Student Data Challenge

Disrupting Healthcare through Machine Learning

Pavlos Protopapas, IACS Scientific Program Director and Lecturer

Kevin Rader, Senior Preceptor in Statistics

Sheila Coveney, IACS Program Manager

Marouan Belhaj, IACS Visiting Researcher

18/19 January 2018

IACS Institute for Applied Computational Science
HARVARD SCHOOL OF ENGINEERING AND APPLIED SCIENCES

1. **Introduction**
2. Data
3. Rules
4. Evaluation
5. Conclusion

Overview

Input: health insurance company, containing information on utilization, payments, and submitted charges organized by doctor.

Task: assign, in an unsupervised fashion, a risk score to each doctor.

Challenge: being able to assign a risk score as high as possible to "malicious" doctors, while keeping the risk score of genuine doctors as low as possible.

1. Introduction

2. Data

3. Rules

4. Evaluation

5. Conclusion

Data

Train data: https://goo.gl/wNCWi7

Data

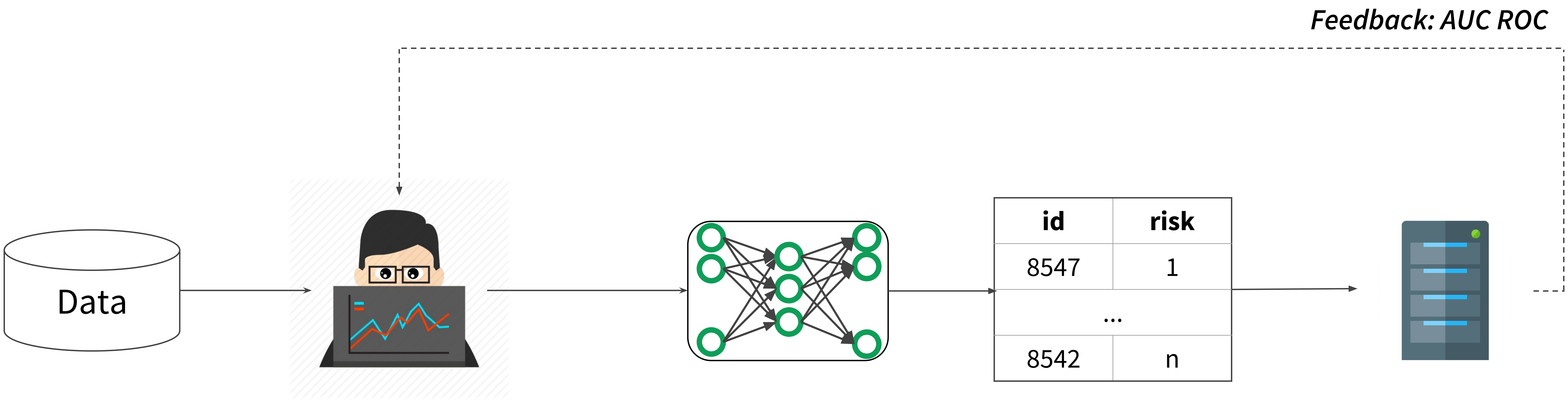
Doctor Identifier	Unique identifier
Provider Type	Anesthesiology, Neurology
Number of Services	Total provider services
Number of Beneficiaries	Total beneficiaries receiving the provider services
Total Submitted Charge Amount	Total charges that the provider submitted for all services
Total Allowed Amount	Allowed amount for all provider services. Sum of the amount the Insurance pays, the deductible and coinsurance amounts that the beneficiary is responsible for paying
Total Payment Amount	Total amount paid after deductible
Total Standardized Payment Amount	Standardization removes geographic differences in payment rates

Data

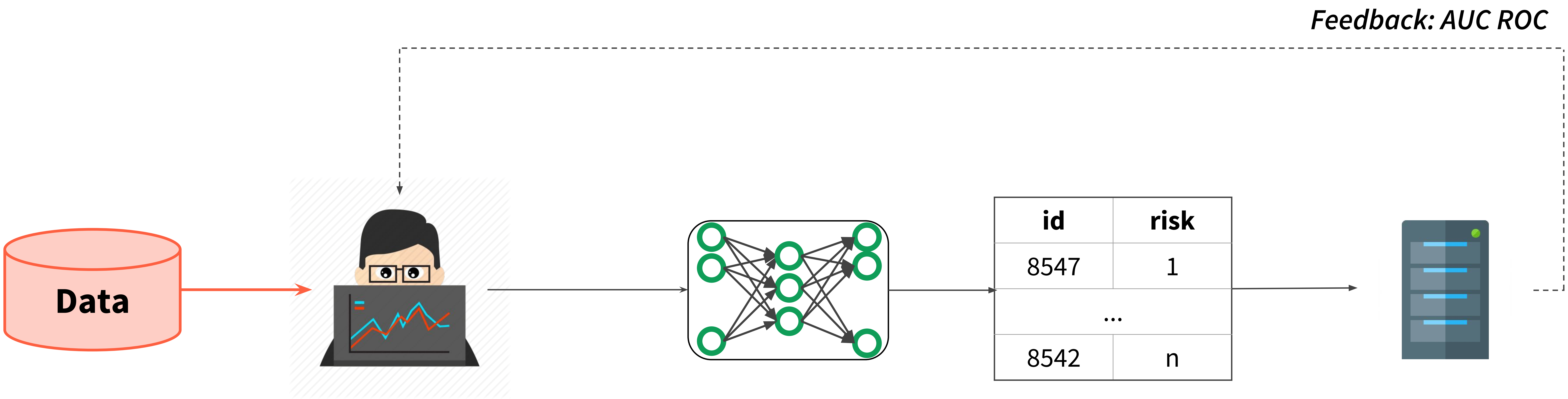
Number of Drug Services	Total drug services
Total Drug Submitted Charge Amount	As for total charges, just for drugs
etc…	Same here
Average HCC Risk Score of Beneficiaries	Average Hierarchical Condition Category (HCC) risk score of beneficiaries
Percent of "X"	Percent of patients with disease "X"

