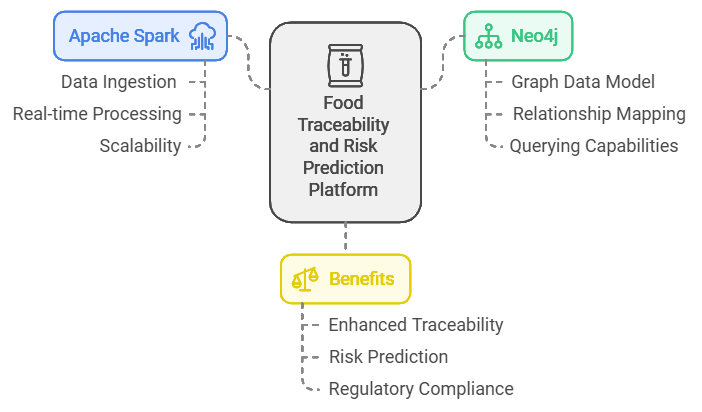
### **1. Title**

**Food Traceability and Risk Prediction Platform Using Apache Spark and Neo4j**

### **2. Introduction**



This project aims to develop a traceability platform for the Food supply chain using advanced analytics and machine learning models. By leveraging Apache Spark for data processing and Neo4j for representing supply chain entities in a graph structure, the platform will enhance traceability, improve supply chain transparency, and provide predictive analytics to manage quality and minimize risks, such as spoilage and logistical issues.

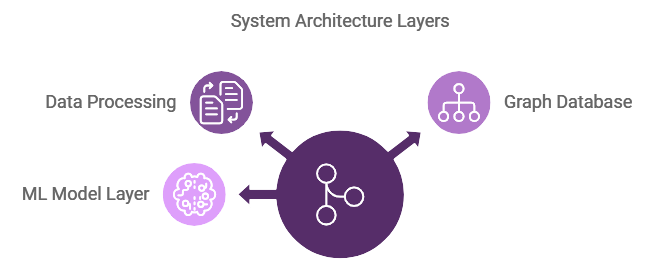
### **3. Key Business Terms**

* **Traceability**: Ability to track the origin and journey of Food products from farm to consumer.
* **Spoilage Risk**: The probability of Food becoming unsuitable for consumption due to poor storage, transport, or environmental factors.
* **Predictive Analytics**: Use of data and machine learning models to forecast potential risks and optimize processes.
* **Clustering and Classification**: Grouping and categorizing supply chain entities to better understand patterns, quality levels, and risk factors.
* **Recommendation System**: An algorithm that suggests optimal practices, routes, and partners in the supply chain based on past data and performance.

### **4. Design & Architecture**

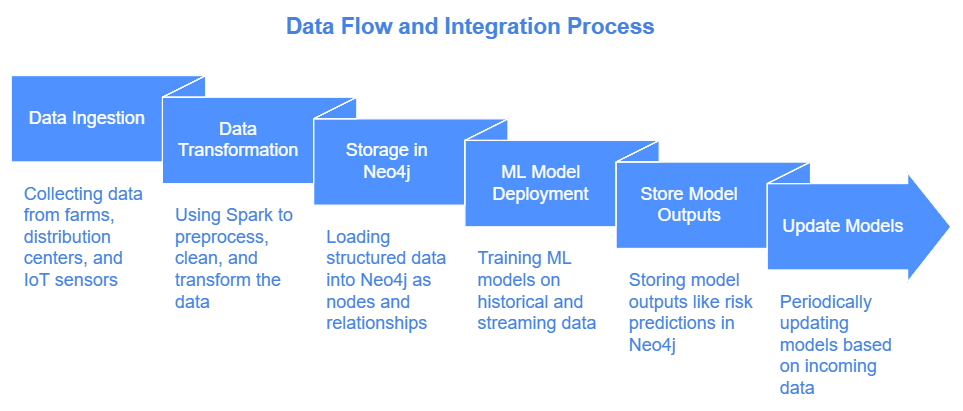
#### **4.1 System Architecture**

The platform architecture includes three main layers: **Data Processing**, **Graph Database**, and **Machine Learning Models**.



