OFFICE OF CHEMICAL SECURITY

# AI-DRIVEN INTEGRATED FRAMEWORK FOR DATA ANOMALY ASSESSMENT

**Dennis Park, PhD**
September 19, 2024

# Agenda

- Background
- Problem statement, research questions, and research approach
- Review on AI & data quality dimensions (4 sections)
- Conceptual integrated framework
- Discussion

# Background

1. EO 14110, “Safe, Secure, and Trustworthy development and use of AI”, signed on Oct. 2023.

2. DHS AI Policy 139-06, “Acquisition and Use of Artificial Intelligence and Machine Learning Technologies by DHS Components”, signed on Aug. 2023.

3. CISA published the roadmap for AI to be aligned with national AI strategy on Nov. 2023 as directed by EO 14110 and DHS AI Policy 139-06.

4. NSA for AISC in collaboration with Cybersecurity Agencies like CISA, FBI, Canadian, Australian, and New Zealand published “Deploying AI Systems Securely.”

# Background

5. CISA directives mandate maintaining data quality on the data assets.

6. CISA directives also outline 7 dimensions and metrics to ensure high data quality.

7. The Office of Chemical Security has a regulatory data asset and is making efforts to maintain it to its highest data quality.

- Question:
  - Given these data assets, is there anyway that AI can help to address data quality issues within those 7 dimensions?

# Problem Statement

While AI techniques have been widely researched and applied in various domains, there remains a gap in understanding their effectiveness in assessing the entry of bad data into systems. Additionally, the integration of data quality dimensions within AI-driven approaches has received limited attention.

# Research Questions & Goals/Objectives

RQ 1. How do various AI techniques compare in their ability to assess the entry of bad data?

Goal/Objective: Conduct a high-level comparison of various AI techniques regarding their ability to assess the entry of bad data.

RQ 2. How can AI techniques be integrated to measure the entry of bad data across CISA's data quality dimensions (e.g., accuracy, completeness, consistency)?

Goal/Objective: Explore an integrated approach to utilizing AI techniques for measuring the entry of bad data across CISA's data quality dimensions.

# Research Approach

To address RQ 1,

- Define the scope of AI techniques;
- Examine benefits and limitations of those AI techniques within the scope of data quality assessment;
- Select one or two AI techniques that may be suitable for data quality assessment implementation.

# Research Approach

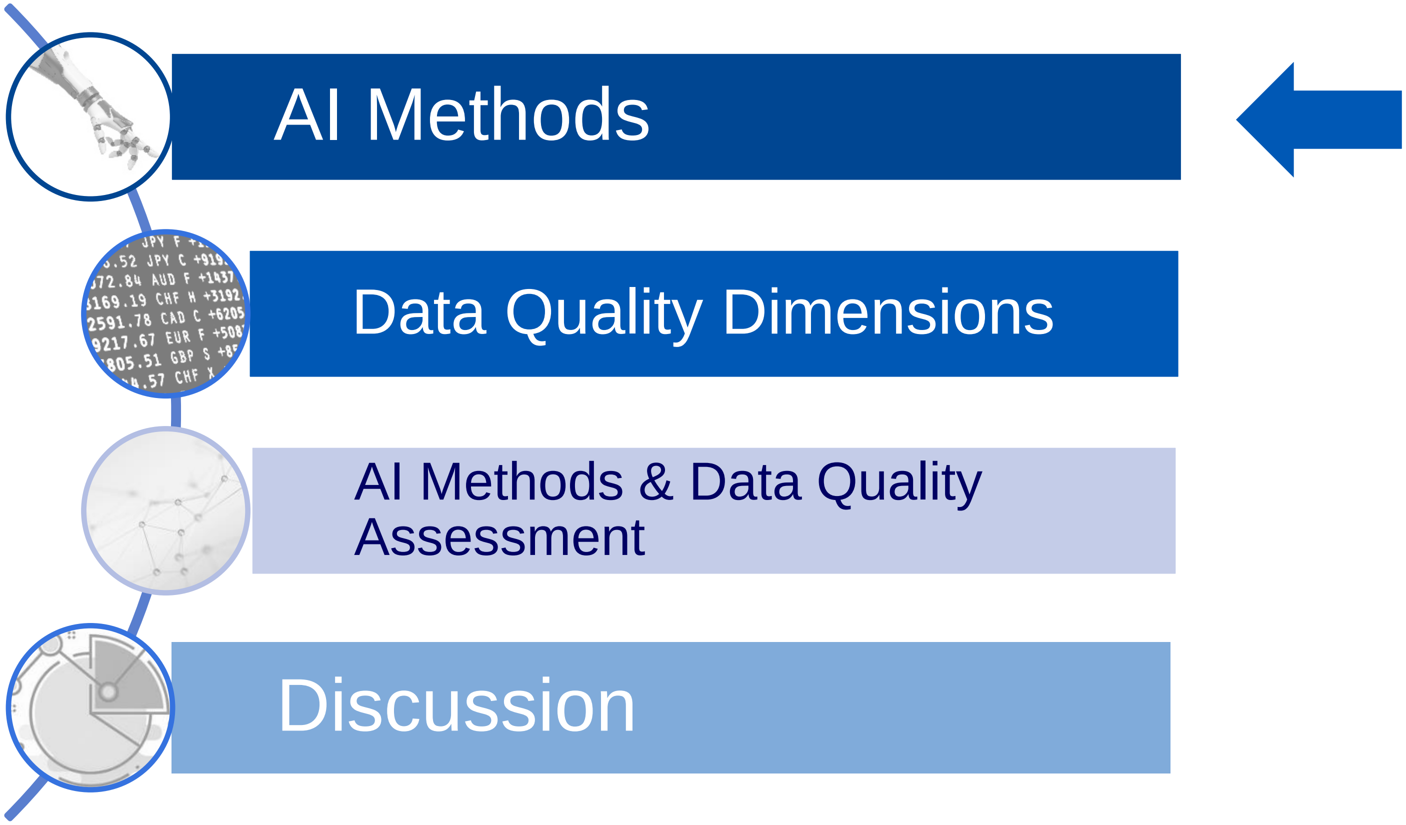
To address RQ 2,

- Identify the scope of data quality dimensions;
- Demonstrate how the entry of bad data can be assessed in an integrated fashion, involving one or two identified AI techniques as candidates;
- Demonstrate the steps or procedures to apply those selected AI techniques from RQ 1 to the data quality dimensions;
- Suggest the visualization of the end results.

# Disclaimer

This presentation will NOT cover the cost specific aspects of AI / ML acquisition or the technical details of AI / ML technologies. Instead, it will focus on how AI techniques can help identify data issues to improve data quality and demonstrate the value that AI technologies can provide when acquired from vendors.

# Presentation: Four Key Sections



AI Methods

Data Quality Dimensions

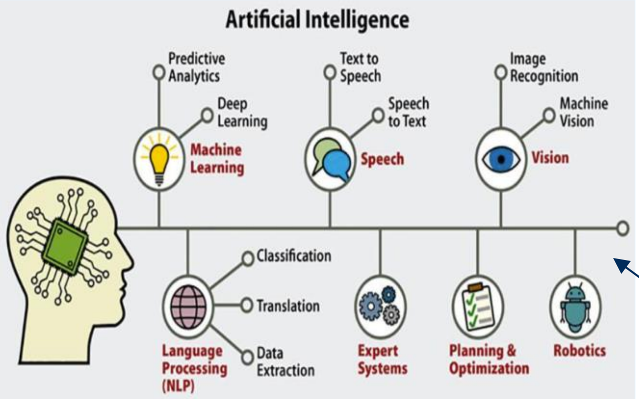
AI Methods & Data Quality Assessment

Discussion

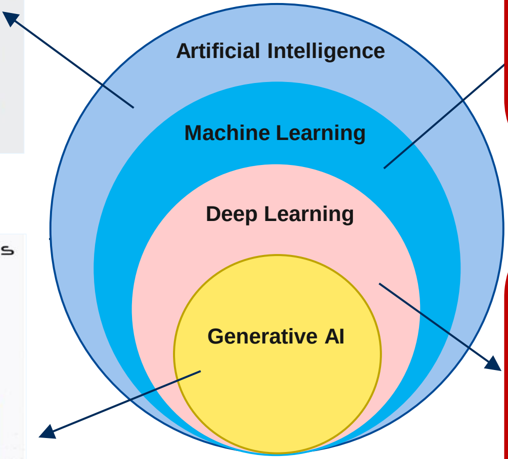
# AI Methods

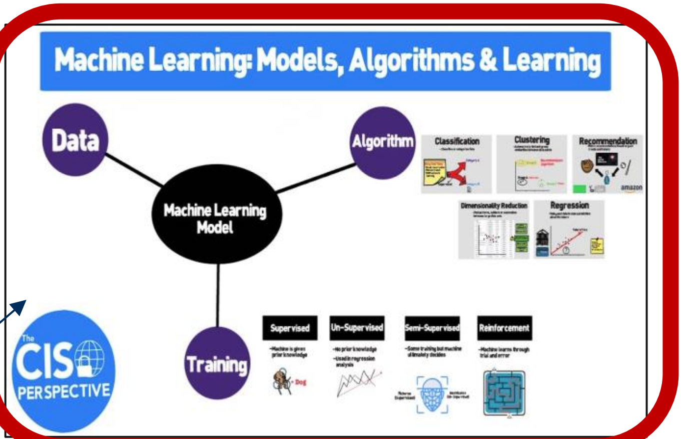
- Subset of the Agena
  - AI Ecosystem
  - Definition and Types of Machine Learning
  - Models of Supervised & Unsupervised Learning
  - (Scope) Eight (8) AI Methods for Anomaly Detection
  - (Comparison & Results) Evaluation of Outcomes from AI Methods