1. Introduction
2. Data
3. **Rules**
4. Evaluation
5. Conclusion

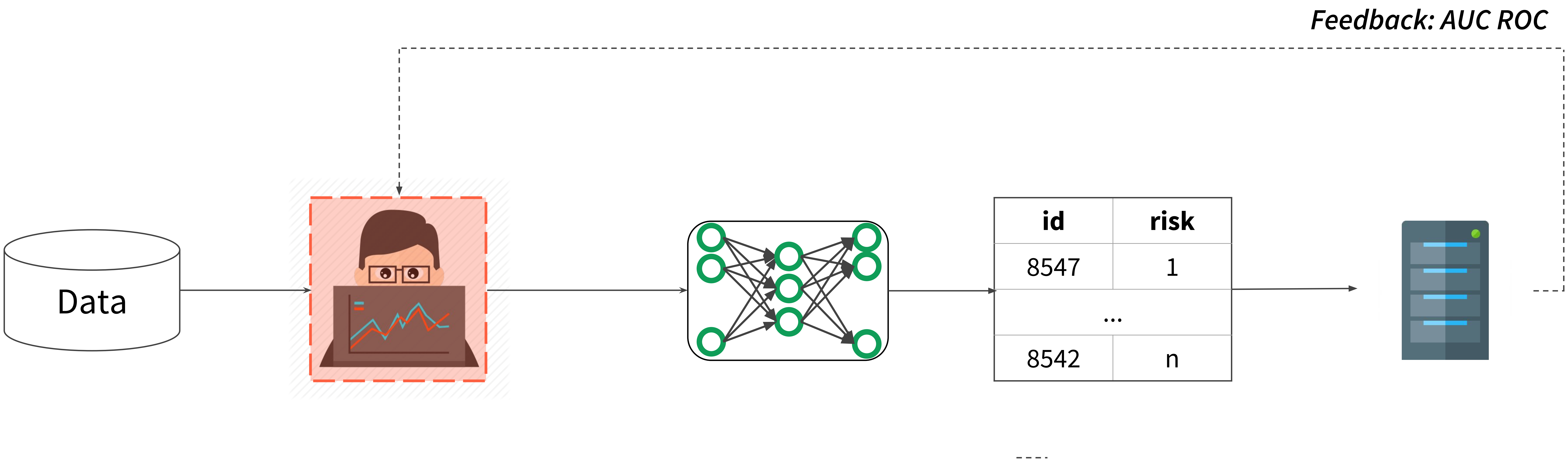
High-level interaction



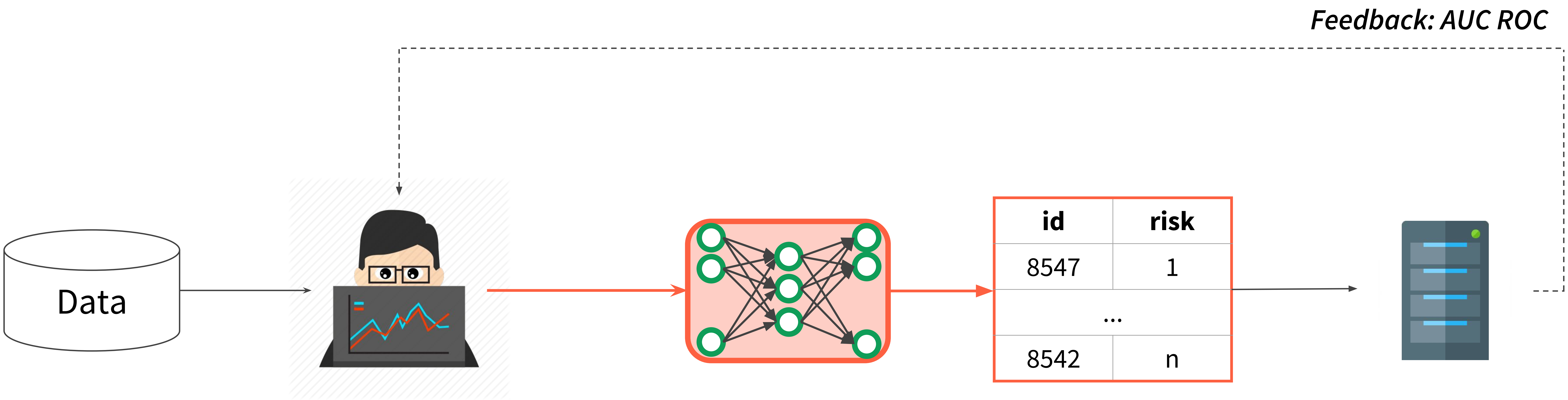
High-level interaction



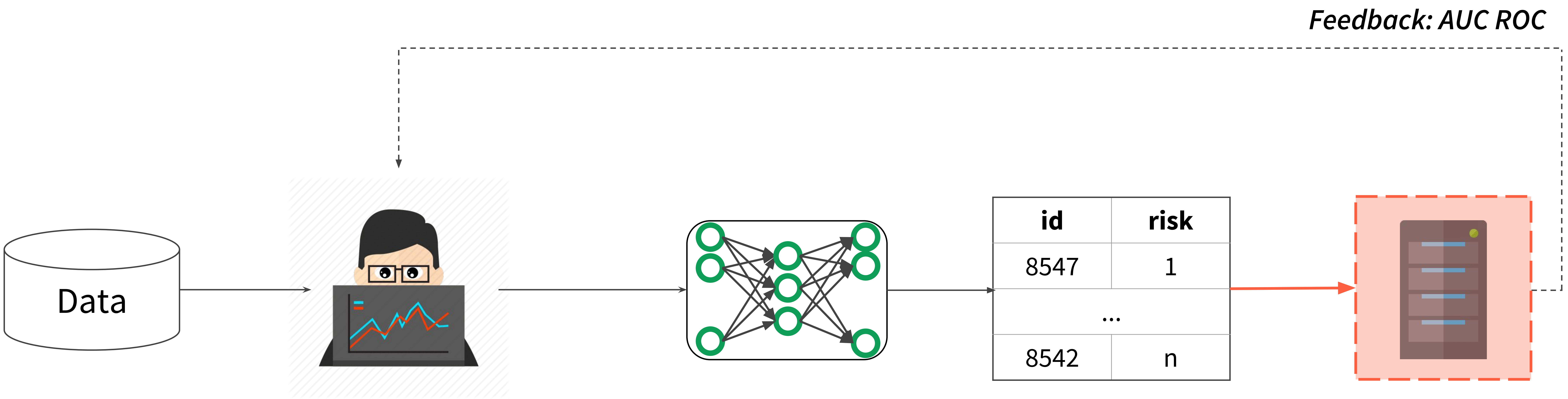
High-level interaction



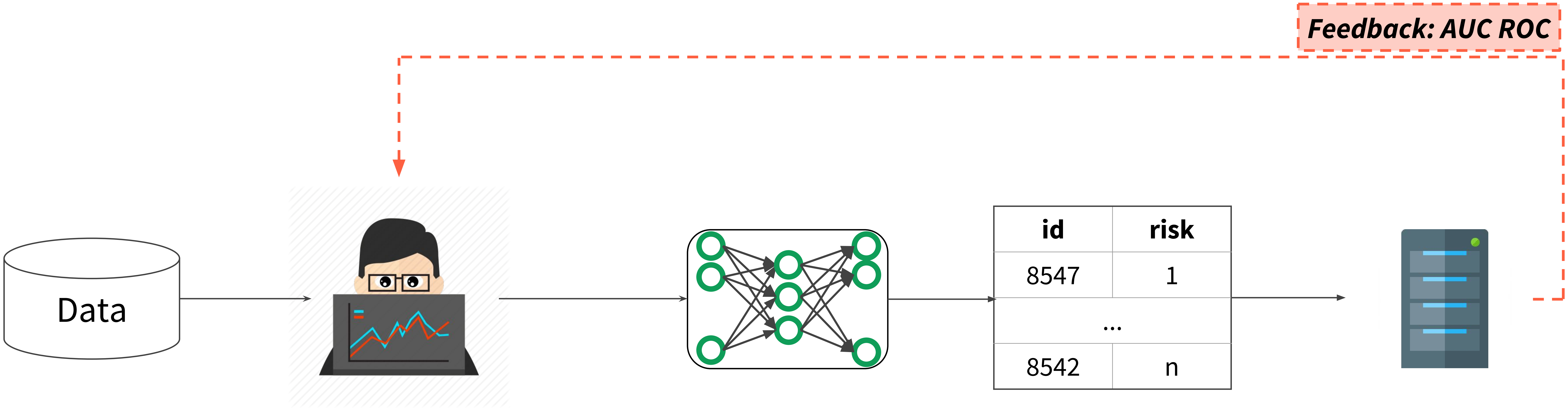
High-level interaction



High-level interaction



High-level interaction



Live

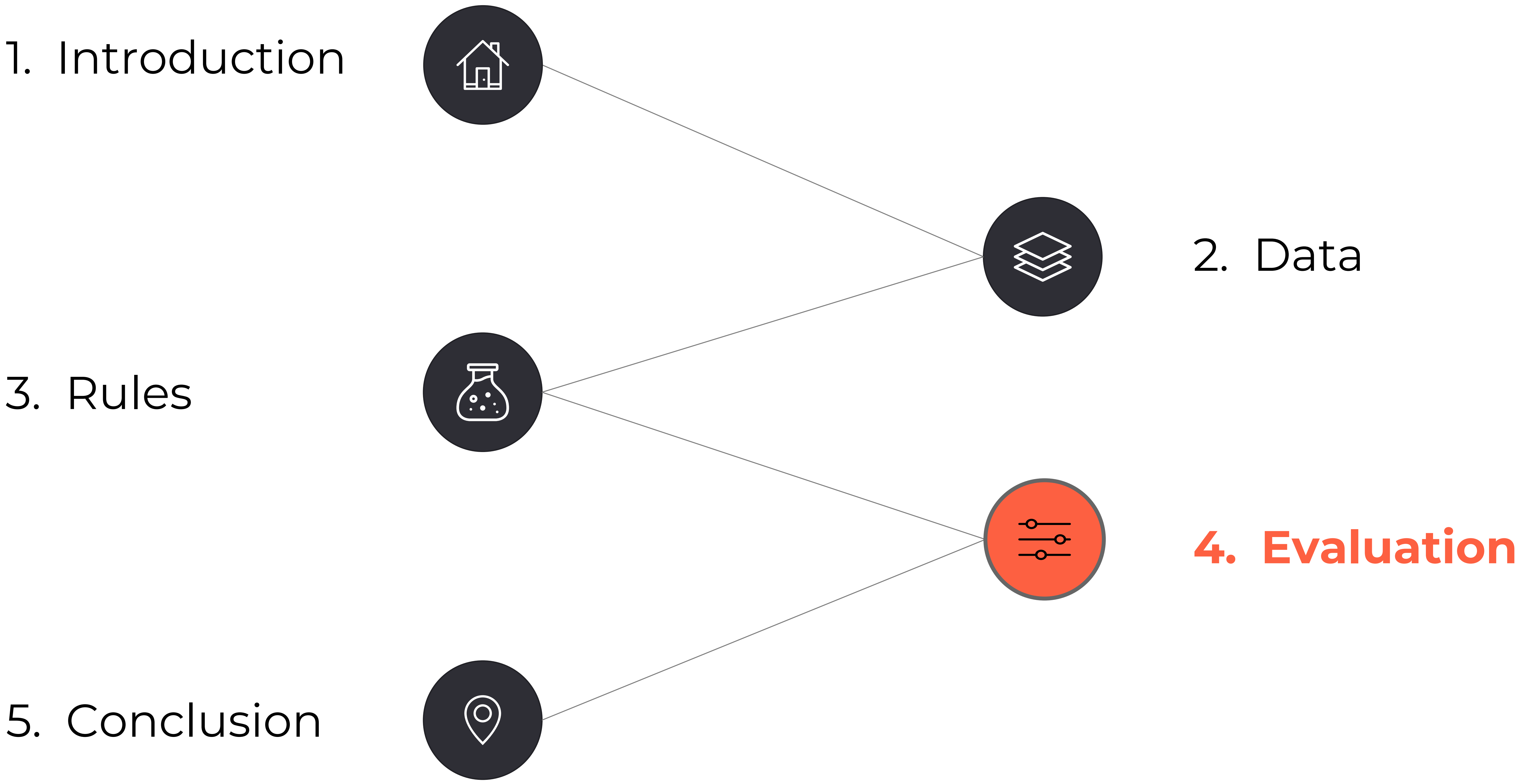
Main rules

One account per team.

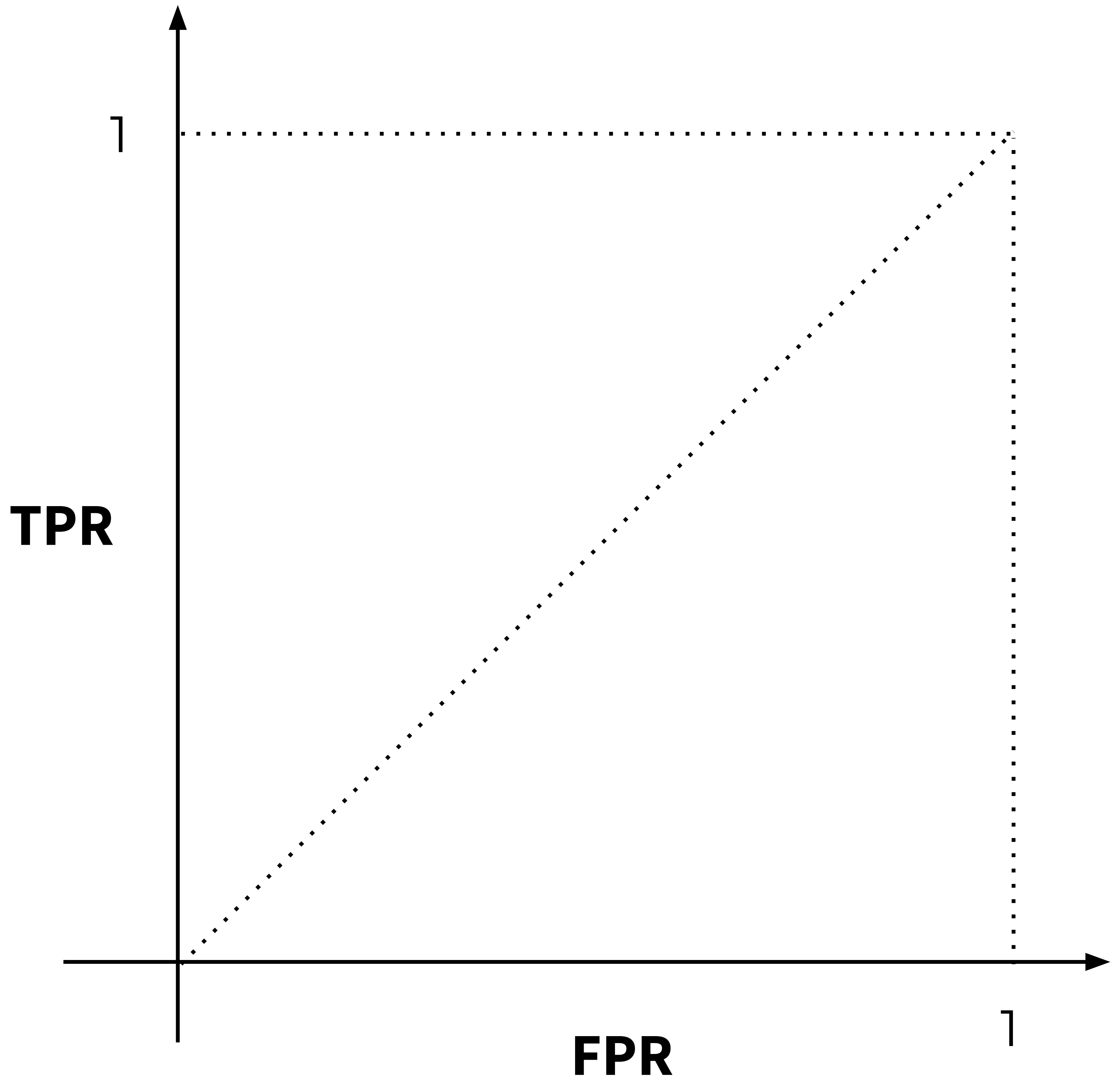
Maximum 3 submissions per hour.

Example submission

```
Doctor Identifier,Risk
53549,-0.128
46612,2543
10648,12
66390,235
53960,99
91381,-87
...
```

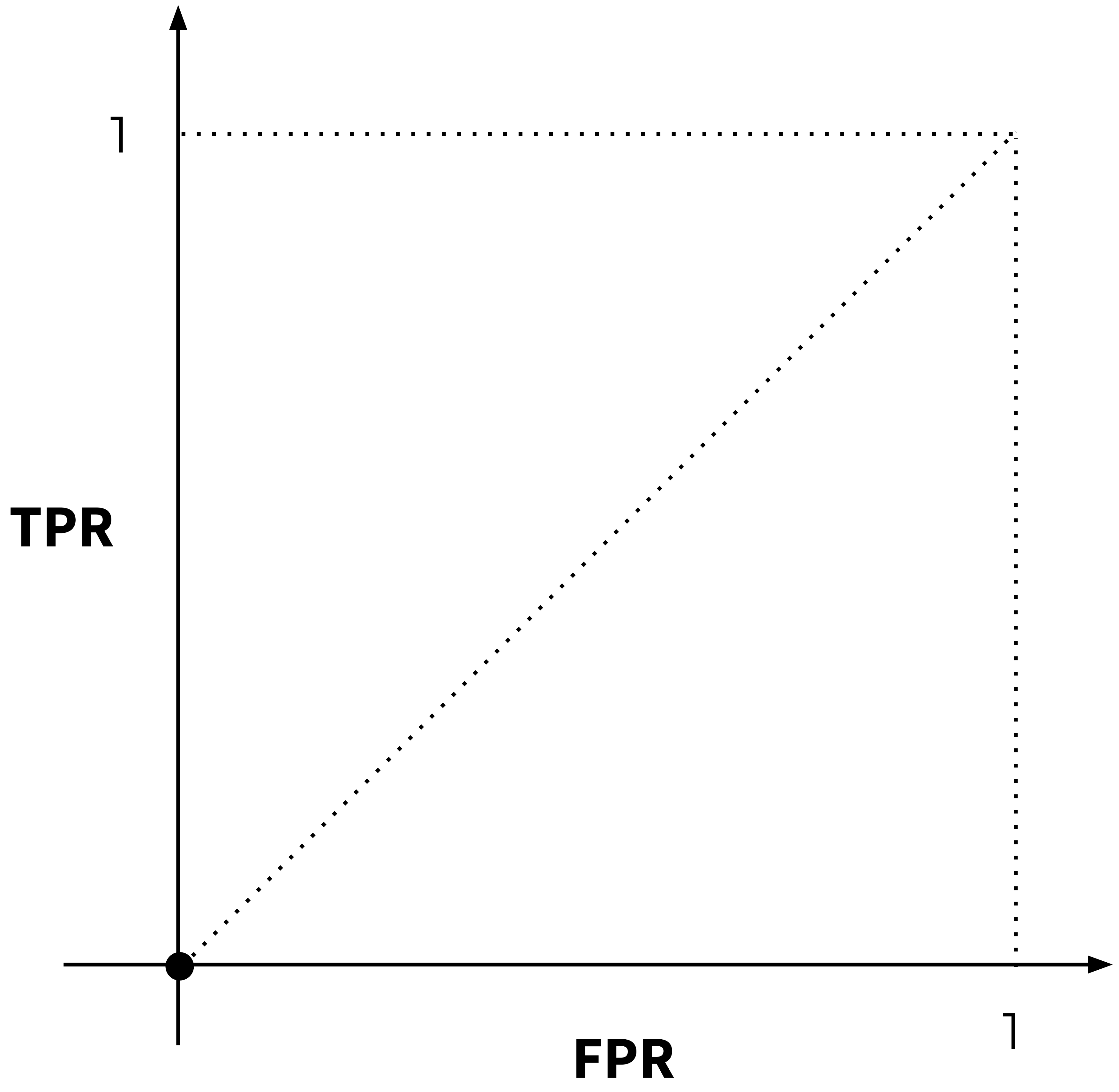

1. Introduction
2. Data
3. Rules
4. Evaluation
5. Conclusion

Evaluation: AUC ROC



Ranking	Risk score	Class
1	10.3	Fraud
2	8.6	Genuine
3	2.1	Fraud
4	1.3	Genuine
5	0.2	Genuine
6	0.1	Genuine

