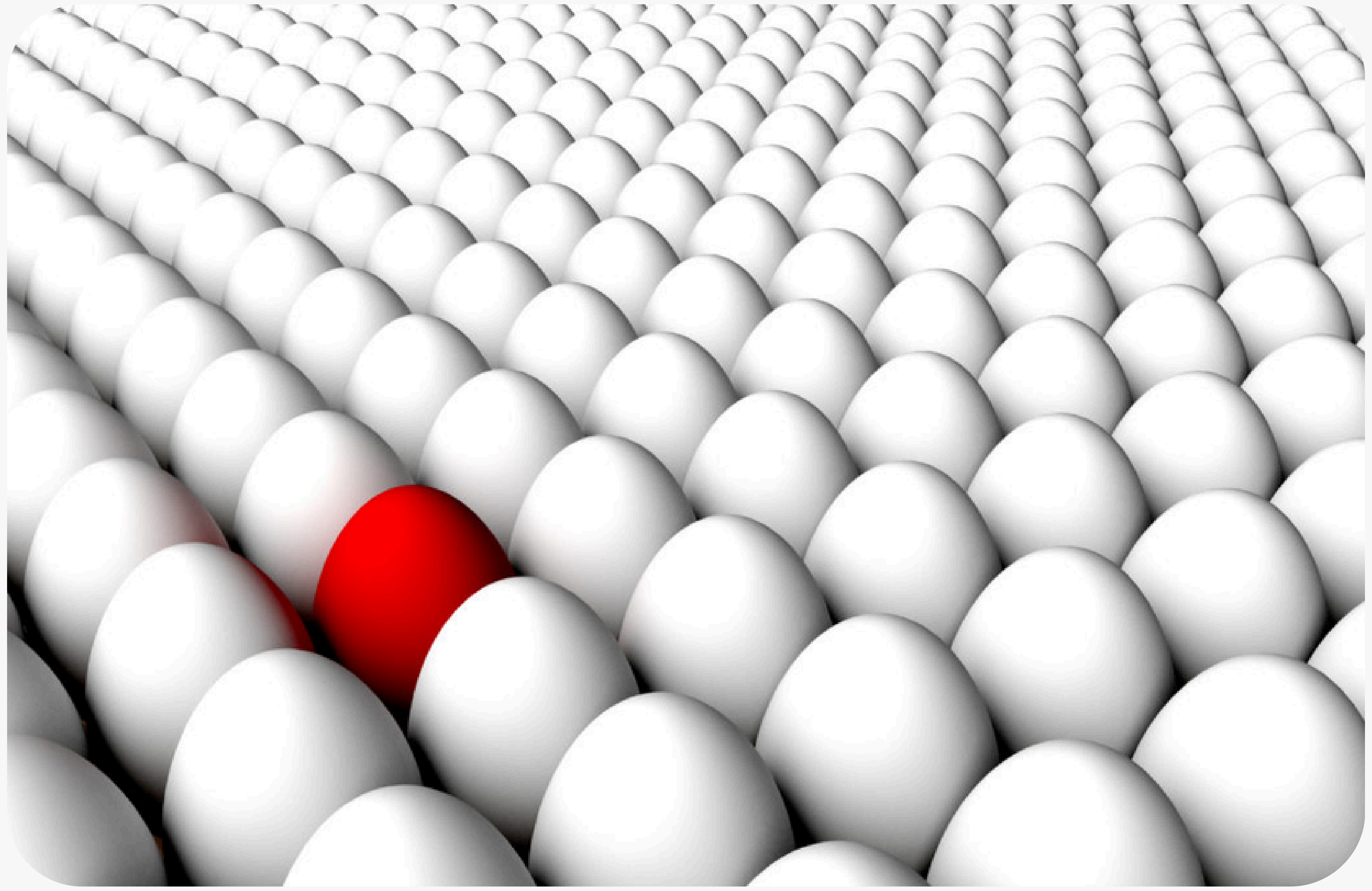
ANOMALY DETECTION

WORKSHOP 5/10
SUMMER 2024
SESSION 3

Lecturer: Reza Shokrazd

Agenda

- Introduction to Anomaly Detection
- Types of Anomalies and Data Patterns
- Statistical Methods for Anomaly Detection
- Machine Learning Techniques
- Deep Learning Approaches
- Applications Across Industries
- Evaluation Metrics and Validation
- Challenges and Limitations
- Tools and Libraries Overview
- Future Trends and Research Directions