# AI Ecosystem

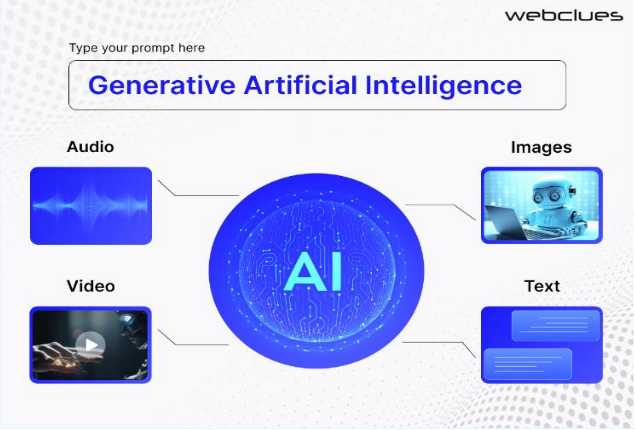


Emulate human behaviors



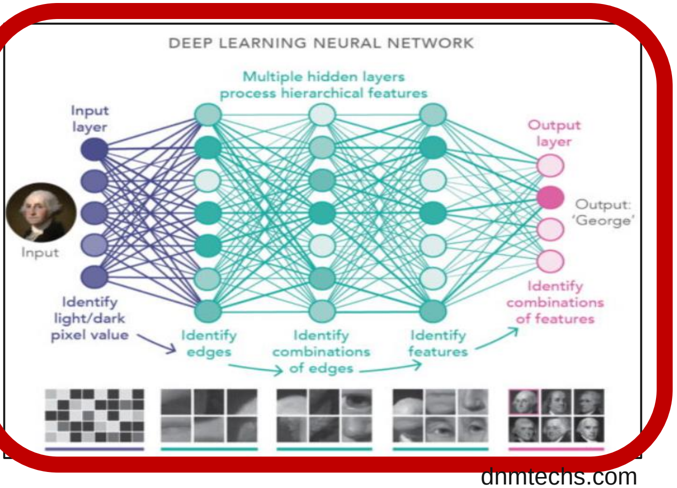


Algorithms, training, data patterns



webcluesinfotech.com

Generate contents

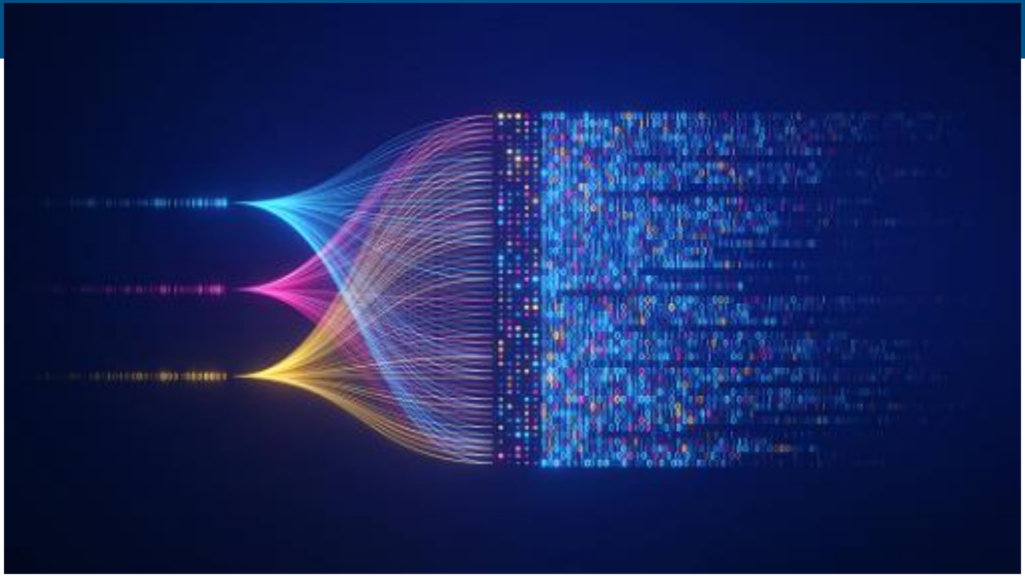

DEEP LEARNING NEURAL NETWORK
Multiple hidden layers process hierarchical features
Input layer
Output layer
Input
Output: 'George'
Identify light/dark pixel value
Identify edges
Identify combinations of edges
Identify features
Identify combinations of features

dnmtechs.com

Neural networks, multi-layers

**Office of Chemical Security**
September 19, 2024

# What’s Machine Learning (ML)?



- When there is a specific job for a computer to do and a way to check how well it does the job, we say the computer is learning if it gives better results over time. This process is called *Machine Learning*.
- *Machine Learning* can be classified into two main categories according to the properties of the given data.

# Types of Machine Learning (ML)

- Supervised learning
    - We know the target value ($Y$) for the given Data. Using the given data ($X$), the goal is to find out the relationship between $X$ and $Y$, and the function $Y = f(x)$ that can represent it. This includes Classification and Regression.
- Unsupervised learning
    - The given data is not given with a target value ($Y$). However, the goal is to find out the feature $f(x)$ that data ($X$) itself has. Using the given data ($X$), we estimate the density of individual objects, find clusters of objects, or infer associations between variables.
- Semi-supervised learning & reinforcement learning

# Models of Supervised & Unsupervised Learning

- **Supervised Learning**
- Classification
  - Logistic Regression
  - Decision Trees
  - Support Vector Machines (SVM)
  - K-Nearest Neighbors (KNN)
  - Random Forests
  - Neural Networks
  - Naïve Bayes
- Regression
  - Linear Regression
  - Polynomial Regression
  - Support Vector Regression
  - Decision Trees and Random Forests Regression
  - Neural Networks

**Unsupervised Learning**

- Clustering
- Association
- Dimensionality Reduction
- Anomaly Detection
- Neural Network-based Approaches

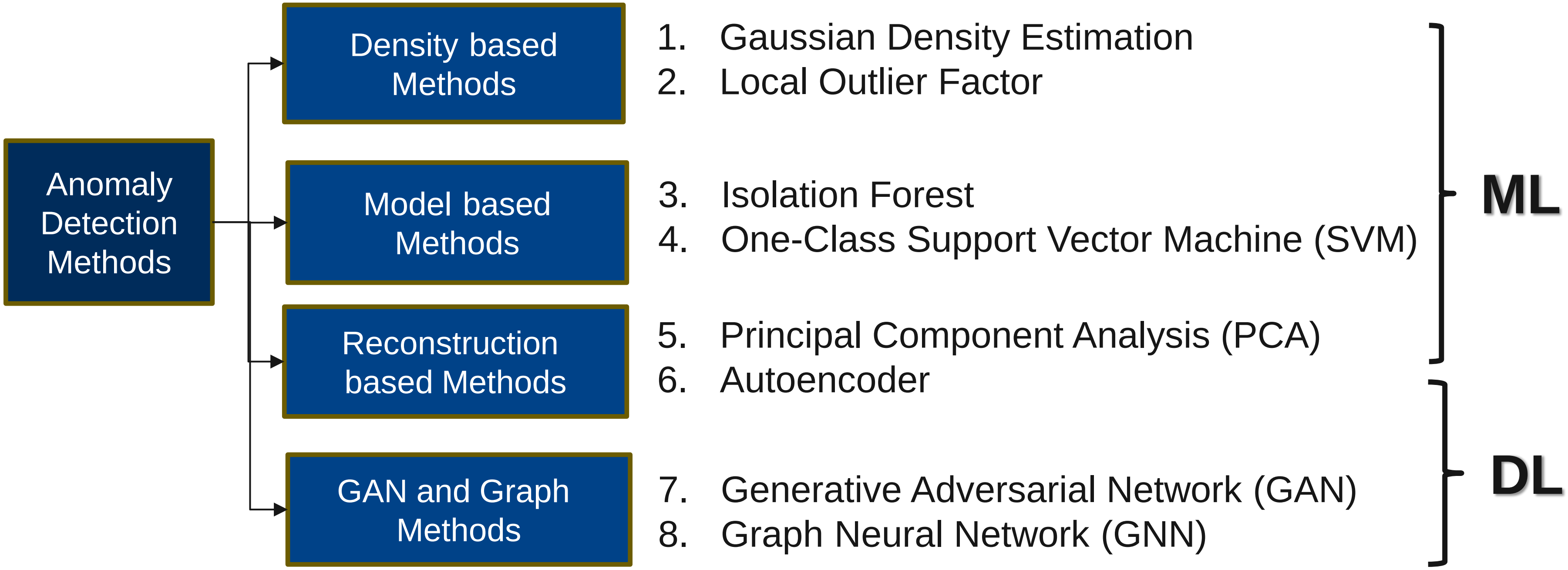
**Anomaly Detection** is the process of identifying rare events or outliers that deviate significantly from the norm in a dataset. These anomalies can signal potential issues or interesting patterns in the data.

