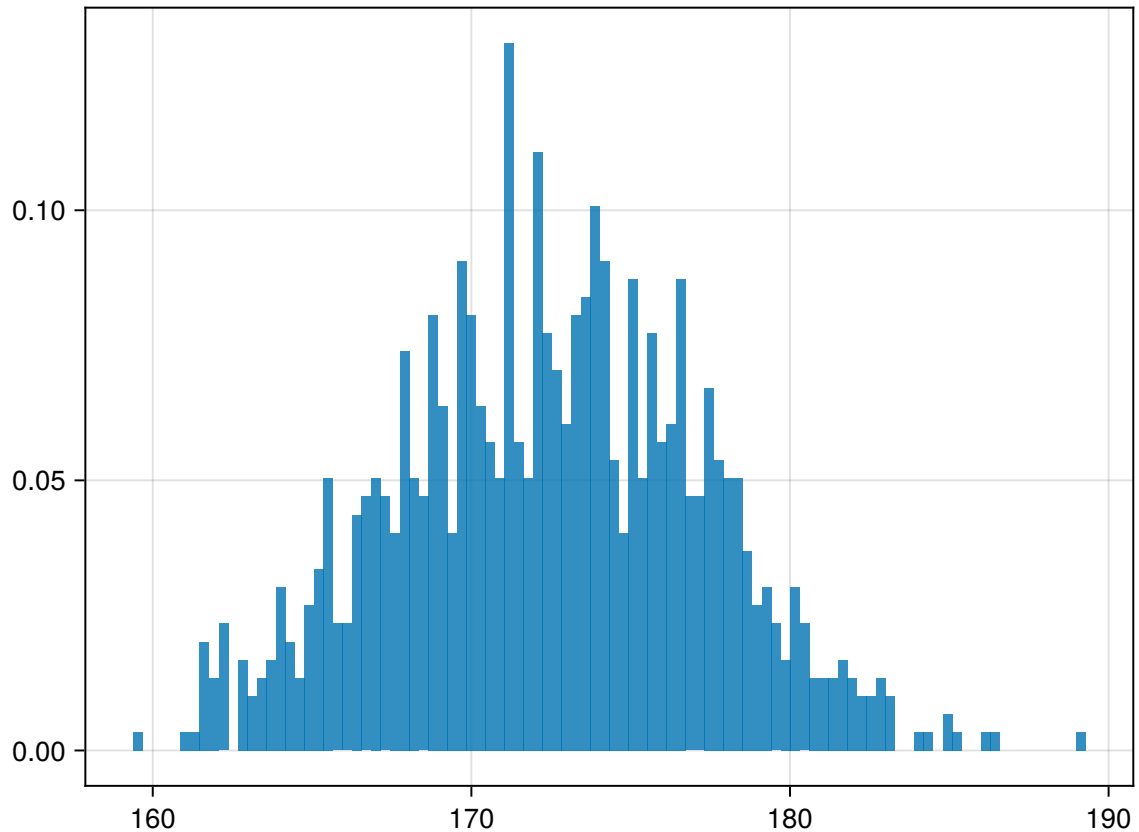
We are given a set of data points

$$\{x_i | x_i \in \mathbb{R}, i \in \{1, \ldots, n\}\}.$$

We want to identify anomalies in or with respect to them .

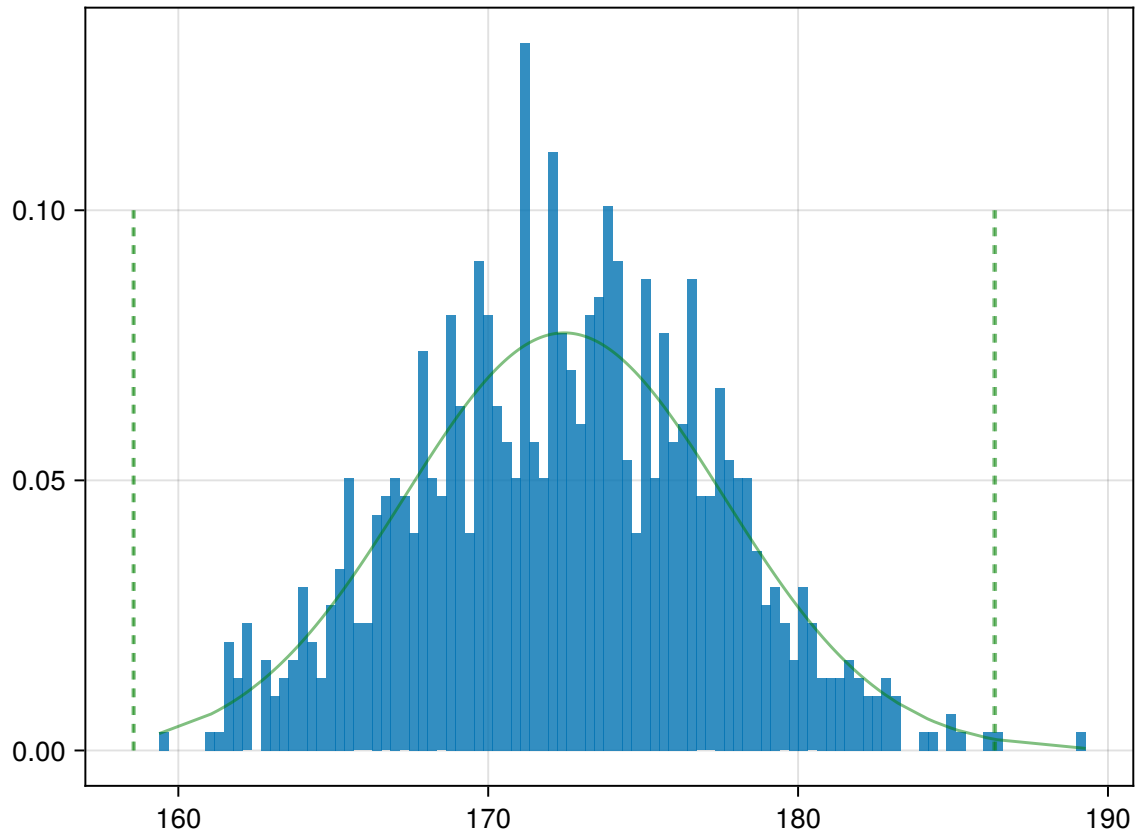
Height of people

159.38
161.06
161.27
161.51
161.52
⋮
185.01
185.16
186.06
186.41
189.28



Fitting known distribution (Normal)

Set thresholds to some quantile, here to 0.0035.

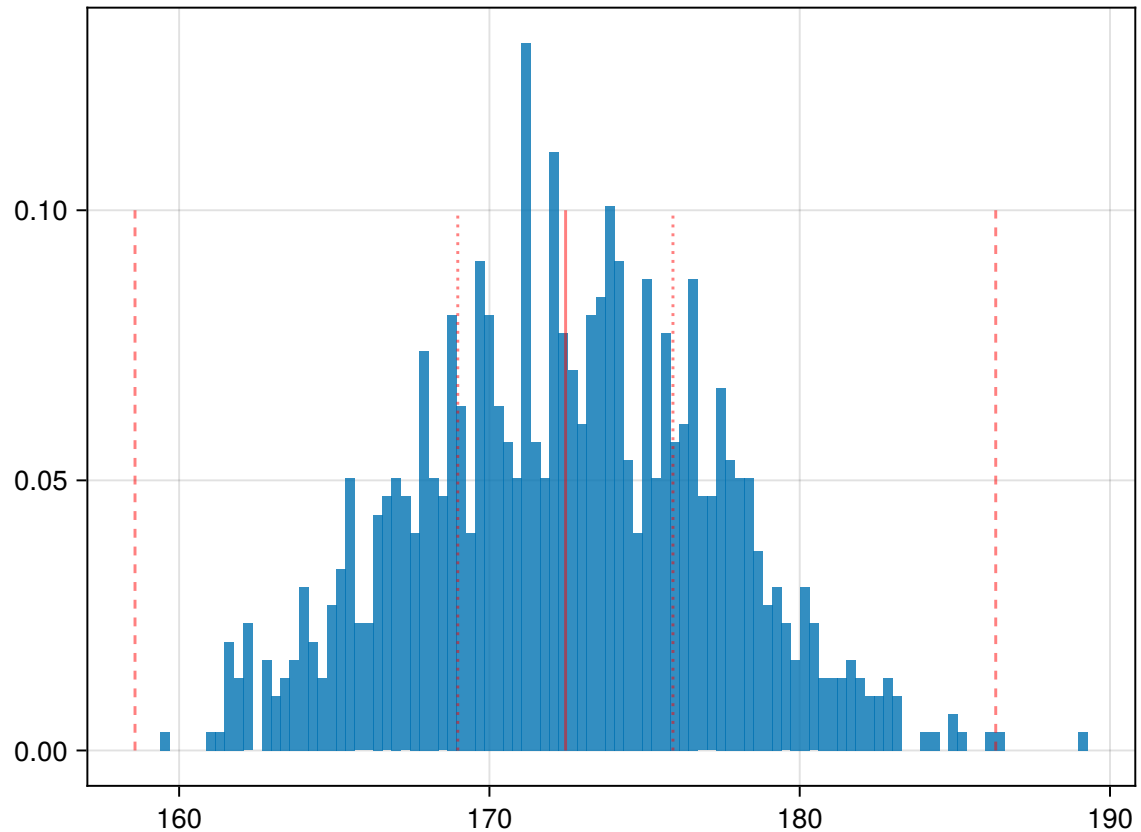


Tukey method — BoxPlot

Set thresholds to

$$\mathrm{md}(x) \pm 1.5\mathrm{IQR}(x)$$

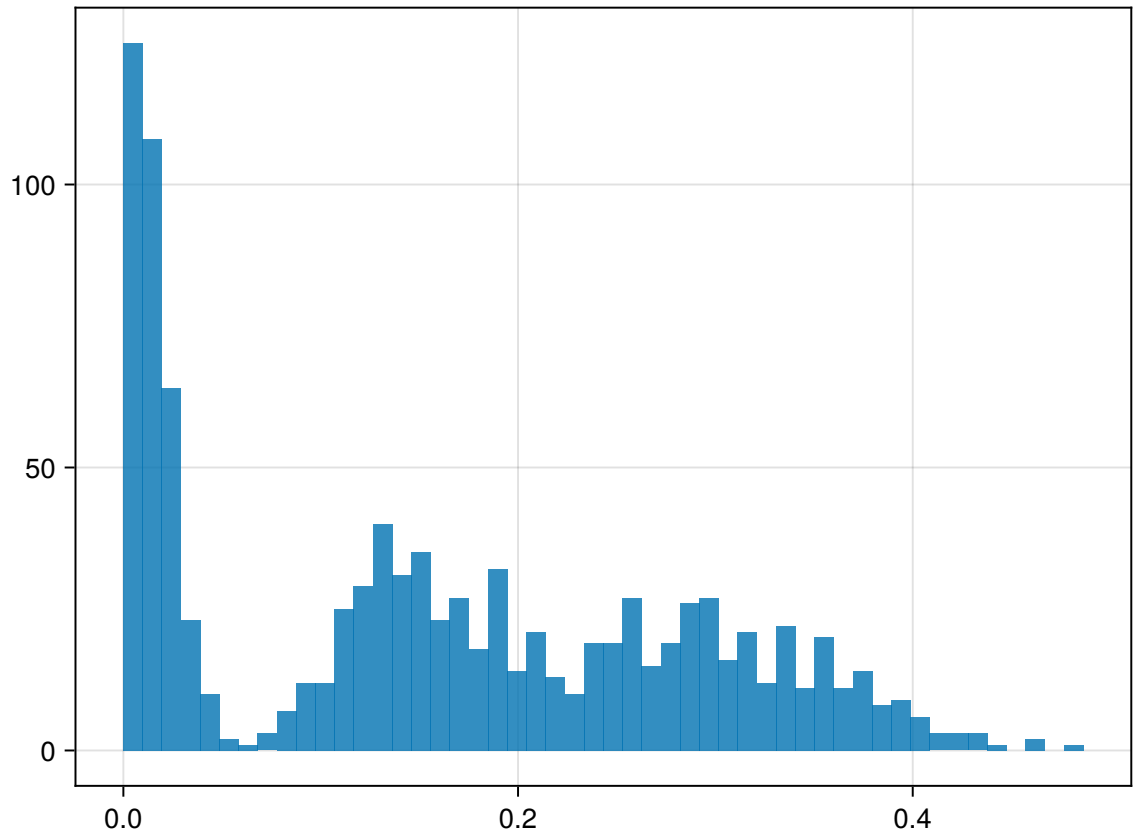
which corresponds to 0.0035 quantile of normal.