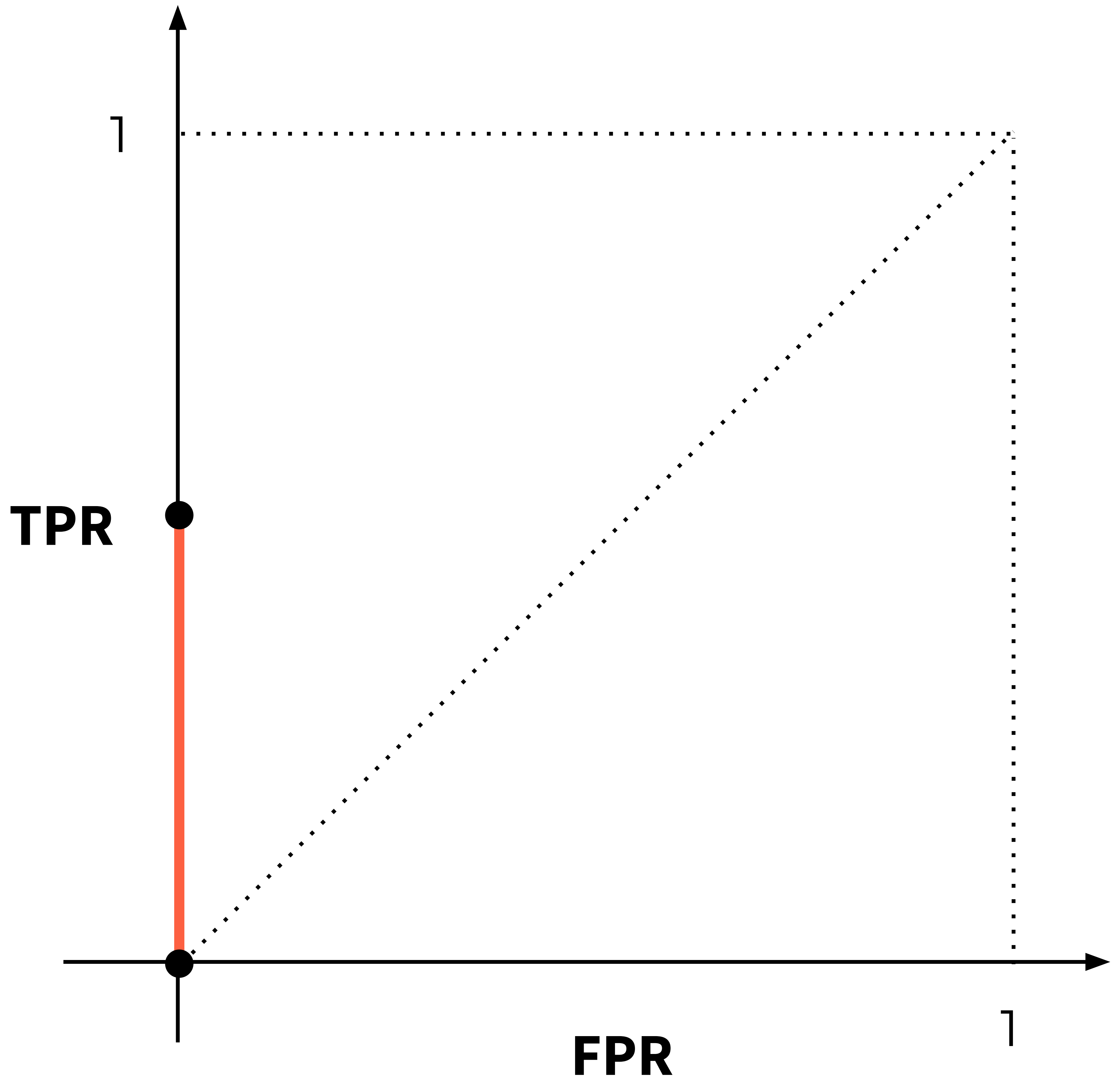
Evaluation: AUC ROC



TPR = 0.0
FPR = 0.0

Ranking	Risk score	Class
1	10.3	Fraud
2	8.6	Genuine
3	2.1	Fraud
4	1.3	Genuine
5	0.2	Genuine
6	0.1	Genuine

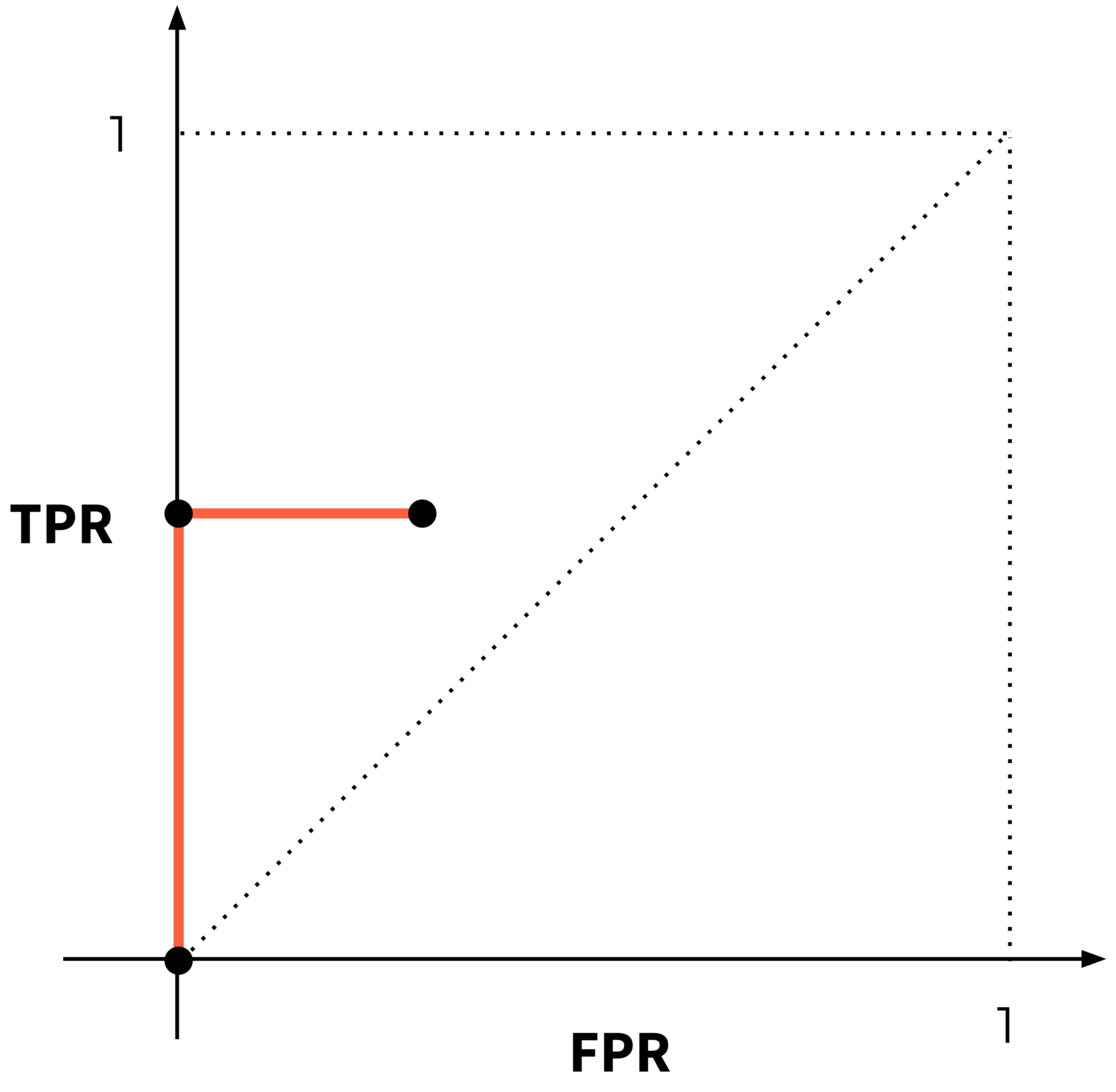
Evaluation: AUC ROC



TPR = 0.5
FPR = 0.0

Ranking	Risk score	Class
1	10.3	Fraud
2	8.6	Genuine
3	2.1	Fraud
4	1.3	Genuine
5	0.2	Genuine
6	0.1	Genuine

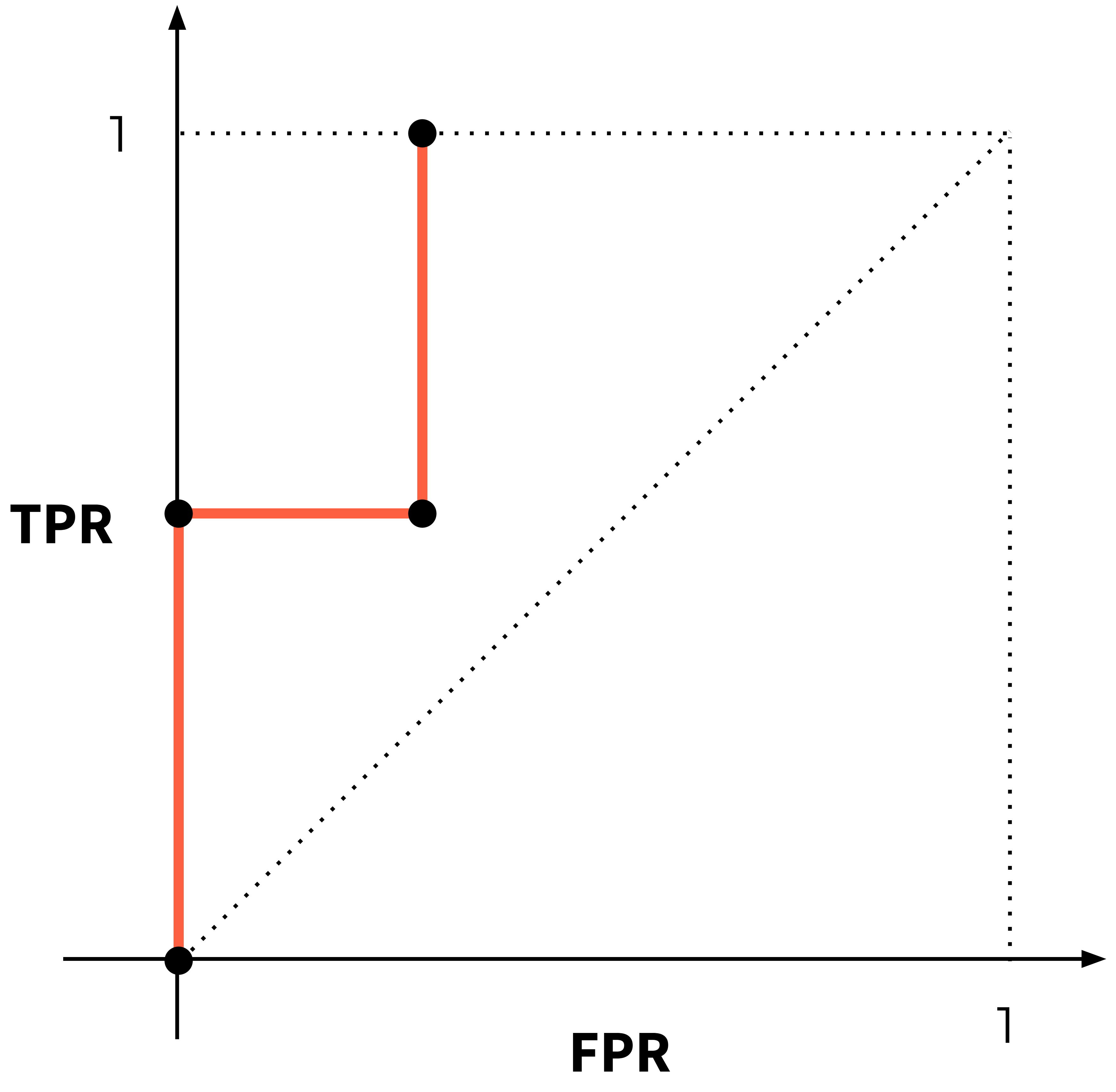
Evaluation: AUC ROC



TPR = 0.5
FPR = 0.25

Ranking	Risk score	Class
1	10.3	Fraud
2	8.6	Genuine
3	2.1	Fraud
4	1.3	Genuine
5	0.2	Genuine
6	0.1	Genuine

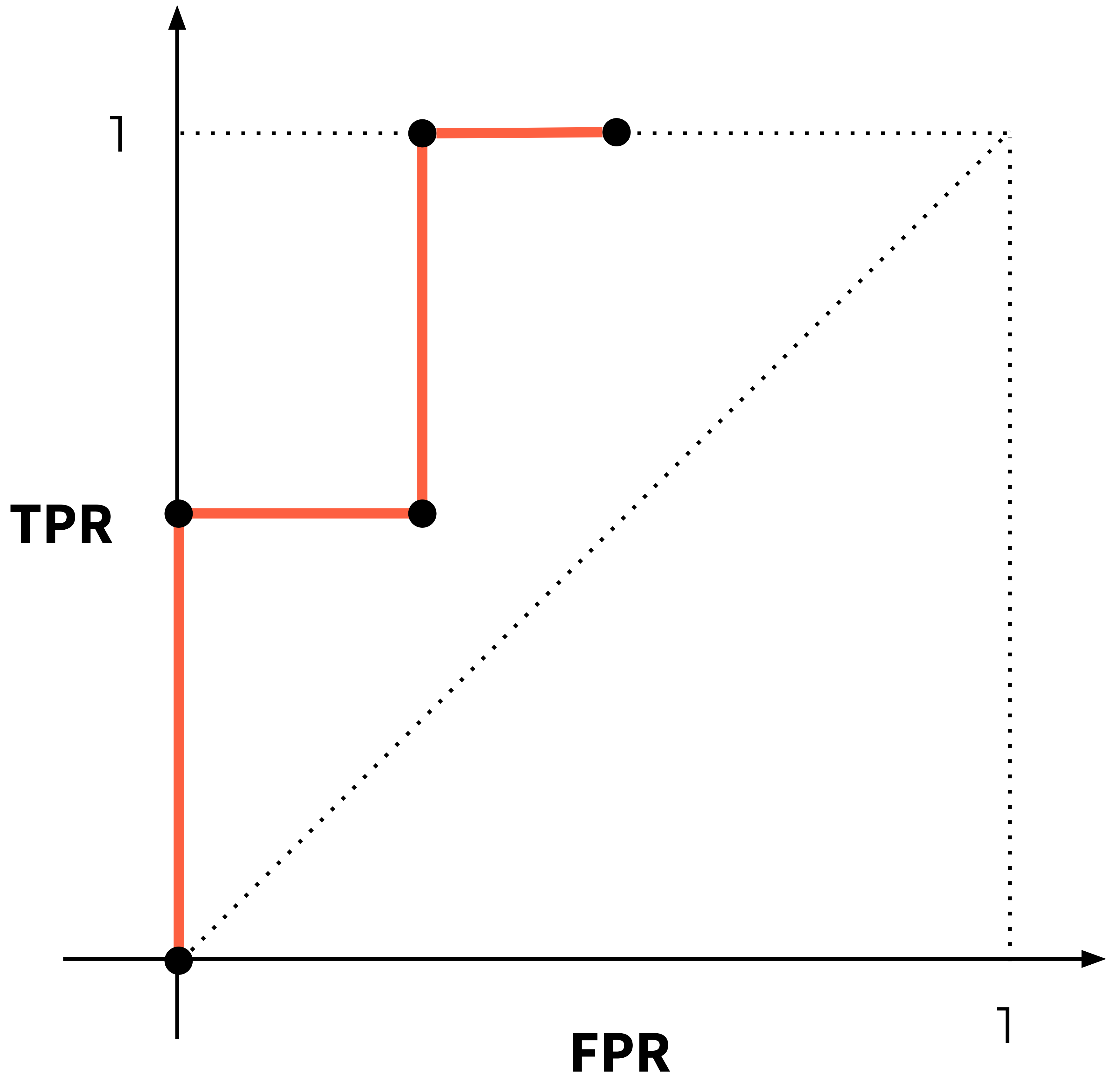
Evaluation: AUC ROC



TPR = 1.0
FPR = 0.25

Ranking	Risk score	Class
1	10.3	Fraud
2	8.6	Genuine
3	2.1	Fraud
4	1.3	Genuine
5	0.2	Genuine
6	0.1	Genuine

