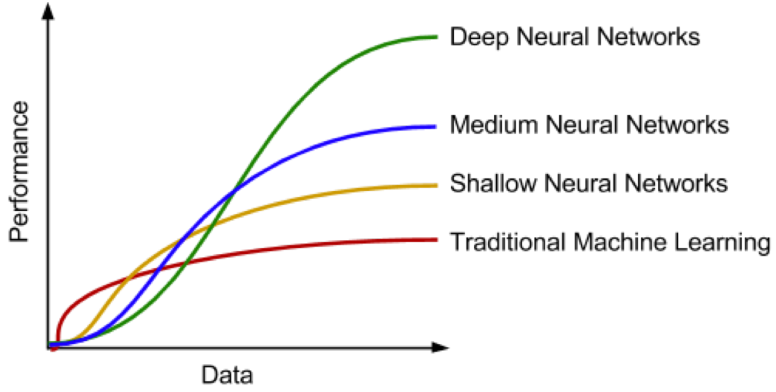
# Anomaly Detection

Hunter

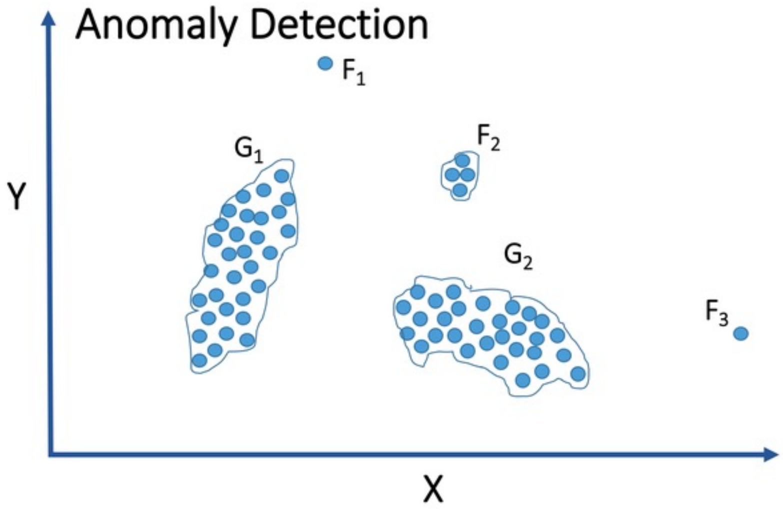
09/03/2023

# Agenda

- Anomaly detection problems
    - Definition
    - Use cases
- Anomaly detection algorithms
    - nearest-neighbor based
    - Clustering based
    - Statistical
    - Classifier based
    - Tree based algorithms
- Deep learning in anomaly detection
    - Autoencoder
    - Deep Belief Networks (DBNs)
    - LSTM
- Anomaly detection on large scale data