Applications:

A. Cybersecurity
B. Fraud Detection
C. Healthcare

# AI Methods for Anomaly Detection

## Existing ML-based and DL-based AI Techniques



Anomaly Detection Methods

Density based Methods

1. Gaussian Density Estimation
2. Local Outlier Factor

Model based Methods

3. Isolation Forest
4. One-Class Support Vector Machine (SVM)

Reconstruction based Methods

5. Principal Component Analysis (PCA)
6. Autoencoder

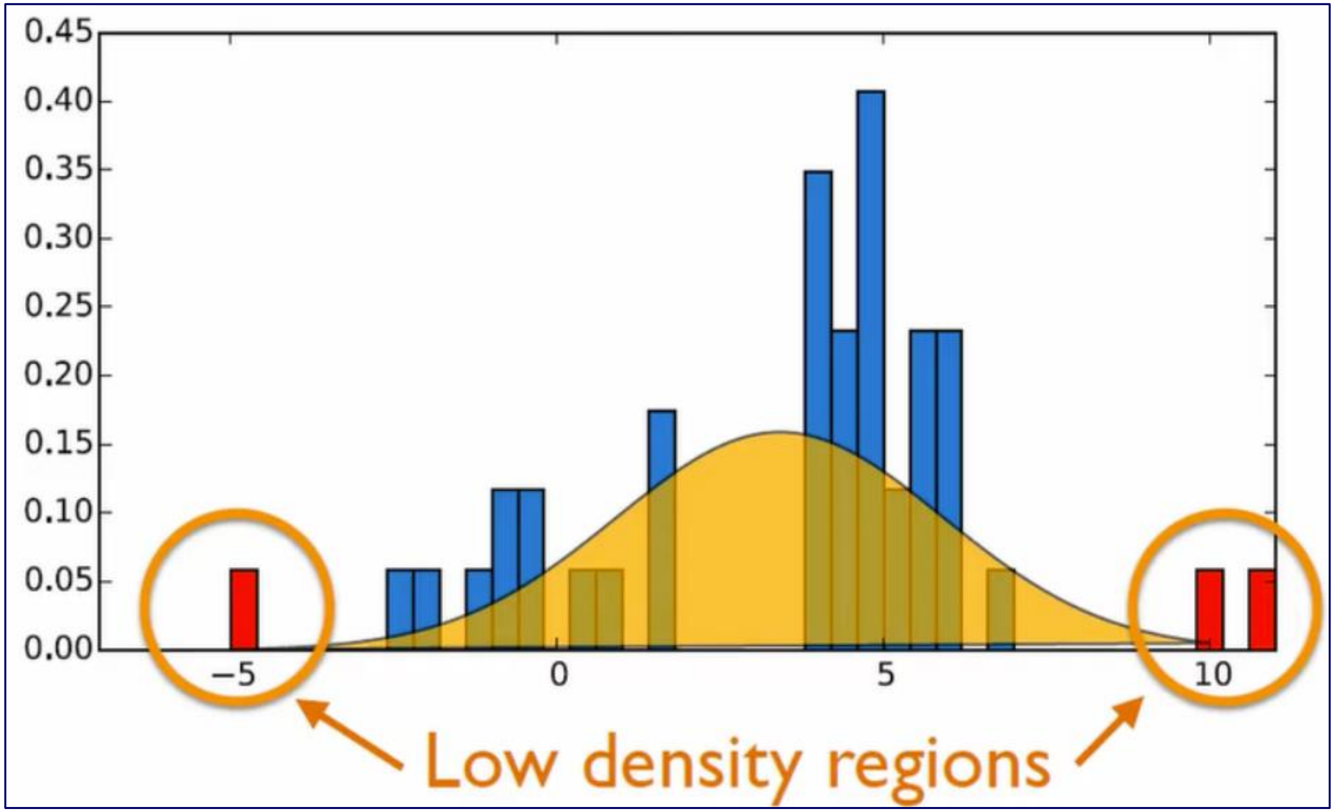
GAN and Graph Methods

7. Generative Adversarial Network (GAN)
8. Graph Neural Network (GNN)

ML

DL

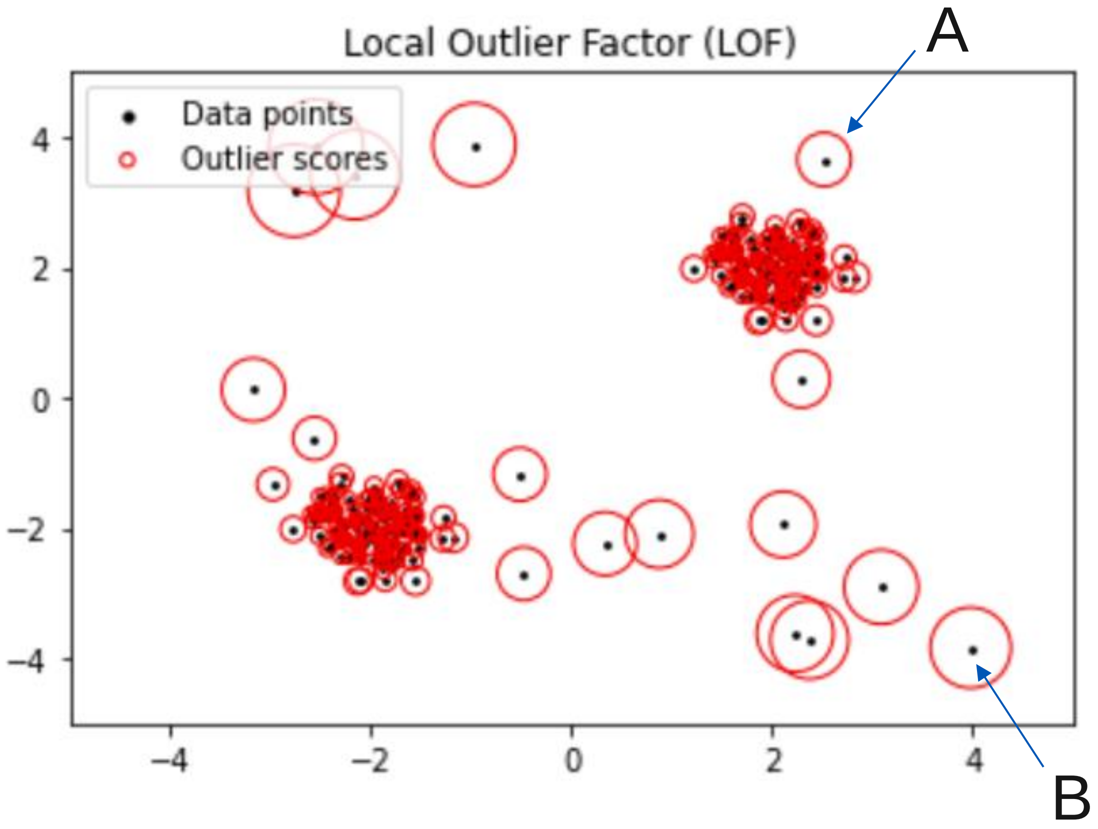
# ML Method #1 – Density Gaussian Density Estimation



- Carl F. Gauss in 1837
- Bell curve with two elongated tails
- Work well if data distribution follows the bell curve
- Ex) Monitoring daily temperature

Source(s):
https://yupsung.blogspot.com/2021/02/gaussian-density-estimation.html
https://towardsdatascience.com/the-basics-of-anomaly-detection

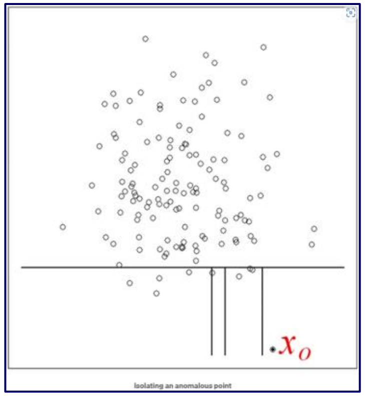
# ML Method #2 – Density Local Outliers Factor (LOF)

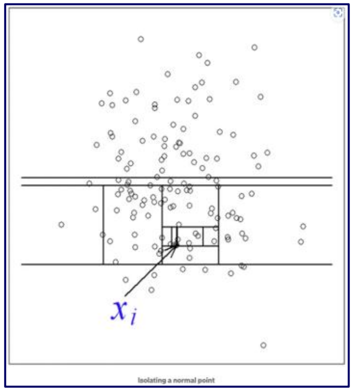


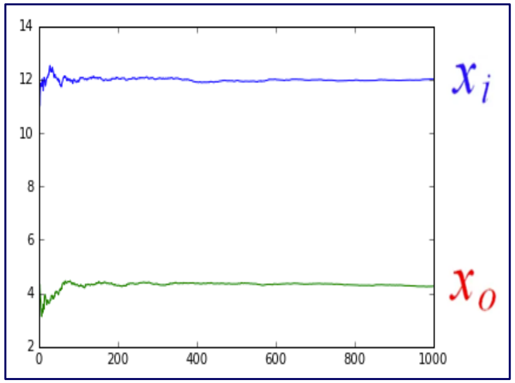
- Data points classified by a score of LOF.
- Based on the concept of "local" density.
- The locality is chosen by a value set to a "k"-nearest.
- Ex) Login attempts with lower density are classified as outliers.