Tukey, John W, Exploratory Data Analysis, 1977

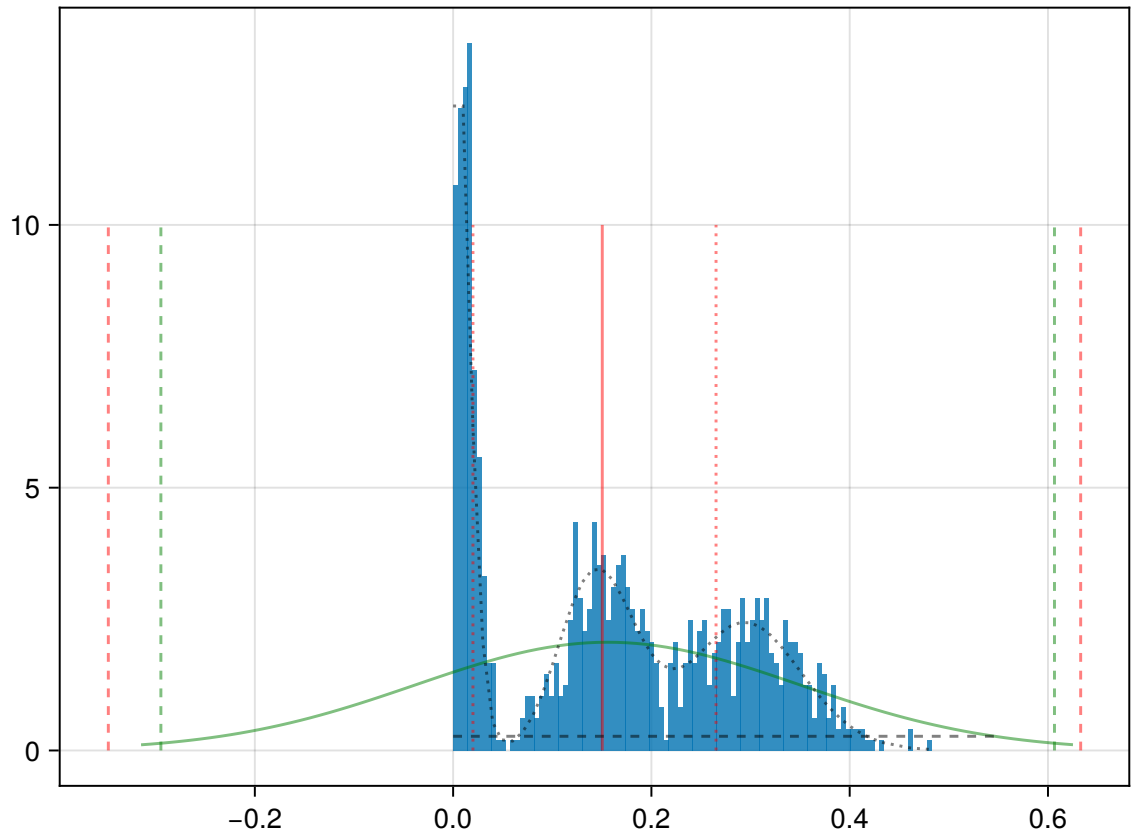
Distribution of round-trip time of packets

The distribution has three modes corresponding to rtt to local, Prague, US servers.



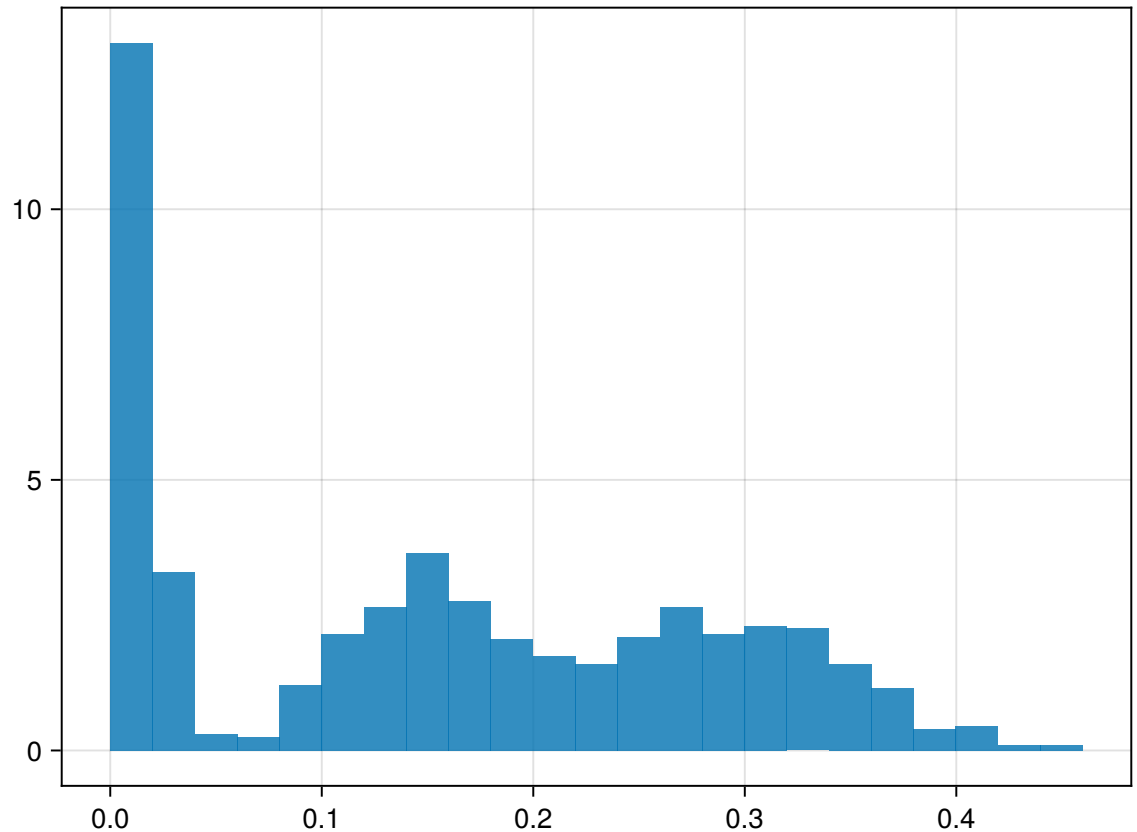
Tukey and Normal

Tukey's method and method assuming Normal distribution fails.



Approximate the distribution by histogram

How to determine the threshold?



Setting the threshold on anomaly

Assume that p is a probability density function and α is desired false positive rate.
Then x is deemed as an anomaly when $p(x) \leq \beta$ with

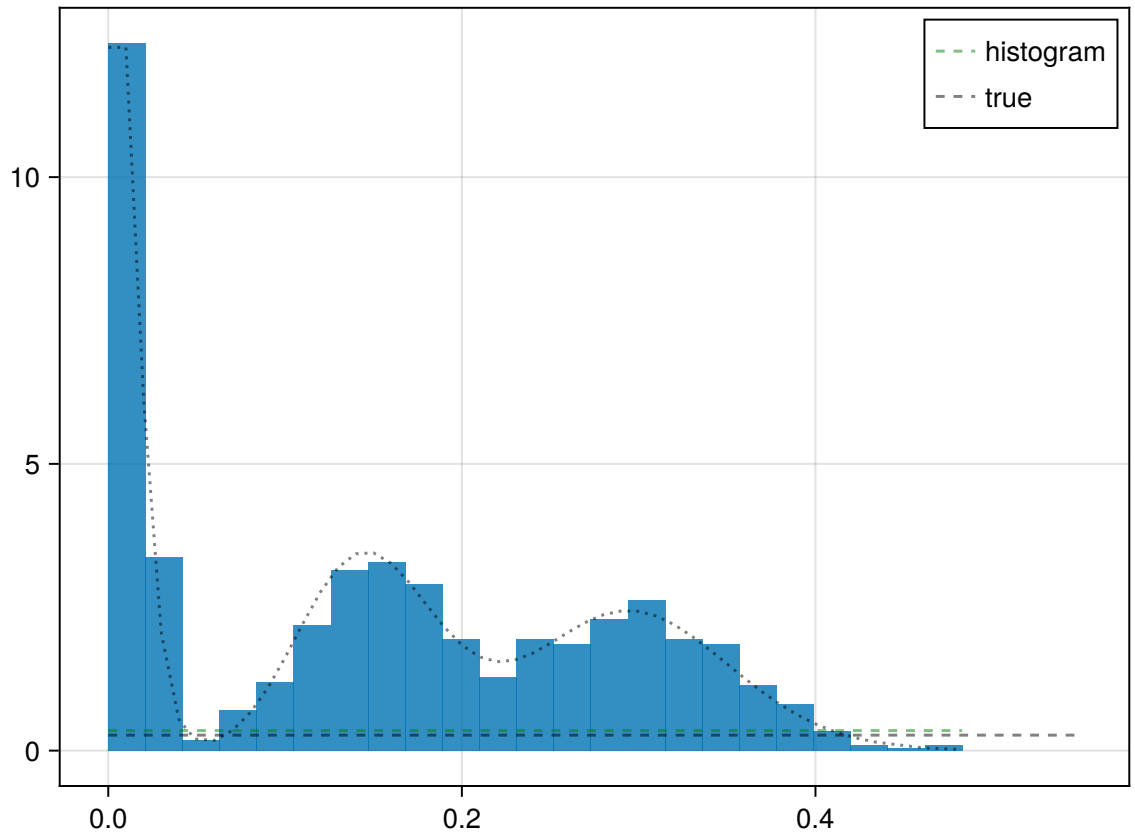
$$\beta = \arg\max_{\beta \in \mathbb{R}} \int_{\mathscr{X}} \mathrm{I}[p(x) \leq \beta] p(x) dx \leq \alpha,$$

where $\mathrm{I}[\cdot]$ denotes Iverson brackets evaluating to 1 if the argument is true and zero otherwise.

Number of bins by Sturge

Sturge's formula

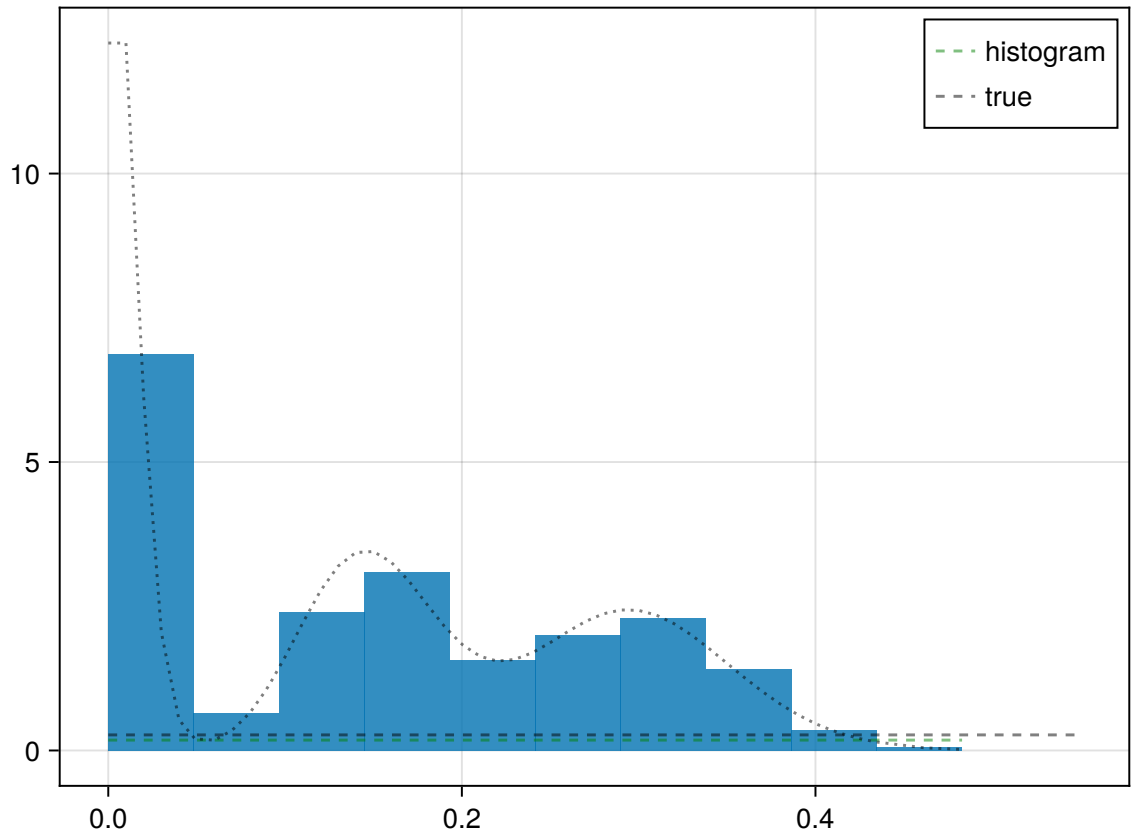
$$1 + 3.222 * \log(n)$$



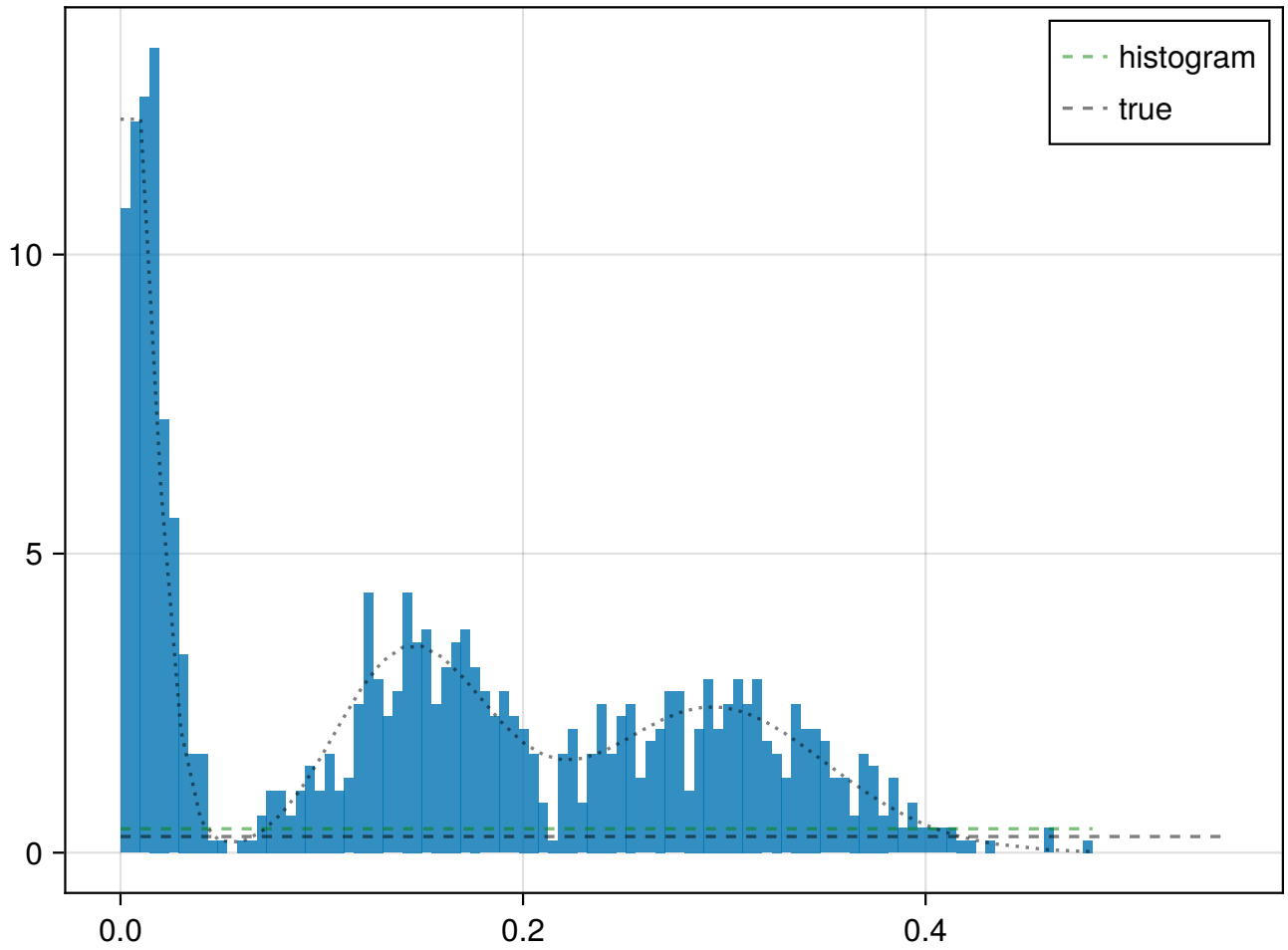
$n = 1000$ yields to 23 bins.