# Anomaly Detection

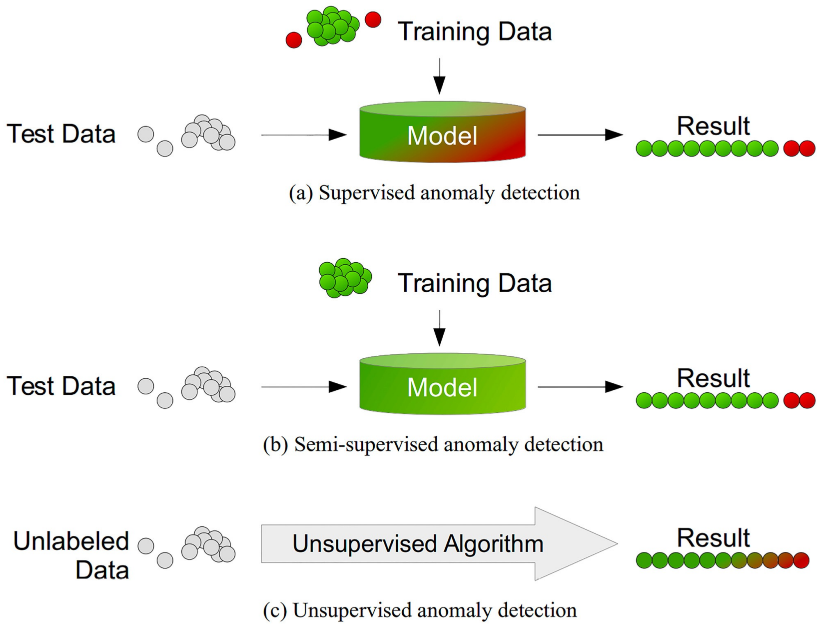



- **Definition**
  - Anomalies are **different** from the norm with respect to their features
  - They are **rare** in a dataset compared to normal instances.
- **Use Cases**
  - **Intrusion detection:** network-based intrusion, host-based intrusion, behavior analysis, commercial intrusion system
  - **Fraud detection:** misuse of a system, suspicious events, financial transactions
  - **Medical applications:** patient monitoring, IOT, medical image analysis
  - **Specialized Applications:** surveillance camera data, energy consumption anomalies, mobile communication monitored

# Traditional Anomaly Detection Algorithms

# Anomaly detection types

- Supervised anomaly detection:
  - Fully labeled data, Unbalanced data
  - Decision tree, SVM, ANN
  - Model output: label
- Semi-supervised anomaly detection:
  - Training data are normal data without anomalies.
  - One-class SVM, autoencoders, Guassian mixture models, kernel density estimation
  - Model output: score
- Unsupervised anomaly detection:
  - Not require labels, use distance or densities to estimate anomalies.
  - Nearest-neighbor based, clustering based, statistical algorithms, subspace techniques, neural networks, SVM.
  - Model output: score