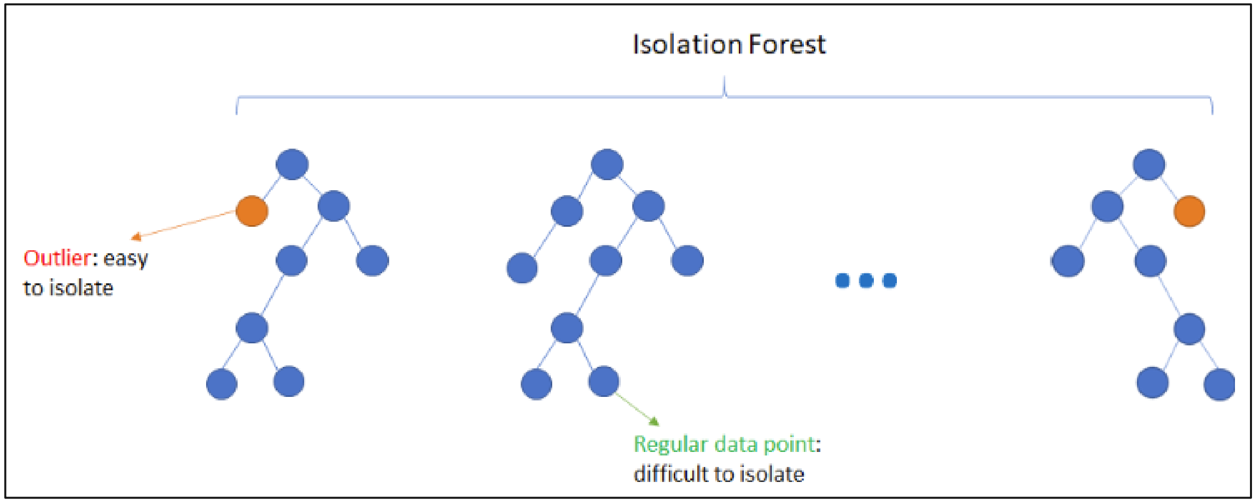
Source(s):
https://campus.datacamp.com/courses/anomaly-detection-in-python/distance-and-density-based-algorithms?ex=14

# ML Method #3 – Model Isolation Forest (aka IF or IForest)





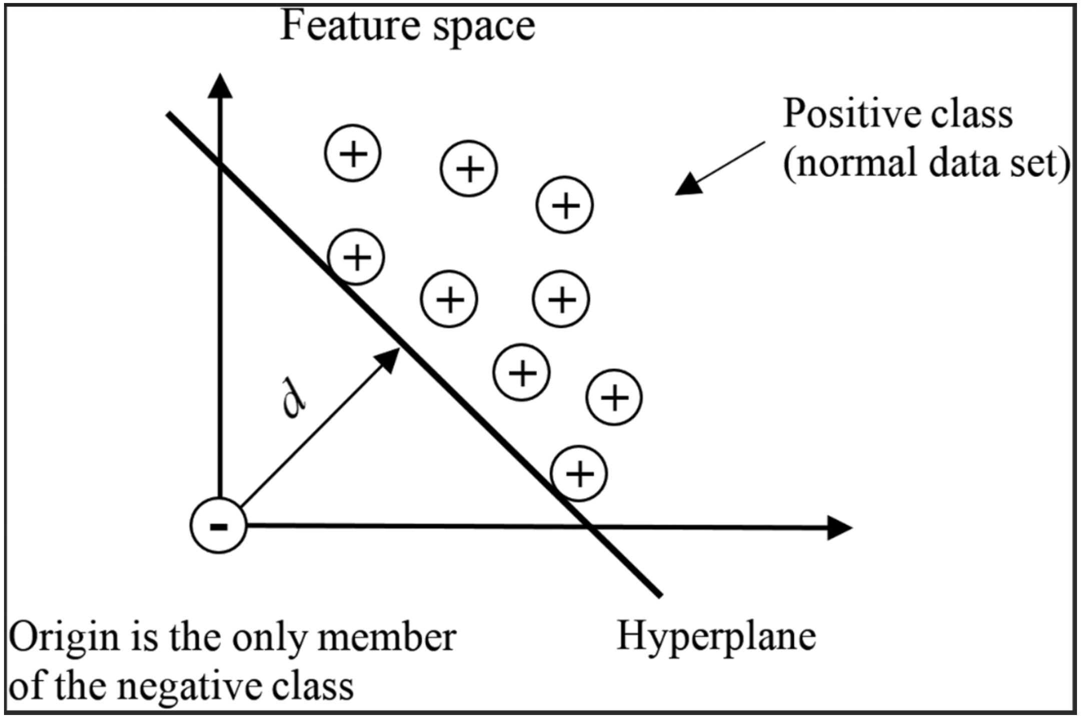




- Forest -> Trees
- Random Trees
- Ensemble Techniques
- Decision Trees
- Build the tree until it reaches a leaf in isolation.
- Ex) Financial transactions: Amount of money spent, location the money spent.

Source(s):
Dairi, Abdelkader at el. (2022, November). Efficient driver drunk detection by sensors: A manifold learning-based anomaly detector. DOI:10.1109/ACCESS.2022.3221145.

# ML Method #4 – Model One-Class Support Vector Machine (SVM)


Feature space
Positive class (normal data set)
+
d
-
Origin is the only member of the negative class
Hyperplane

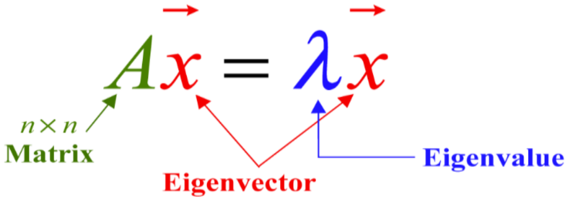
$$\min_{w \in F, \xi \in \mathbb{R}^{\ell}, \rho \in \mathbb{R}} \quad \frac{1}{2}\|w\|^2 + \frac{1}{\nu \ell}\sum_i \xi_i - \rho$$

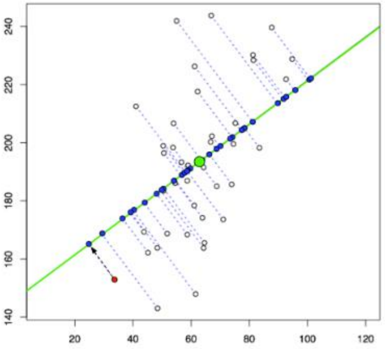
$$\text{subject to} \quad (w \cdot \Phi(\mathbf{x}_i)) \geq \rho - \xi_i, \quad \xi_i \geq 0.$$

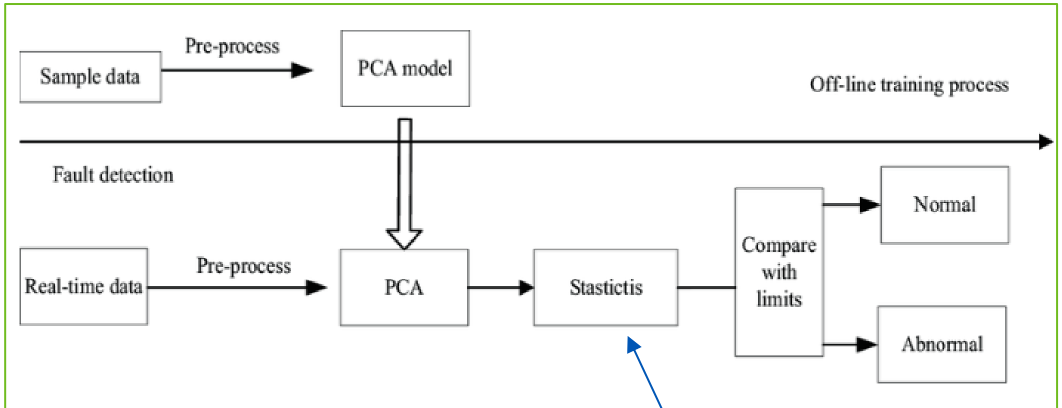
- One of the SVM models
- Only one class (normal dataset)
- The other class (negative) is the origin.
- Find the hyperplane.
- d (distance) is determined by nu ($\nu$) in calculations.
- Ex) Novelty Detection

Source(s):
Scholkopf, Bernhard et al. (n.d.). Support Vector Method for Novelty Detection.
https://machinelearninginterview.com/topics/machine-learning/what-is-one-class-svm-how-to-use-it-for-anomaly-detection/

# ML Method #5 – Reconstruction
# Principal Component Analysis (PCA)







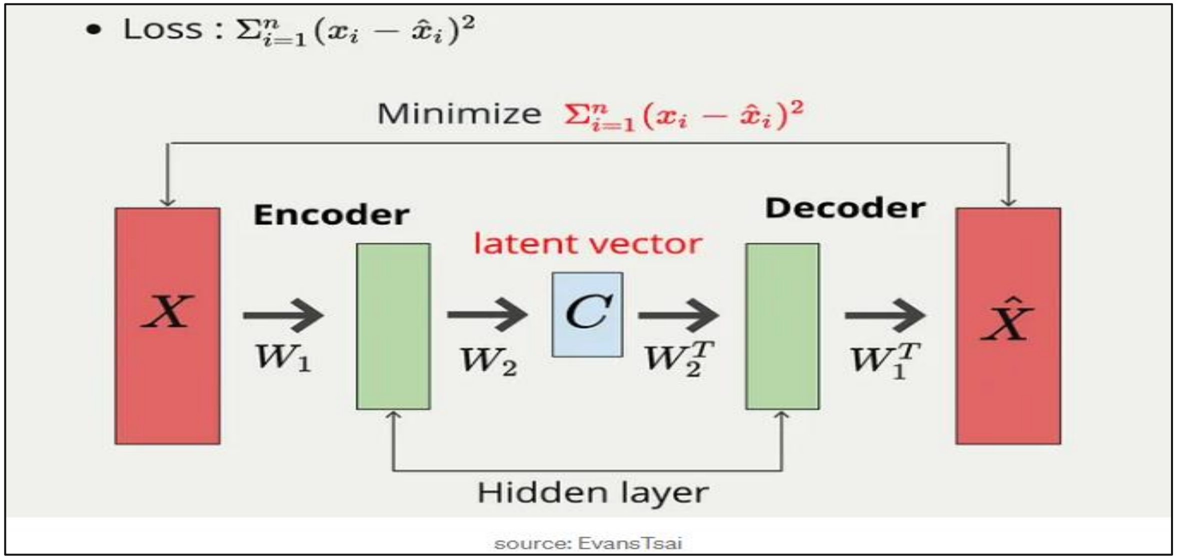
Abnormal detection based on PCA. PCA, principle component analysis.

Thresholding or distance from the mean

- Use eigenvectors (directions of principal components) & eigenvalues (variance of each principal component)
- Use these two to select the components that captures the most variance in the data.
- Use them as a PCA model.

