U.S. FOOD & DRUG ADMINISTRATION

Modernizing Post-Market Quality Surveillance Through Application of Advanced Analytics

Alex Viehmann, *Division Director*
Nandini Rakala, Ph.D., *Data Scientist & Visiting Associate*

Office of Quality Surveillance, Office of Pharmaceutical Quality
CDER | US FDA

Post-Market Reports | SBIA Webinar | May 24, 2023

Outline

Introduction to the Risk Problem

Post-Market Quality Surveillance using Advanced Analytics

Risk-based Predictive Prioritization

Quality Signal Detection and Topic Modeling

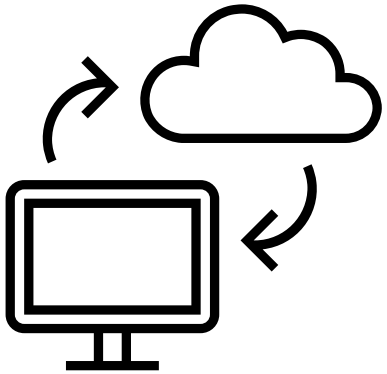
Decision-Making using Identified Signals

Summary

Learning Objectives

- ❑ Define the risk problem
- ❑ Describe the machine learning process cycle
- ❑ Application of advanced analytics: case study 1
- ❑ Application of advanced analytics: case study 2
- ❑ Describe the process of proactive risk-based decision-making

Introduction to the Risk Problem



- **The Office of Quality Surveillance's mission** is to 'turn intelligence into insights and actions to promote the availability of quality medicines for the American public'.
- Increasing volume of data and variety of intelligence available to OQS requires utilizing **objective** approaches for achieving **comprehensive quality surveillance**.
- Relying **solely** on **human intervention** increases **risk** of not detecting **potential quality issues**.
- Develop a systematic approach for performing effective **Risk-based Prioritization** and **Signal Detection**
- Application deployed for **Field Alert Reports** (FAR) data

Introduction to the Risk Problem Cont.

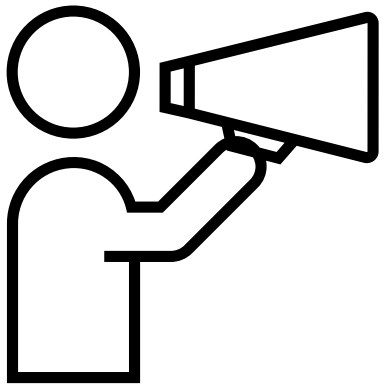
- **Identify Changes** in reporting habits to objectively **prioritize risks**
- Improve **detectability** of risks as part of OQS's oversight strategy.
- Promote effective and timely **decision-making**
  - Utilizing **Quality Risk Management** principles
  - Multi-disciplinary teams.
- Enable effective **Knowledge Management** which decreases **uncertainty.**

Integrating Advanced Analytics into Quality Surveillance

Objective: To enhance quality surveillance framework through integration of predictive analytics and AI-based machine learning techniques.

Why?

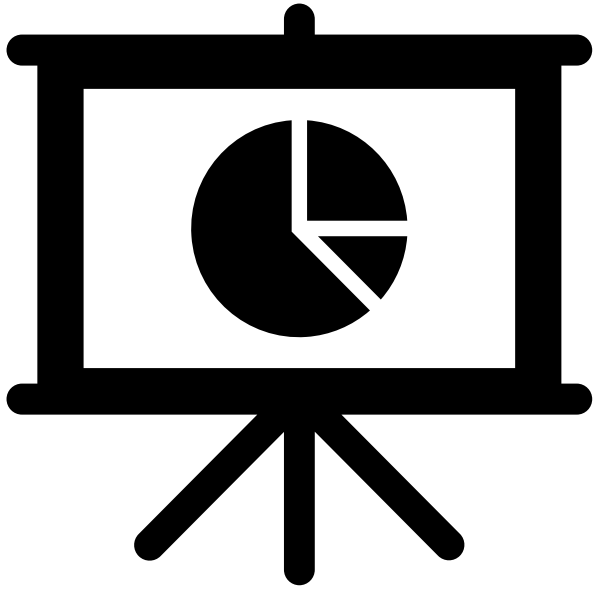
- Going beyond knowing what has happened to **Predict** what will happen
- Growing **Volume** and **Variety** of Data/Parameters
- **Advancements** in **Technology**
- Competitive **Advantage** with **Optimization**
- Integrate **Human Intelligence** with **Machine Intelligence**