Histogram — number of bins

Too few bins might oversmooth.

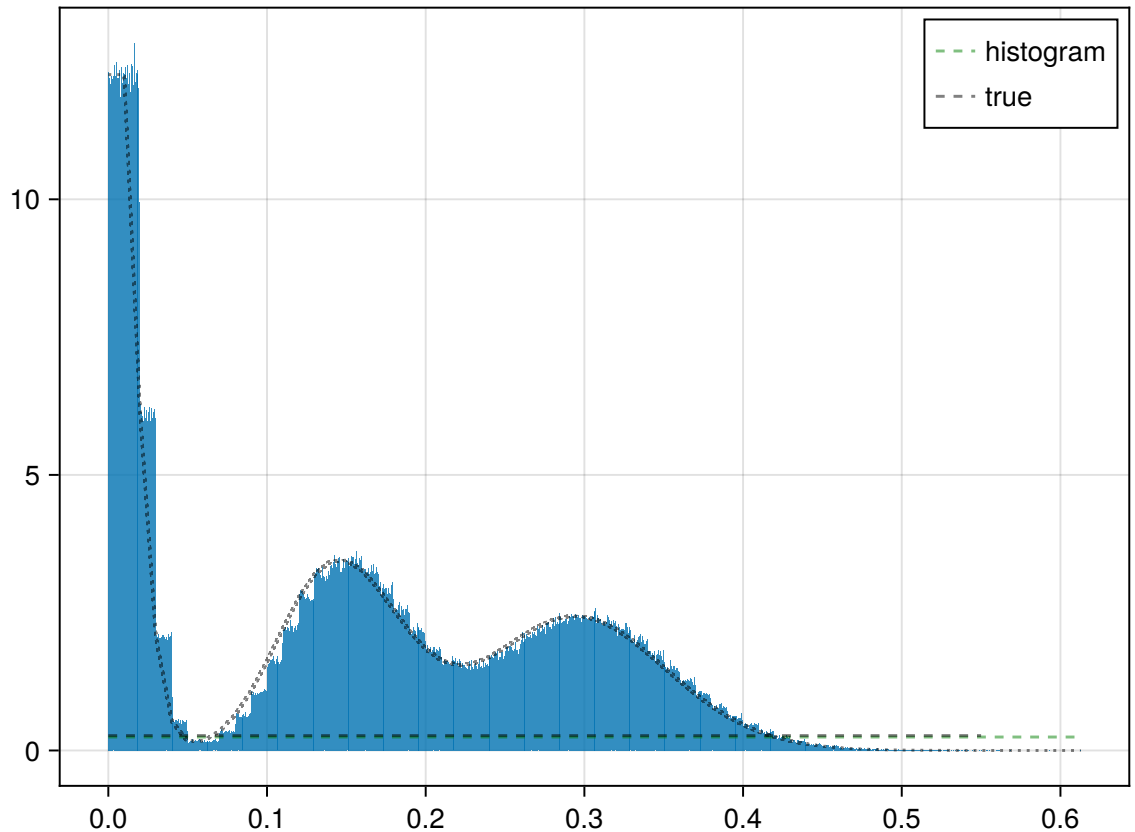


Histogram — too many bins might overfit.



Histogram — number of bins

With a lot of data,
it works great.
100 000 data
points, 1000 bins.



Kernel density estimate / Parzen window estimate

The idea is that each data point defines a probability distribution, which results in a mixture.

$$p(x) = \frac{1}{N} \sum_{i=1}^{n} h(x - x_i|\sigma),$$

where h is a some probability distribution function.

For example Gaussian distribution $h(x|\sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{x^2}{2\sigma^2}}$).

The issue is computational complexity and setting bandwidth σ.

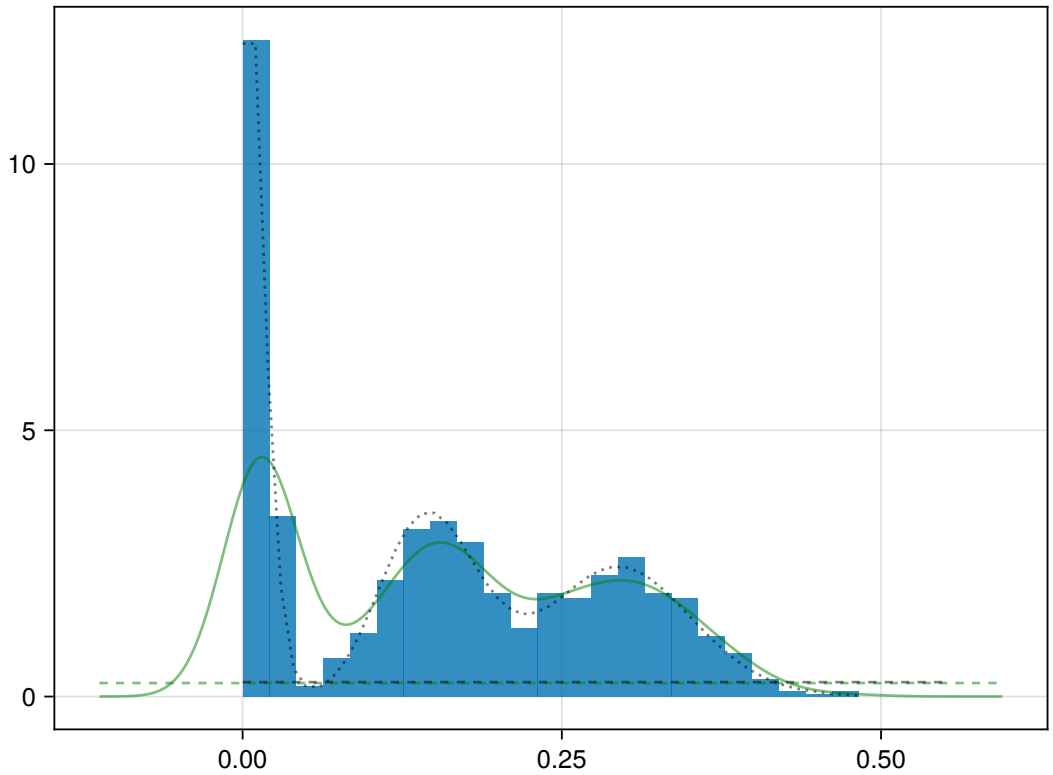
Kernel density estimate / Parzen window estimate

Silverman's rule set

$$\sigma = 0.9 * \hat{\sigma} * n^{-0.2}$$

where

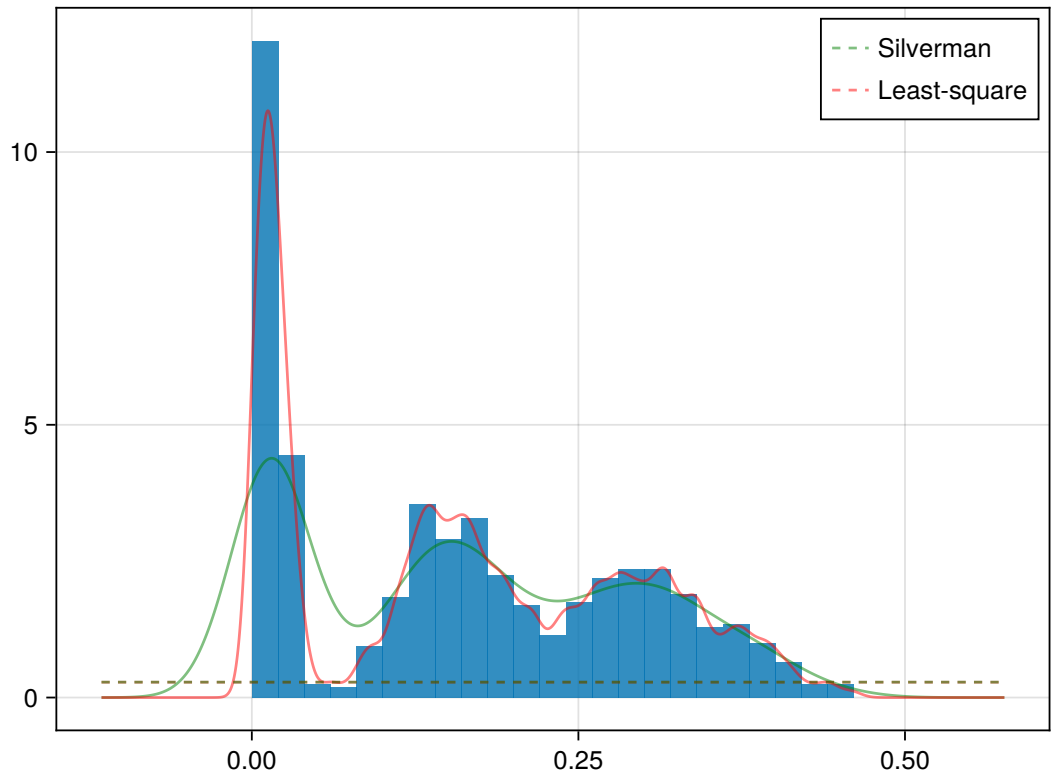
$$\hat{\sigma} = \min\left(\text{std}(x), \frac{\text{IQR}(x)}{1.34}\right)$$



B. W. Silverman, "Algorithm AS 176: Kernel Density Estimation Using the Fast Fourier Transform", 1982

Kernel density estimate / Parzen window estimate

KDE is a probability density function, set threholds as for histograms.



Summary

- Anomaly detection in 1D of unimodal data is "simple".
- Multi-modal data are difficult, requiring more flexible models.
- We need to know something about the data to get good results.
- Flexible models have hyper-parameters, which needs to be set.
- Setting thresholds is principled for models providing pdf.

Plan

Options

Methods based on

- likelihood,
- distance,
- classification,
- pseudo-likelihood,
- miscelaneous.

Parzen window estimator

Works the same as in 1D, but we need to modify the kernel to work in higher dimensions.

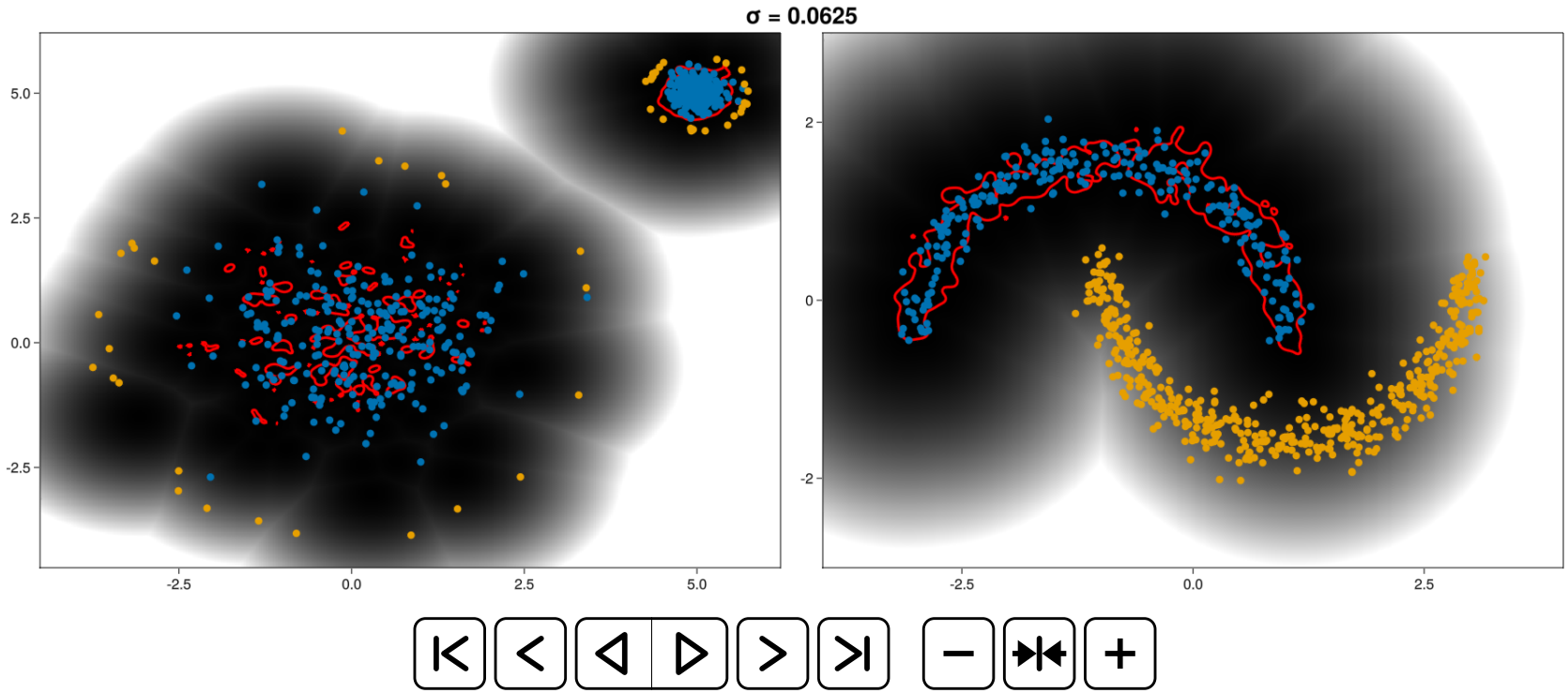
$$p(x) = \frac{1}{n} \sum_{i=1}^{n} h(x - x_i | \sigma),$$

where h is some probability distribution function e.g. Normal.