Source(s):
https://www.datacamp.com/tutorial/pca-analysis-r

# DL Method #6 – Reconstruction Autoencoder



• Loss : $\Sigma_{i=1}^{n}(x_i - \hat{x}_i)^2$
Minimize $\Sigma_{i=1}^{n}(x_i - \hat{x}_i)^2$
Encoder
latent vector
Decoder
$X$
$W_1$
$W_2$
$C$
$W_2^T$
$W_1^T$
$\hat{X}$
Hidden layer

source: EvansTsai

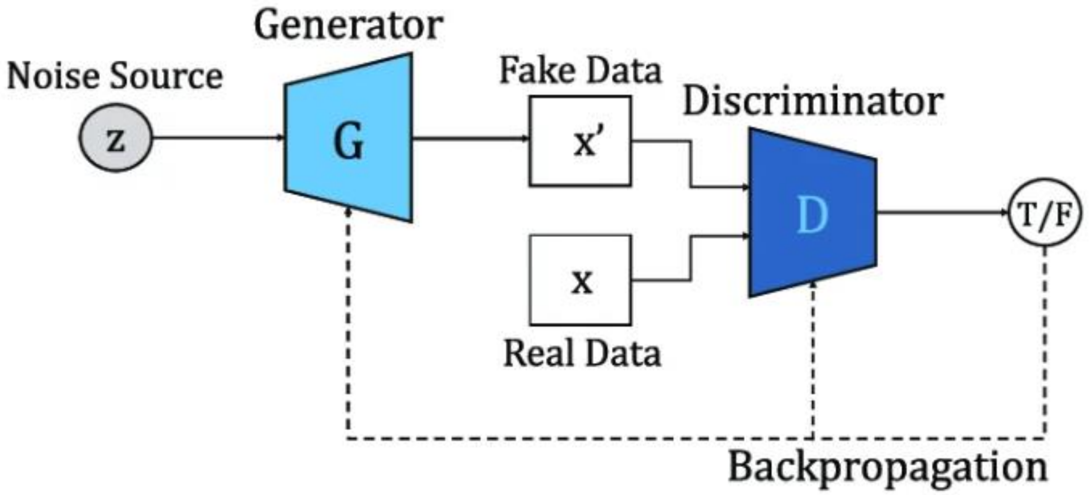
Encoder
Latent vector
Decoder

The difference between the original and the reconstructed ones is the loss, which may be anomalies.

Through training, it forces to capture the most important features.

Source(s):
https://www.geeksforgeeks.org/auto-encoders/

# DL Method #7
## Generative Adversarial Networks (GAN)



Generative:

- Generate fake data

Adversarial (Zero-sum):

- Generator: Fake samples
- Discriminator: Distinguish between fake and real

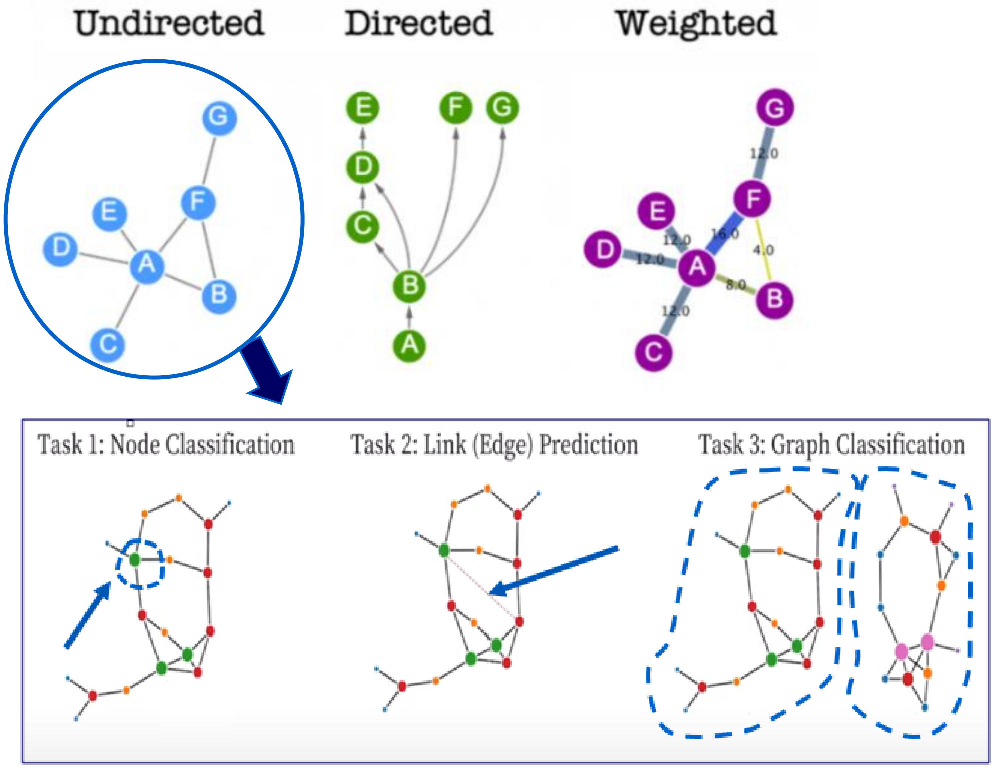
Backpropagation:

- Adjustments

Ex) AnoGAN (Anomaly) Train both G & D on normal data to their equilibrium through many *epochs*.

Source(s): https://medium.com/@sahar.millis/train-a-nn-using-keras-to-fit-the-predator-prey-cycle-using-gan-architecture-73fa69827243

# DL Method #8
# Graph Neural Networks (GNN)


Undirected
Directed
Weighted
G
F
E
D
A
B
C
12.0
16.0
4.0
8.0
Task 1: Node Classification
Task 2: Link (Edge) Prediction
Task 3: Graph Classification

*G = (V, E)*

- V = set of vertices, E = set of edges
- *Undirected graph*:
  - Edge (u,v) = Edge (v,u)
  - No self-loops
- *Directed graph*:
  - Edge (u,v) goes from u to v, notated u -> v
- A *weighted graph*: associates weights with either the edges or the vertices

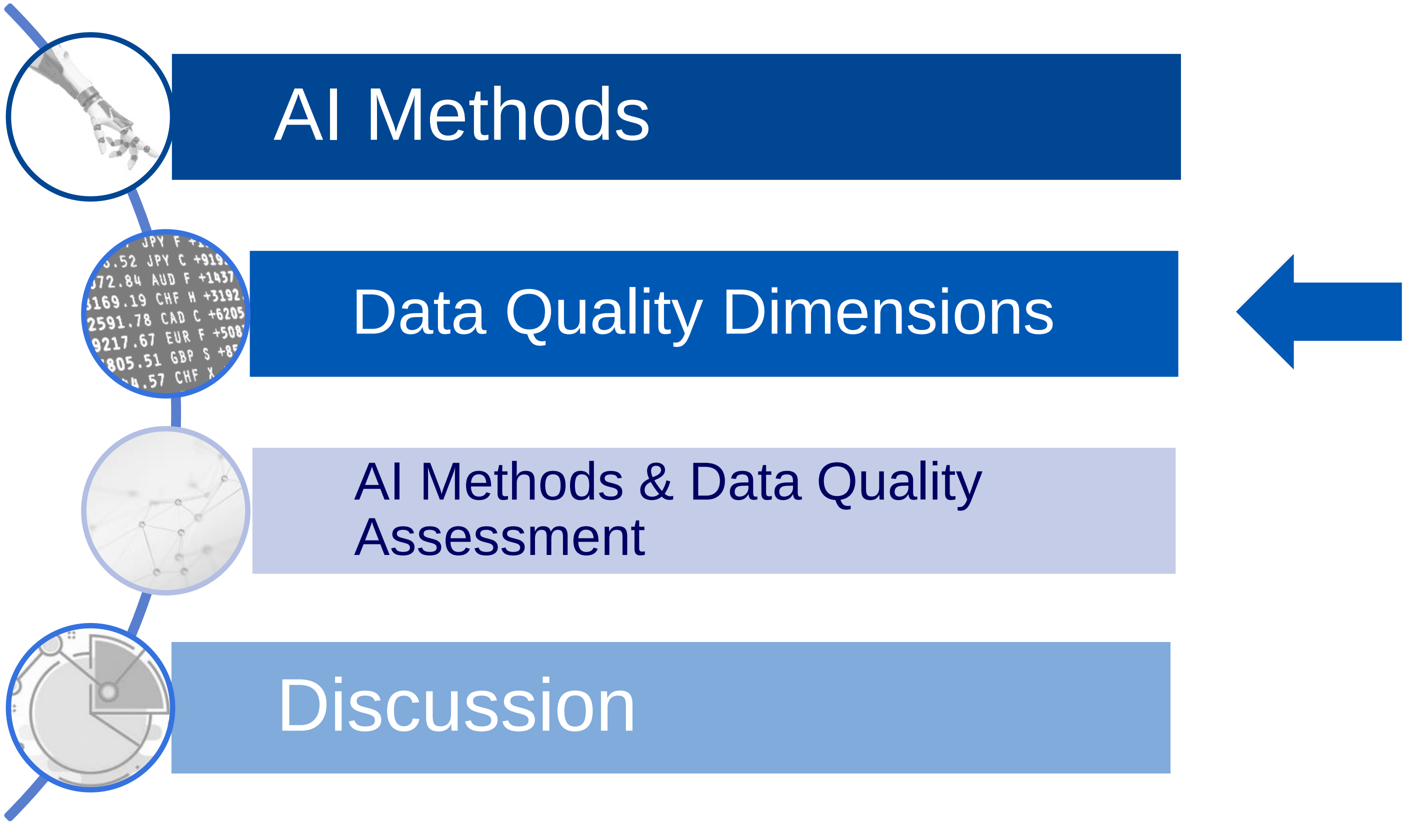
Source(s):
https://studylib.net/doc/11502617/graphs-a-graph-g-=--v--e-
https://www.ebi.ac.uk/training/online/courses/network-analysis-of-protein-interaction-data-an-introduction/introduction-to-graph-theory/graph-theory-adjacency-matrices/
https://www.datacamp.com/tutorial/introduction-to-activation-functions-in-neural-networks

# Evaluation of Outcomes from AI Methods

1. **Complementary Strengths**:
    - **Isolation Forest** is efficient for large datasets and high-dimensional data, making it suitable for quickly identifying anomalies that might indicate bad data entries.
    - **Graph Neural Networks** excel at modeling complex relationships and dependencies within the data, which is particularly useful for detecting referential integrity issues and other relational anomalies.