Risk-based Predictive Prioritization: Hybrid ML-NLP Model

Application of Advanced Analytics: Case Study 1

INTRODUCTION



Objective

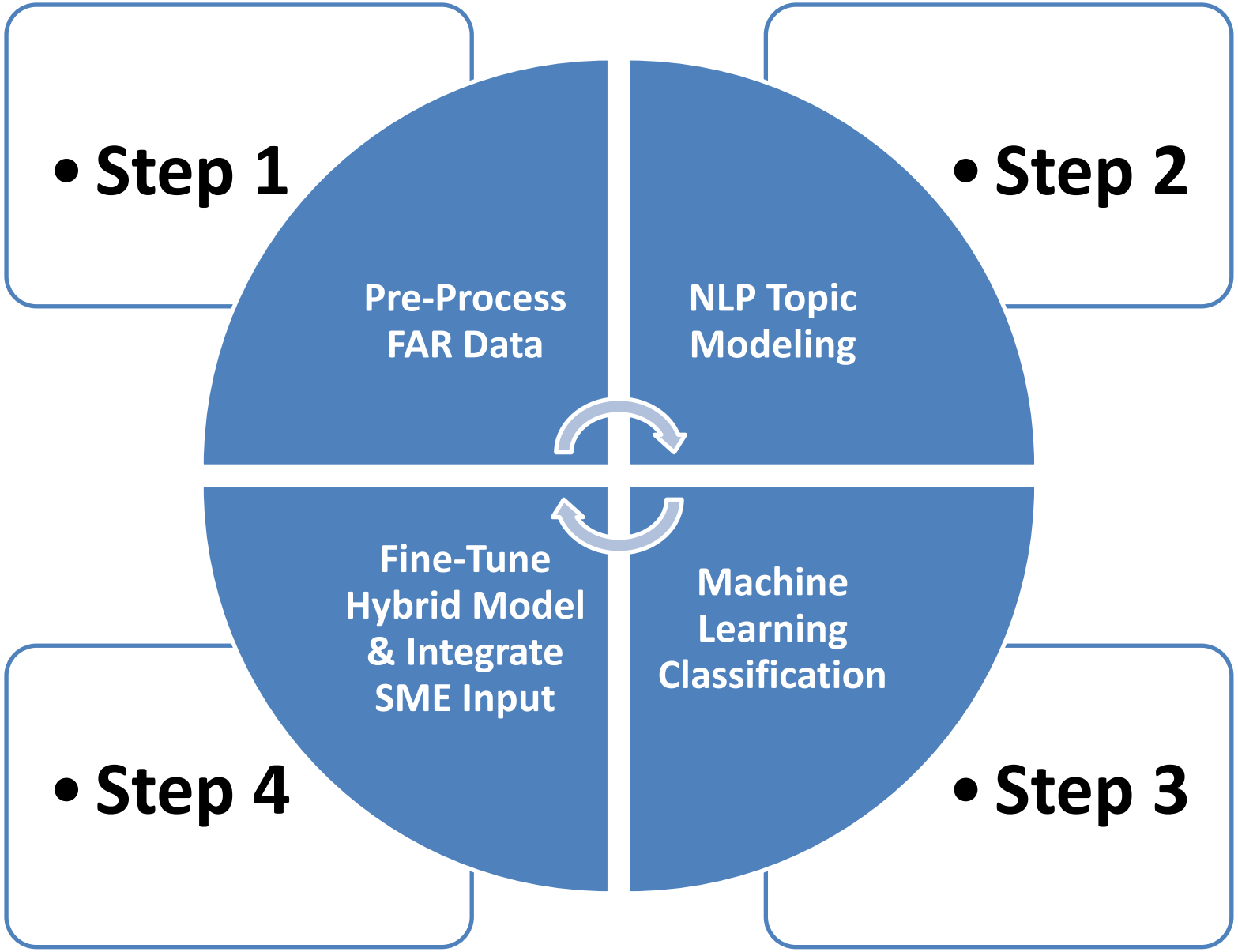
To Prioritize Incoming FAR Using a Data-Driven and Risk-Based Approach.

What is Machine Learning?

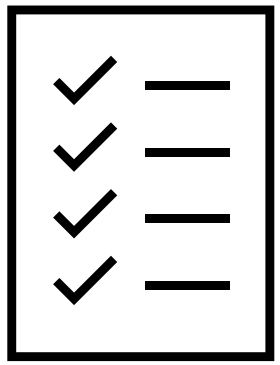
Why is Natural Language Processing needed?

How do we validate the accuracy of the generated models?

Machine Learning Process Cycle



FAR Prioritization: Overview



- **FAR Prioritization** is an innovative AI framework leveraging machine learning and NLP for risk-based prioritization of incoming Field Alert Reports.
- Streamline data pre-processing, develop topic model predictors, and merge with other key indicator variables from FAR data.

- Multi-disciplinary Analytics and SME collaboration by incorporating input on data cleaning, topic-keywords refinement, rare-events tagging, and labeling of risk-based target in a programmatic manner.
- Insights generated by the Hybrid Model are used to proactively inform key indicators for FAR reviews which help prioritize high-risk issues.

FAR DOCUMENT STRUCTURE

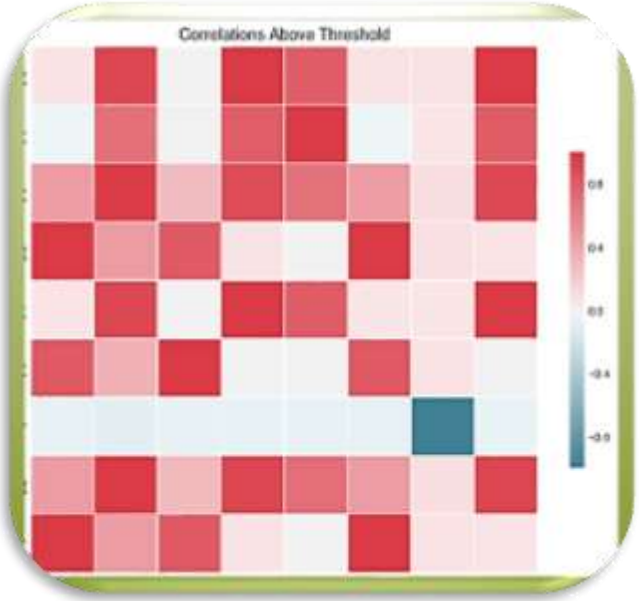
Application (Number/Type)