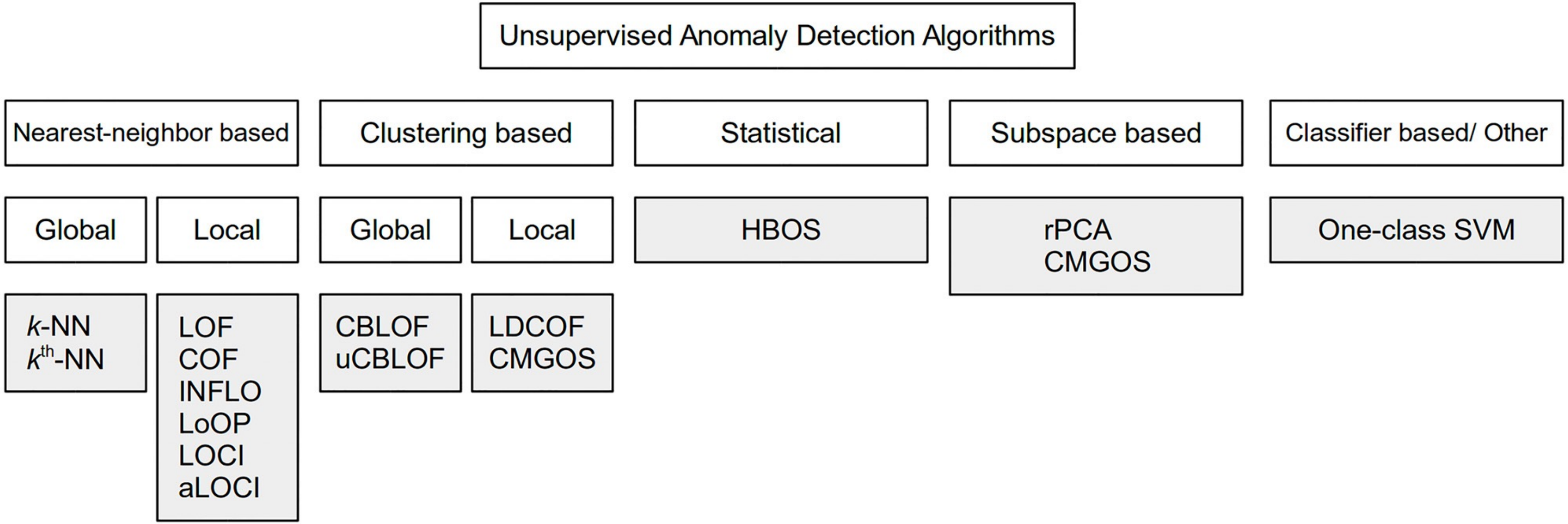



(a) Supervised anomaly detection

(b) Semi-supervised anomaly detection

(c) Unsupervised anomaly detection

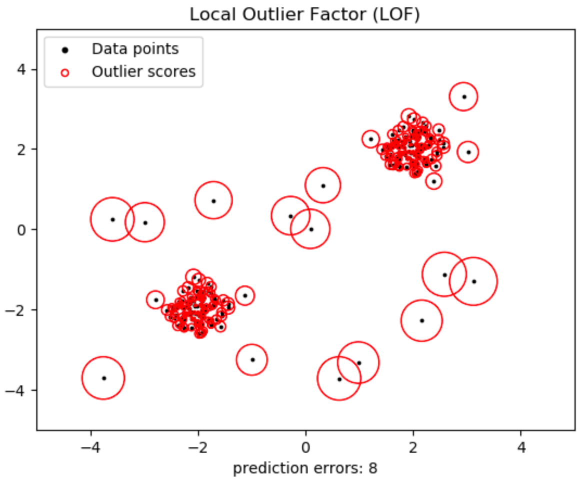
# Unsupervised Anomaly Detection



# Nearest-Neighbor Based

- **k-NN Global Anomaly Detection**
- **Local Outlier Factor (LOF)**
- **Connectivity-Based Outlier Factor (COF)**
- **Influenced Outlierness (INFLO)**
- **Local Outlier Probability (LoOP)**
- **Local Correlation Integral (LOCI)**
- **Approximate Local Correlation Integral (aLOCI).**

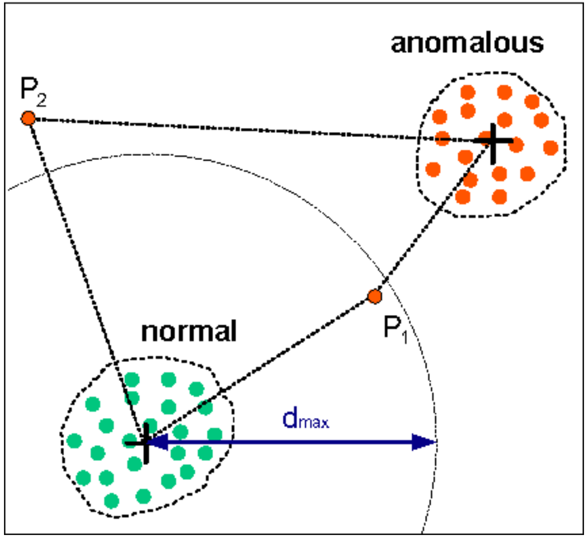
- LOF:
  - Measures the local density
  - Lower density is outlier



# Clustering Based

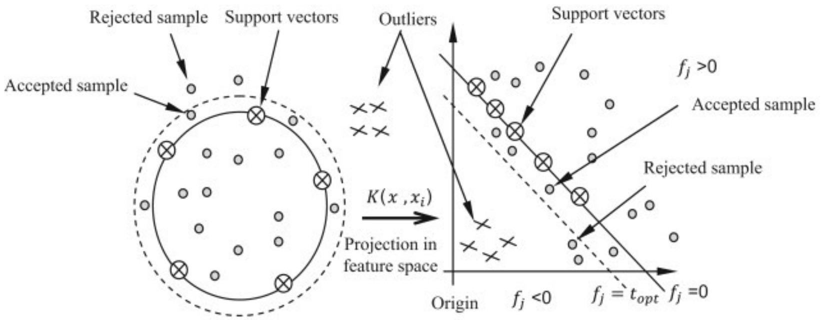
- **Cluster-Based Local Outlier Factor (CBLOF/ uCBLOF)**
- **Local Density Cluster-based Outlier Factor (LDCOF)**
- **Clustering-based Multivariate Gaussian Outlier Score (CMGOS)**

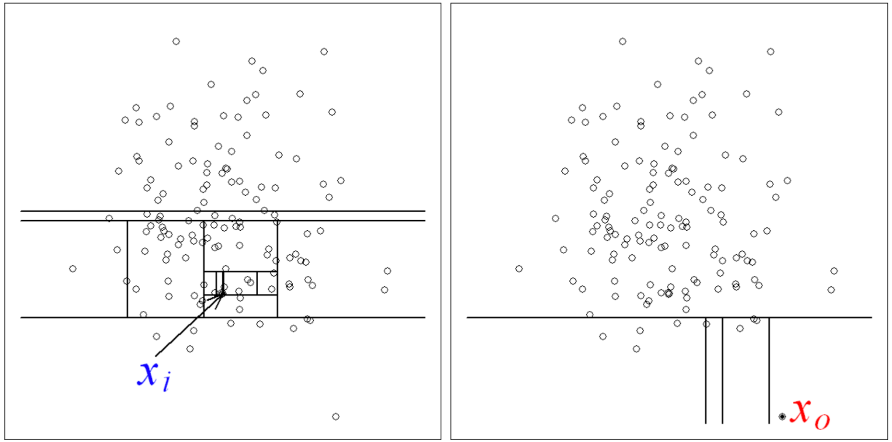
- **K-means**
  - Get the distance between each point and its nearest centroid.
  - The biggest distances are considered as anomaly.