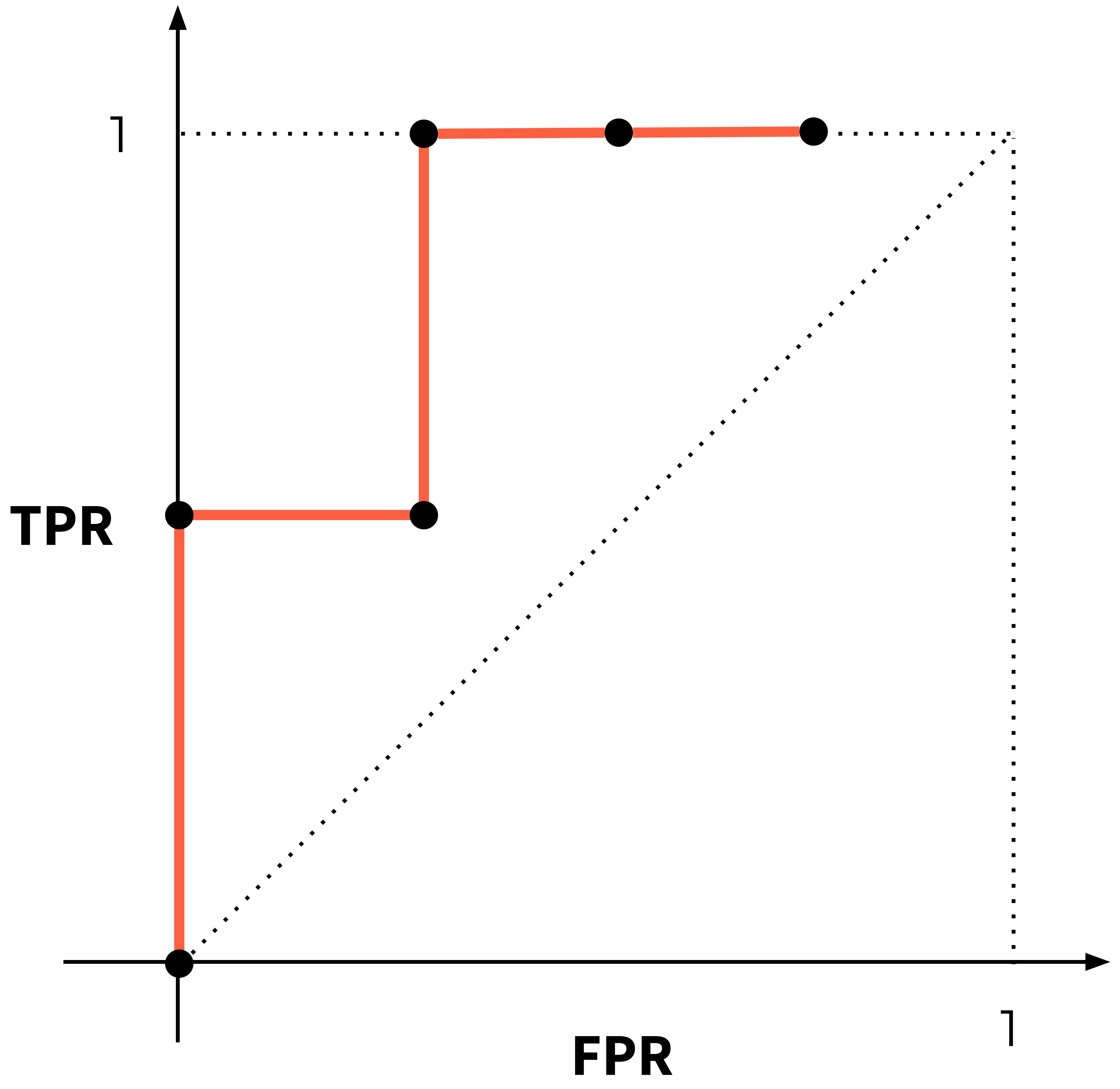
Evaluation: AUC ROC



TPR = 1.0
FPR = 0.5

Ranking	Risk score	Class
1	10.3	Fraud
2	8.6	Genuine
3	2.1	Fraud
4	1.3	Genuine
5	0.2	Genuine
6	0.1	Genuine

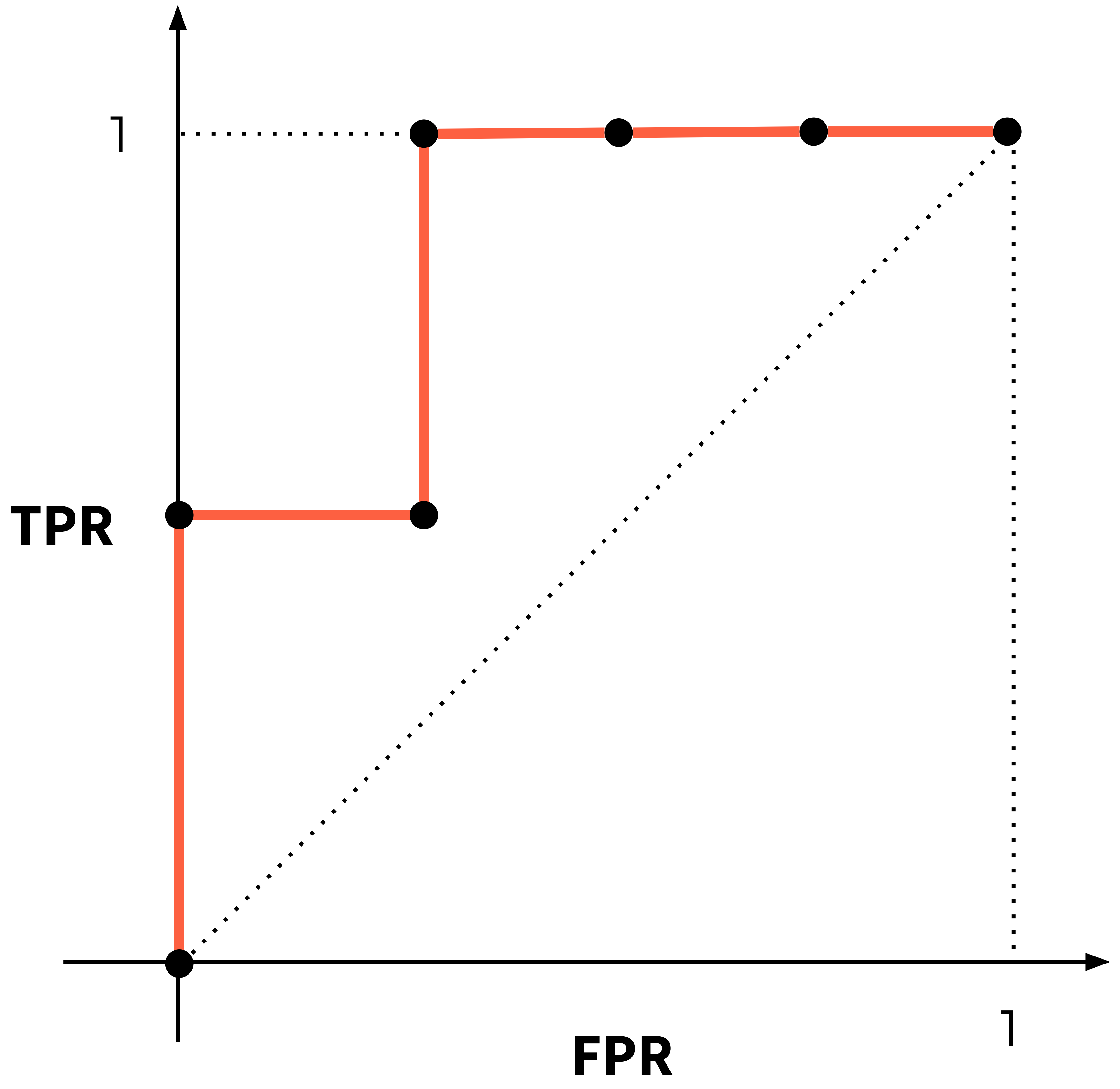
Evaluation: AUC ROC



TPR = 1.0
FPR = 0.75

Ranking	Risk score	Class
1	10.3	Fraud
2	8.6	Genuine
3	2.1	Fraud
4	1.3	Genuine
5	0.2	Genuine
6	0.1	Genuine

Evaluation: AUC ROC



TPR = 1.0
FPR = 1.0

Ranking	Risk score	Class
1	10.3	Fraud
2	8.6	Genuine
3	2.1	Fraud
4	1.3	Genuine
5	0.2	Genuine
6	0.1	Genuine

1. Introduction
2. Data
3. Rules
4. Evaluation
5. **Conclusion**

Questions?