Product Name (Generic/Brand)

Problem (How Discovered/Remarks)

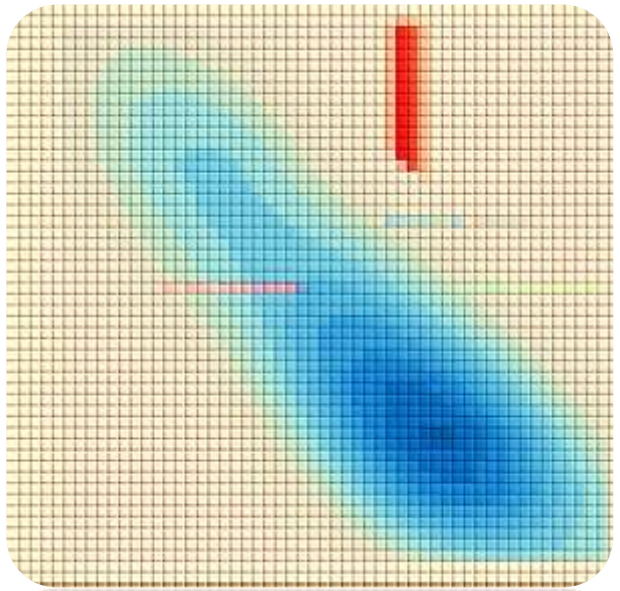
Dosage Form, etc.

FAR Data Pre-Processing



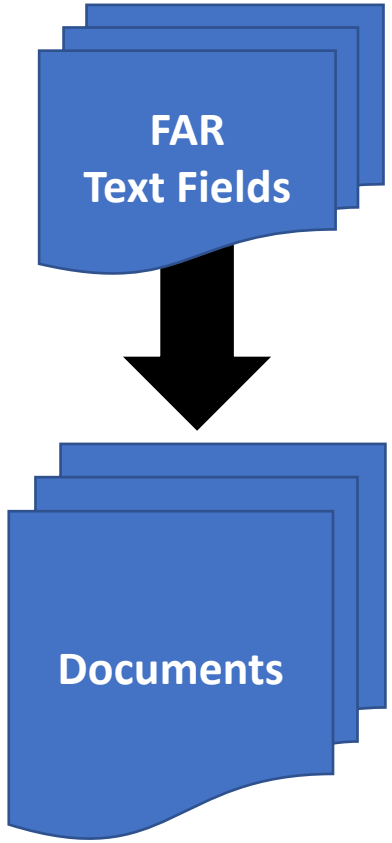
- ❑Initial FAR **Data Cleaning**
  - ❑Variable Distributions
  - ❑Exploratory Data Analysis

- ❑Analyze **Linear** & **Non-Linear** Relationships
  - ❑Missing Completely At Random (MCAR)
  - ❑Missing At Random (MAR)
  - ❑Missing Not At Random (MNAR)
  - ❑Multicollinearity Analysis
  - ❑Feature Engineering
  - ❑Feature Selection



NLP Model 1

Corpus



Document Characteristics

- Application Type
- Product Name
- Dosage Form
- . . .

Document Level Data

Topic Prevalence

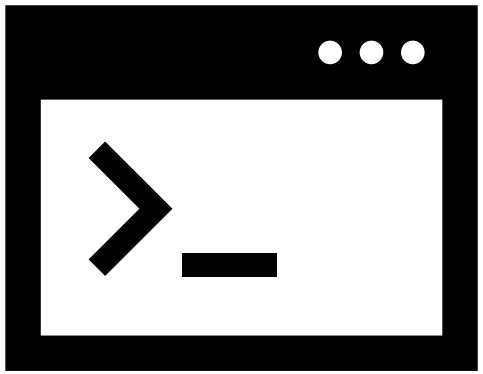
Variables affecting the frequency for which a topic is discussed

Topic Content

Variables affecting the words used within a given topic, or how a topic is discussed

NLP Model 2

- Iterative **Topic Model Generation**
  - Merge/Split/Create New Topics
  - Code Rare-Events
- Topics include **Groupings** of **Keywords** Identified in a Document Collection.
- Top Five Terms with Highest **Relevancy Scores** are used to Identify a Particular Topic.