# Other Based

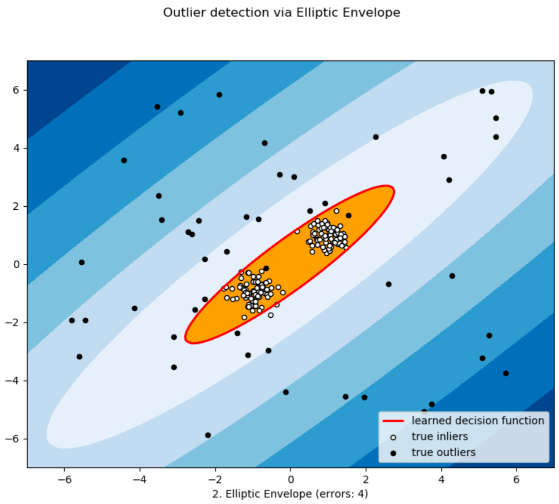
- **One-Class SVM**
  - Only train on ‘normal’ class to build the boundaries.
  - Minimize the radius of hyperball
  - Non-linear kernel (RBF)
- **Isolation Forest**
  - How many nodes to isolated a single point.
  - Using the fewer nodes is the outlier.

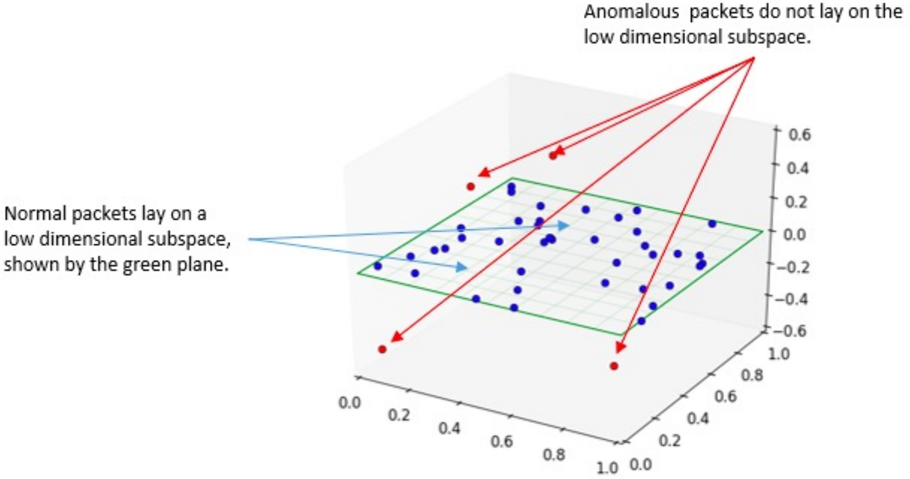


# Other Based

- **Histogram-based Outlier Score (HBOS)**
  - Assume normal distributed
  - Build ellipse based on the distribution
- **Robust Principal Component Analysis (rPCA)**
  - Reduce dimension
  - Measure the distance of each observation from the center of the data for anomaly detection.