The performance depends on the choice of kernel and bandwidth (might be anizotropic).

E. Parzen, On Estimation of a Probability Density Function and Mode, 1962

Parzen window estimator — example



Mixture models

The mixture model is defined as

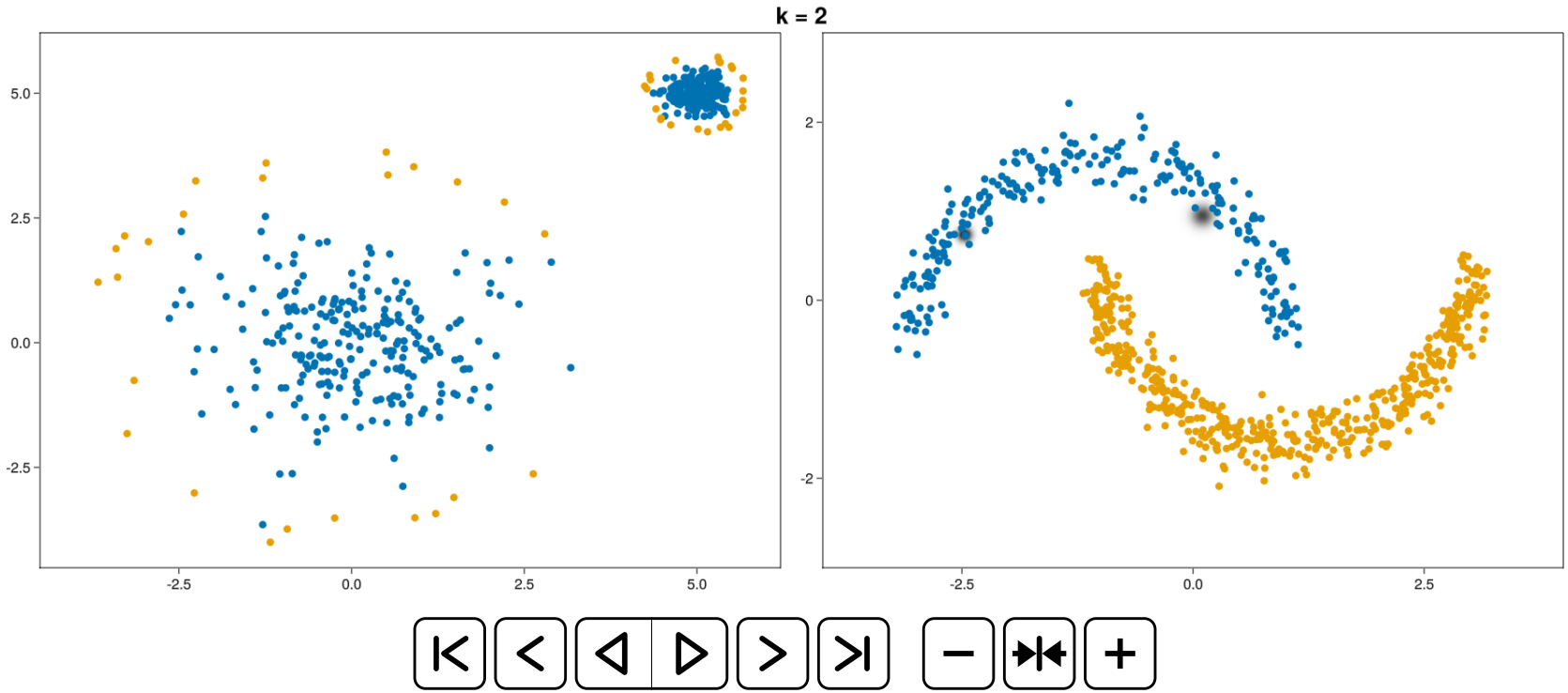
$$p(x|\theta, w) = \sum_{j=1}^{m} w_j p(x|\theta_j),$$

where

- $w_j \geq 0, \sum_{j=1}^{m} w_j = 1$ are weights,
- and $p(x|\theta_j)$ are simple distributions (Normal, Categorical).

The performance depends on the number of components, choice of distribution, and quality of fit (initialization).

Mixture of multivariate Gaussian distributions



Flow models

- Fits a samples to a normal distribution transformed by a bijection
- $p(x) = |f^{-1}(x)| p_z(f^{-1}(x))$
- Masked autoregressive models, flow models
- Real NVP

$$
\begin{aligned}
y_{1:d} &= x_{1:d} \\
y_{d+1:D} &= x_{d+1:D} \odot \exp(s(x_{1:d})) + t(x_{1:d})
\end{aligned}
$$

---

https://lilianweng.github.io/posts/2018-10-13-flow-models/

Example of normalizing flows

Inference

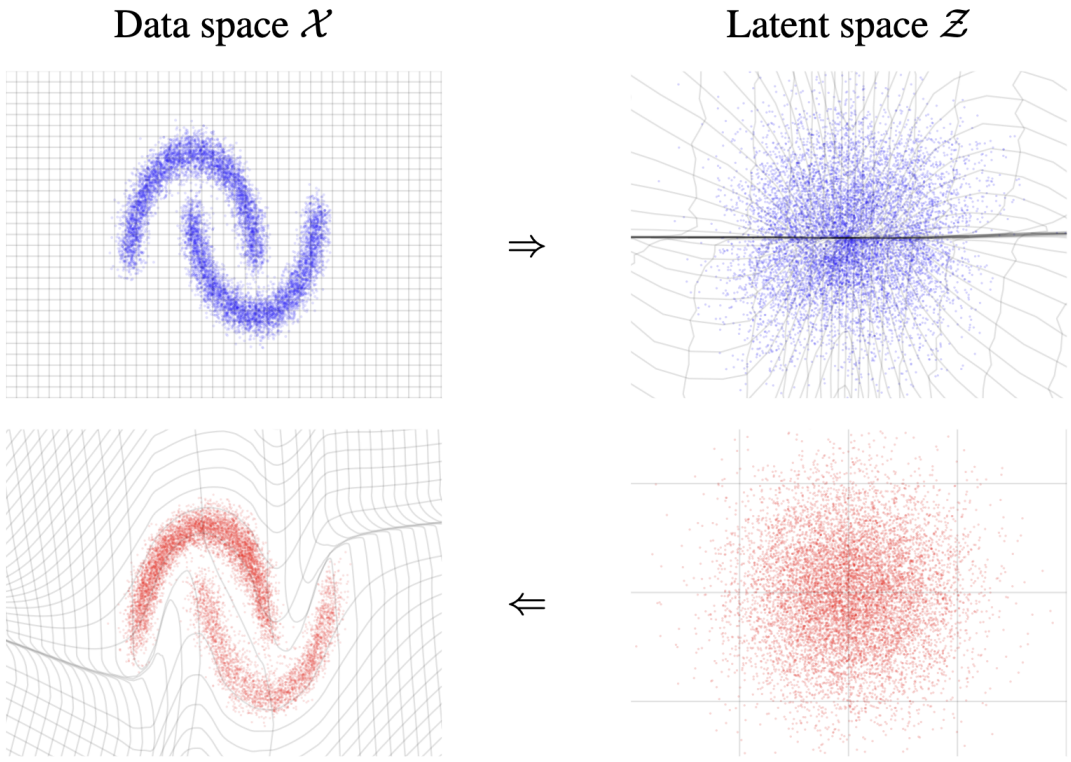
$x \sim \hat{p}_X$

$z = f(x)$

Generation

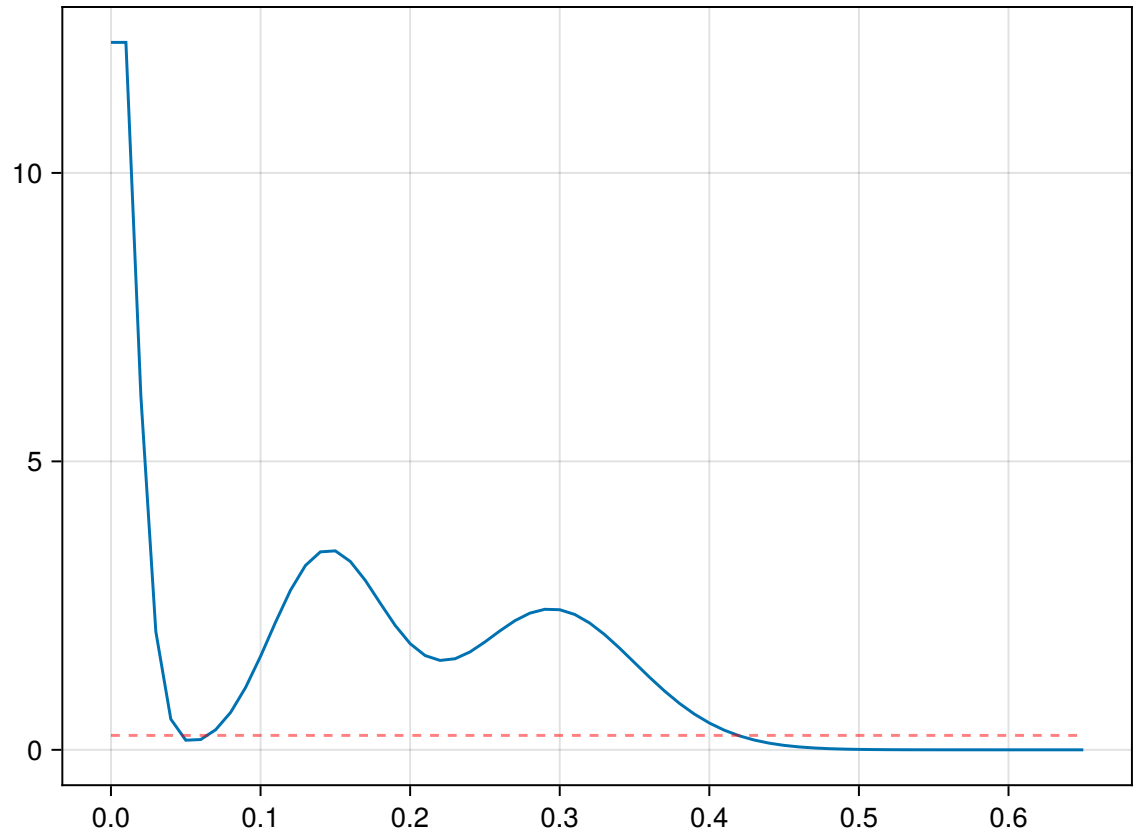
$z \sim p_Z$

$x = f^{-1}(z)$



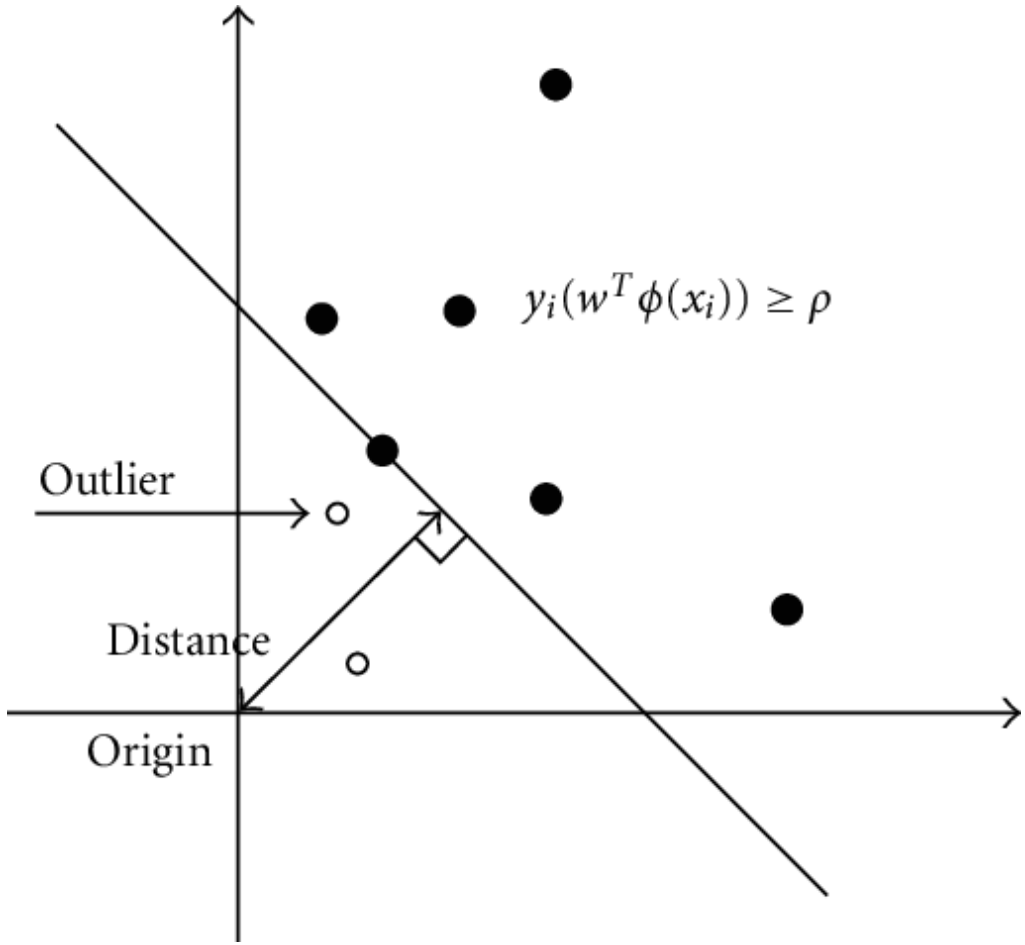
Plan

Do do we have to model density?