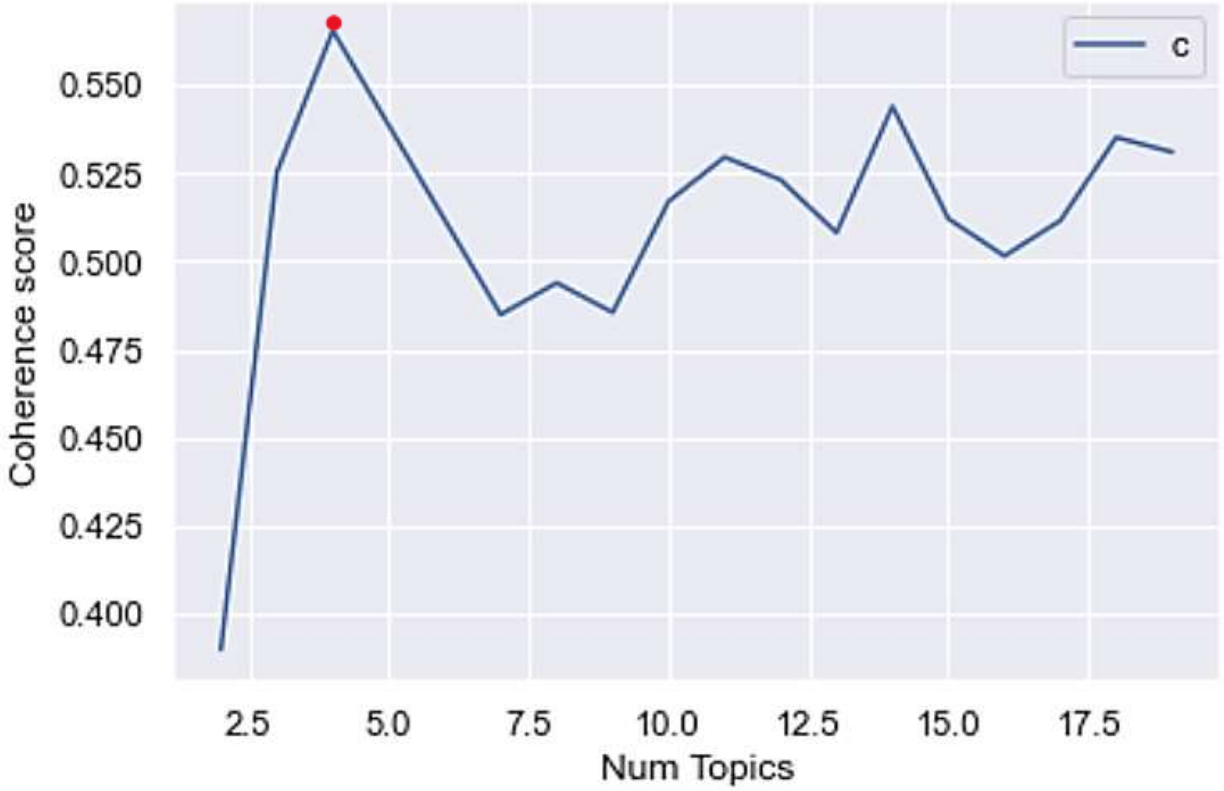
NLP Model 3

Bigram & Trigram Models, Lemmatization, Remove Stopwords

Data Dictionary & Corpus Creation

Optimal Topic Model Generation & Dominant Probability Distributions

Optimized Topic Model Selection

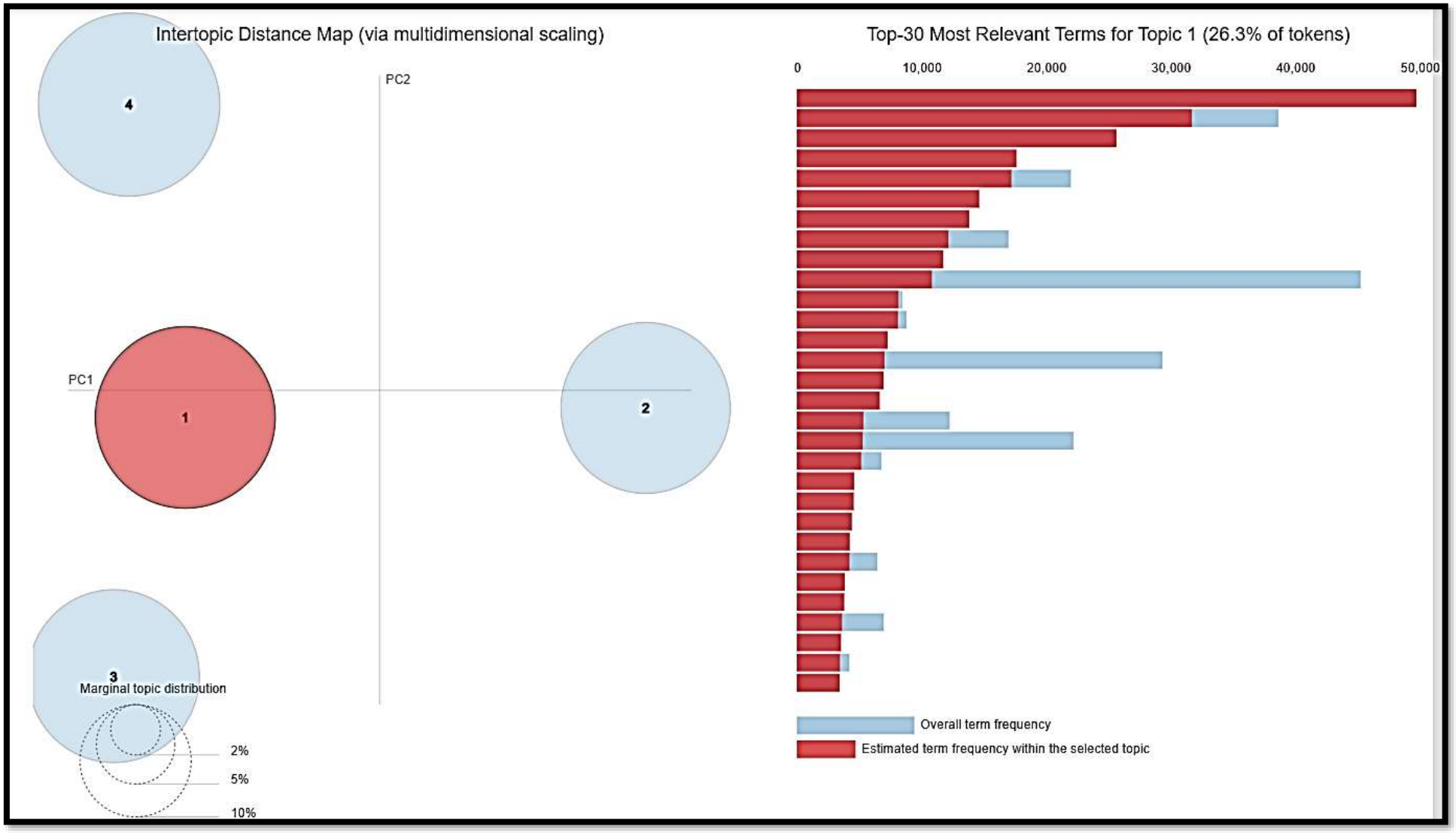


Topic Coherence Score is a Quantitative Measure Evaluating Model Performance of an NLP-Topic Model

Higher Coherence Value => Better NLP Model

```
Num Topics = 2  has Coherence Value of 0.38993
Num Topics = 3  has Coherence Value of 0.52527
Num Topics = 4  has Coherence Value of 0.56538
Num Topics = 5  has Coherence Value of 0.53854
Num Topics = 6  has Coherence Value of 0.51169
Num Topics = 7  has Coherence Value of 0.48491
Num Topics = 8  has Coherence Value of 0.49388