2. **Comprehensive Analysis**:
    - Using both methods allows you to leverage the strengths of each. Isolation Forest can provide a broad, efficient anomaly detection mechanism, while GNNs can offer deeper insights into relational data structures and dependencies.
    - So, one method misses certain anomalies, but the other might catch them, leading to a more comprehensive detection system.

# Presentation: Four Key Sections



AI Methods

Data Quality Dimensions

AI Methods & Data Quality Assessment

Discussion

# Data Quality Dimensions

- Subset of the Agenda
  - Research scope:
    - CISA’s Seven (7) Data Quality Dimensions and Metrics

# CISA's Data Quality Dimensions

- CISA has the data quality directive that encompasses seven (7) key dimensions, and each with specific goals and metrics for measurement.
  - Completeness
  - Uniqueness
  - Referential Integrity
  - Consistency (Value)
  - Integrity
  - Consistency (Format)
  - Accuracy

# Define Data Quality Dimensions

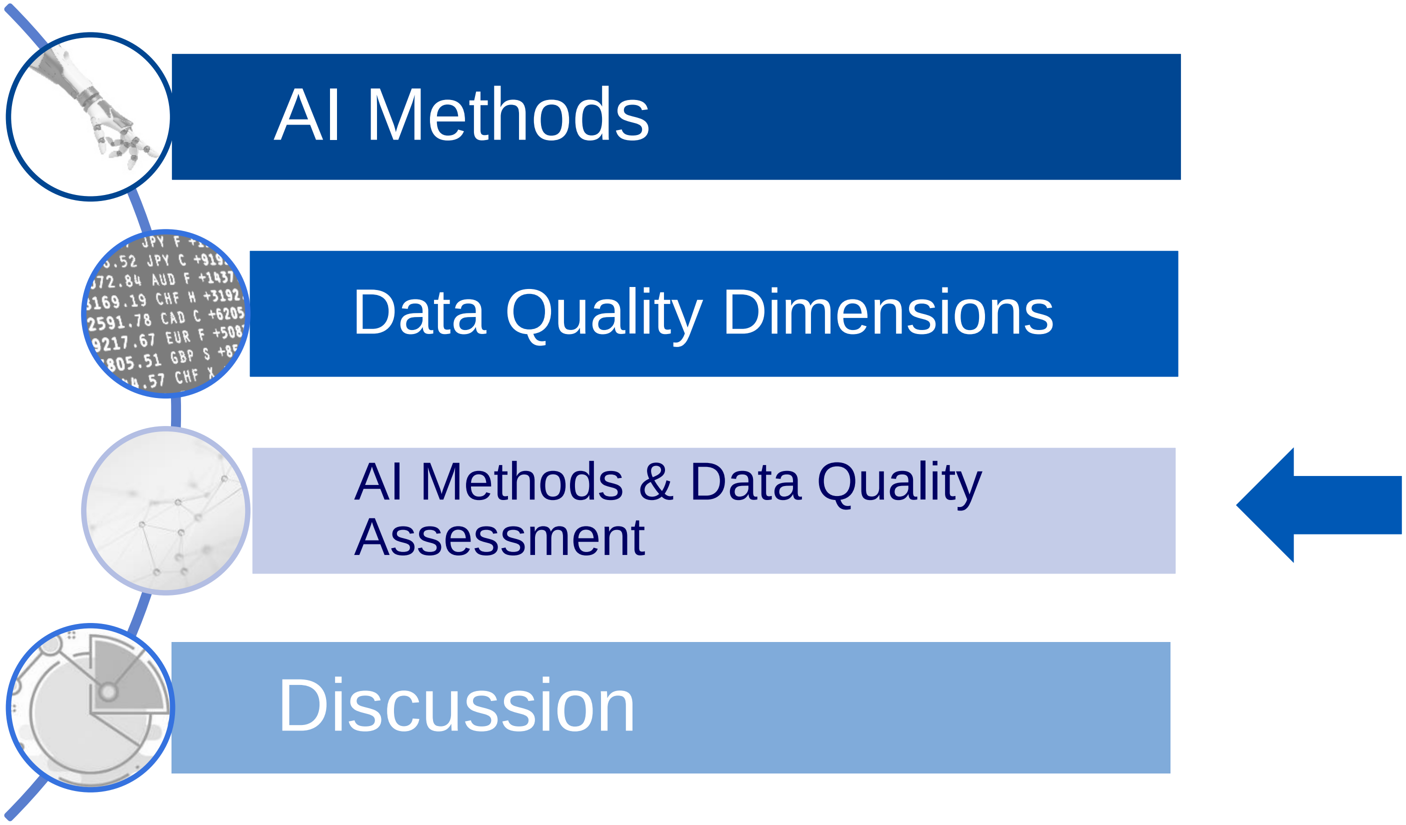
Each dimension has its own measures, metrics, and goals.

| Objective | Business Rule | Dimension | Measures | Metrics | Metric Outcome Goals |
|---|---|---|---|---|---|
| Ensure that database fields that require data are not missing values. | Fields that require **data must contain values** (data may have null, blank, zero, or empty state for required fields if requirement supersedes data creation date). | Completeness | • Total # of complete vales (non-zero, null, black or other empty states).<br>• Total # of values in field | Total complete values / Total # of values in field | 100% |
| Ensure uniqueness. | A data record with specific details appears only once (i.e., **duplicate detection for secondary (non-PK)** fields should exist): | Uniqueness | • # of duplicate entities<br>• Total # of entities within table | # of duplicate entities / Total # of entities within table | >95% |
| Ensure that data requiring relationships to other data are not missing. | Rows of data in tables must exist based on row(s) of data in other table(s) (e.g., **parent/child relationships**). | Referential Integrity | • # of relevant table or field records with proper parent-child relationship as defined by business rules or explicit data logic<br>• Total # of relevant table or field records | # of relevant table or field records with proper parent-child relationship / Total # of relevant table or field records | >99% |
| Ensure reasonable consistency (uniformity) of data. | **Entered values must conform to consistency validation criteria**. (e.g. IPv4 and IPv6 values should not be mixed in the same field; Zipcode vs zip+4; phone numbers with and without extensions) | Consistency | • # of records conforming to validation constraints<br>• Total # of records | # of records conforming to validation constraints/ Total # of records | >99% |

# Define Data Quality Dimensions

| Objective | Business Rule | Dimension | Measures | Metrics | Metric Outcome Goals |
|---|---|---|---|---|---|
| Ensure that only valid values are stored in database fields. | Data in fields must be **limited to a set of valid value**s:<br>e.g.: City, State; Country Name/Country Code; Department and Agency | Integrity | • # of fields conforming to value constraints<br>• Total # of values | # of fields conforming to value constraints / Total # of values | >97% |
| | **Data in fields must be formatted appropriately**:<br>• Phone Number (e.g. (XXX) XXX•XXXX)<br>• Fax (e.g. (XXX) XXX•XXXX)<br>• Ext. (should not include letters)<br>• Email Address<br>• Website<br>• IPv4 address has 4 values separated by periods. | Consistency | • # of values conforming to format standards<br>• Total # of values | # of values conforming to format standards / Total # of values | >97% |
| | **Data in fields must be limited**, based on reasonableness (business rules must match mission/data set requirements)<br><br>e.g.<br>• One phone number per phone number field<br>• One email address per email address field and one @ symbol<br>• Lat/Long coordinates are collected with exactly the required precision for the desired accuracy (e.g. 6 decimals for <1ft, 5 decimals for <4ft, etc.) | Accuracy | • # of fields conforming to accuracy standards<br>• Total # of values | # of fields conforming to accuracy standards / Total # of values | >90% |