* **Data Processing Layer**: Ingests data from sensors, transactional records, and historical data. Uses Apache Spark for data cleaning, preprocessing, and streaming.
* **Graph Database Layer**: Neo4j manages the Food supply chain as a graph structure, with nodes for different entities (e.g., Farm, Processing Plant, Distributor) and edges representing relationships (e.g., SUPPLIES, DELIVERS).
* **ML Model Layer**: Various ML models for prediction, classification, clustering, and recommendation are trained and used to identify potential risks and optimize decision-making.

#### **4.2 Data Flow and Integration**



1. **Data Ingestion**: Collect data from various sources (farms, distribution centers, IoT sensors).
2. **Data Transformation**: Use Spark to preprocess, clean, and transform the data.
3. **Storage in Neo4j**: Load structured data into Neo4j as nodes and relationships.
4. **ML Model Deployment**: Train ML models on historical and streaming data. Store model outputs (e.g., risk predictions) in Neo4j and periodically update based on incoming data.

### **5. Technologies**

* **Apache Spark**: For batch and real-time data processing.
* **Neo4j**: For managing supply chain data in a graph structure.
* **Python**: Language for developing and deploying ML models and data processing scripts.
* **TensorFlow or PyTorch**: For building deep learning models, especially for time series forecasting and classification.
* **Scikit-Learn**: For simpler ML models (e.g., clustering, decision trees).
* **Kafka**: For handling real-time data ingestion if streaming is required.
* **Docker**: For containerizing the platform’s various components for easier deployment and scaling.

### **6. Implementation**

#### **6.1 Data Collection**

* Define data sources for data collection from farms, IoT devices, distribution points, and retail outlets.

#### **6.2 Data Preprocessing**

* Use Apache Spark to clean and transform raw data into a structured format suitable for Neo4j and ML models.
* Create Spark jobs for batch and real-time processing.

#### **6.3 Graph Database Design**

* Define Neo4j nodes and relationships based on the supply chain entities (e.g., Farm, Processing Plant, Distributor, Retailer).

### **Graph Database Design**

The updated design will focus on the traceability of onions rather than eggs:

1. **Define Neo4j Nodes and Relationships Based on Supply Chain Entities**
   * **Nodes:**
     + **Farm**
       - Attributes: farm\_id, location, type (e.g., organic, conventional), owner, certification\_details
     + **Manufacturer**
       - Attributes: manufacturer\_id, name, facility\_location, processing\_capacity, certifications
     + **Processing Plant**
       - Attributes: plant\_id, name, location, handling\_capacity, processing\_methods
     + **Distributor**
       - Attributes: distributor\_id, name, distribution\_centers, coverage\_area
     + **Retailer**
       - Attributes: retailer\_id, name, location, store\_type (e.g., supermarket), storage\_conditions
     + **Consumer**
       - Attributes: consumer\_id, location, purchase\_date, feedback
     + **Logistics Provider**
       - Attributes: provider\_id, name, fleet\_size, transport\_type (e.g., refrigerated), track\_and\_trace\_capability
     + **Warehouse**
       - Attributes: warehouse\_id, name, location, storage\_capacity
   * **Relationships:**
     + (Farm) -[SUPPLIES]-> (Manufacturer)
       - Represents the supply of raw onions from farms to manufacturers.
     + (Manufacturer) -[PROCESSES]-> (Processing Plant)
       - Represents the processing of onions for packaging and distribution.
     + (Processing Plant) -[PACKAGES\_FOR]-> (Distributor)
       - Represents the packaging of onions specifically for distributors.
     + (Distributor) -[STORES\_IN]-> (Warehouse)
       - Represents the storage of onions by distributors in warehouses for further distribution.
     + (Warehouse) -[SHIPS\_VIA]-> (Logistics Provider)
       - Represents the role of logistics providers in transporting onions from warehouses to retailers.
     + (Logistics Provider) -[DELIVERS\_TO]-> (Retailer)
       - Represents the delivery of onions to retailers.
     + (Retailer) -[SELLS\_TO]-> (Consumer)
       - Represents the sale of onions to end consumers.
     + (Logistics Provider) -[MONITORED\_BY]-> (IoT Sensor)
       - Represents the integration of IoT sensors to monitor transportation conditions (e.g., temperature, humidity) for onions.

