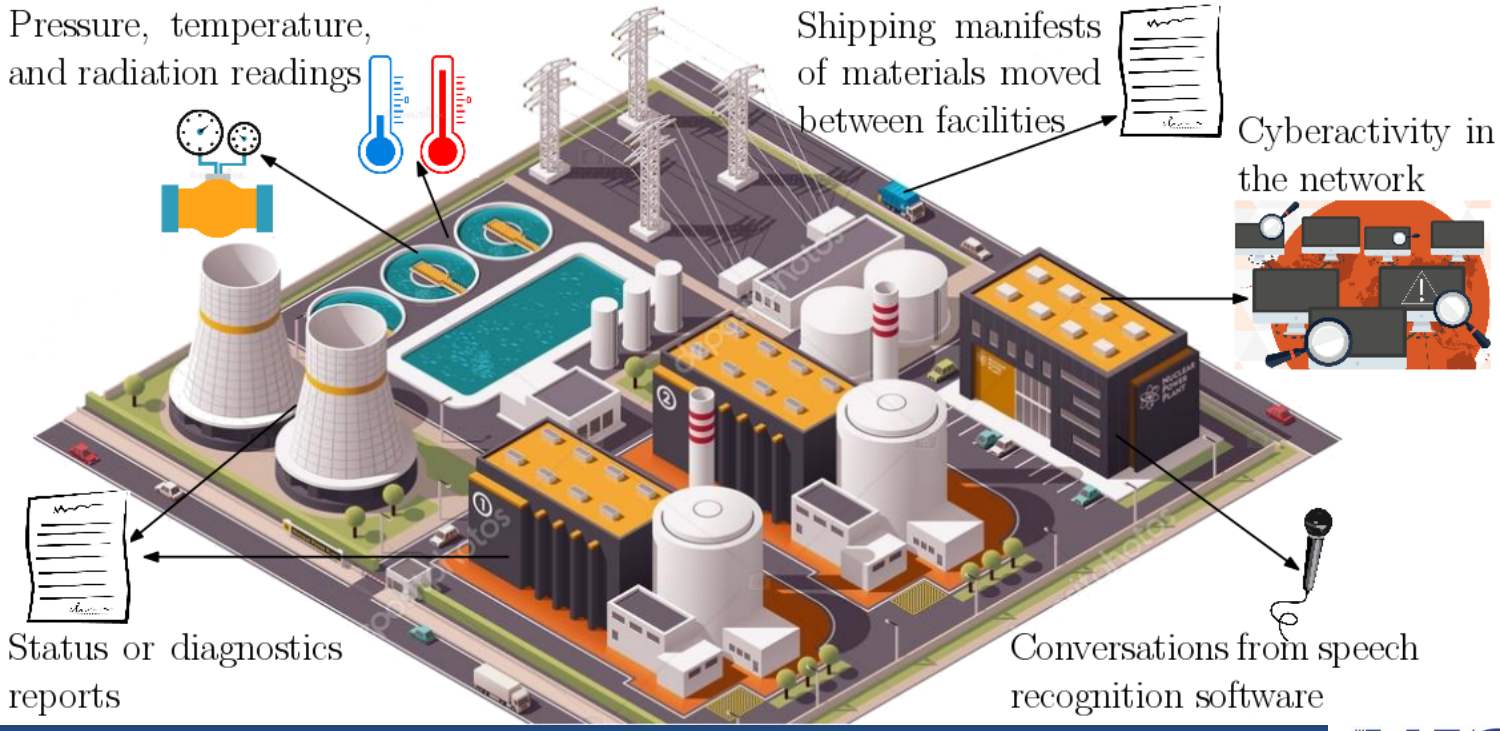
# SEQUENTIAL DETECTION OF UNUSUAL DOCUMENTS

Elizabeth Hou, Alfred Hero
University of Michigan

# Motivation

- The facilities in a nuclear fuel cycle produce many modalities of data
- One common modality is documents, which can be collected sequentially in time (e.g. every week, month)
- Documents can take many forms