Do not model distribution, focus on decision boundary.

One-class support vector machines



Finds the hyper-plane separating the data from the origin with the highest margin, allowing at most ν misclassified points.

One-class support vector machines

training:

$$\arg \min_{w \in \mathbb{R}^n, \rho} \frac{1}{2} \sum_{i,j=1}^{n,n} \alpha_i \alpha_j k(x_i, x_j) - \rho + \frac{1}{\nu N} \sum_{i=1}^{N} \xi_i$$

subject to

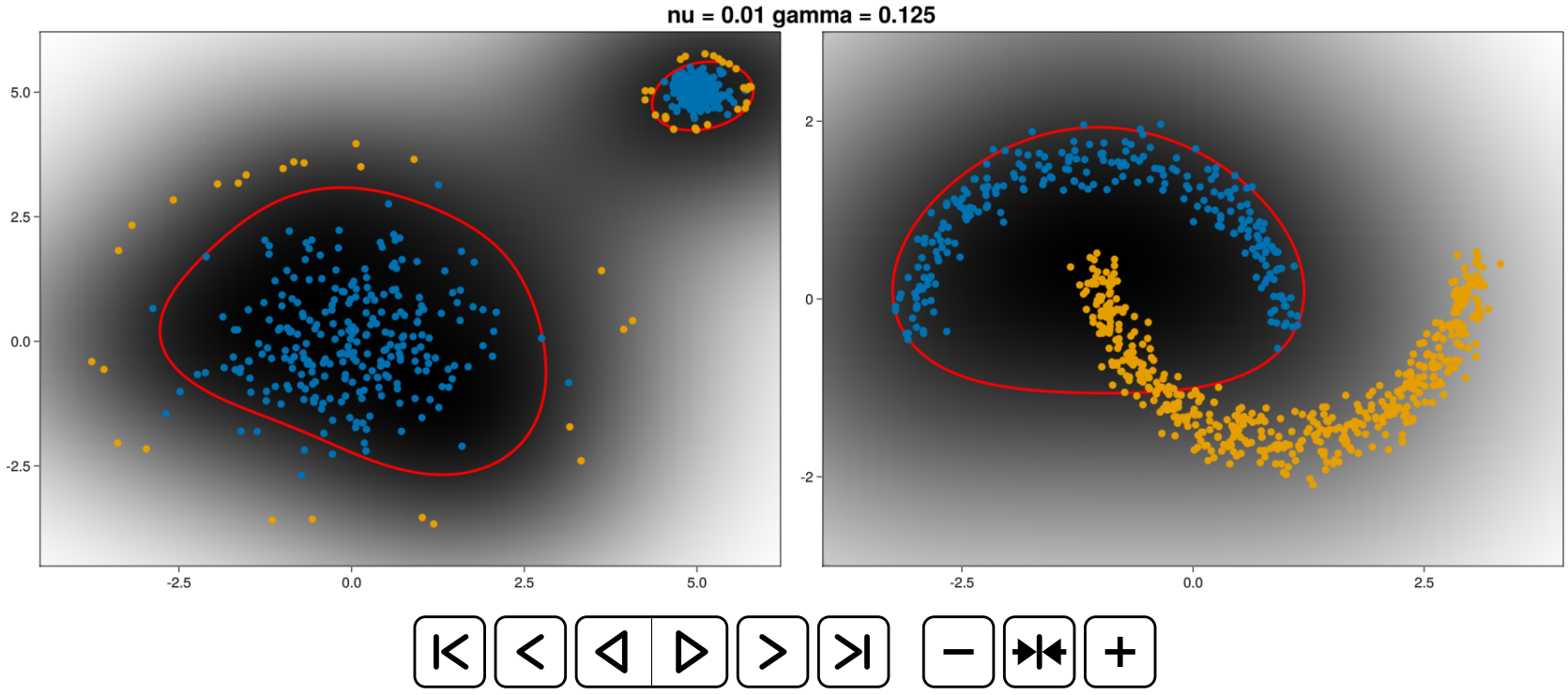
$$\sum_{j=1}^{n} \alpha_i k(x_j, x_i) \geq \rho - \xi_i, \quad \xi_i \geq 0.$$

prediction:

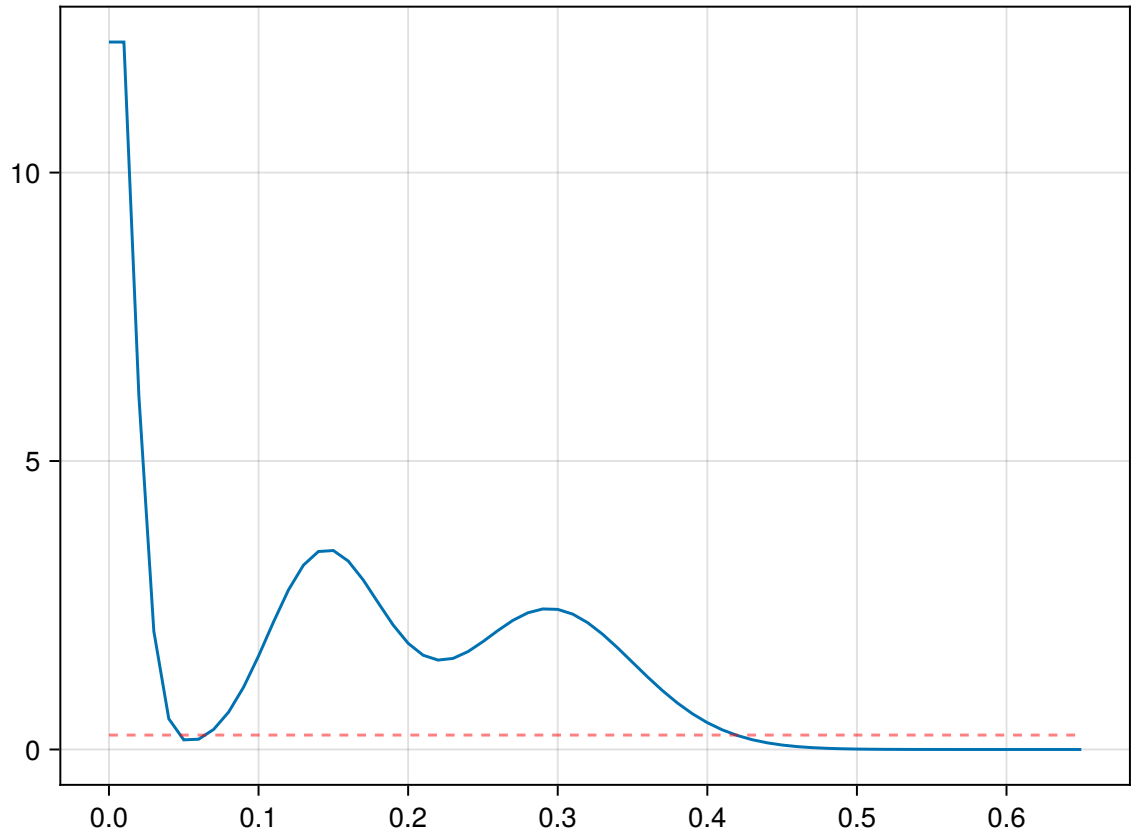
$$f(x) = \sum_{i=1}^{N} \alpha_j k(x_i, x) - \rho > 0$$

$k(\cdot, \cdot)$ is a kernel function, e.g. $k(x, x') = e^{-\gamma \|x - x'\|^2}$.

One-class support vector machines

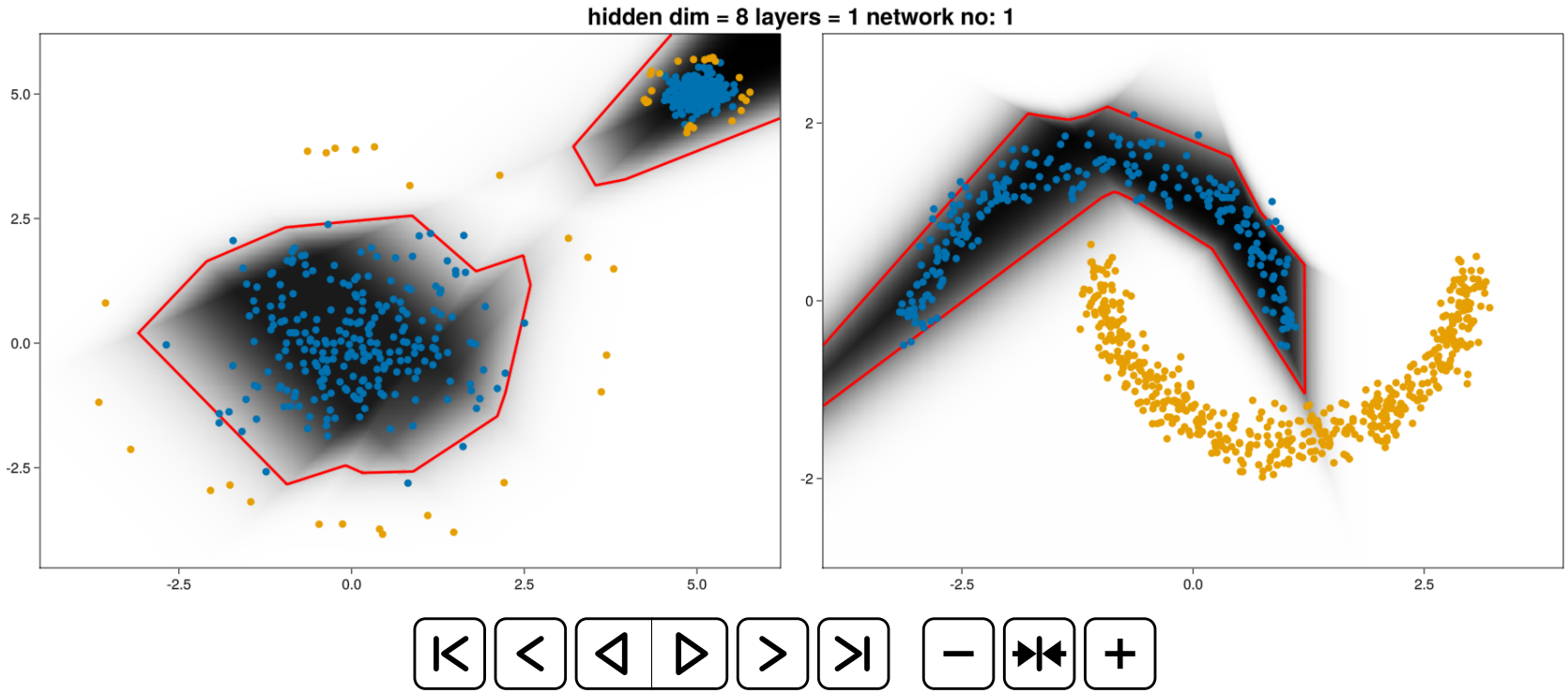


Density detection as classification



Classify the normal samples against the baseline measure (noise).

Density detection as classification



Plan

K-nearest neighbor — motivation

Outliers are far from points / they have "empty" neighbourhood.

S. Ramaswamy, R. Rastogi, K. Shim, Efficient algorithms for mining outliers from large data sets, 2000

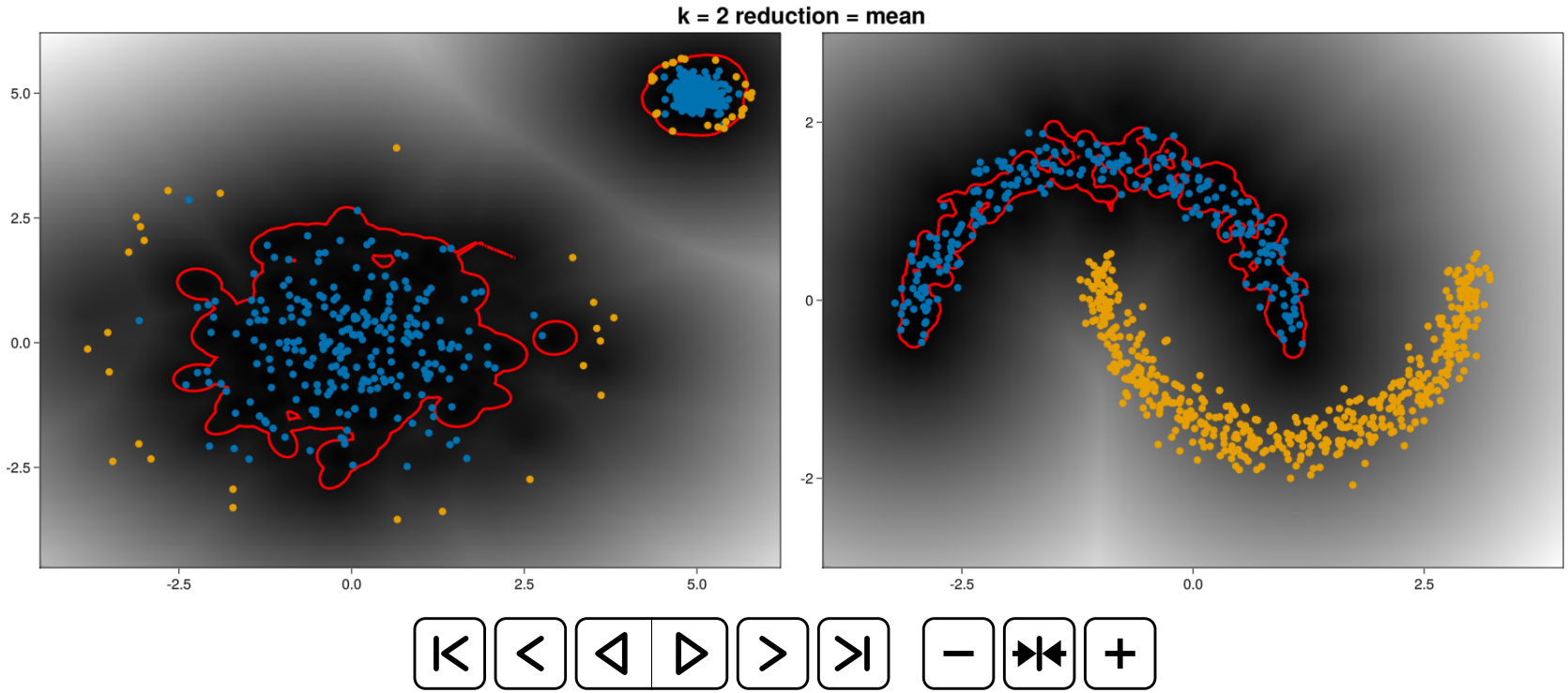
K-nearest neighbor — calculation

1. For sample $\{x_i\}_{i=1}^N$ calculate its distance to k^{th} nearest neighbor.
2. Return fraction p of samples as outliers.