### **6.4 Machine Learning Models**

#### **Predictive Analytics Models**

These models help predict potential risks in the supply chain, such as spoilage or transportation delays.

* **Time Series Forecasting Models:**
  + **ARIMA, SARIMA:** Useful for predicting temperature, humidity, or demand based on historical data. These models can forecast conditions that may affect the quality of onions, such as fluctuations in warehouse temperatures or seasonal demand changes.
  + **LSTM (Long Short-Term Memory) Networks:** A type of recurrent neural network (RNN) that captures long-term dependencies in time-series data. It can help predict conditions impacting onion quality, such as trends in storage temperature or transportation delays.
  + **Prophet:** Suitable for seasonality and trend forecasting, which can predict potential disruptions in the availability or demand for onions.

#### **Anomaly Detection Models**

These models identify abnormal conditions that could indicate risks to the quality and safety of onions.*Using biosensors we can detect harmful bacteria at different stages in the supply chain (e.g., at farms, processing plants, or retail outlets).*

* **Isolation Forest:** Detects outliers by isolating each point in the dataset. It is useful for spotting unusual environmental conditions, such as unexpected temperature spikes during storage or transportation.
* **One-Class SVM (Support Vector Machine):** Classifies data as either normal or anomalous. This model can identify unusual sensor readings that may indicate a risk to onion quality, such as excessive humidity.
* **Autoencoders:** Neural networks that detect anomalies in high-dimensional data, making them suitable for identifying abnormal patterns across multiple sensors monitoring onion batches.

#### **Risk Prediction using Classification Models**

These models predict the likelihood of spoilage or other quality-related risks.

*Biosensors for Real-Time Monitoring: Use biosensors dataset to monitor environmental conditions like humidity, temperature, and microbial contamination in storage facilities. These sensors can send real-time data to the traceability platform, improving the system's ability to detect risks.*

*Bioremediation Solutions: Using microorganisms to clean up any contaminants in the environment (e.g., soil or water) where onions are grown, thus reducing the initial risk of contamination.*

*Development of advanced molecular techniques, such as PCR (Polymerase Chain Reaction) and CRISPR-based diagnostics, for quickly detecting pathogens like E. coli in onions. These methods can be integrated with the traceability platform to flag contaminated batches in real time.*

* **Logistic Regression:** A simple and interpretable model suitable for binary risk predictions, such as predicting whether an onion batch is at risk of spoilage (yes/no).
* **Random Forest, Gradient Boosting (e.g., XGBoost):** Tree-based ensemble models that classify risk levels (e.g., low, medium, high) based on factors such as transportation conditions, storage durations, and temperature control.
* **CatBoost:** Handles categorical data effectively without extensive preprocessing, ideal for mixed data types like batch ID, supplier location, and transport method.
* *Use biological data, such as microbial profiles, that can be integrated with the machine learning models for predictive analytics. This can improve the accuracy of models in predicting spoilage, contamination risks, and other quality issues.*

#### **Classification Models**

These models help sort batches, identify quality grades, or flag defective products.