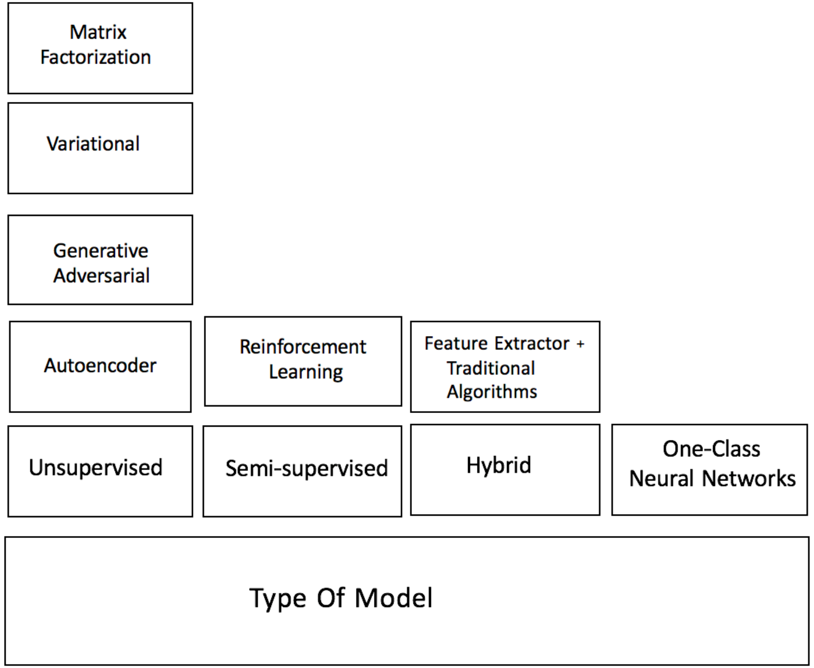
# Computational Complexity

- **$O(n^2)$:** All nearest neighbor based algorithms except LOCI.
- **$O(n^3)$:** LOCI
- **Faster than $O(n^2)$:** Clustering based
- **Faster than clustering:** HBOS
- **Depends on support vectors:** One-class SVM
- **$O(d^2n+d^3)$:** rPCA

# Deep Anomaly Detection Algorithms

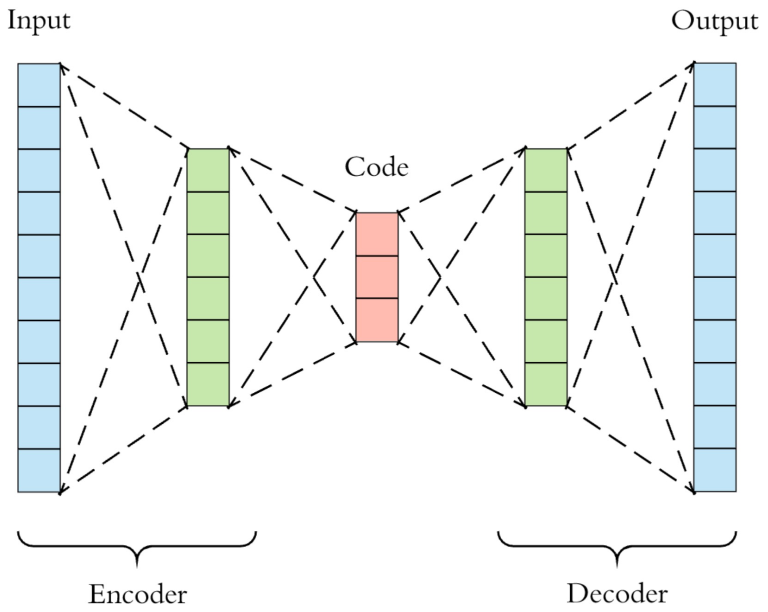
# Deep Anomaly Detection (DAD)

- **Advantages of DAD**
  - Traditional algorithms fail to capture complex structures in image (e.g. medical images) and sequence datasets.
  - Traditional methods is hard to scale to large scale data to find outliers.
  - DAD can automatically capture features. Thus, eliminate the need of developing manual features.
  - As the data size increases, the DAD outperforms traditional anomaly detection algorithms.