The performance depends on the choice of distance (e.g L_p), izotropiness of the space, choice of k.

S. Ramaswamy, R. Rastogi, K. Shim, Efficient algorithms for mining outliers from large data sets, 2000

K-nearest neighbor — example



S. Ramaswamy, R. Rastogi, K. Shim, Efficient algorithms for mining outliers from large data sets, 2000

Local outlier factor — motivation

Outliers have low density with respect to its k neighborhood.

M. M. Breunig, H.-P. Kriegel, R. T. Ng, and J. Sander, Lof: Identifying density-based local outliers, 2000.

Local outlier factor — calculation

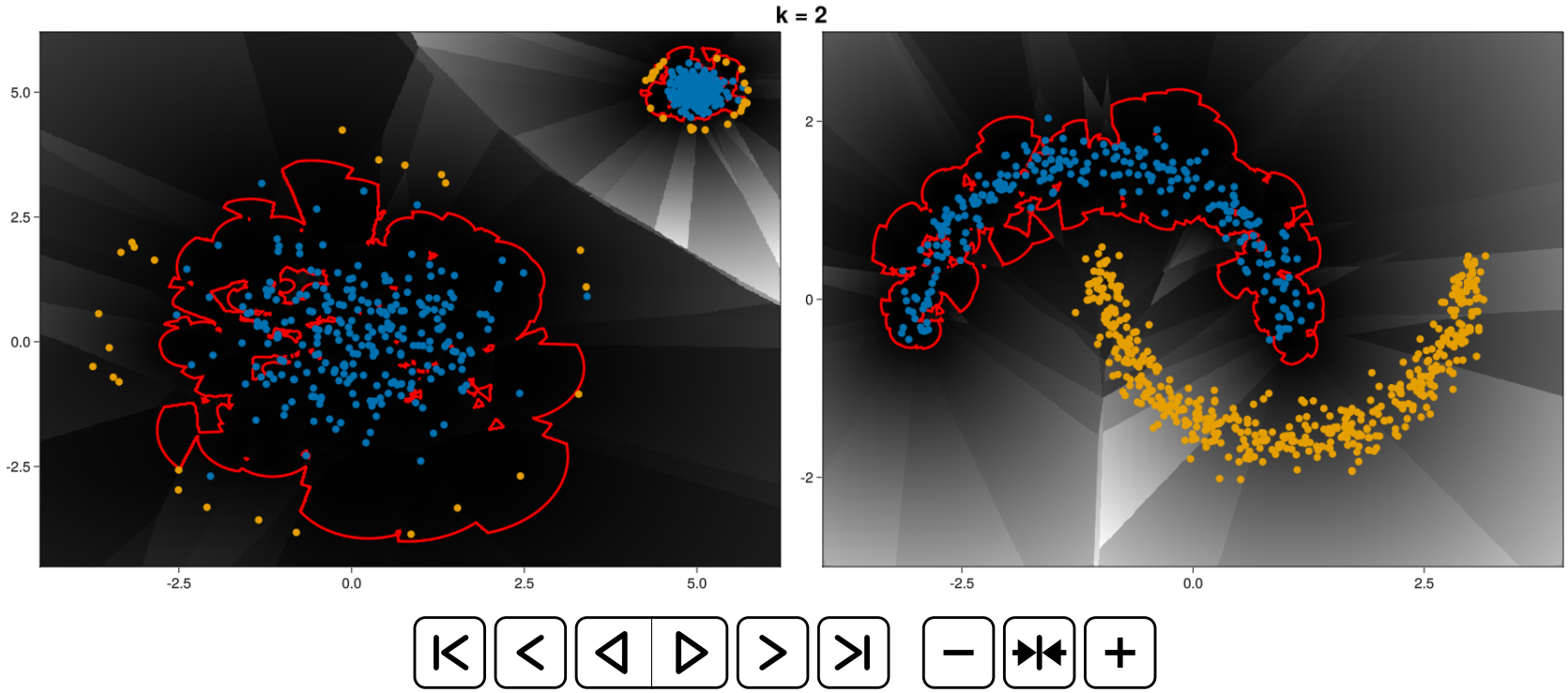
1. For every $\{x_i\}_{i=1}^N$ estimate the local density, $\mathrm{ld}_k(x_i)$, as an inverse of a distance to k nearest neighbor.
2. Compare local density of x with its k nearest neighbors $\mathscr{N}_k(x)$,

$$\mathrm{lof}(x) = \frac{1}{k} \sum_{x_i \in \mathscr{N}_k(x)} \frac{\mathrm{ld}_k(x_i)}{\mathrm{ld}_k(x)}.$$

3. $\mathrm{lof}(x) \geq 2$ indicates outlier

M. M. Breunig, H.-P. Kriegel, R. T. Ng, and J. Sander, Lof: Identifying density-based local outliers, 2000.

Local outlier factor — example



M. M. Breunig, H.-P. Kriegel, R. T. Ng, and J. Sander, Lof: Identifying density-based local outliers, 2000.

What is great on distance-based methods?

Distance is defined for many types of spaces, for which might not be easy to defined probability distribution.

- strings (edit distance),
- sets (Jaccard and Haussdorf distance),
- trees (tree edit distance),
- multi-sets (partial wasserstein distance),
- probability distributions (Total-Variation, Wasserstein distance),
- JSONs (this department).

What is bad on distance-based methods?

They are expensive, as they have both

- high computational complexity,
- and high storage complexity.

Plan

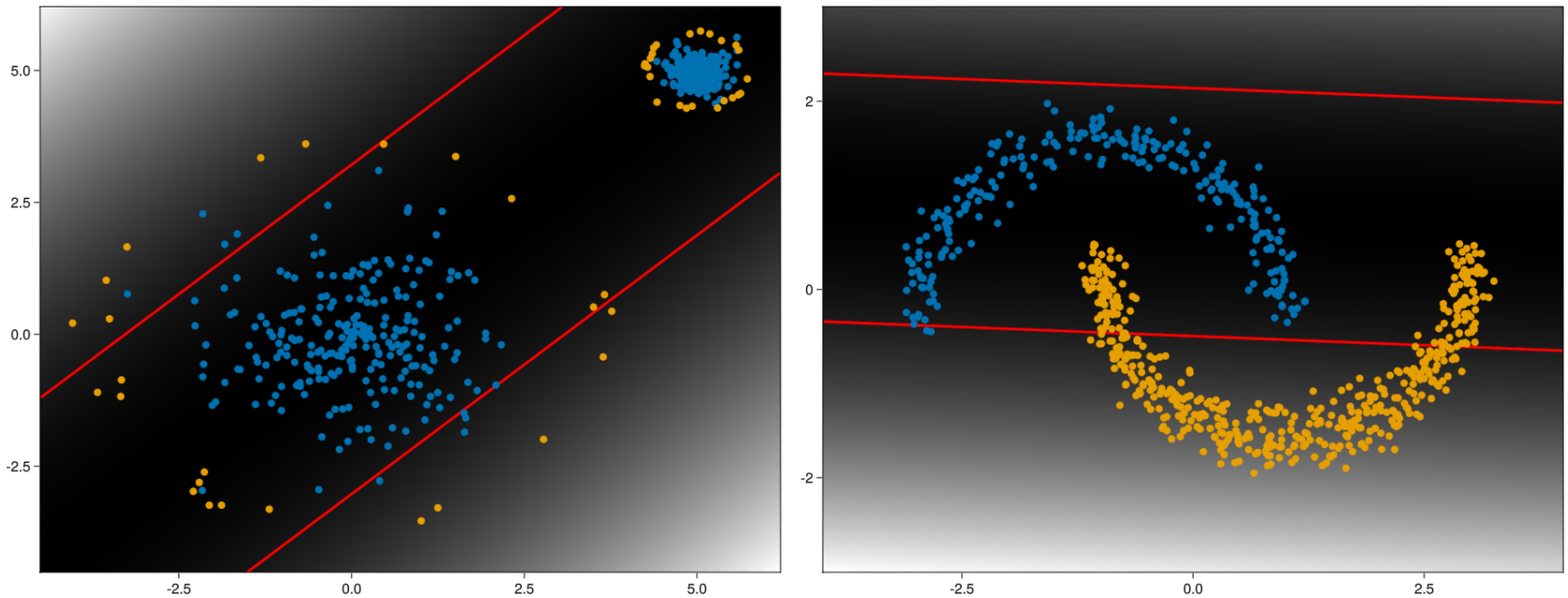
Principal component analysis

- Assumes the data are located on a hyperplane.
- Finds projection P of the data on the hyperplane explaining most variance.
- Computes the reconstruction error as

$$\|x^{\mathrm{T}}\mathrm{P}\mathrm{P}^{\mathrm{T}} - x\|^2$$

- Data points with low reconstruction error are located on the hyperplane and therefore considered normal.
- Data points with high reconstruction error are considered outliers.