* **Support Vector Machine (SVM):** Effective for binary or multi-class classification, such as categorizing onions based on quality grades (e.g., Grade A, B).
* **Decision Trees and Random Forests:** Suitable for hierarchical classification of onion batches based on attributes like size, weight, and quality. These models provide interpretability for identifying specific factors affecting quality issues.
* **Convolutional Neural Networks (CNNs):** Useful for image classification during visual inspections, such as detecting signs of rot, mold, or other defects in onions. A CNN can analyze images captured during sorting and flag defective onions.
* **Naive Bayes:** Can classify onions based on categorical attributes (e.g., organic, conventional). It's fast and performs well when features are independent, which is often the case with product attributes.

#### **Clustering Models**

Clustering helps group similar batches, identify quality patterns, or segment distributors based on performance.

* **K-Means Clustering:** Groups onions, farms, or batches based on similar characteristics (e.g., freshness, size, location). Useful for segmenting suppliers and identifying quality patterns among batches.
* **DBSCAN (Density-Based Spatial Clustering of Applications with Noise):** Finds clusters of arbitrary shapes, making it ideal for segmenting onion suppliers or distributors by geographical regions. It can help detect regional quality issues or high-risk zones.
* **Hierarchical Clustering:** Allows a hierarchical structure in clustering, which can categorize farms based on multiple factors like region, size, and compliance history.
* **Gaussian Mixture Models (GMM):** A probabilistic clustering model suitable for grouping batches with mixed distributions, such as varying freshness levels in different regions or under different storage conditions.

#### **Recommendation Models**

Recommendation models support supply chain decision-making by suggesting optimal routes, suppliers, or practices based on historical data and performance.

* **Collaborative Filtering (Matrix Factorization, SVD):** Provides recommendations based on patterns in historical data. For example, it can suggest preferred logistic providers or routes to minimize spoilage during onion transport.
* **Content-Based Filtering:** Recommends suppliers or distributors based on attributes such as certification, proximity, and transport reliability. This model helps in selecting suppliers that meet quality and freshness requirements.
* **Association Rule Learning (e.g., Apriori, FP-Growth):** Analyzes patterns among different conditions (e.g., temperature ranges or transport durations that correlate with better quality retention). It is useful for optimizing storage conditions or identifying risky combinations of transport and storage conditions.
* **Deep Neural Networks (DNNs):** Can be used for complex recommendation systems that factor in dynamic elements like real-time sensor data and historical quality incidents. DNNs provide more adaptive and granular recommendations for supply chain decisions.

### **5. Supply Chain Optimization Models**

These models help optimize supply chain processes by predicting delays, managing inventory, and minimizing waste, which is crucial for perishable items like onions.

***DNA Fingerprinting and Whole Genome Sequencing (WGS)****: genomic tracing to identify the genetic signature of pathogens. WGS can help trace the contamination back to specific batches or even the original farm source.*

***Strain Typing:*** *Techniques can differentiate between strains of pathogens to establish whether cases of E. coli are linked to the same source, aiding the investigation and narrowing down the origin of contamination.*

***CRISPR-Based Gene Editing****: Gene editing techniques can be used to create onion varieties that are more resilient to environmental stressors like drought or temperature changes, ensuring better storage and longer shelf life, thus reducing spoilage risk*

***Microbiome Analysis:*** *By analyzing the microbiome of onion batches at different stages, biotechnology can help identify microbial patterns associated with spoilage or contamination, enabling early intervention*

#### **Optimization Algorithms**

* **Linear Programming (LP):**
  + LP can be used to minimize costs or maximize efficiency in various supply chain decisions. For onions, LP can optimize the routing of shipments to minimize transportation costs while considering constraints like delivery deadlines, temperature control, and storage capacity.
  + An objective function might be formulated to minimize the total cost of transportation, considering factors such as distance, delivery time, and refrigeration needs, subject to constraints like vehicle capacity and warehouse limits.
* **Genetic Algorithms (GA):**
  + GA can optimize complex problems in the supply chain, such as route planning for delivering onions to multiple retailers while considering varying delivery windows and traffic patterns.
  + It can also be used to find the optimal combination of suppliers that balance cost, quality, and reliability, helping to ensure a consistent supply of high-quality onions while minimizing costs.