# Presentation: Four Key Sections



AI Methods

Data Quality Dimensions

AI Methods & Data Quality Assessment

Discussion

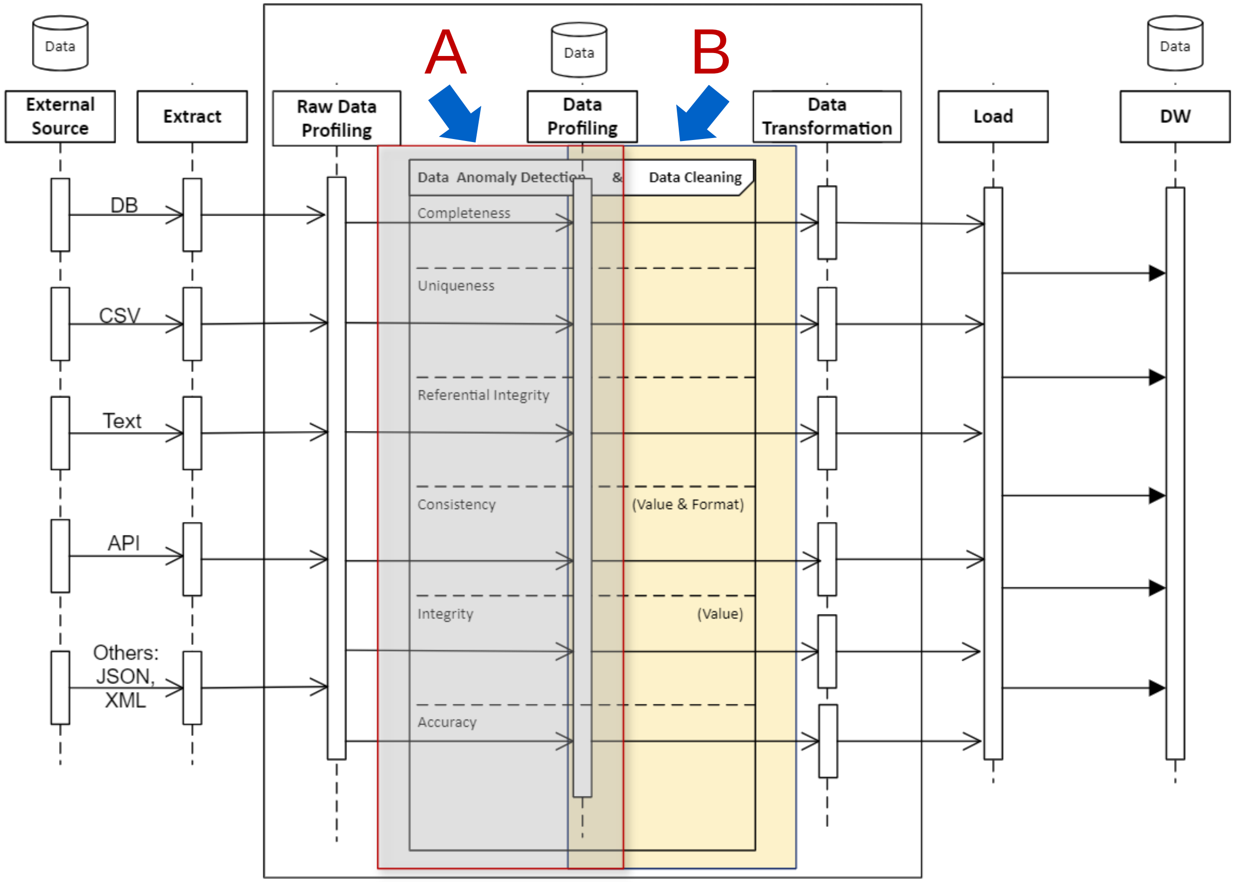
# AI Methods & Data Quality Assessment

- Subset of the Agenda
  - Integrated ETL Framework
  - Isolation Forest for Data Anomaly Detection
  - GNN for Data Anomaly Detection
  - Integration into ETL Pipeline
  - Conceptual Data Quality Status Visualization

# Integrated ETL Framework

- The purpose of this framework is to study an integrated way of applying AI models to data quality assessment for each dimension.
- AI models:
  - Isolation Forest for the dimensions of completeness, uniqueness, consistency (values and formats), accuracy
  - Graph Neural Network for the dimensions of referential integrity and integrity.

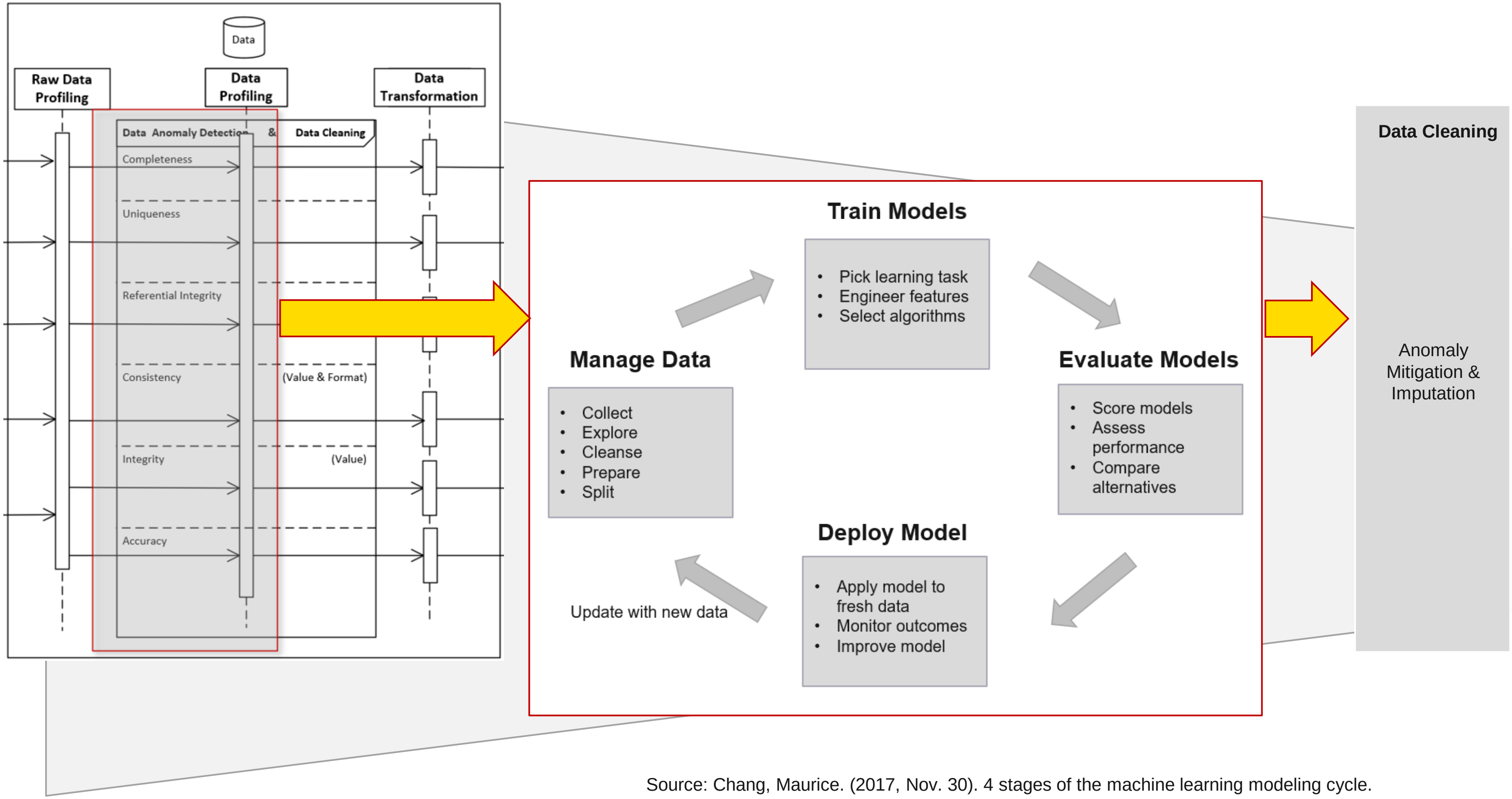
# Integrated ETL Framework



**Data Profiling:** Examining data contents, structure, and quality

**Data Transformation:** Creating data into a consistent format so that it gets loaded into a central repository.

# Machine Learning Modeling Cycle



### Train Models

- Pick learning task
- Engineer features
- Select algorithms

### Evaluate Models

- Score models
- Assess performance
- Compare alternatives

### Deploy Model

- Apply model to fresh data
- Monitor outcomes
- Improve model

Update with new data

### Manage Data

- Collect
- Explore
- Cleanse
- Prepare
- Split

**Data Cleaning**

Anomaly Mitigation & Imputation

Source: Chang, Maurice. (2017, Nov. 30). 4 stages of the machine learning modeling cycle. Retrieved (8/30/2024) from: https://www.linkedin.com/pulse/4-stages-machine-learning-ml-modeling-cycle-maurice-chang/

# Isolation Forest for Data Anomaly Detection

## Steps (Python)

1. **Extract**:
   - Load your dataset into a DataFrame using pandas.
2. **Transform**:
   - Preprocess the data (by handling missing values and normalizing features if necessary)
   - Convert the cleaned data into a NumPy array.
3. **Isolation Forest Application**:
   - Train / fit the Isolation Forest model on your dataset to detect anomalies.
   - Use the model to predict anomaly scores for each data point.
   - Identify anomalies based on the scores.

# GNN for Data Integrity and Consistency

## Steps (Python)

**1. Extract**:

- Load your dataset and create a graph representation where nodes represent data entities and edges represent relationships.

**2. Transform**:

- Construct a graph using libraries like torch_geometric.

**3.GNN Application**:

- Define and train a GNN model to learn embeddings from the graph.
- Use the embeddings to detect inconsistencies and integrity issues.

# Integration into ETL Pipeline

## Steps (Python)

**Extract**:

- Load your dataset into a DataFrame.

1. **Transform**:
    - Apply preprocessing steps.
    - Use Isolation Forest to detect anomalies.
    - Construct a graph and train a GNN for integrity checks.
2. **Load**:
    - Save the processed data and quality metrics.

