Anomaly Detection Overview

Identify rare items/events that differ from normal patterns

🎯 Types of Anomalies

Detection Approaches

Popular Anomaly Detection Algorithms

Key methods for identifying outliers and anomalies

📊 Statistical Methods

- **Z-Score**
  Measures standard deviations from mean. Points beyond 3σ are anomalies.
- **IQR (Interquartile Range)**
  Uses quartiles Q1 and Q3. Outliers beyond Q1-1.5×IQR or Q3+1.5×IQR.
- **Gaussian Distribution**
  Models normal data, identifies points in low-probability regions.

🎯 Density-based Methods

- **LOF (Local Outlier Factor)**
  Measures local density deviation. Fast training, effective for local anomalies.
- **DBSCAN**
  Clustering algorithm that identifies points not belonging to any cluster.
- **K-Means Clustering**
  Points far from cluster centers are considered anomalies.

📏 Distance-based Methods

- **K-Nearest Neighbors (KNN)**
  Detects anomalies based on distance to k-nearest neighbors.
- **Mahalanobis Distance**
  Accounts for correlation between features in multivariate data.

🌲 Ensemble Methods

- **Isolation Forest**
  Isolates anomalies using decision trees. Works well on most datasets.
- **Random Forest**
  Can be adapted for anomaly detection in classification tasks.

⚠ Imbalanced data distribution

⚠ Defining normal behavior

⚠ Lack of labeled anomalies

⚠ Evolving patterns over time

🚀 Advanced Methods & Learning Approaches

Modern techniques and learning paradigms for anomaly detection

🧠 Machine Learning Methods

- **One-Class SVM**
  Creates hypersphere around normal data, separating anomalies.
- **Autoencoders**
  Neural networks that reconstruct data. High reconstruction error indicates anomalies.
- **LSTM Networks**
  Captures temporal patterns in time-series data for sequence anomaly detection.

📐 Dimensionality Reduction

- **PCA (Principal Component Analysis)**
  Reduces dimensions, highlights influential features and outliers.
- **t-SNE**
  Non-linear technique for visualizing high-dimensional data and clusters.
- **UMAP**
  Fast dimensionality reduction preserving global structure.

🎓 Learning Paradigms

Supervised

Requires labeled data with normal and anomalous examples

SVM, Neural Networks, Random Forest

Unsupervised

No labels needed. Most common approach in practice

Isolation Forest, LOF, K-Means, Autoencoders

Semi-supervised

Trained on normal data only, detects deviations

One-Class SVM, Novelty Detection

💡 Key Insight

Algorithm choice depends on data characteristics, computational resources, and whether labels are available. Isolation Forest and LOF are popular starting points due to their effectiveness and ease of use.

Implementation & Popular Tools

Python libraries and practical considerations

Python Libraries

Scikit-learn
Isolation Forest, One-Class SVM, LOF, DBSCAN

PyOD
Python Outlier Detection - 40+ algorithms specialized for anomaly detection

TensorFlow/PyTorch
Deep learning models: Autoencoders, LSTM, Variational Autoencoders

PyCaret
Low-code ML library with built-in anomaly detection module

Implementation Tips

Data Preprocessing
Normalize/standardize features for distance-based methods

Feature Engineering
Create relevant features that capture domain knowledge

Threshold Tuning
Adjust contamination rate based on expected anomaly percentage

Model Comparison
Test multiple algorithms to find best fit for your data

Interpretability
Use explainable methods when stakeholder understanding is critical

Performance Considerations

Speed
LOF: Fast training
Isolation Forest: Moderate

Accuracy
Depends on data characteristics & tuning

Scalability
Consider distributed algorithms for big data

Adaptability
Retrain models for evolving patterns

Best Practice

Start with simple statistical methods (Z-score, IQR) for baseline performance, then progress to more sophisticated algorithms (Isolation Forest, One-Class SVM) if needed. Always validate results with domain experts.