Train data: https://goo.gl/wNCWi7

Test data and website: coming soon…

Slides: https://goo.gl/SmukXk

Anomaly detection techniques

Univariate distribution

Clustering-based

Nearest-Neighbour

Subspace-based

One-class SVM

Isolation Forest

Probabilistic

Anomaly detection techniques

Clustering-based

Subspace-based

Isolation Forest

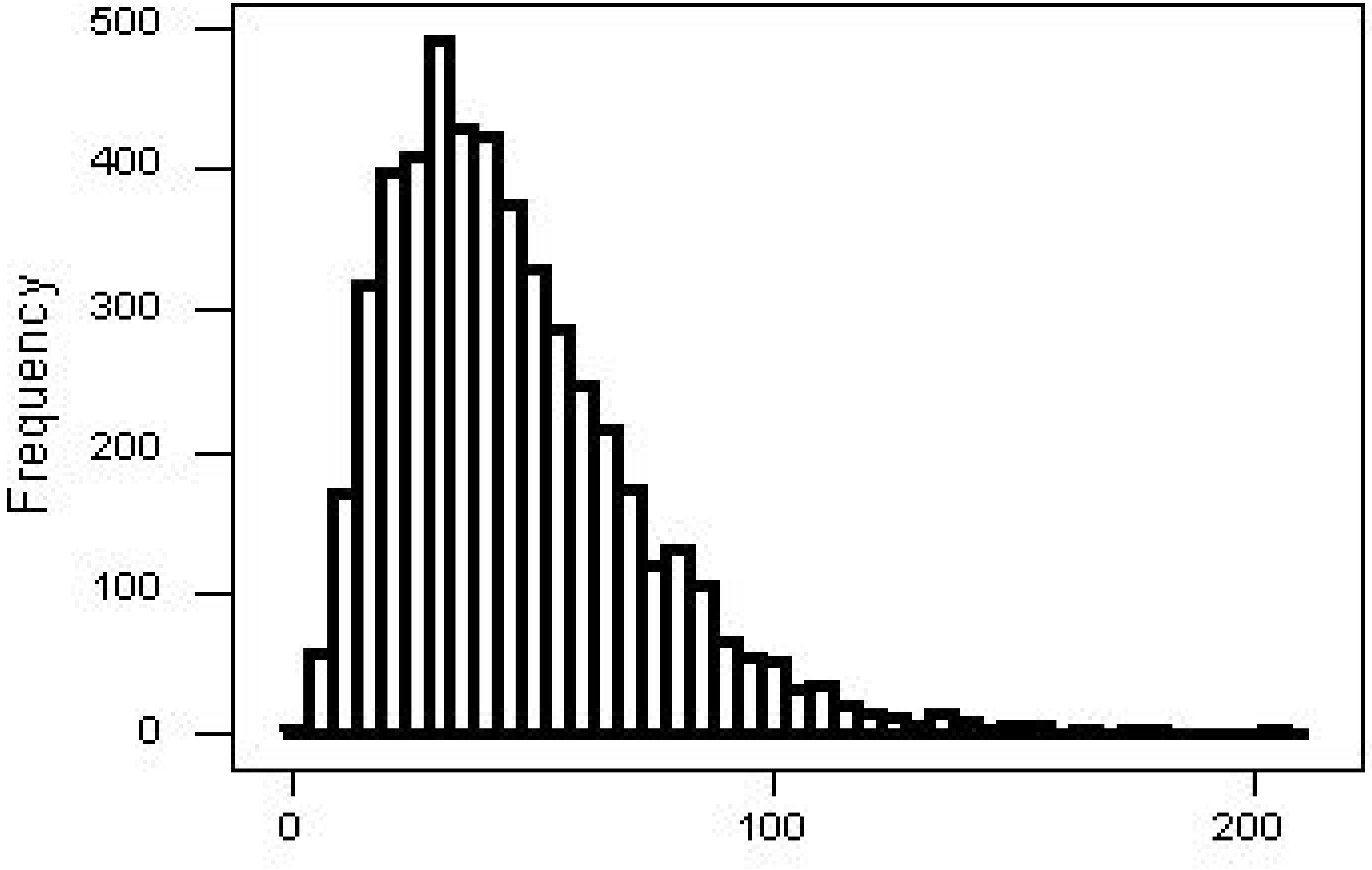
Univariate distribution

Nearest-Neighbour

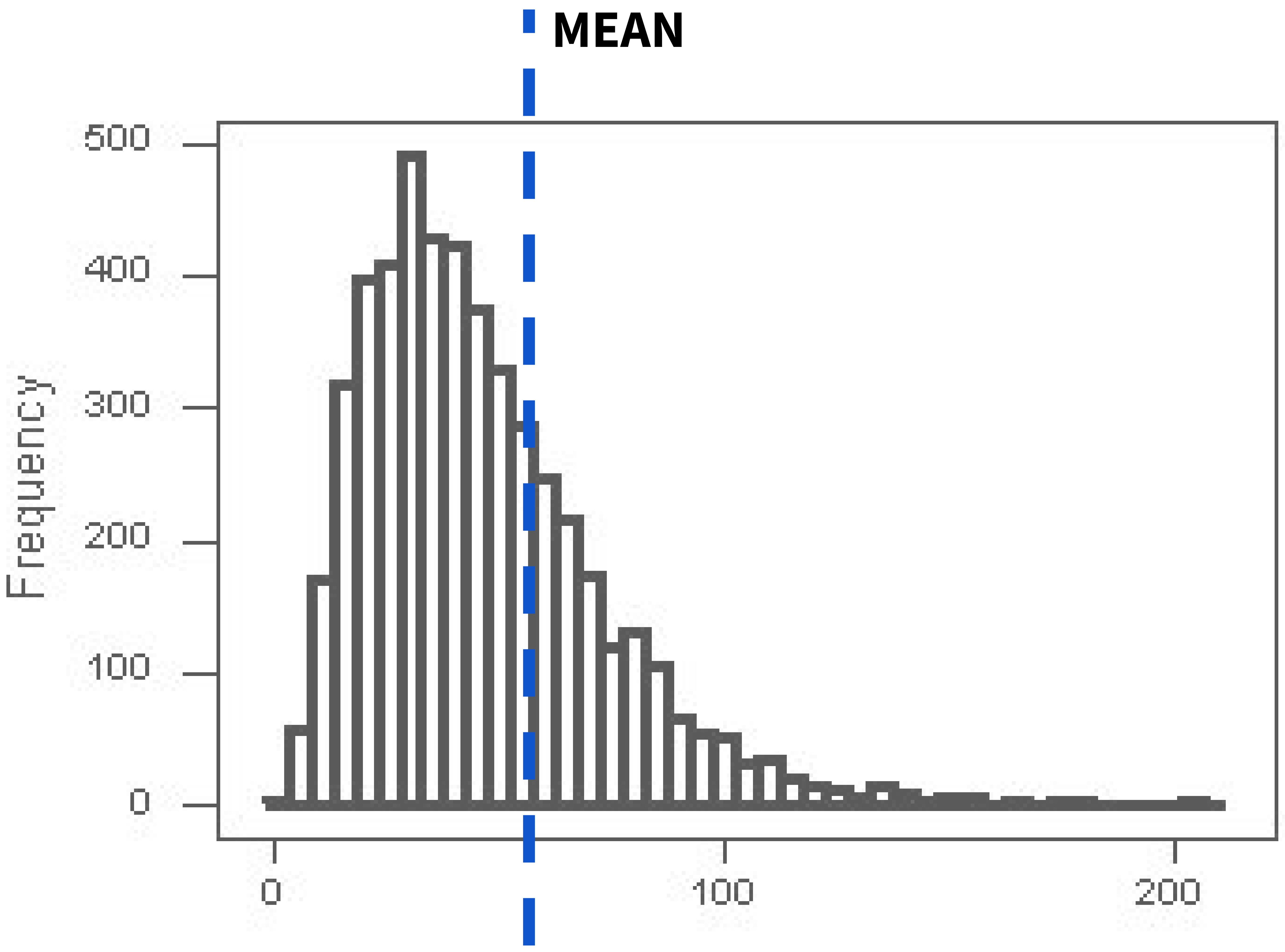
One-class SVM

Probabilistic

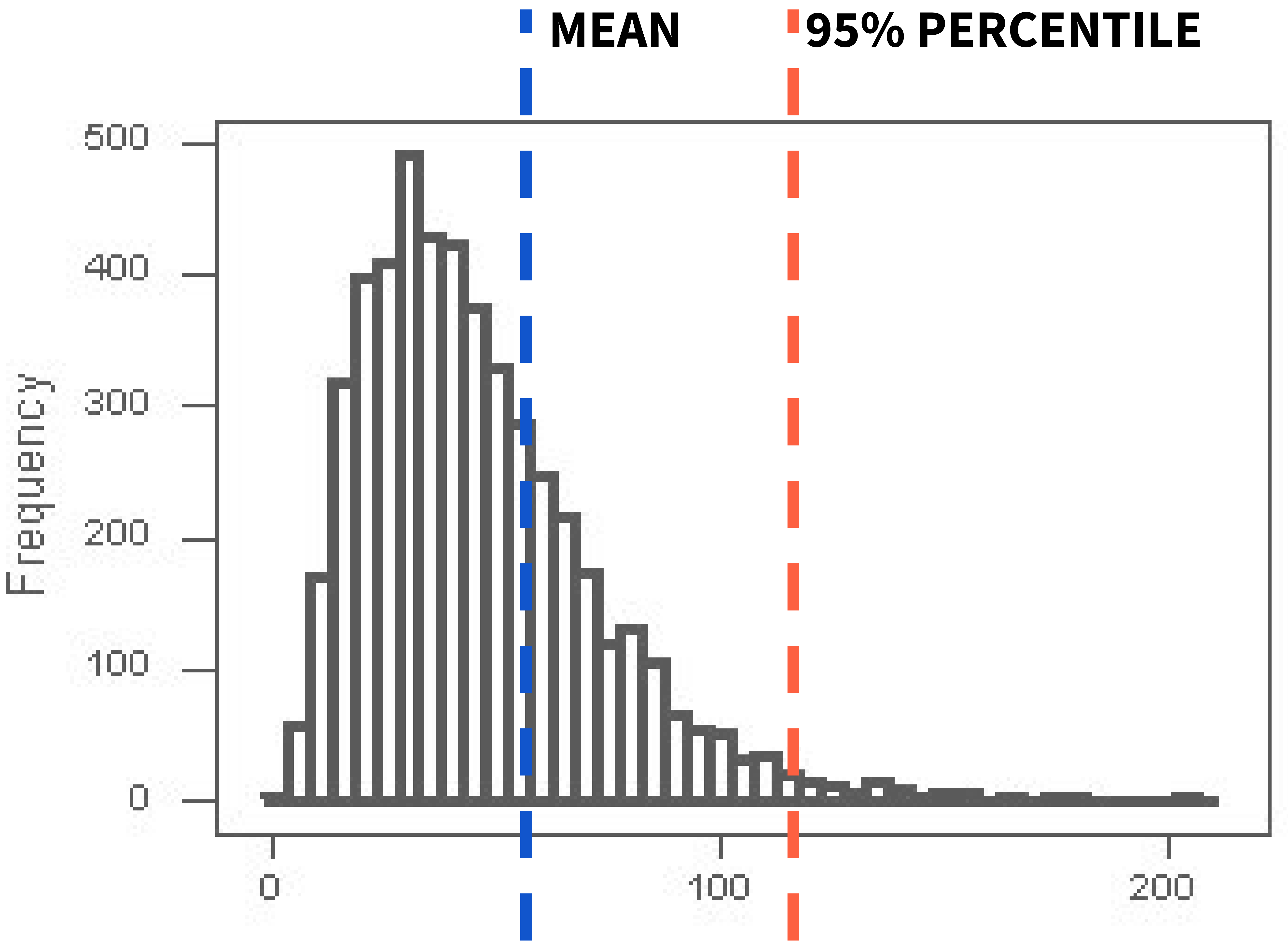
Histogram-based Outlier Score



Histogram-based Outlier Score



Histogram-based Outlier Score



Anomaly detection techniques

Univariate distribution

Clustering-based

Nearest-Neighbour

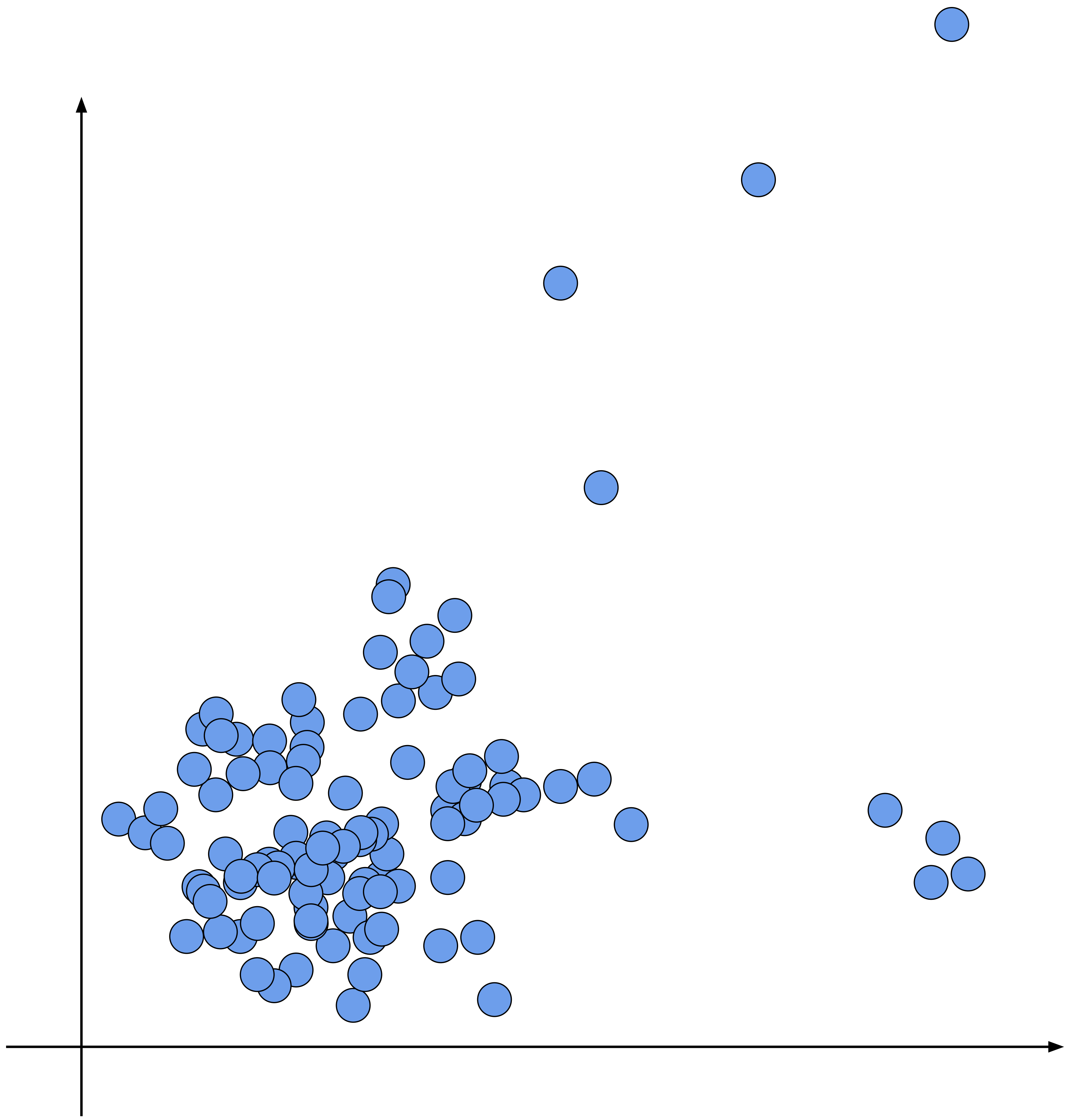