Num Topics = 9  has Coherence Value of 0.48551
Num Topics = 10  has Coherence Value of 0.517
Num Topics = 11  has Coherence Value of 0.52953
Num Topics = 12  has Coherence Value of 0.52313
Num Topics = 13  has Coherence Value of 0.50812
Num Topics = 14  has Coherence Value of 0.54391
Num Topics = 15  has Coherence Value of 0.51204
Num Topics = 16  has Coherence Value of 0.50151
Num Topics = 17  has Coherence Value of 0.51155
Num Topics = 18  has Coherence Value of 0.53512
Num Topics = 19  has Coherence Value of 0.53096
```

Inter-Topic Distance Map



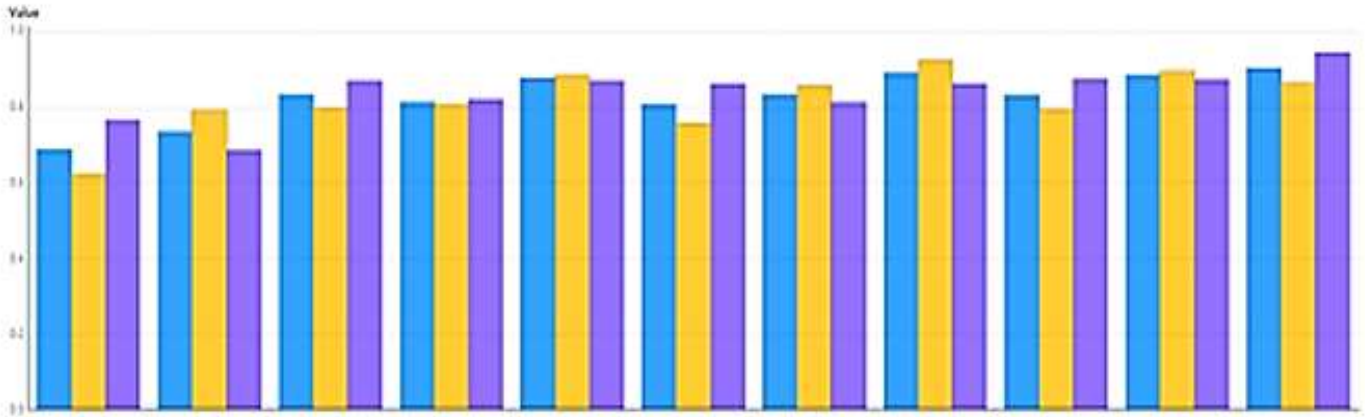
- **2D Space Visualization:** Topic-Keyword Distributions
- **Area of Each Circle:** Proportional to Frequency of Words within Topic Cluster