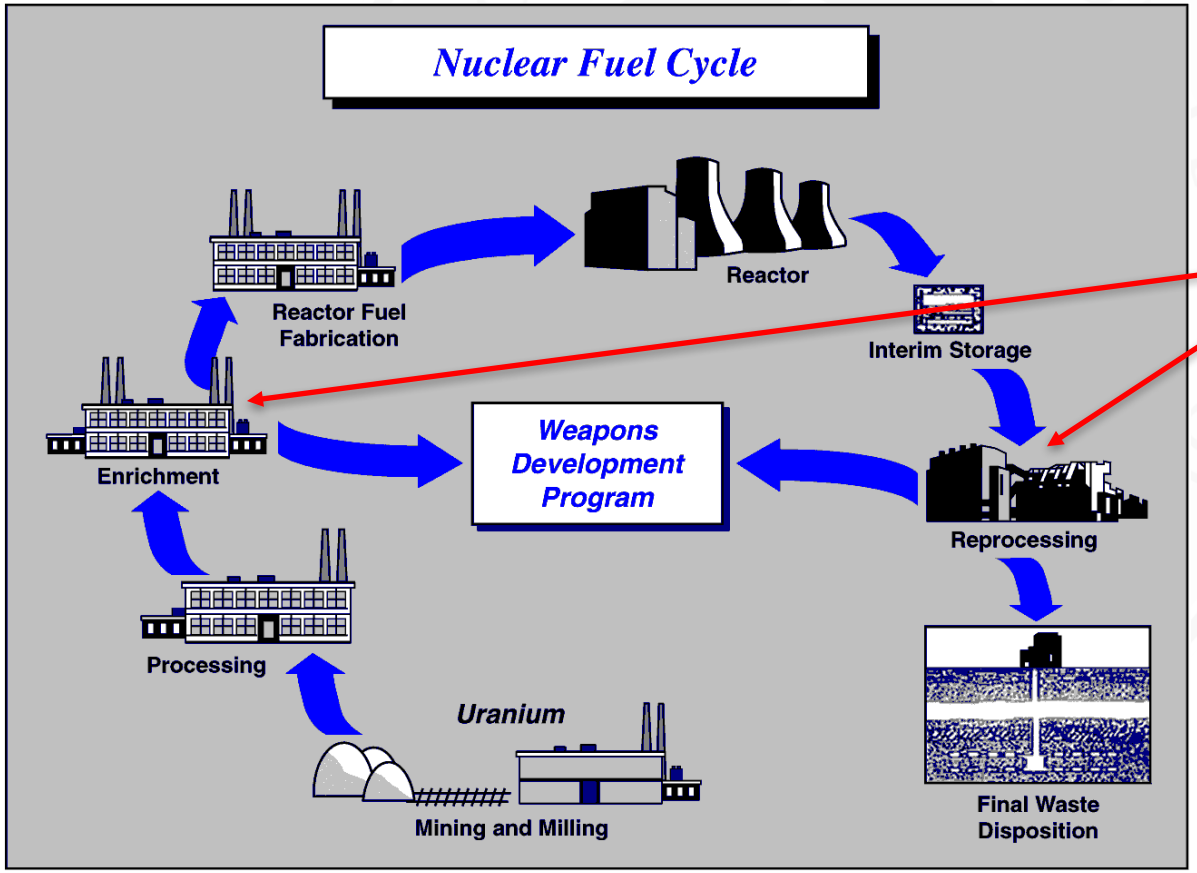


# Motivation

- Identifying documents indicating an unusual transaction between facilities can be helpful in identifying a diversion of nuclear materials
- While all diversions will produce anomalous documents, not all anomalous documents will necessarily indicate a diversion
- Some anomalous documents could be due to another rare event such as a transaction that only occurs once a year
  - Announcement for the annual plant shutdown