Subspace-based

One-class SVM

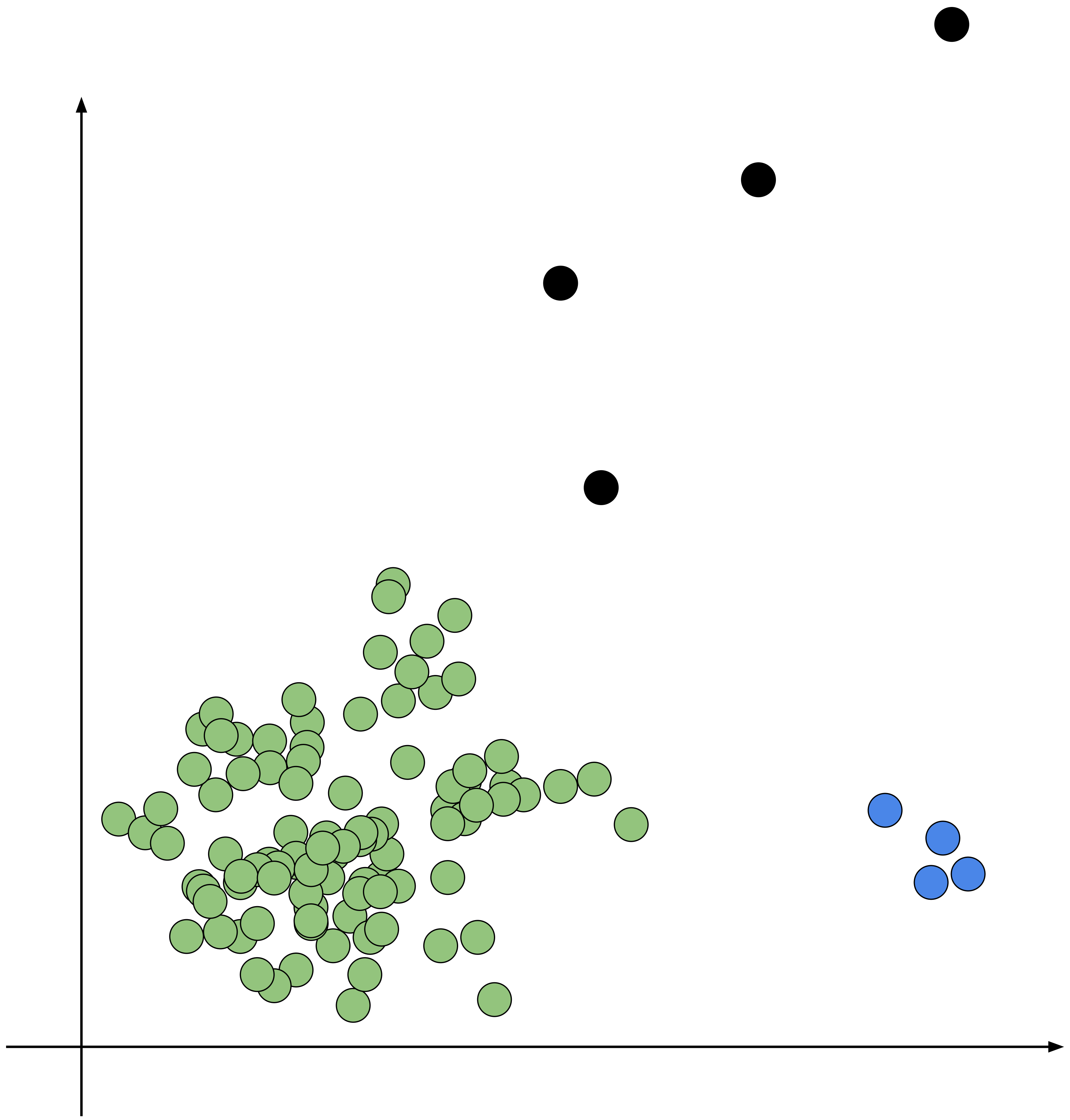
Isolation Forest

Probabilistic

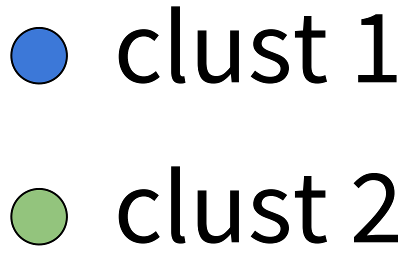
Iterative DBSCAN + uCBLOF



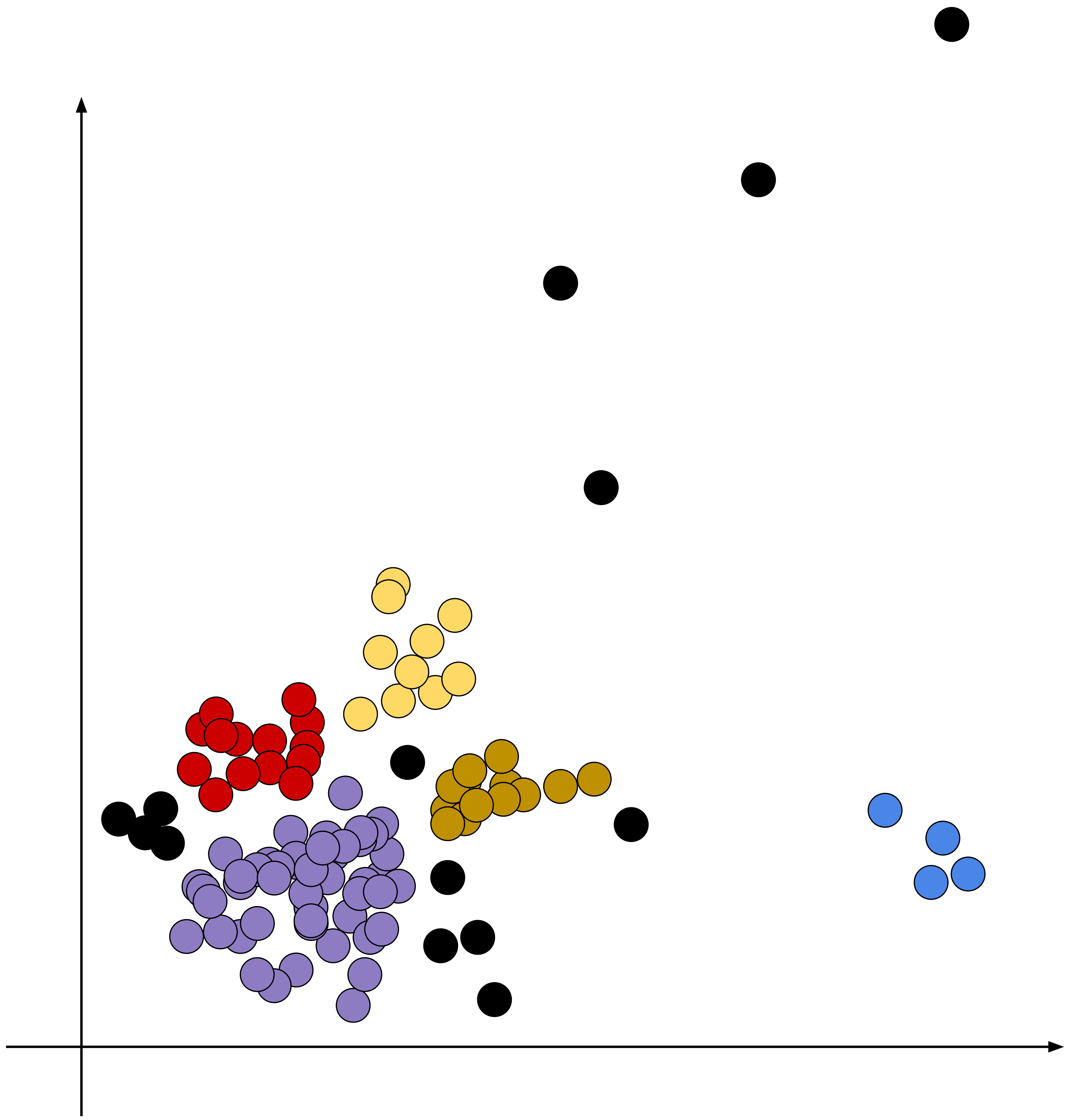
Iterative DBSCAN + uCBLOF



1. *choose large ε and cluster*



Iterative DBSCAN + uCBLOF



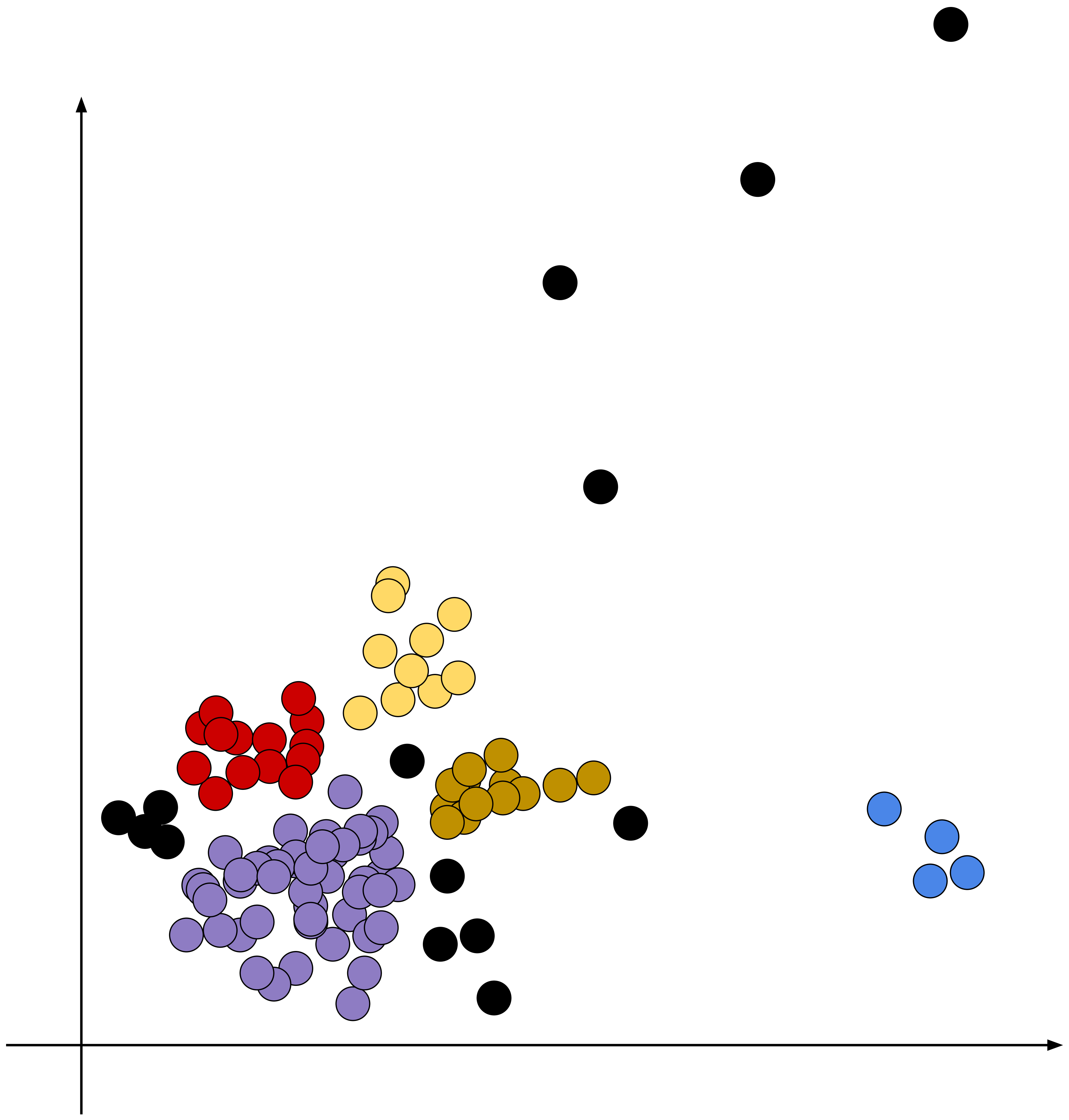
1. *choose large ε and cluster*

   - clust 1
   - clust 2 → clust 3, clust 4, clust 5, clust 6

2. *take largest cluster and repeat with smaller ε*

   - clust 3
   - clust 4
   - clust 5 → ...
   - clust 6

Iterative DBSCAN + uCBLOF



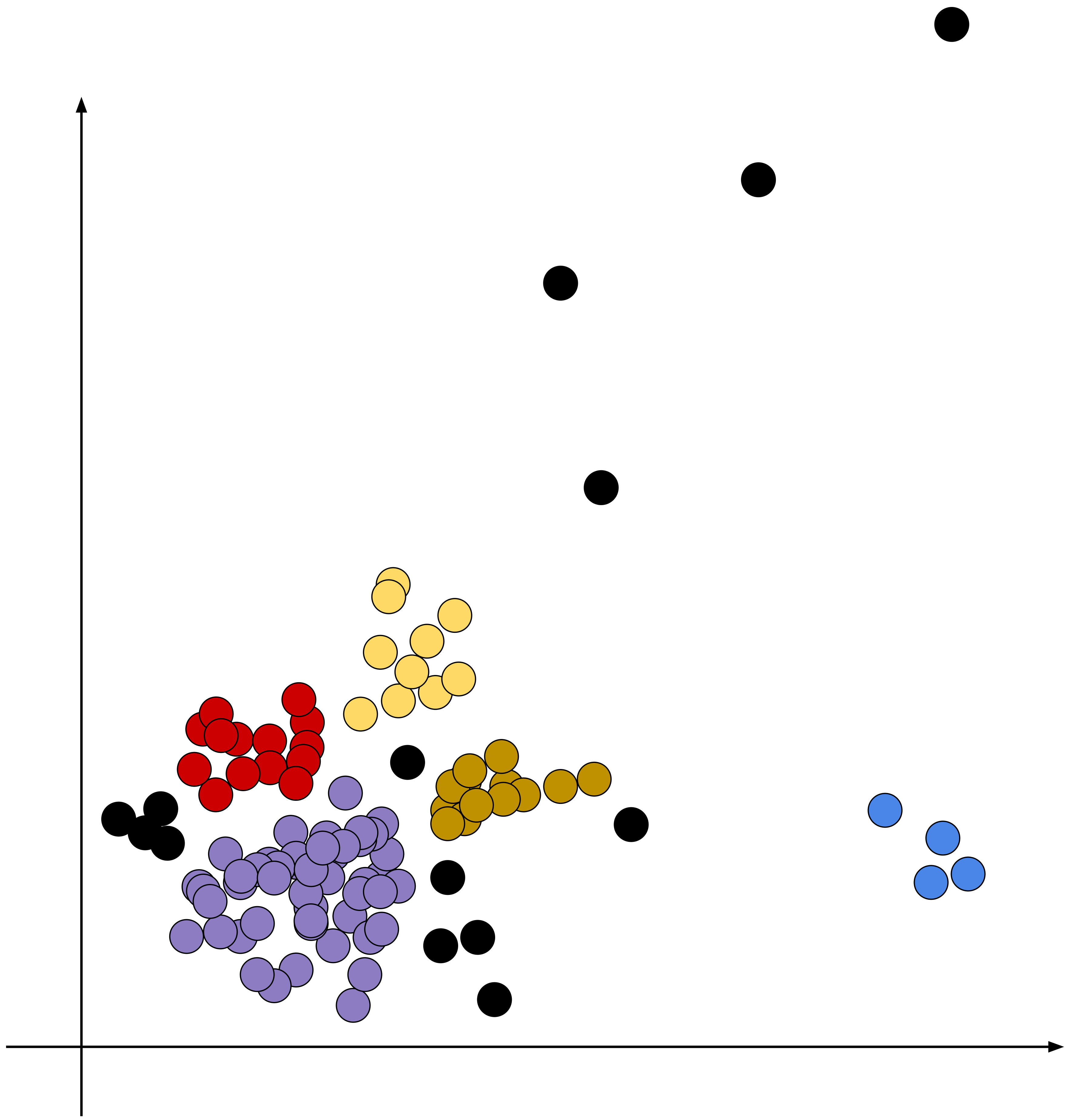
1. *choose large ε and cluster*
2. *take largest cluster and repeat with smaller ε*

clust 1

clust 2 → clust 3, clust 4, clust 5, clust 6

clust 5 → ...