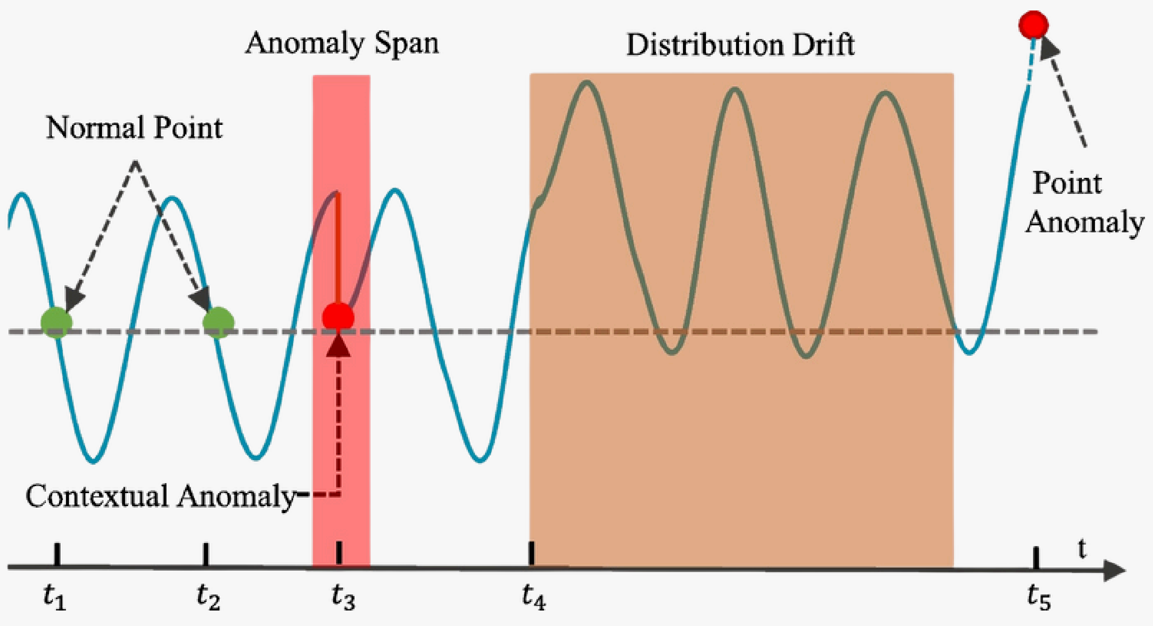
1. Dimensionality Reduction

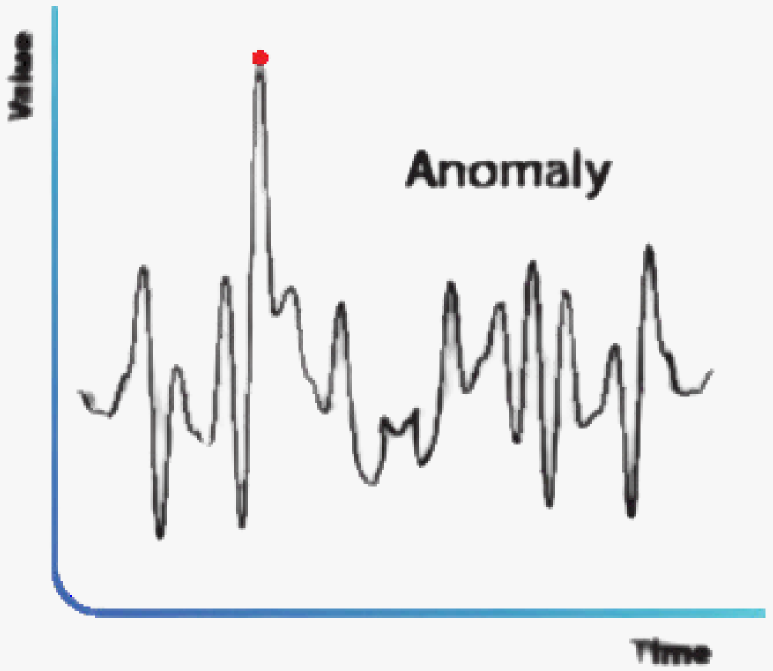
- Identifying data points that deviate from normal patterns.
- **Importance:** Essential for uncovering errors, fraud, or unexpected events.
- **Applications:** Utilized in finance, healthcare, cybersecurity, and manufacturing.
- **Types:** Includes point, contextual, and collective anomalies.
- **Challenges:** Difficulties due to noise, high dimensionality, limited labels.