Model Performance & Interpretability

Model Comparison

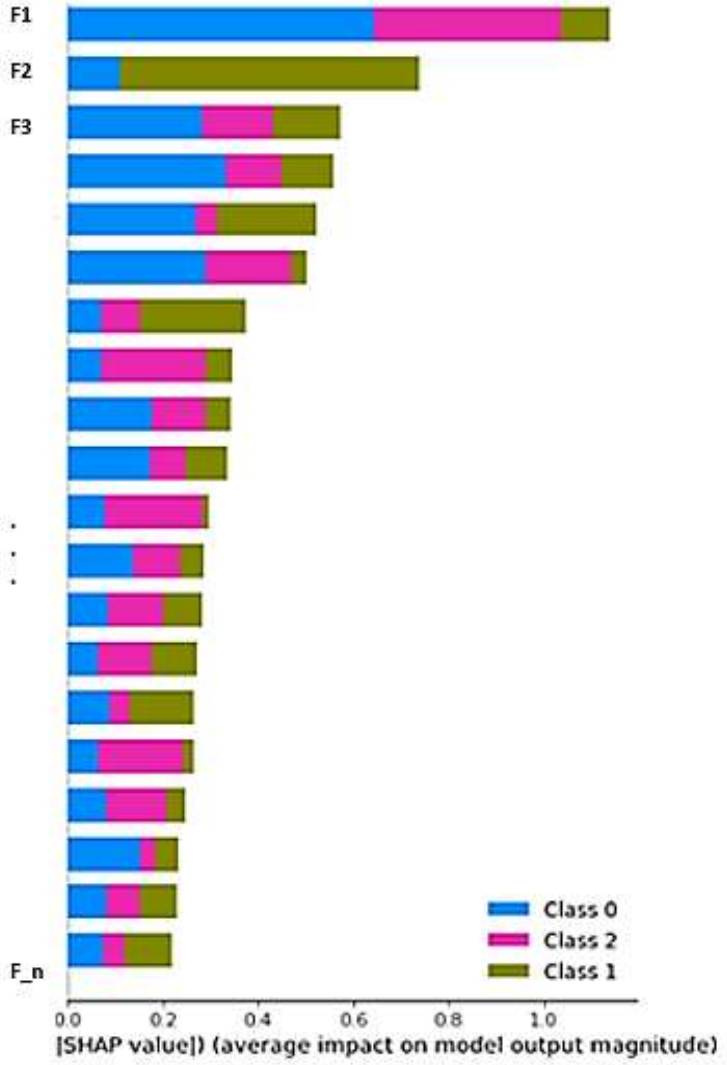
Champion	Name	Algorithm Name	Area Under ROC	Misclassification Rate
✓	Gradient Boosting	Gradient Boosting	0.9492	0.1553
	Ensemble	Ensemble	0.9448	0.1662
	Forest	Forest	0.9415	0.1844
	Neural Network	Neural Network	0.9266	0.1949
	Forward Logistic Regression	Logistic Regression	0.9118	0.2463
	Decision Tree	Decision Tree	0.8989	0.1836
	Stepwise Logistic Regression	Logistic Regression	0.5000	0.4687

Diagnostic Metrics for Automatically Generated Categories

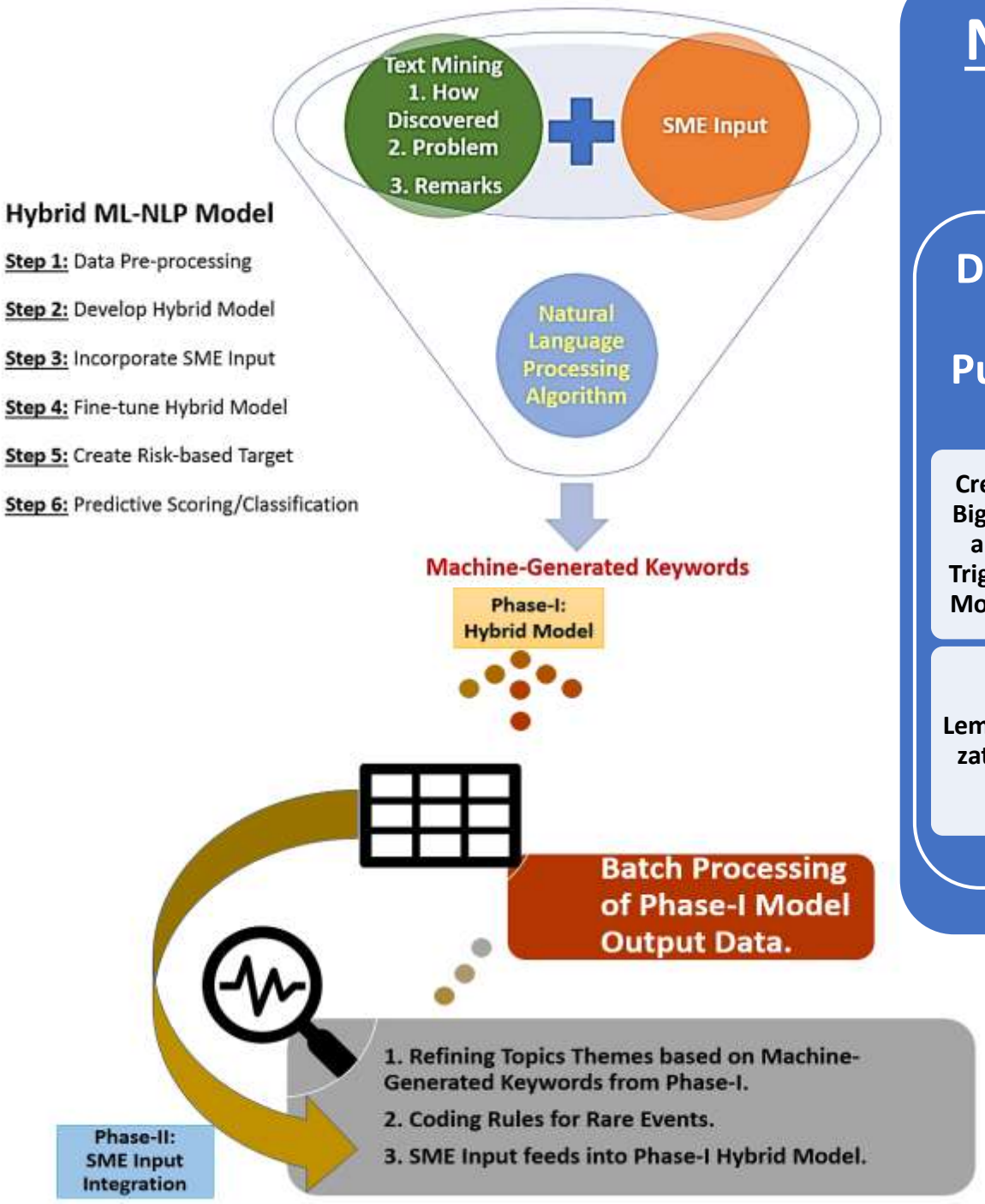


X-axis: Metrics on F-Measure, Precision, & Recall for text categories respectively.