Iterative DBSCAN + uCBLOF



1. *choose large ε and cluster*

- clust 1
- clust 2

2. *take largest cluster and repeat with smaller ε*

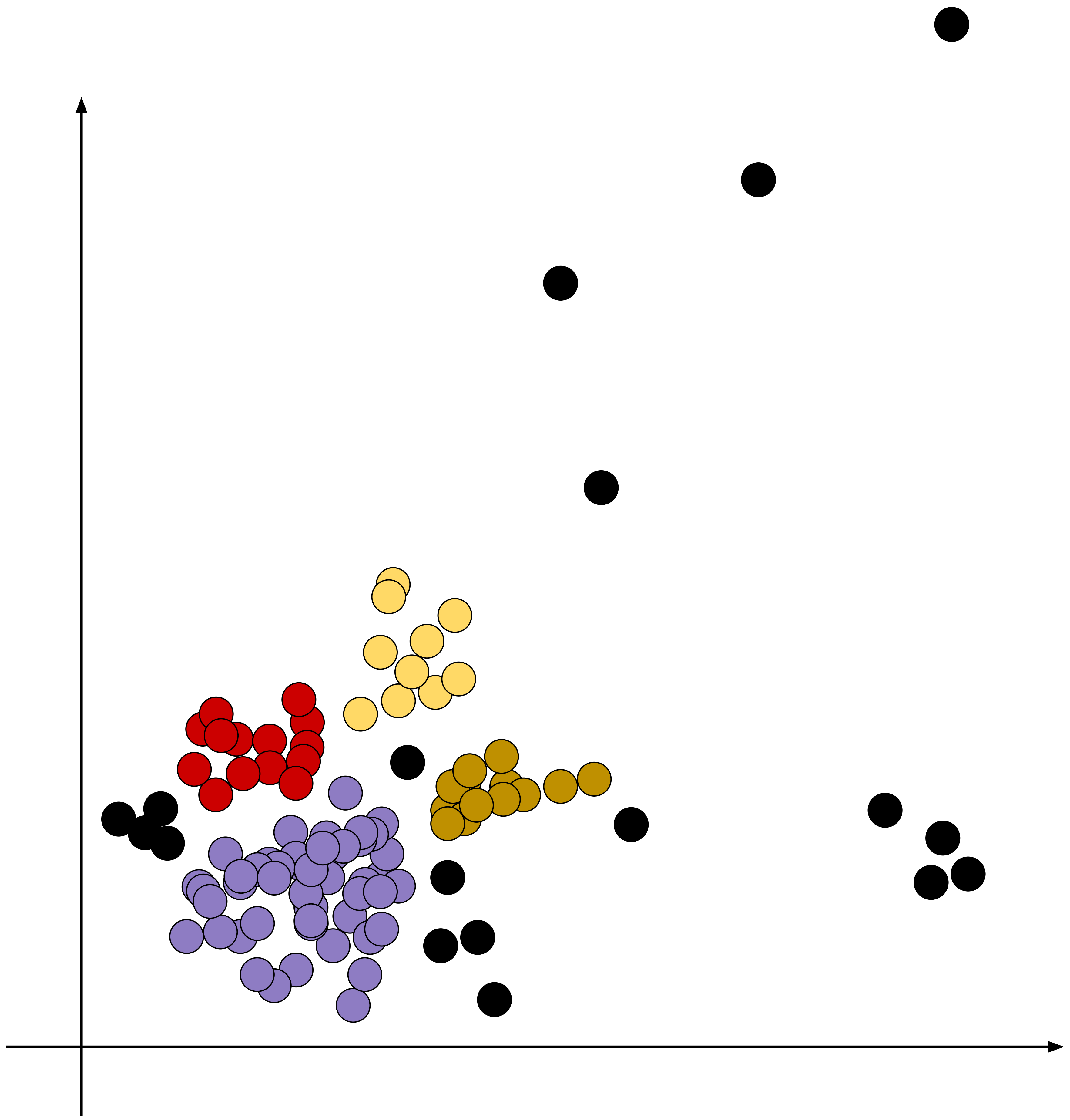
clust 2 → clust 3, clust 4, clust 5, clust 6

clust 5 → ...

3. *order by # of elements*

clust 5 > clust 6 > clust 4 > clust 3 > clust 1

Iterative DBSCAN + uCBLOF



1. *choose large ε and cluster*
2. *take largest cluster and repeat with smaller ε*

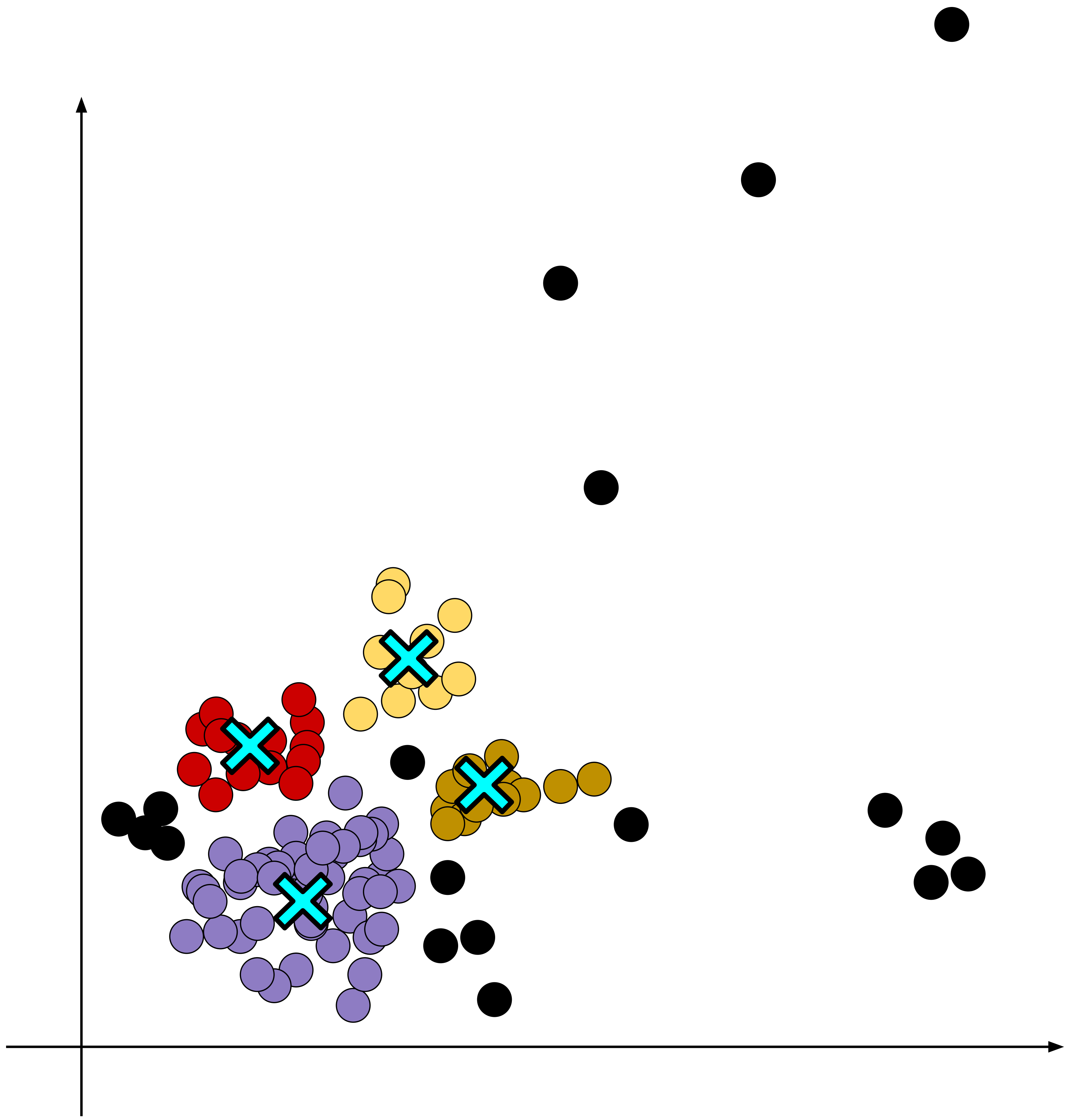
clust 1
clust 2 → clust 3, clust 4, clust 5, clust 6 → ...

3. *order by # of elements*

clust 5 > clust 6 > clust 4 > clust 3 > ~~clust 1~~

4. *keep clusters with 90% of data points*

Iterative DBSCAN + uCBLOF



1. *choose large ε and cluster*

2. *take largest cluster and repeat with smaller ε*

clust 1

clust 2 → clust 3, clust 4, clust 5, clust 6 → …

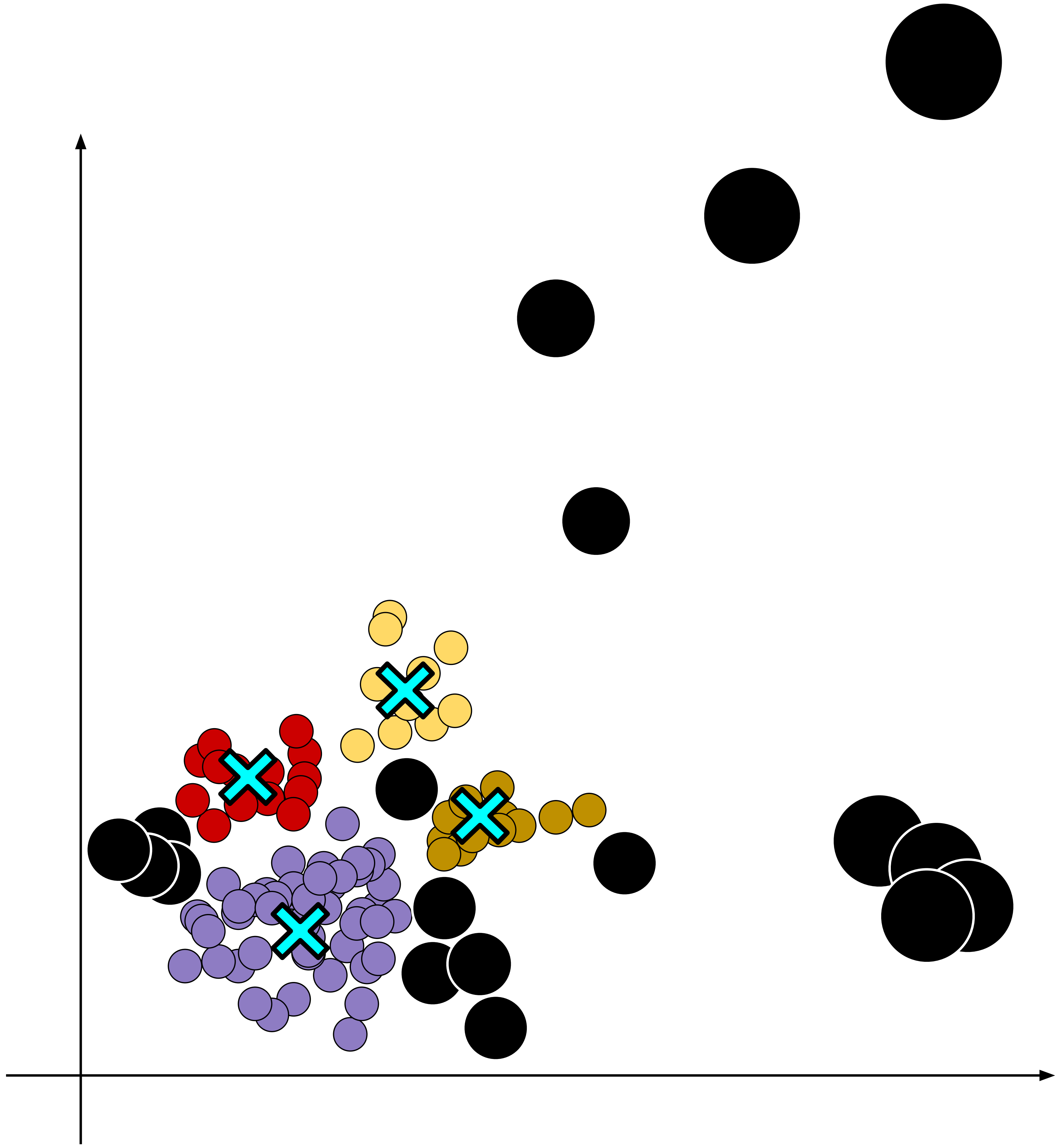
3. *order by # of elements*

clust 5 > clust 6 > clust 4 > clust 3 > ~~clust 1~~

4. *keep clusters with 90% of data points*

5. *risk scores = distance between points and closest centroid*

Iterative DBSCAN + uCBLOF



1. *choose large ε and cluster*

clust 1

clust 2 →

2. *take largest cluster and repeat with smaller ε*

clust 3

clust 4

clust 5 → ...