- **Example ETL Pipeline Pseudo-Code:**

# Integration into ETL Pipeline

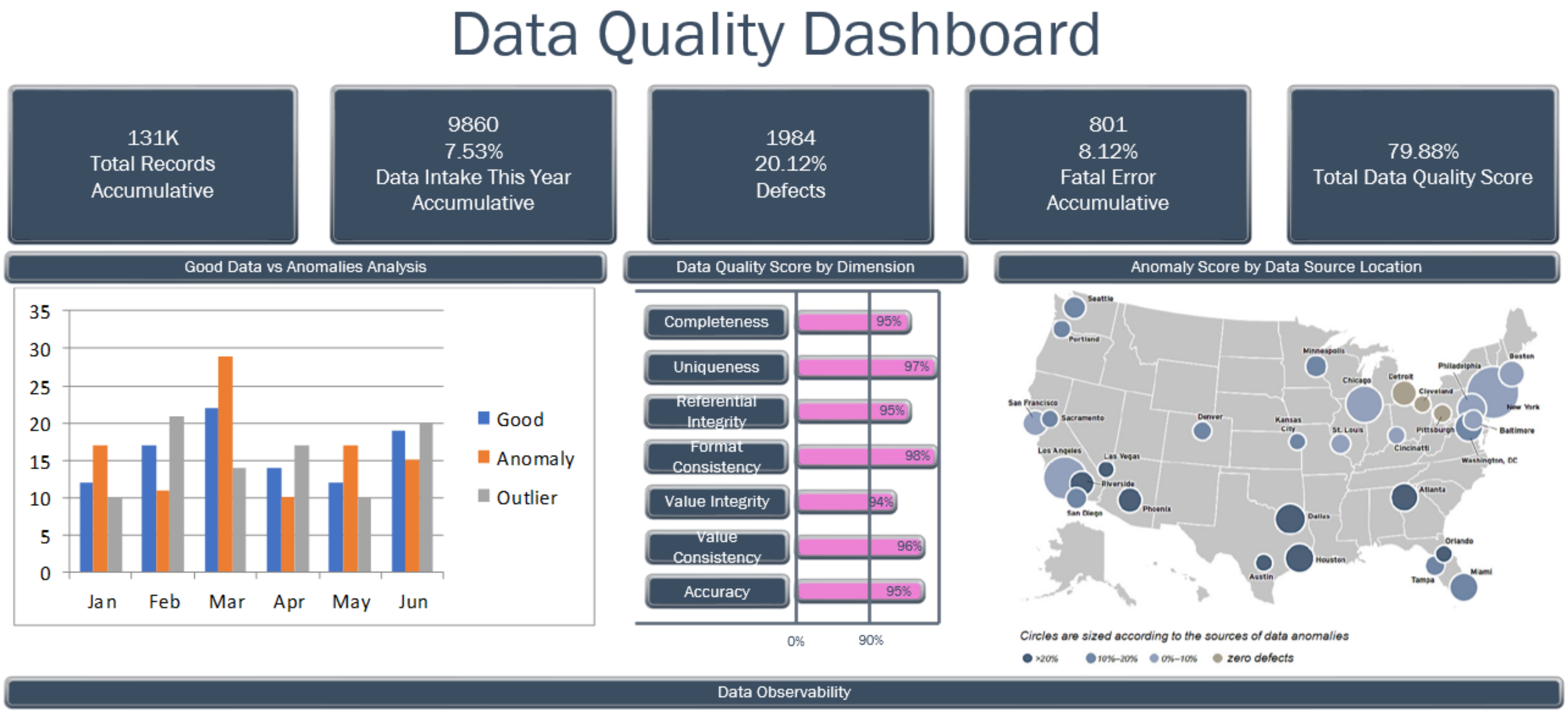
## Pseudo-Code (Python)

```
def etl_pipeline():
    # Extract
    data = pd.read_csv('your_dataset.csv')

    # Transform
    # Preprocess data
    iso_forest = IsolationForest(contamination=0.1)
    data['anomaly_score'] = iso_forest.fit_predict(data)
    anomalies = data[data['anomaly_score'] == -1]

    # Graph Neural Network for integrity checks
    edge_index = torch.tensor([[0, 1, 1, 2],
                               [1, 0, 2, 1]], dtype=torch.long)
    x = torch.tensor([[1], [2], [3]], dtype=torch.float)
    graph_data = Data(x=x, edge_index=edge_index)
    model = GNN()
    model.train()
    for epoch in range(200):
        optimizer.zero_grad()
        out = model(graph_data)
        loss = F.nll_loss(out, torch.tensor([0, 1, 0], dtype=torch.long))
        loss.backward()
        optimizer.step()
    embeddings = model(graph_data).detach().numpy()
```

```
# Load
    # Save processed data and quality metrics
    data.to_csv('processed_data.csv')
    return data, anomalies, embeddings

# Run ETL pipeline
processed_data, anomalies, embeddings = etl_pipeline()
```

# Data Quality Status Visualization

## Conceptual Dashboard



Data Quality Dashboard

131K
Total Records
Accumulative

9860
7.53%
Data Intake This Year
Accumulative

1984
20.12%
Defects

801
8.12%
Fatal Error
Accumulative

79.88%
Total Data Quality Score

Good Data vs Anomalies Analysis

Data Quality Score by Dimension

| | |
|---|---|
| Completeness | 95% |
| Uniqueness | 97% |
| Referential Integrity | 95% |
| Format Consistency | 98% |
| Value Integrity | 94% |
| Value Consistency | 96% |
| Accuracy | 95% |

Anomaly Score by Data Source Location

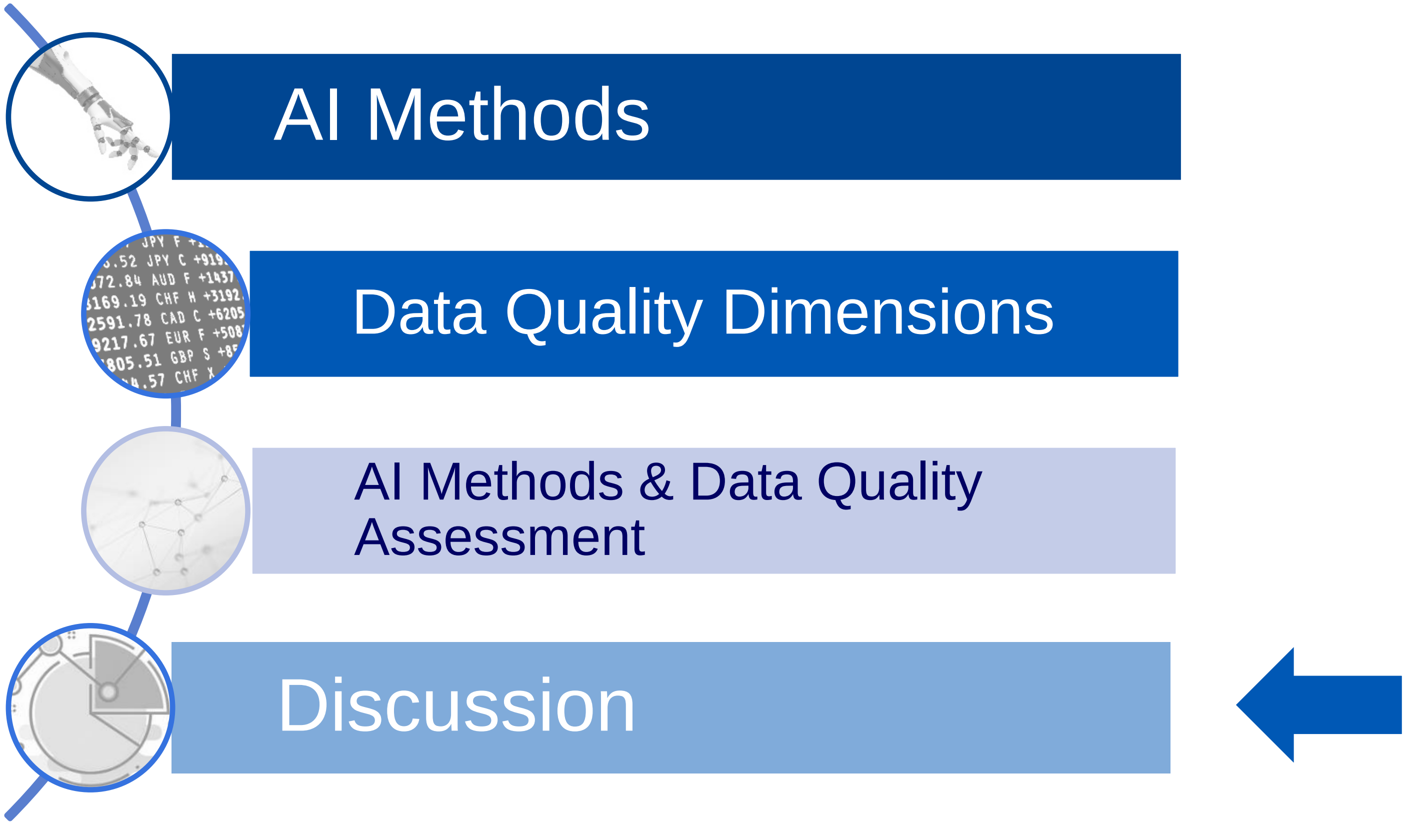
*Circles are sized according to the sources of data anomalies*

● >20% ● 10%–20% ● 0%–10% ● zero defects

Data Observability

*Illustration Only*

# Presentation: Four Key Sections



AI Methods

Data Quality Dimensions

AI Methods & Data Quality Assessment

Discussion

# Discussion – Summary

## RQ 1 – AI Works:

- This research chose eight feasible AI-based anomaly detection methods as tools to examine data quality check.
- This research performed to identify the strengths and weaknesses of those eight AI techniques for the use of data quality check.
- This research adopted two AI techniques, Isolation Forest and Graph Neural Network for data quality check.

# Summary

## RQ 2 – Data Quality Check

- This research applied CISA’s 7 Data Quality Dimensions.
- This research developed the semi-pseudo code by applying IForest and GNN to address those 7 data quality dimensions.
- This research also demonstrated how the entry of bad data can be assessed in each of those 7 dimensions in an integrated fashion, using a proposed framework.
- This research illustrated a data quality dashboard that depicts the end results of the assessment.

# Summary

- Invaluable insights about ETL pipeline
- Anomaly mitigation and imputation
- SQL, Python, and other programming languages and databases
- Requires solid requirements or use cases that can result in more sensible, practical, and productive methods.

# References

- NSA AISC. (2024, April 15). Joint Guidance on Deploying AI Systems Securely. https://www.cisa.gov/news-events/alerts/2024/04/15/joint-guidance-deploying-ai-systems-securely.
- EO 14110. (2023, October 30).Safe, Secure and Trustworthy Development and Use of AI. https://www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/

# Thank You!

## Questions?

Dennis Park, PhD
**OCS/Mission Technology Data Management**
**Email:** Dennis.Park@mail.cisa.dhs.gov