Principal component analysis



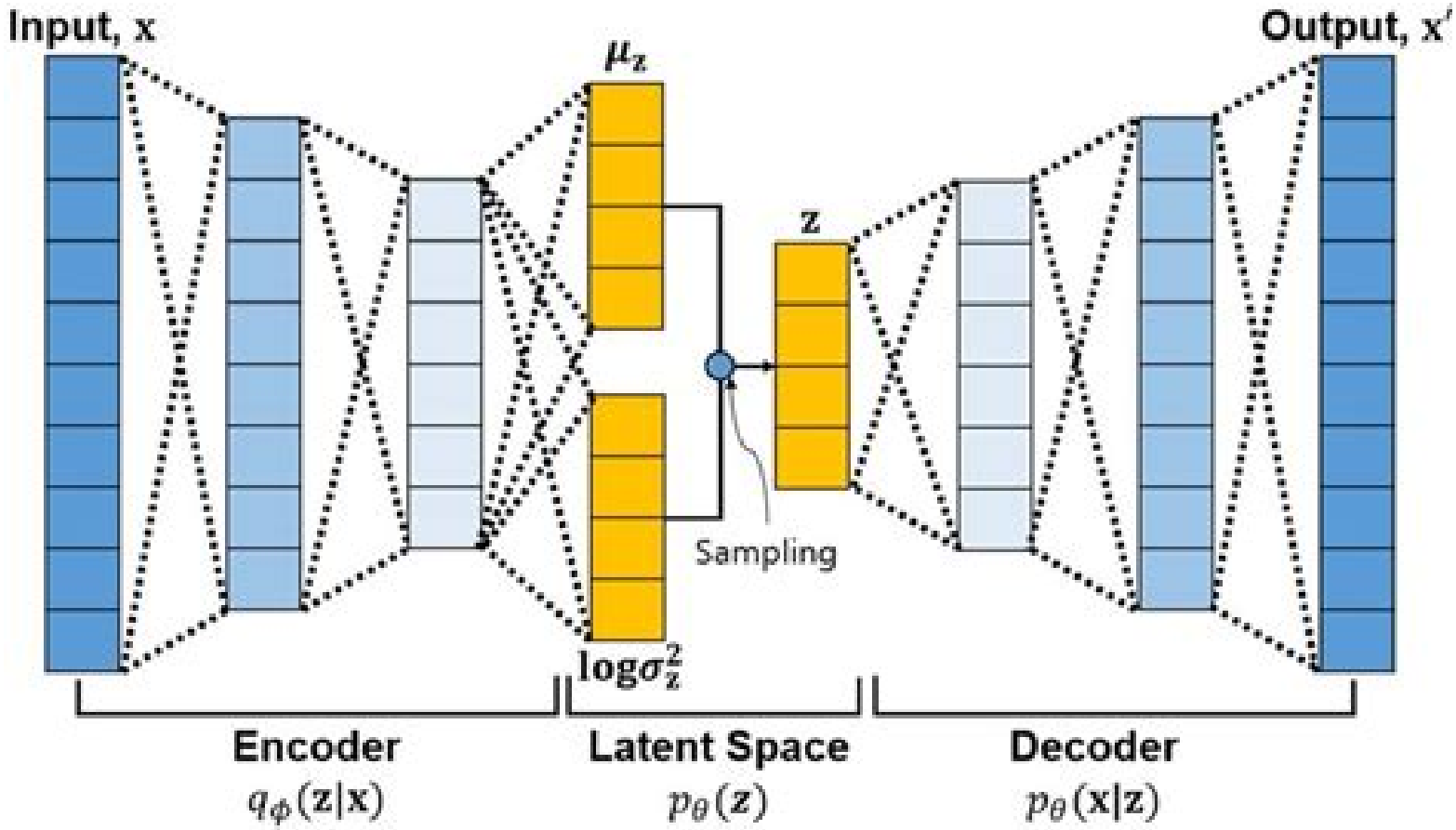
(Variational) autoencoder

- Recall PCA method's reconstruction error

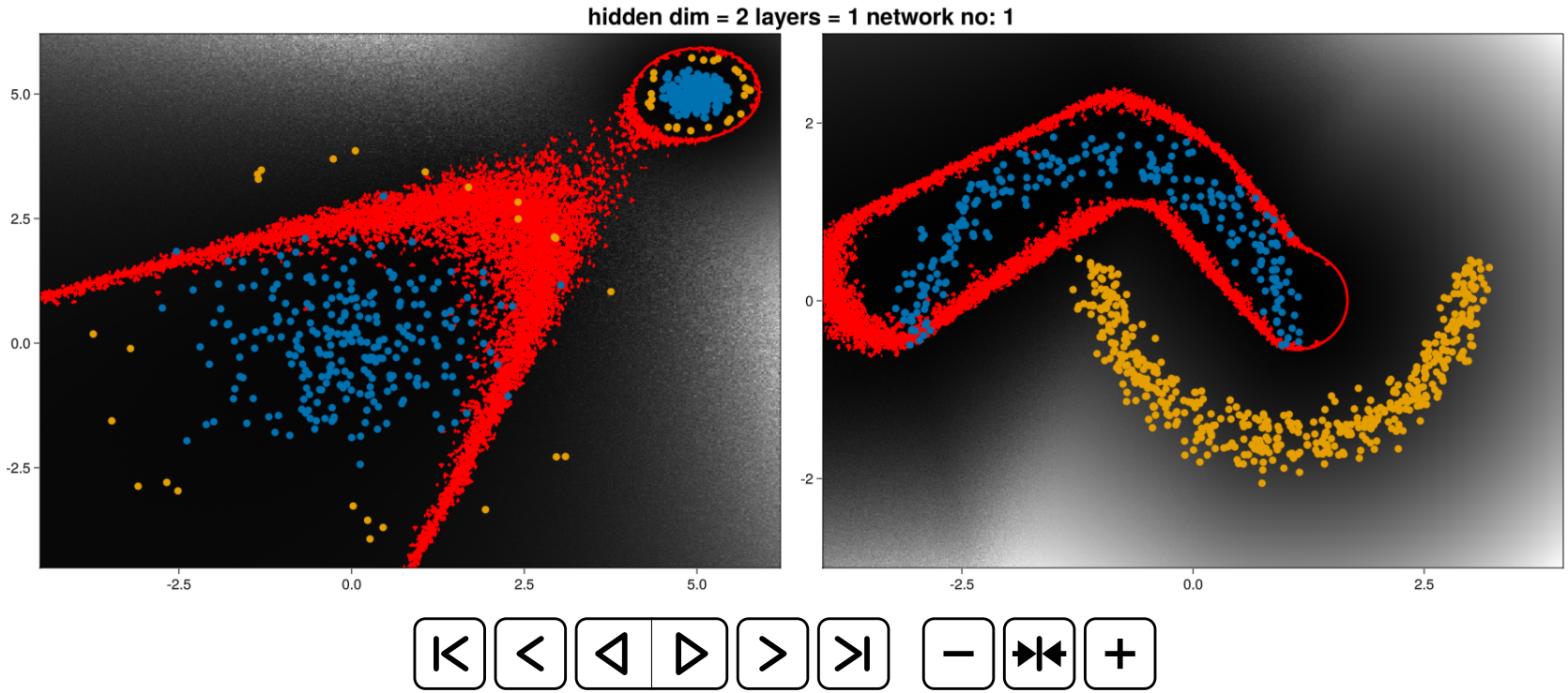
$$\|x^{\mathrm{T}}\mathrm{P}\mathrm{P}^{\mathrm{T}} - x\|^2$$

- View P as an encoder $\mathrm{enc}(x)$ and P^{T} as a decoder $\mathrm{dec}(x)$.
- Then the reconstruction error is $\|dec(enc(x)) - x\|^2$.
- $\mathrm{enc}(x)$ and $\mathrm{dec}(x)$ are some functions (neural networks)
- Variational autoencoder adds regularization on latent $D_{\mathrm{KL}}(enc(x)\|N(0, I))$

Variation autoencoder



Variational autoencoder



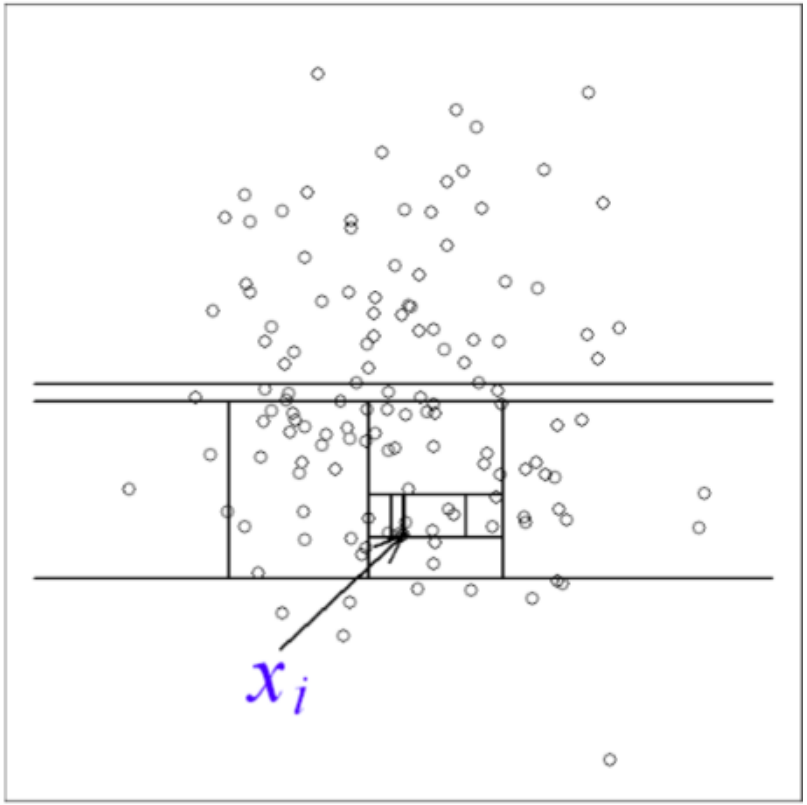
Plan

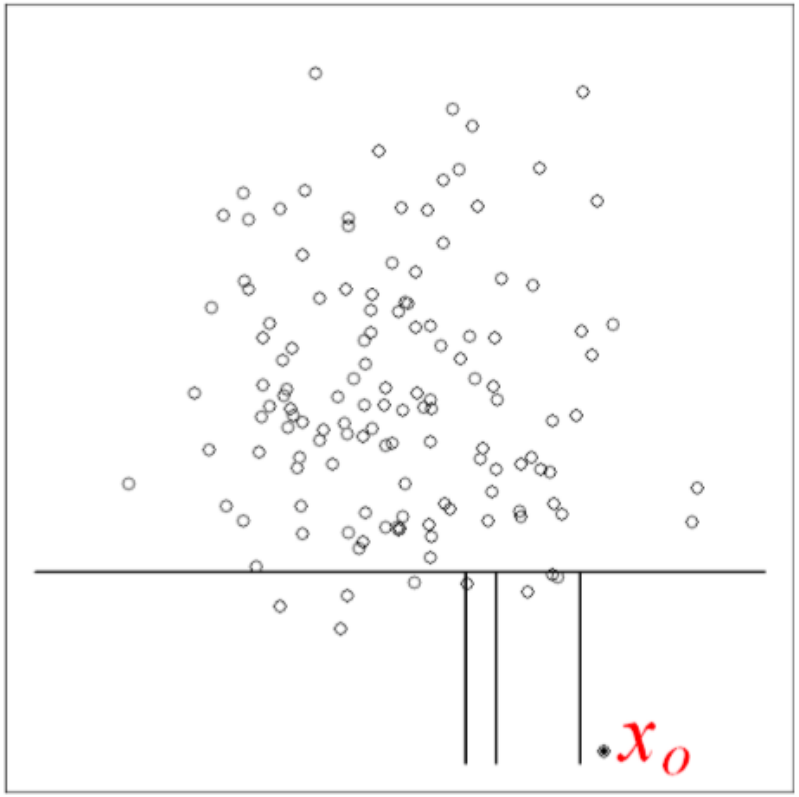
Isolation Forest — motivation

- Create a random tree by randomly selecting feature and splitting point, but alway separate at least one data-point.
- Anomalous points are close to the root of the tree.

F. T. Liu, K. M. Ting, Z. H. Zhou, Isolation Forest, 2008

Isolation Forest — Example





Isolation Forest — calculation

The anomaly score a sample x is defined as

$$s(x) = 2^{-\frac{E(h(x))}{c(n)}},$$

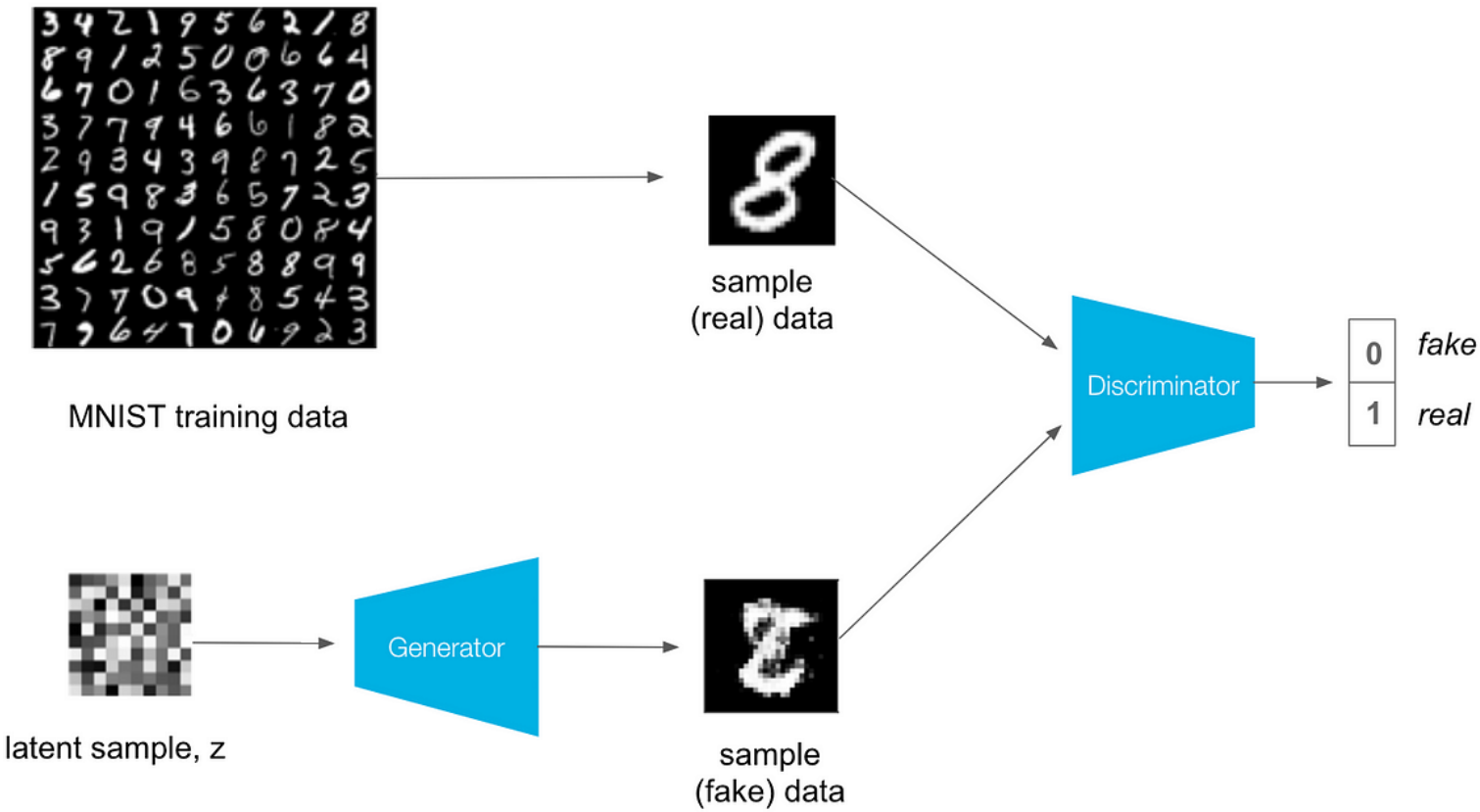
where

- $h(x)$ is depth of list containing x
- $c(n)$ is the average path length of unsuccessful search in binary search tree with n items