clust 6

3. *order by # of elements*

clust 5 > clust 6 > clust 4 > clust 3 > ~~clust 1~~

4. *keep clusters with 90% of data points*

5. *risk scores = distance between points and closest centroid*

Anomaly detection techniques

Univariate distribution

Clustering-based

Nearest-Neighbour

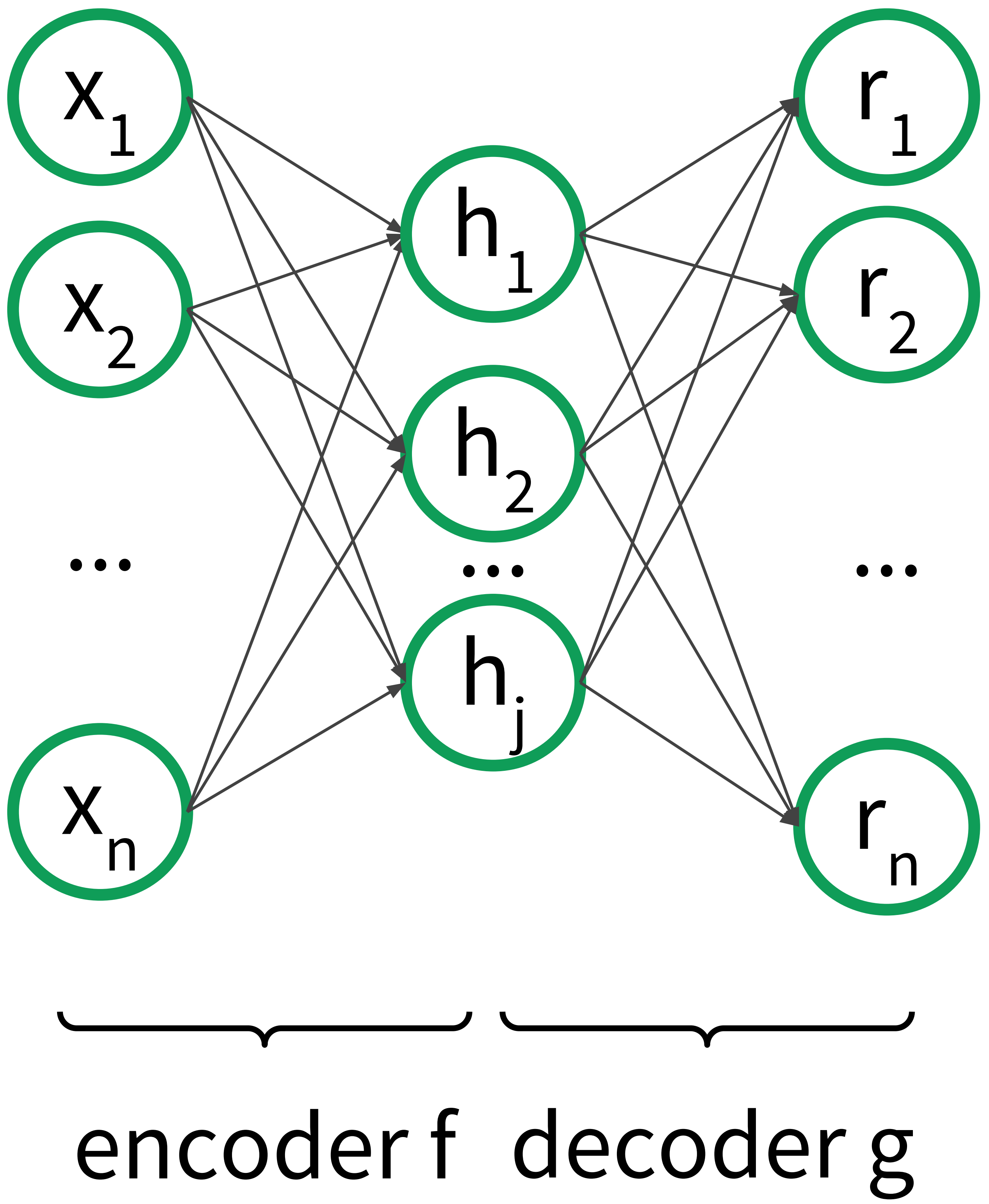
Subspace-based

One-class SVM

Isolation Forest

Probabilistic

Autoencoder



How it works?

NN approximates identity function producing output as similar as possible to given input.

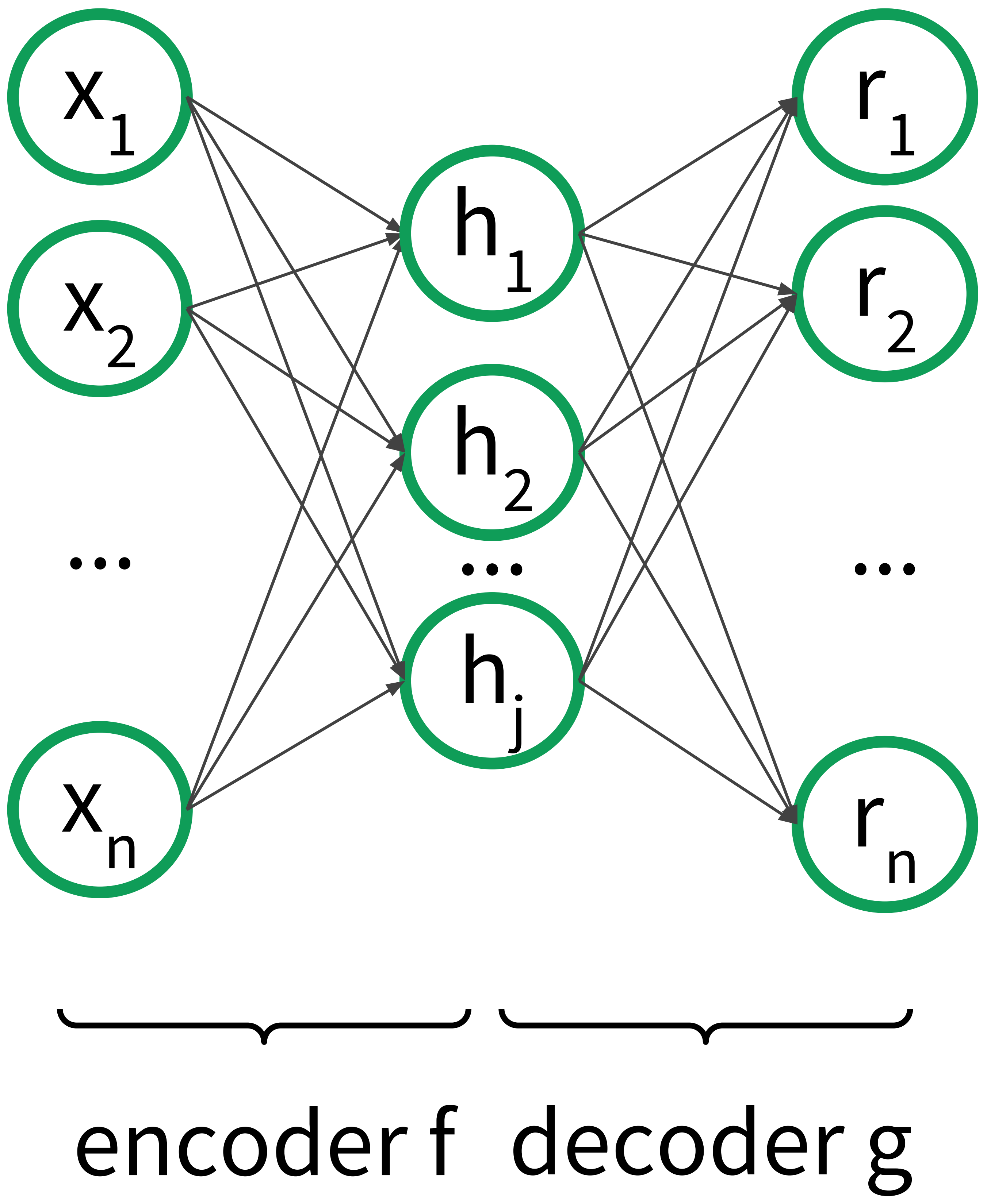
What do we learn?

Compression in hidden layer force learning of data low dimensional representation.

Why is it useful?

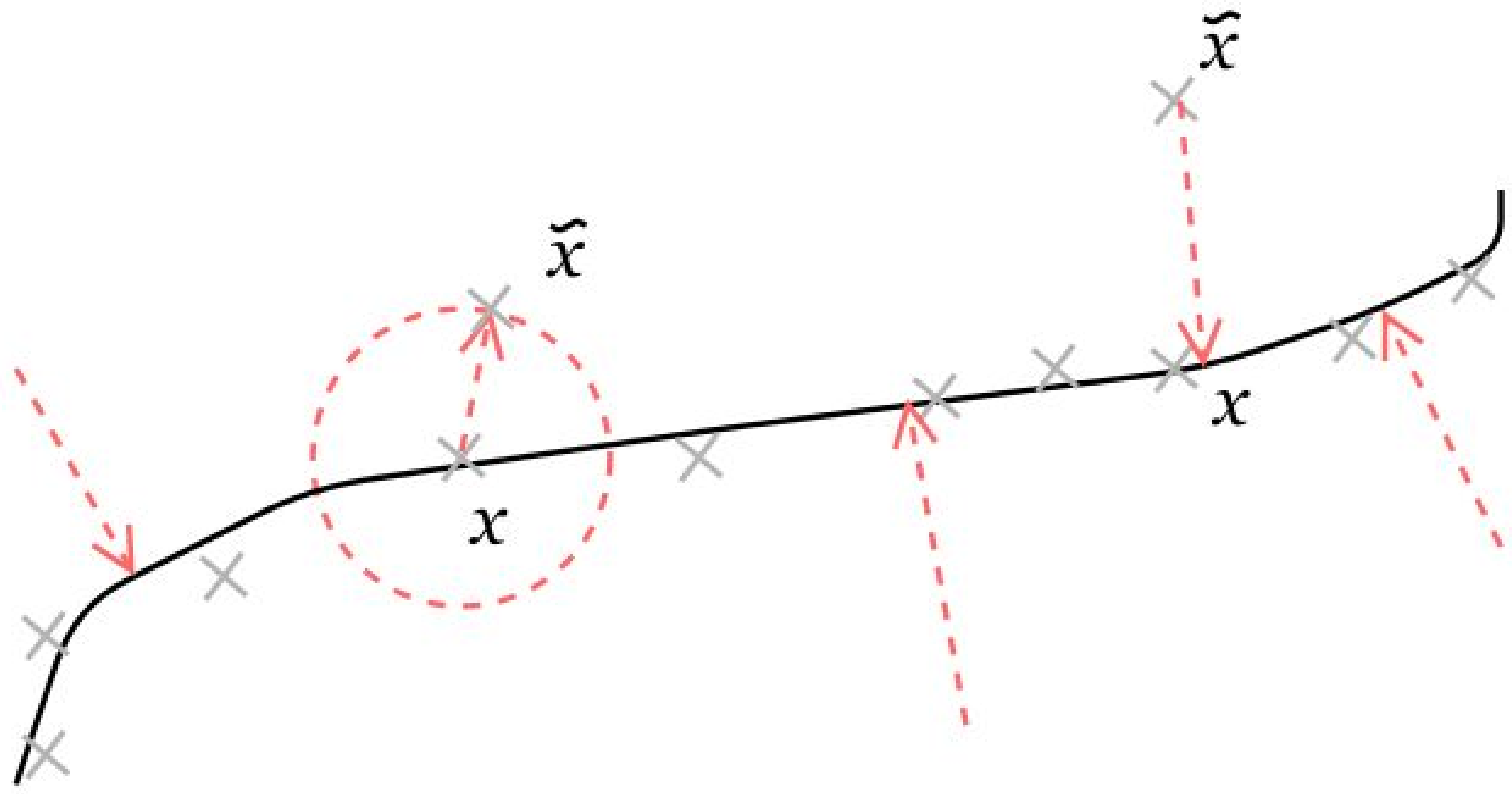
Anomalies, supposed to present a rare features' distribution, are poorly reconstructed.

Autoencoder

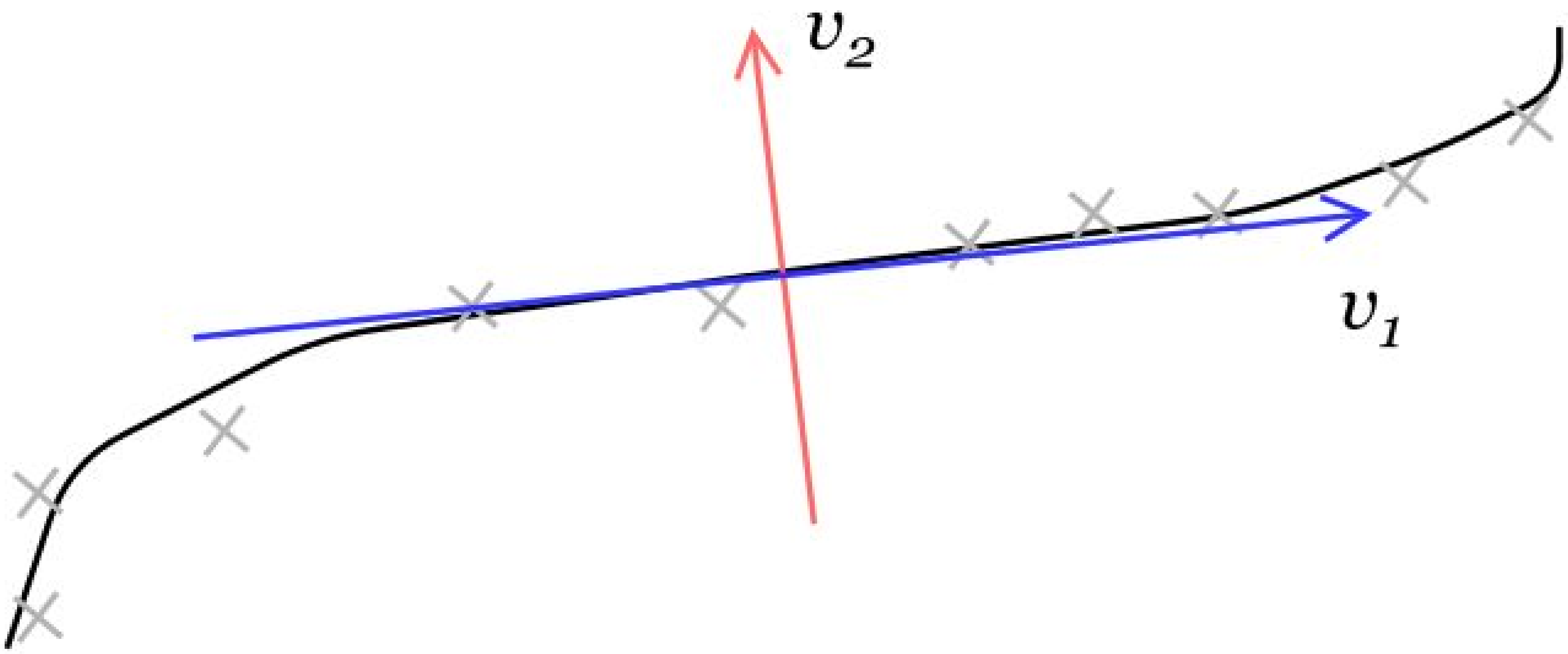


- $\mathcal{J}_{AE}(\theta) = \sum_{x \in D_n} L(x, g(f(x))) = \sum_{x \in D_n} (r_i - x_i)^2$
- $\mathcal{J}_{DAE}(\theta) = \sum_{x \in D_n} \mathbb{E}_{\tilde{x} \sim q(\tilde{x}|x)} [L(x, g(f(\tilde{x})))]$
- $\mathcal{J}_{CAE}(\theta) = \sum_{x \in D_n} (L(x, g(f(x))) + \lambda \| J_f(x) \|_F^2)$

DAE explained



CAE explained



Anomaly detection techniques