#### **Inventory Management Models**

* **Economic Order Quantity (EOQ):**
  + EOQ helps determine the optimal order quantity that minimizes total inventory costs, including ordering and holding costs. For onions, EOQ can be adjusted based on factors like shelf life, storage conditions (e.g., temperature and humidity), and demand variability.
  + The EOQ formula is given by: EOQ=2DSHEOQ = \sqrt{\frac{2DS}{H}}EOQ=H2DS​​ where DDD is the annual demand for onions, SSS is the ordering cost per order, and HHH is the holding cost per unit per year.
* **Reorder Point (ROP):**
  + ROP models determine the inventory level at which a new order should be placed to avoid stockouts. For onions, this model accounts for the lead time, demand during the lead time, and safety stock to handle unexpected demand fluctuations or delivery delays.
  + The ROP formula can be given by: ROP=(Lead Time Demand)+Safety StockROP = (Lead \, Time \, Demand) + Safety \, StockROP=(LeadTimeDemand)+SafetyStock
  + For onions, factors like spoilage rates and seasonal demand spikes would influence safety stock calculations.

#### **Vehicle Routing Problem (VRP) Solutions**

* **Ant Colony Optimization (ACO):**
  + ACO can solve VRP scenarios where multiple delivery vehicles need to be routed efficiently. This approach is suitable for distributing onions from warehouses to retail locations, considering constraints such as vehicle capacity and delivery time windows.
  + The algorithm uses a probabilistic technique inspired by the behavior of ants searching for food, which helps identify optimal paths for delivery routes to minimize total travel time or distance while ensuring timely delivery of perishable goods like onions.
* **Simulated Annealing (SA):**
  + SA can be used to optimize delivery schedules for onions to multiple destinations. It starts with a random solution and iteratively improves it by exploring neighboring solutions, mimicking the process of annealing in metallurgy.
  + For onions, SA can help optimize routes while considering variables like refrigeration availability and traffic patterns, aiming to reduce spoilage and transportation costs.

#### **Inventory Optimization using Time Series Forecasting**

* **Seasonal Stock Adjustment:**
  + Models like **ARIMA** or **Prophet** can predict seasonal demand variations for onions, helping adjust stock levels proactively. These models forecast demand spikes or dips based on historical sales data, allowing for dynamic stock management.
  + Predictions can be integrated with EOQ and ROP models to update inventory policies in real-time, reducing the risk of spoilage during low-demand periods and ensuring sufficient stock during high-demand seasons.

#### **Waste Minimization Strategies**

* **Perishable Goods Allocation Models:**
  + Models like **First-Expire, First-Out (FEFO)** are implemented to prioritize the distribution of onions nearing the end of their shelf life. This approach reduces waste by ensuring older inventory is shipped out before newer stock.
  + **Predictive Analytics** can identify which batches of onions are more likely to spoil based on environmental conditions and historical data. This information can be used to optimize stock rotation practices.
* **Multi-Echelon Inventory Optimization:**
  + This approach manages inventory across different levels of the supply chain (e.g., warehouses, distribution centers, retail outlets). Multi-echelon optimization can be used to balance stock levels, ensuring that onions are available where needed while minimizing total inventory costs.
  + Advanced techniques like **Stochastic Inventory Models** can account for demand uncertainty and lead-time variability in multi-echelon networks.

These models enhance the efficiency of onion supply chains by optimizing logistics, managing inventory effectively, and minimizing waste, contributing to a more transparent and resilient food traceability system.

#### **6.5 Integration and Deployment**

* Integrate Neo4j and Spark to allow seamless data flow between data processing and the graph database.
* Containerize the application components using Docker for scalability.
* Deploy ML models and periodically update predictions based on real-time data ingestion.