Note: The plots displayed here are based on dummy data for illustration purposes only, and don't necessarily represent actual data values/figures.



Comprehensive Quality Surveillance: Lifecycle Approach Based-On Machine Learning and Natural Language Processing



NLP Topic Modeling

Training, Selection, and Validation

Data Preprocessing: Remove Duplicates, Clean Punctuation, and Lower Case

- Create Bigram and Trigram Models
- Lemmatization

Refine and Remove Stop-Words Iteratively

- Data Dictionary and Corpus Creation. Optimal Topic Word Cloud Clusters Generation
- Dominant Topic Probability Distribution across entire Population of Reports

Challenge Question 1

‘Shapley Value’ is a mathematical method for model interpretation derived from which of the following concepts?

a. Game Theory
b. Topic Modeling
c. Non-Linear Programming
d. No Idea

Quality Signal Detection and Topic Modeling of Post-Market Surveillance Reports

Application of Advanced Analytics: Case Study 2

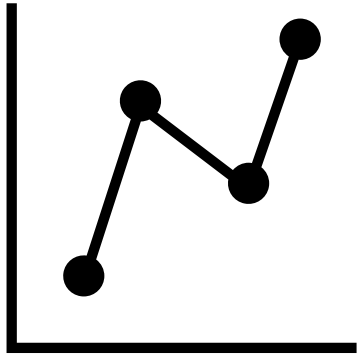
Signal Detection Methodology

- **Statistical Process Control (SPC)**

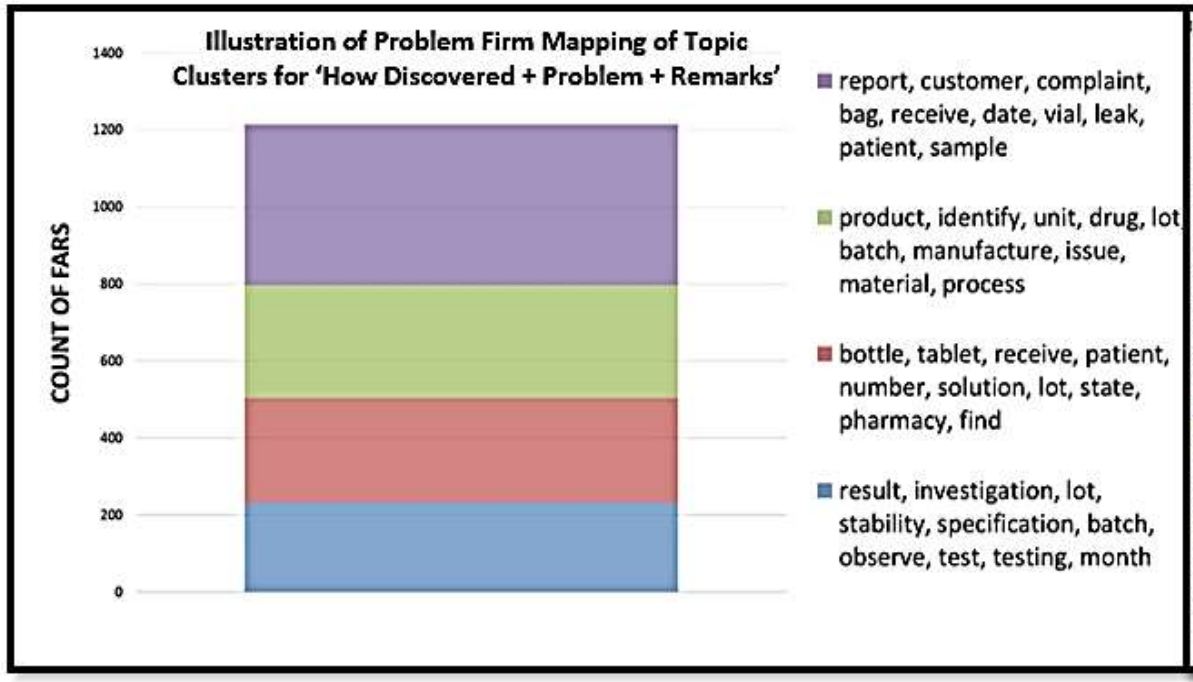
- **Traditional U** and **Laney U Prime Attribute Control Charts**
  - Account for false positives, over-dispersion, under-dispersion, varying sample sizes
  - Out Of Control (OOC) points flagged: intersection of u and u' charts
  - Historical range: 12- and 9-month baseline data

- **Escalation of Signals:** sites flagged with repeat OOC points

- **Natural Language Processing**
  - Predictive insights into problem clusters
  - Recurring topic themes for detected signals