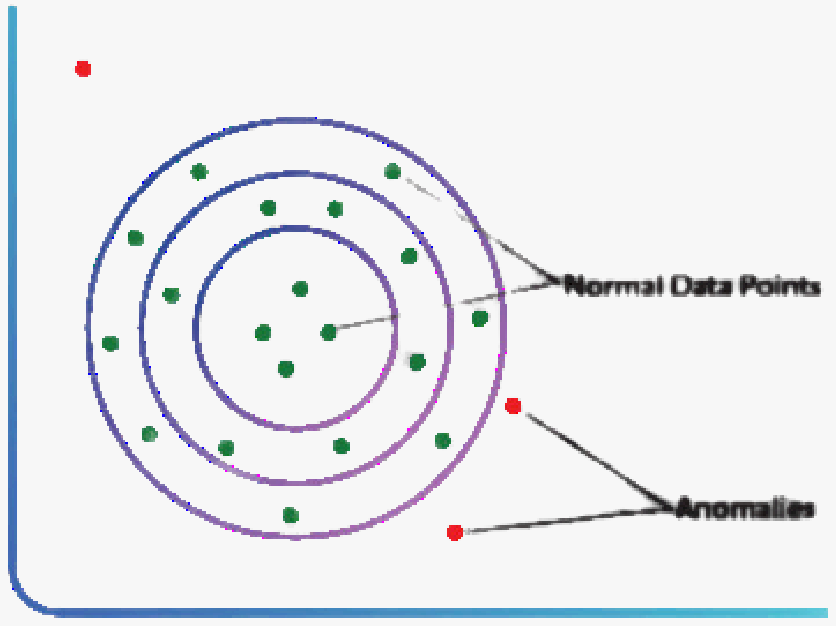
Anomaly detection identifies patterns deviating from expected behavior.

2. Types of Anomalies and Data Patterns

- Point anomalies (single instances)
- Contextual anomalies (context-dependent)
- Collective anomalies (group deviations)
- Temporal anomaly patterns
- Spatial data anomalies







2. Types of Anomalies and Data Patterns

- **Point Anomalies:** Deviations in single data points
  - Example: A sudden spike in network traffic | Popular Model: Isolation Forest
- **Contextual Anomalies:** Data that is anomalous in a specific context
  - Example: High temperature readings during winter months | Popular Model: LSTM Neural Networks
- **Collective Anomalies:** Anomalies found in a group of related data points
  - Example: Unusual patterns in a sequence of transactions | Popular Model: DBSCAN Clustering
- **Temporal Anomalies**: Anomalies occurring over time
  - Example: Sudden drop in stock market indices | Popular Model: Time Series Decomposition
- **Spatial Anomalies:** Anomalies based on geographical or spatial data
  - Example: Unusual disease outbreak in a region | Popular Model: Spatial Scan Statistics

3. Statistical Methods for Anomaly Detection

- **Z-Score Method:** Measures how far a data point is from the mean in standard deviations
- **Statistical Hypothesis Testing:** Determines if a data point significantly deviates from the population
- **Control Charts:** Monitors process metrics over time to detect shifts or trends
- **Density-Based Methods:** Identifies anomalies based on data density in the feature space
- **Probabilistic Models:** Uses probability distributions to model normal behavior
- **Regression Analysis:** Predicts expected values and flags significant deviations
- **PCA:** Reduces dimensionality to identify outliers in transformed space
- **Time-Series Decomposition:** Separates data into trend, seasonality, and residual components
- **Mahalanobis Distance:** Measures distance considering data covariance
- **Exponential Smoothing Models:** Uses weighted averages of past observations for forecasting