### **7. Testing**

#### **7.1 Unit Testing**

* Conduct unit tests on individual components like data preprocessing, ML model functions, and Cypher queries.

#### **7.2 Integration Testing**

* Test end-to-end data flow from ingestion, processing, storage in Neo4j, and prediction with ML models to ensure all components work cohesively.

#### **7.3 Performance Testing**

* Evaluate system performance to ensure it can handle high data volumes, especially for real-time data ingestion and querying in Neo4j.

#### **7.4 Model Evaluation**

* Use metrics like accuracy, F1-score, precision, and recall for classification models.
* Evaluate clustering models using silhouette scores and adjust the parameters accordingly.
* Measure time series model accuracy using metrics such as RMSE and MAPE.

### **8. Reports**

#### **8.1 Traceability Reports**

* **Purpose:** Provide a visual representation of the entire supply chain journey of each batch of onions, from farm to retailer, including details about every processing, storage, and transportation step.
* **Content:** Show the path taken by recalled batches of yellow onions, highlighting entities involved (e.g., Taylor Farms, distributors, restaurants like McDonald’s). The report will include timelines and locations, enabling investigators to quickly identify where the onions may have become contaminated.
* **Focus Area:** The report will emphasize the batches supplied to McDonald's and other affected food service customers, showing all steps before the voluntary recall was issued.

#### **8.2 Quality Risk Analysis**

* **Purpose:** Assess the risk levels for batches of onions based on factors such as spoilage predictions, environmental conditions, and transport data.
* **Content:** Include risk scores for the batches involved in the recall, evaluating spoilage risk or contamination likelihood based on conditions such as temperature and humidity during storage and transportation.
* **Investigation Aid:** Use risk scores to prioritize the analysis of batches that may have higher contamination risks, aiding in determining whether slivered onions supplied to McDonald’s were the likely source of the E. coli outbreak.

#### **8.3 Anomaly Detection Report**

* **Purpose:** Identify any anomalies in environmental conditions (e.g., temperature, humidity) throughout the onion supply chain that could contribute to quality issues or contamination risks.
* **Content:** Highlight unusual readings from IoT sensors during the storage or transport of recalled onion batches. For example, sudden temperature spikes in warehouses or during transportation could indicate potential quality degradation.
* **Suggested Actions:** Recommend actions to address detected anomalies, such as increasing inspections or adjusting storage conditions for other batches from the same supplier to prevent further contamination.

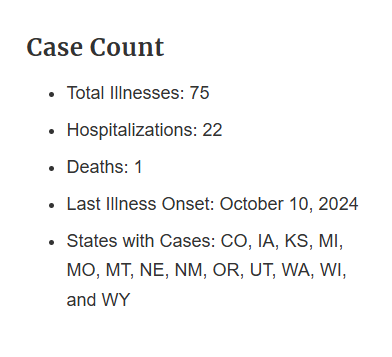
#### **8.4 Supplier and Distributor Performance Reports**

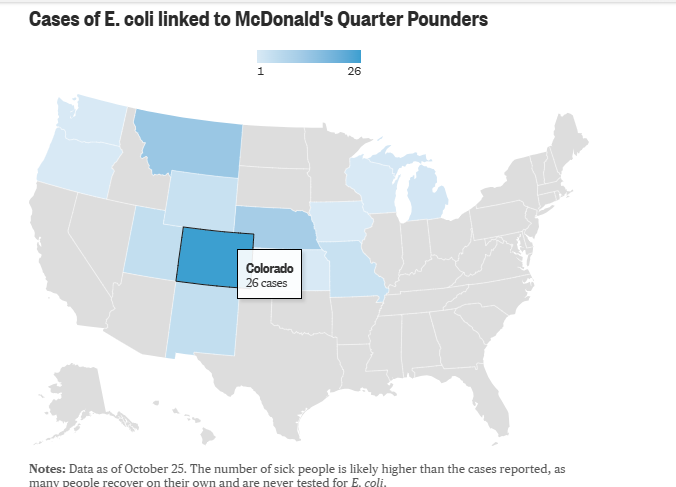
* **Purpose:** Evaluate the performance of suppliers and distributors based on key metrics like quality consistency, delivery time, and customer feedback.
* **Content:** Analyze performance data for Taylor Farms and other distributors who handled the recalled onions, identifying any patterns in quality issues. The report will assess compliance with quality standards across different distribution centers.
* **Ranking:** Rank suppliers and distributors based on their performance history, allowing McDonald’s and other food service customers to make informed decisions about sourcing onions in the future.