NLP Topic Cluster Mapping



Frequencies

Level	Count	Prob
Adhesion Lacking	3	0.00265
Admin Device Malfunction	38	0.03360
Ampoule Shatter	2	0.00177
Appearance Not As Expected	20	0.01768
Bag Port Defect	3	0.00265
Bag Seal Failure	6	0.00531
Building and/or Facility for Manufacturing Operations Not Adequate	1	0.00088
Caking / Clumping	5	0.00442
Closure Defect	6	0.00531
Closure Type Wrong	1	0.00088
Cloudy / Turbid	12	0.01061
Coating Cracked	1	0.00088
Coating Mottled	1	0.00088
Container Cracked	11	0.00973
Container Defect	15	0.01326
Container Leaks	160	0.14147
Container Seal Defect	11	0.00973
Container Type Wrong	1	0.00088
Contamination Chemical	13	0.01149
Contamination Microbial	9	0.00796
Contamination Mold	7	0.00619
Contamination Suspected	26	0.02299
Content Uniformity Questioned	1	0.00088
Cracked Dosage Form	9	0.00796
Crystal Formation	5	0.00442
N Missing	79	
142 Levels		

Signals Flagged

OOC Points | Trends | First Time Reporters

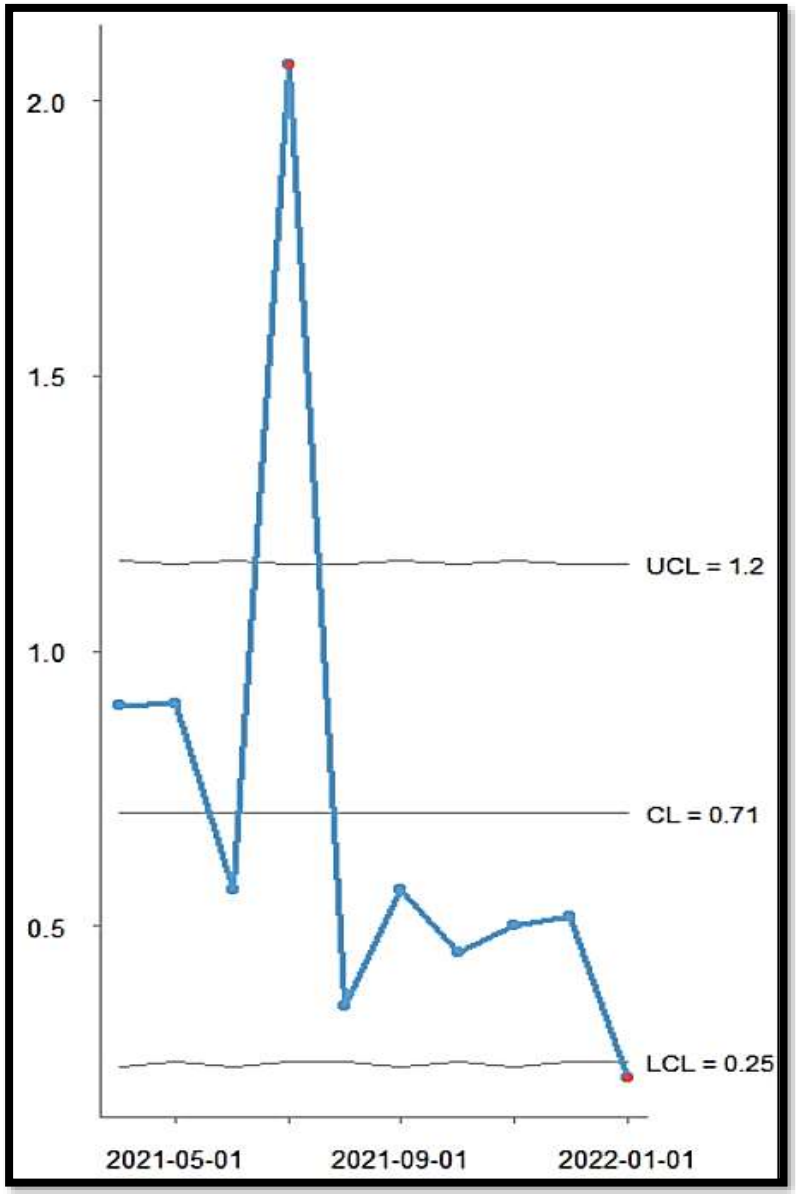
Proactive Risk-Based Decision-Making Using Identified Signals

Risk-Based Decision-Making

Out of Control FEI Signals Flagged

Natural Language Processing Algorithm

Prioritize High-Risk Issues



Challenge Question 2

Which of the following Statistical Attribute Control Chart is appropriate for correction of large sample sizes (> 1000)?

a. p-chart
b. np-chart
c. c-chart
d. u’-chart

SUMMARY

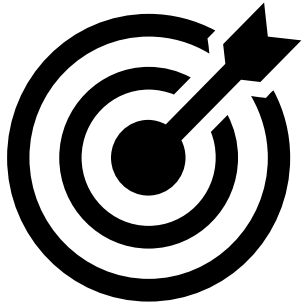
Comprehensive Quality Surveillance through **Life-Cycle Approach**

- ❑ Inform Key Indicator Variables
- ❑ Proactively Prioritize Reviews
- ❑ Allocate Optimal Resources

What Happened → What will Happen → What will make it Happen

Let's work together to promote public health through improving global pharmaceutical product quality…

WE ARE A TEAM!



Thank You All!
Any Questions?

Alex Viehmann, *Division Director*

Nandini Rakala, Ph.D., *Data Scientist & Visiting Associate*

Office of Quality Surveillance, Office of Pharmaceutical Quality

CDER | US FDA