4. Machine Learning Techniques

- **K-Nearest Neighbors (KNN):** Anomalies are distant from their nearest neighbors.
- **Support Vector Machines (SVM):** Identifies data outside the normal data boundary.
    - One-Class SVM
- **Isolation Forest:** Isolates anomalies using random partitioning.
- **Local Outlier Factor (LOF):** Measures local deviation of density.
- **Clustering Algorithms:** Anomalies don't fit into any cluster.
- **Neural Networks:** Learns patterns; anomalies deviate from learned norms.
- **Gaussian Mixture Models (GMM):** Models data as a mixture of Gaussians.
- **Bayesian Networks:** Uses probabilistic models for anomaly detection.

5. Deep Learning Approaches

- **Autoencoders:** Learn to reconstruct input; anomalies have higher reconstruction error.
- **Recurrent Neural Networks (RNNs):** Capture temporal patterns in sequential data.
- **Generative Adversarial Networks (GANs):** Generate data to model normal patterns; anomalies deviate significantly.
- **Convolutional Neural Networks (CNNs):** Extract spatial features for anomaly detection in images.
- **Variational Autoencoders (VAEs):** Probabilistic modeling of data distribution for anomalies.
- **Deep Belief Networks (DBNs):** Layered networks that learn hierarchical representations.
- **Attention Mechanisms:** Focus on relevant parts of data sequences.
- **Graph Neural Networks (GNNs):** Model relational data for anomaly detection in graphs.
- **Hybrid Models:** Combine deep learning with other techniques for improved detection.

6. Applications Across Industries

- **Finance:** Fraud detection in transactions
- **Healthcare:** Anomaly detection in patient data
- **Manufacturing:** Identifying defects in production processes
- **Cybersecurity:** Intrusion and threat detection
- **Retail:** Detecting anomalies in sales and inventory
- **Energy:** Monitoring irregularities in consumption patterns
- **Telecommunications:** Network fault and outage detection
- **Transportation:** Analyzing anomalies in traffic patterns
- **Social Media:** Spam and bot activity detection
- **Environmental Monitoring:** Identifying anomalies in sensor data

7. Evaluation Metrics and Validation

- Precision and Recall
- F1-Score
- Receiver Operating Characteristic (ROC) Curve
- Area Under the Curve (AUC)
- Confusion Matrix Analysis
- Mean Squared Error (MSE)
- Threshold Selection Methods
- Cross-Validation Techniques
- Handling Imbalanced Datasets
- Importance of Ground Truth Data

8. Challenges and Limitations

- High Dimensionality of Data
- Lack of Labeled Anomaly Data
- Imbalanced Datasets
- Defining "Normal" vs. "Anomalous"
- Real-time Processing Constraints
- Data Privacy and Security Concerns
- Noise and Outliers in Data
- Scalability Issues
- Computational Complexity
- Model Interpretability

9. Tools and Libraries Overview

- scikit-learn
- TensorFlow
- Keras
- PyTorch
- Statsmodels
- PyOD (Python Outlier Detection)
- ELKI
- RapidMiner
- Apache Spark MLlib
- MATLAB Toolboxes

11. When AD Models Are the Best Choice

- **Z-Score Method (Statistical Approach):**
  - Best when you have univariate, normally distributed data and need a simple, quick method for detecting outliers based on statistical thresholds.
- **Isolation Forest:**
  - Ideal for high-dimensional datasets where anomalies are rare and significantly different in terms of feature values from the normal instances.
- **Local Outlier Factor (LOF):**
  - Optimal when anomalies are local and you need to consider the density of data points, making it suitable for detecting anomalies in clusters.
- **One-Class Support Vector Machine (OCSVM):**
  - Suitable when you have data with only the normal class available and want to identify novelties that differ from this class.
- **Autoencoders (Neural Networks):**
  - Best for complex, high-dimensional data where you can learn a compressed representation and detect anomalies based on reconstruction error.
- **LSTM Autoencoders:**
  - Ideal for sequential or time-series data where capturing temporal dependencies is crucial for identifying anomalies.