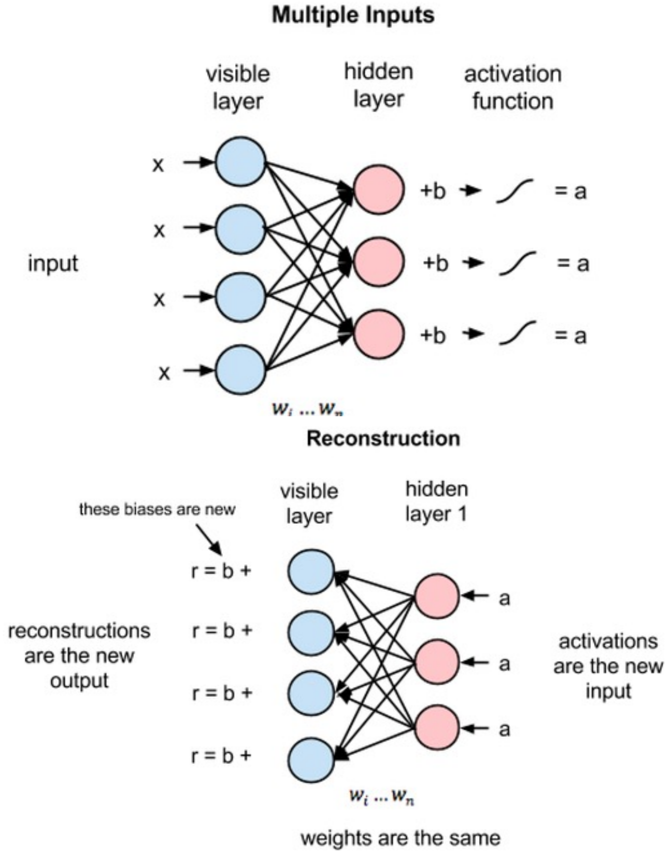
# Autoencoder

- Autoencoders are the core of all Unsupervised DAD models
  - Use encoder to transfer input to a hidden layer (code)
  - use decoder to reconstruct the code as output
  - the optimization is to reduce the error between input and output.
  - Anomalies are the data that have high difference between input and output
- Deep Hybrid Model:
  - Use deep neural networks mainly autoencoder to extract features.
  - Input features to traditional anomaly detection algorithms to detect outliers



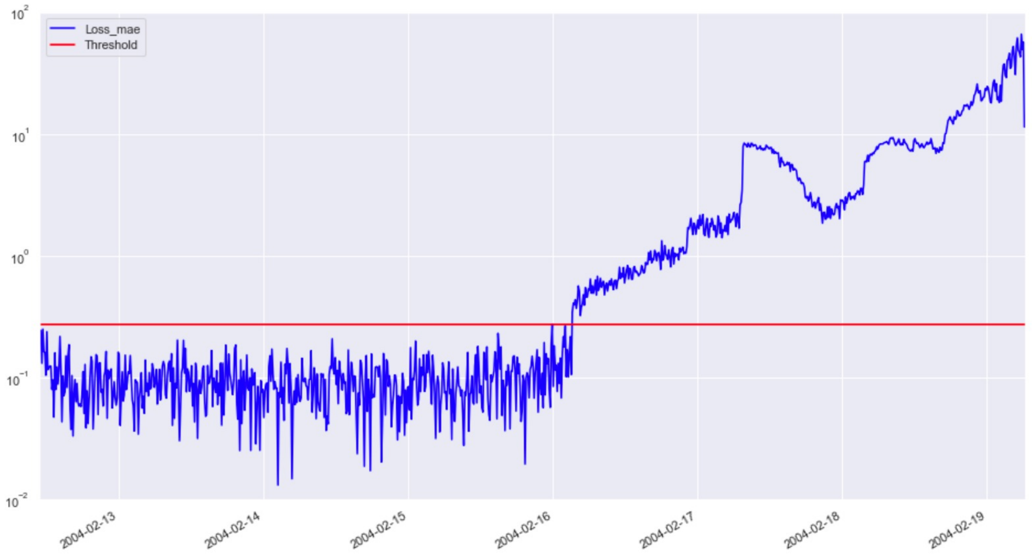
# Restricted Boltzmann Machine

- Forward: the sum of Xs multiply by weights, plus a bias, apply the activation function.
- Backward: use activations as input, multiply by same weights, plus a new bias
- Calculate the reconstruction error to identify anomalies
- Deep belief network (DBN) is a network consists of several middle layers of Restricted Boltzmann machine (RBM) and the last layer as a classifier.