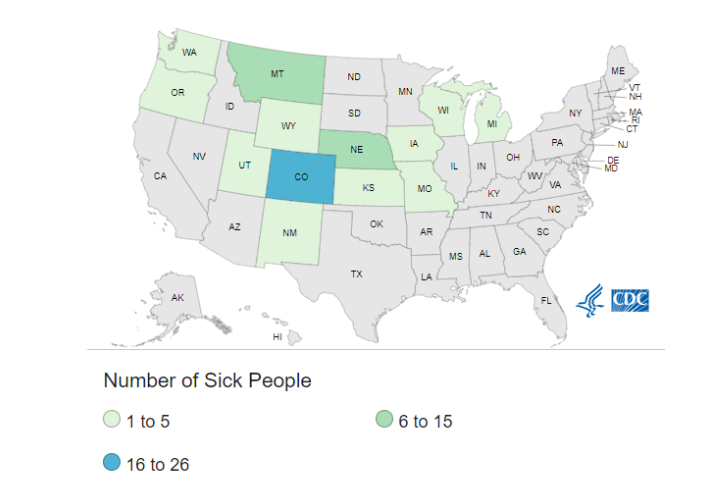
#### **8.5 Outbreak Status and Recall Impact Report**

* **Purpose:** Track the status of the ongoing E. coli outbreak investigation and measure the impact of the recall.
* **Content:** Provide an update on the number of reported cases, hospitalization data, affected states, and the steps taken by Taylor Farms and McDonald’s to address the situation. Include details about the investigation's progress and any new findings related to the potential source of contamination.
* **Recall Effectiveness:** Evaluate the effectiveness of the recall measures taken, including the removal of slivered onions from McDonald's menus in affected states and notifications sent to other food service customers.

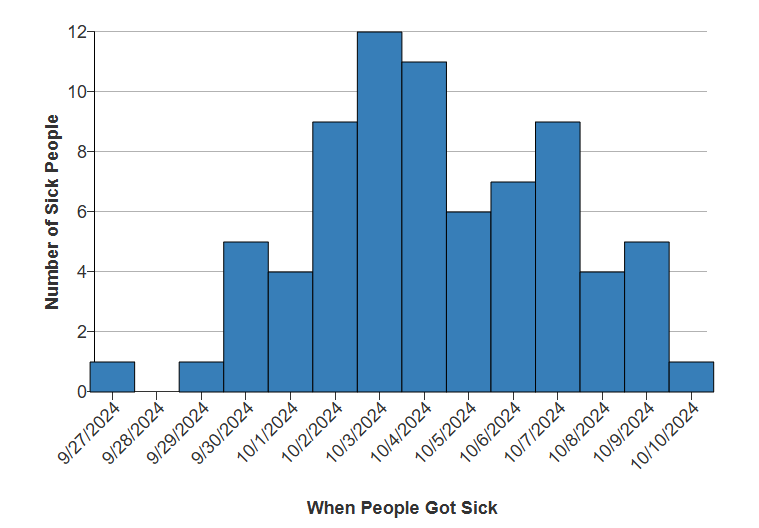
#### **8.6 Visualisations**



****

****





References

<https://www.youtube.com/watch?v=K-s4hr87994>

<https://www.youtube.com/watch?v=NqsGIjWjiCQ>