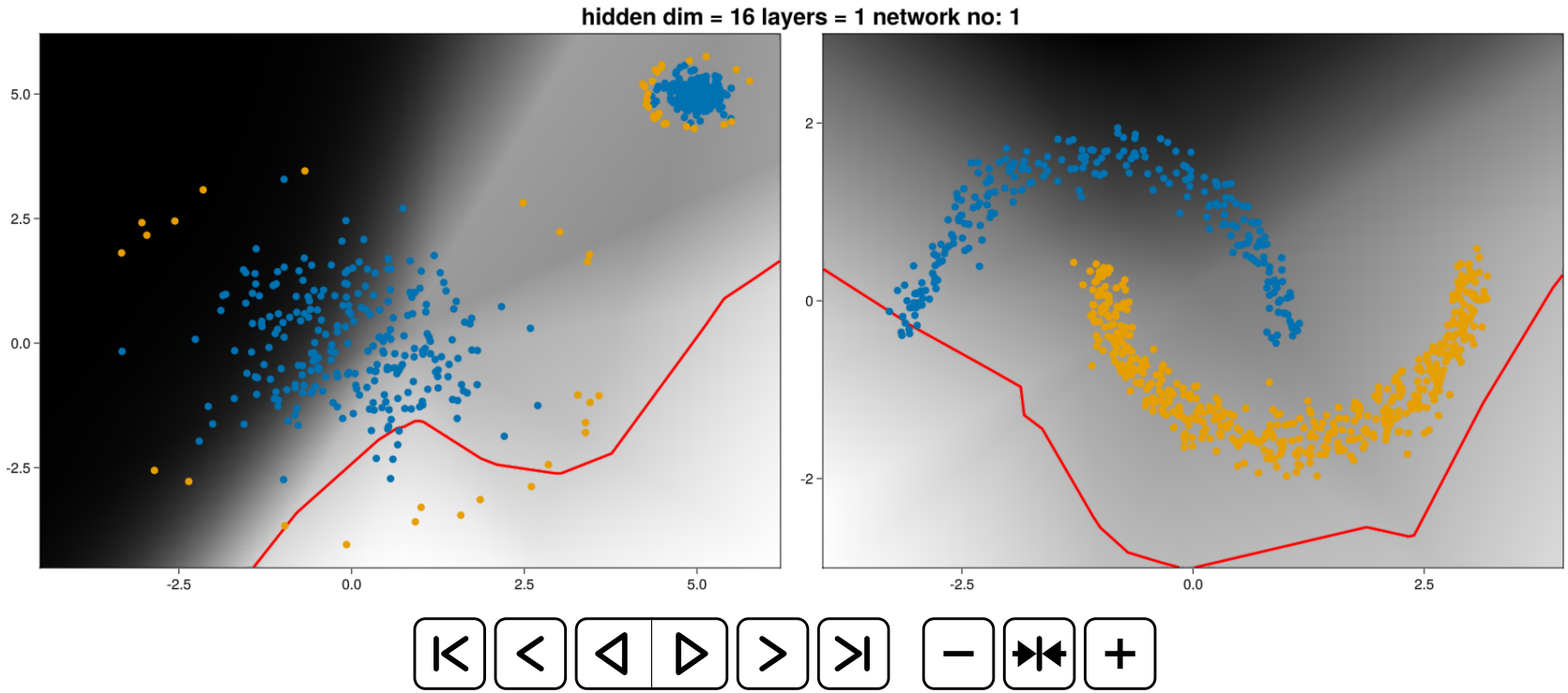
$$c(n) = 2H(n-1) - 2\frac{n-1}{n}$$

- $H(i) \approx ln(i) + 0.5772156649$

Generative adversarial networks



Generative adversarial networks



Plan

Why we have so many methods?

dataset	aae	avae	gano	vae	wae	abod	hbos	if	knn	loda	lof	orbf	osvm	pidf	maf	rnvp	sptn	fmgn	gan	mgal	dsvd	vaek	vaeo
aba	0.91	0.87	0.89	0.92	0.91	**0.93**	0.75	0.87	**0.93**	0.84	0.90	**0.93**	**0.93**	0.89	0.91	0.90	0.91	0.78	0.80	0.62	0.82	0.91	0.90
ann	0.83	0.83	0.80	0.84	0.86	0.78	0.89	0.78	0.78	0.69	0.80	0.77	**0.99**	0.93	0.85	0.86	0.87	0.81	0.74	0.65	0.65	0.81	0.81
arr	0.75	0.76	0.77	0.73	0.75	0.74	0.77	0.78	0.74	0.77	0.73	0.75	**0.81**	0.75	0.76	0.77	0.74	0.74	0.73	0.55	0.72	0.71	0.79
bcw	0.99	0.95	0.94	0.98	0.99	0.94	0.97	0.97	0.93	0.94	0.94	0.93	0.99	0.91	0.99	0.98	0.95	**1.00**	0.99	0.64	0.83	0.93	0.99
blt	0.94	0.84	0.90	0.95	0.94	0.87	0.88	0.90	0.89	0.82	0.91	0.94	0.89	0.87	**0.96**	0.93	0.94	0.71	0.58	0.82	0.94	0.93	0.94
bts	0.98	0.85	0.99	0.99	0.97	0.99	0.98	0.98	0.98	0.95	0.96	0.98	**1.00**	0.77	0.98	0.99	0.99	0.96	0.95	0.68	0.96	0.97	0.97
crd	0.65	0.61	0.69	0.62	0.62	0.56	0.50	0.69	0.61	0.74	0.67	0.69	**0.90**	0.64	0.60	0.51	0.50	0.67	0.66	0.72	0.86	0.63	0.83
eco	**0.90**	0.86	0.84	0.86	0.85	0.87	0.81	0.83	0.88	0.77	0.80	0.87	0.89	0.84	**0.90**	0.85	0.88	0.85	0.87	0.58	0.76	0.86	0.85
gls	0.86	0.77	0.77	0.78	**0.87**	0.79	0.62	0.52	0.71	0.51	0.81	0.80	0.78	0.40	0.73	0.74	0.78	0.86	0.84	0.65	0.67	0.73	0.72
hab	0.96	0.97	0.87	0.93	**0.98**	0.95	0.92	0.93	0.95	0.95	0.96	0.95	0.95	0.97	0.96	0.95	0.96	0.76	0.81	0.68	0.96	0.95	0.93
har	0.99	0.69	0.99	**1.00**	**1.00**	0.76	0.84	0.71	0.76	0.81	0.97	0.79	**1.00**	0.47	**1.00**	**1.00**	0.96	**1.00**	0.99	0.58	0.84	0.48	**1.00**
htr	0.96	0.93	0.94	0.96	0.95	0.95	0.96	0.95	0.95	0.95	0.95	0.95	**0.97**	0.94	0.95	0.95	0.95	0.96	0.96	0.61	0.92	0.95	0.96
ion	0.97	0.98	0.98	0.97	0.97	0.98	0.78	0.92	0.98	0.87	0.96	0.98	0.98	0.90	0.98	**0.99**	0.97	0.90	0.80	0.68	0.97	0.96	0.97
irs	0.88	0.83	0.97	0.96	0.92	0.97	0.99	0.89	0.94	**1.00**	0.88	0.92	0.93	0.99	0.79	0.80	0.93	**1.00**	0.99	0.73	0.29	0.88	0.92
iso	0.74	0.70	0.79	0.76	0.77	0.64	0.55	0.60	0.77	0.55	0.82	0.76	**0.84**	0.60	0.71	0.70	0.60	0.78	0.78	0.50	0.62	0.81	0.81
kdd	0.96	0.95	0.99	**1.00**	**1.00**	0.99	0.99	**1.00**	**1.00**	0.90	0.98	**1.00**	**1.00**	0.99	0.99	**1.00**	0.99	**1.00**	**1.00**	0.02	0.24	0.99	0.99
lbr	0.71	0.64	0.74	0.64	0.75	0.65	0.58	0.55	**0.78**	0.56	0.70	**0.78**	**0.78**	0.55	0.73	0.77	0.55	0.77	**0.78**	0.51	0.58	**0.78**	**0.78**
ltr	0.78	0.78	0.77	0.80	0.80	0.68	0.56	0.62	0.80	0.59	**0.83**	0.80	0.81	0.60	0.76	0.75	0.67	0.76	0.74	0.48	0.65	0.82	0.82
mam	0.88	0.89	0.89	0.88	0.88	0.85	0.84	0.88	0.88	0.89	0.85	0.89	**0.91**	0.86	0.87	0.89	0.88	0.78	0.82	0.75	**0.91**	0.90	0.90
mgc	0.94	0.91	0.89	**0.97**	0.95	0.94	0.83	0.90	0.94	0.82	0.93	0.94	0.94	0.91	0.96	0.96	0.96	0.85	0.84	0.55	0.81	0.89	0.90
mlt	**0.99**	0.98	0.98	**0.99**	**0.99**	0.91	0.73	0.87	0.98	0.74	0.98	0.98	**0.99**	0.83	0.98	**0.99**	0.94	**0.99**	**0.99**	0.47	0.74	**0.99**	**0.99**
mnb	0.89	0.90	0.88	0.89	0.91	0.81	0.91	0.81	0.86	0.92	0.70	0.87	**0.94**	0.82	0.90	0.90	0.86	0.73	0.83	0.67	0.85	0.85	0.88
pen	0.97	0.95	0.98	0.99	0.99	0.99	0.77	0.96	0.99	0.90	**1.00**	0.99	0.99	0.95	0.98	0.98	0.99	0.96	0.92	0.59	0.86	0.98	0.99
pgb	0.98	0.98	0.98	0.98	0.98	0.97	0.88	0.97	0.98	0.96	0.98	0.98	0.98	0.96	**0.99**	**0.99**	0.98	0.75	0.73	0.59	0.97	**0.99**	**0.99**
pim	0.85	0.78	0.81	0.85	0.85	0.83	0.81	0.83	0.84	0.81	0.82	0.84	**0.89**	0.78	0.86	0.85	0.84	0.78	0.81	0.61	0.81	0.78	0.78
prk	0.76	0.60	0.73	0.72	0.81	0.75	0.55	0.66	0.80	0.55	0.70	0.74	**0.88**	0.45	0.72	0.71	0.74	0.78	0.79	0.64	0.72	0.79	0.80
sat	0.98	0.87	0.96	0.92	0.94	0.96	0.95	0.94	0.97	0.90	0.98	0.96	**0.99**	0.95	0.91	0.93	0.84	0.97	0.97	0.74	0.82	0.95	0.97
scc	0.96	0.96	0.89	0.98	0.98	0.90	0.82	0.92	0.97	0.86	**0.99**	0.98	0.98	**0.99**	**0.99**	0.96	0.90	0.96	0.97	0.59	0.87	0.96	0.97
seg	0.92	0.91	0.94	0.93	0.92	0.95	0.86	0.90	**0.96**	0.93	0.94	0.95	0.95	0.93	0.92	0.92	0.93	0.89	0.89	0.60	0.72	0.94	0.95
sei	0.72	0.73	0.74	0.72	0.72	0.74	0.73	0.70	0.74	0.70	0.65	0.72	**0.77**	0.74	0.73	0.73	0.74	0.68	0.71	0.56	0.74	0.73	0.73
sht	0.94	0.99	0.99	0.99	0.99	**1.00**	0.93	0.98	**1.00**	0.90	**1.00**	**1.00**	**1.00**	0.99	**1.00**	0.99	**1.00**	0.85	0.87	0.65	0.93	**1.00**	**1.00**
snr	0.67	0.65	0.76	0.65	0.69	0.64	0.49	0.55	0.64	0.52	0.85	0.64	0.84	0.50	0.65	0.66	0.58	0.81	0.81	0.57	0.47	0.74	**0.86**
sph	0.37	0.35	0.52	0.27	0.52	0.35	0.30	0.35	0.50	0.47	0.40	0.47	**0.82**	0.28	0.26	0.30	0.28	0.74	0.80	0.55	0.50	0.50	0.80
spm	0.76	0.80	0.78	0.85	0.87	0.77	0.82	0.82	0.78	0.63	0.81	0.77	**0.94**	0.84	0.86	0.85	0.83	0.91	0.91	0.54	0.60	0.54	0.81
vhc	0.73	0.65	0.73	0.73	0.73	0.74	**0.77**	0.69	0.73	0.70	0.60	0.72	0.72	0.70	0.75	**0.77**	0.74	0.65	0.70	0.57	0.70	0.63	0.71
wf1	0.75	0.71	0.78	0.80	0.87	0.72	0.87	0.83	0.83	0.81	0.75	0.79	**0.95**	0.85	0.75	0.75	0.77	0.85	0.84	0.63	0.75	0.82	0.92
wf2	0.79	0.73	0.75	0.78	0.84	0.73	0.86	0.84	0.84	0.82	0.76	0.79	**0.94**	0.85	0.74	0.79	0.77	0.87	0.84	0.60	0.75	0.79	**0.94**
wne	**1.00**	0.98	0.96	0.99	0.98	0.95	0.92	0.91	0.98	0.83	0.97	0.98	0.99	0.73	0.98	0.95	0.96	0.97	0.92	0.61	0.94	0.95	0.99
wrb	0.73	0.73	0.81	0.85	0.85	0.80	0.87	0.81	0.82	0.72	0.76	0.82	0.85	**0.91**	0.78	0.82	0.81	0.82	0.78	0.56	0.57	0.80	0.80
yst	0.72	0.70	0.66	0.74	0.73	0.66	0.53	0.63	0.66	0.67	0.68	0.68	**0.75**	0.60	0.72	0.72	0.67	0.62	0.65	0.65	0.70	0.54	0.64
σ_1	0.03	0.05	0.03	0.03	0.02	0.02	0.02	0.02	0.02	0.03	0.02	0.02	0.01	0.02	0.02	0.03	0.02	0.03	0.05	**0.13**	0.04	0.03	0.02
σ_{10}	0.02	**0.06**	0.02	0.01	0.02	0.00	0.01	0.00	0.01	0.01	0.01	0.03	0.01	0.01	0.01	0.01	0.01	0.03	0.03	0.04	0.03	0.01	0.01
rank	8.8	13.4	10.2	8.0	6.3	11.3	14.4	14.2	8.0	15.6	11.3	8.1	**2.8**	14.1	8.8	8.7	10.3	11.3	11.8	21.4	16.0	10.9	6.7

Is anomaly detection useful in practice?

- Pure anomaly detection is a myth.
- You need little bit of labels to set hyper-parameters.
- Basic methods frequently works great.
- You need to manually label data.
- Work in tandem with supervised learning.