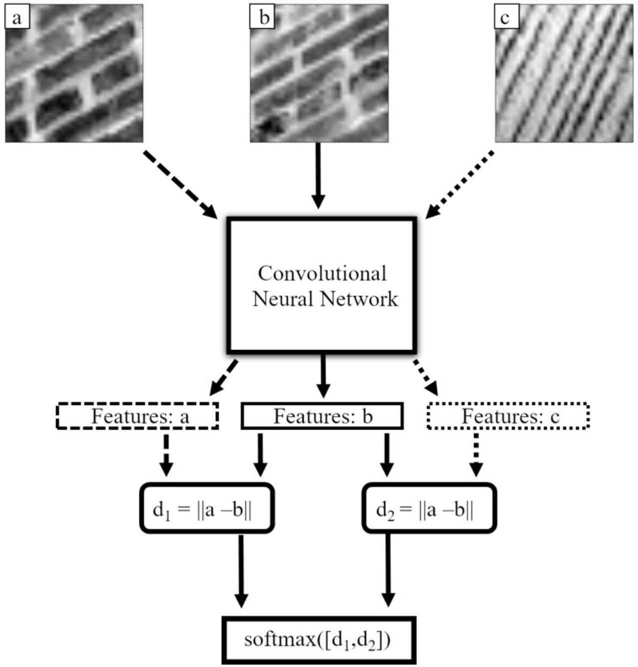
# LSTM

- Apply to time series data.:
    - Use LSTM and normal data to build a prediction model
    - Predict next few steps in the time series
    - Use the error in prediction as anomaly score to identify anomalies.



# CNN

- Apply to image data:
  - Extracting features from normal images and calculating the mean feature values.
  - Calculate the feature distance between training data and new data.
  - The data that have the largest distance are anomalies.



# Anomaly Detection on Large Scale Data

# DAD Applications on Large Scale Data

- Fraud Detection: Detecting a deliberate act of deception to access valuable resources
- Intrusion Detection: Identifying malicious activity in a computer-related system
- Medical Anomaly Detection: Detecting prohibited drug name mention and fraudulent health-care transactions
- Social Networks Anomaly Detection: capturing irregular often unlawful behavior pattern of individuals within a social network
- Internet Of Things (IoT) Big-data Anomaly Detection: identifying fraudulent, faulty behavior of massive scales of interconnected devices
- Log-Anomaly Detection: indicating the reasons and the nature of the failure of a system
- Video Surveillance : monitoring designated areas of interest in order to ensure security
- Industrial Damage Detection: detecting the damage of wind turbines, power plants, and high-temperature energy systems

# Different Data Types

**Sequential Data**

- Data:
    - Video
    - Speech
    - Protein Sequence
    - Timeseries
    - Text
- DAD model:
    - LSTM
    - RNN
    - CNN

**Non-Sequential Data**

- Data:
    - Image
    - Sensor
- DAD models:
    - CNN
    - AE

# References


- Borghesi, A., Bartolini, A., Lombardi, M., Milano, M., & Benini, L. (2019, July). Anomaly detection using autoencoders in high performance computing systems. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 33, pp. 9428-9433).
- Breunig, M. M., Kriegel, H. P., Ng, R. T., & Sander, J. (2000, May). LOF: identifying density-based local outliers. In *Proceedings of the 2000 ACM SIGMOD international conference on Management of data* (pp. 93-104).
- Chalapathy, R., & Chawla, S. (2019). Deep learning for anomaly detection: A survey. *arXiv preprint arXiv:1901.03407*.
- Chandola, V., Banerjee, A., & Kumar, V. (2009). Anomaly detection: A survey. *ACM computing surveys (CSUR)*, *41*(3), 1-58.
- Fiore, U., Palmieri, F., Castiglione, A., & De Santis, A. (2013). Network anomaly detection with the restricted Boltzmann machine. *Neurocomputing*, *122*, 13-23.
- Goldstein, M., & Uchida, S. (2016). A comparative evaluation of unsupervised anomaly detection algorithms for multivariate data. *PloS one*, *11*(4), e0152173.
- Guerbai, Y., Chibani, Y., & Hadjadji, B. (2015). The effective use of the one-class SVM classifier for handwritten signature verification based on writer-independent parameters. *Pattern Recognition*, *48*(1), 103-113.
- Liu, F. T., Ting, K. M., & Zhou, Z. H. (2008, December). Isolation forest. In *2008 Eighth IEEE International Conference on Data Mining* (pp. 413-422). IEEE.
- Malhotra, P., Vig, L., Shroff, G., & Agarwal, P. (2015, April). Long short term memory networks for anomaly detection in time series. In *Proceedings* (Vol. 89). Presses universitaires de Louvain.
- Münz, G., Li, S., & Carle, G. (2007, September). Traffic anomaly detection using k-means clustering. In *GI/ITG Workshop MMBnet* (pp. 13-14).
- Paffenroth, R., Kay, K., & Servi, L. (2018). Robust pca for anomaly detection in cyber networks. *arXiv preprint arXiv:1801.01571*.
- Staar, B., Lütjen, M., & Freitag, M. (2019). Anomaly detection with convolutional neural networks for industrial surface inspection. *Procedia CIRP*, *79*(1), 484-489.