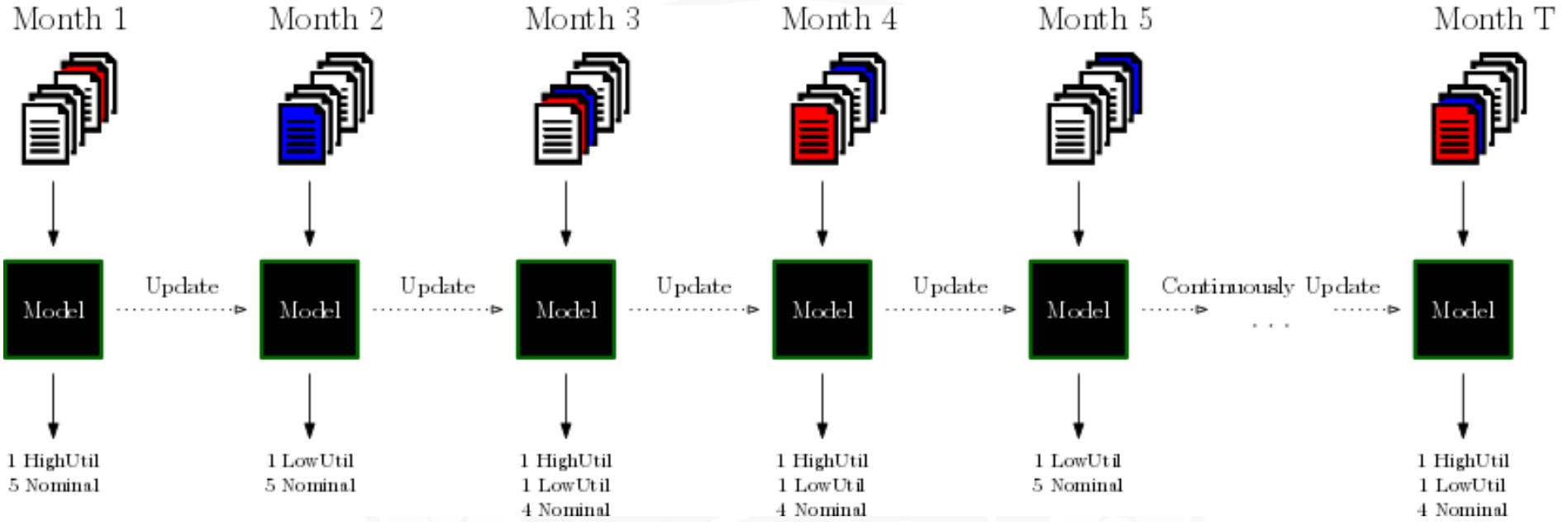
# Motivation

*Objective of IAEA safeguards: "...the timely detection of any diversion of significant quantities of nuclear material and to deter diversion by creating the risk of early detection." Leonard Weiss, Bulletin of Atomic Scientists, vol. 47, no. 4, 1991.*



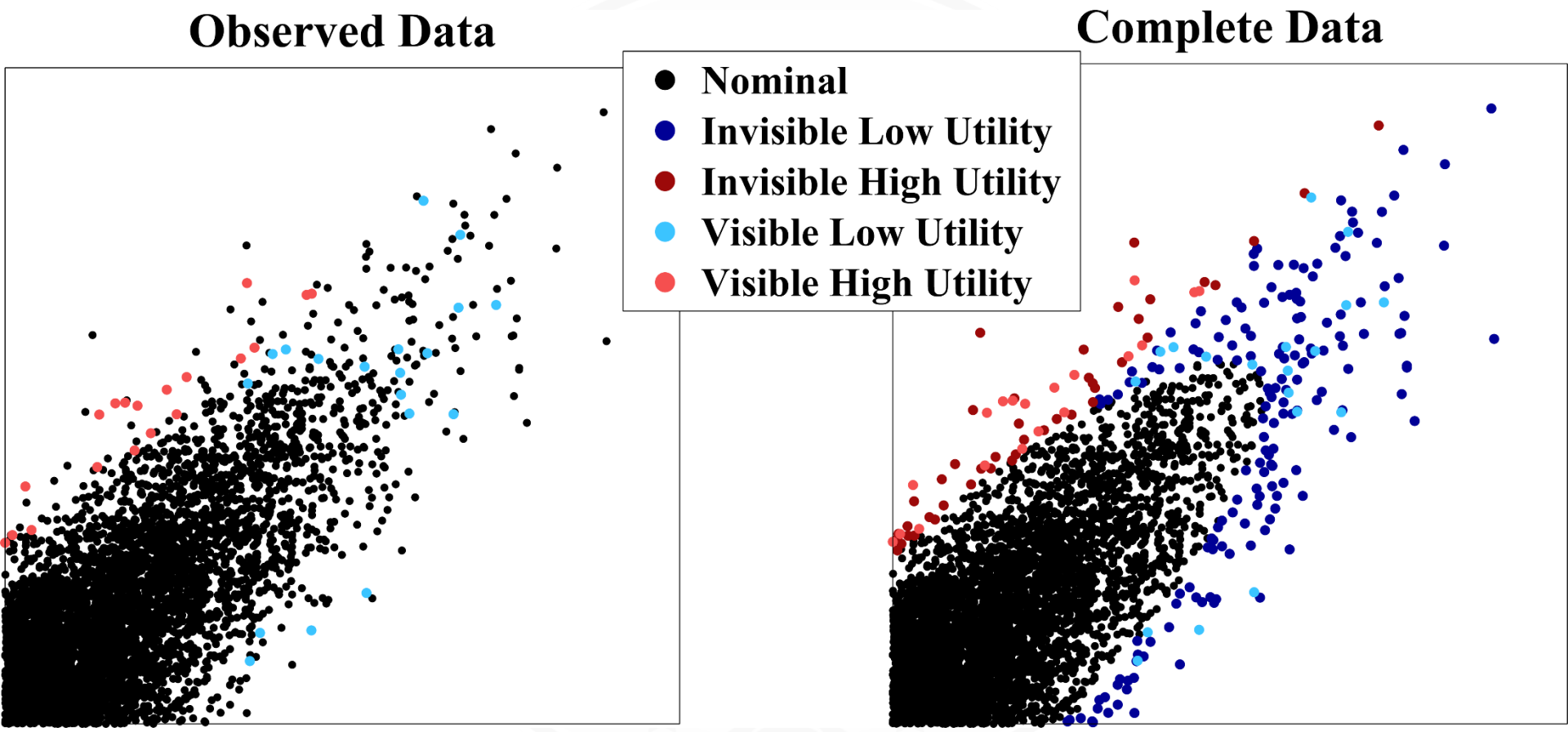

At the end of every month, documents from these facilities are collected and analyzed for unusual transactions

# Data Stream of Documents



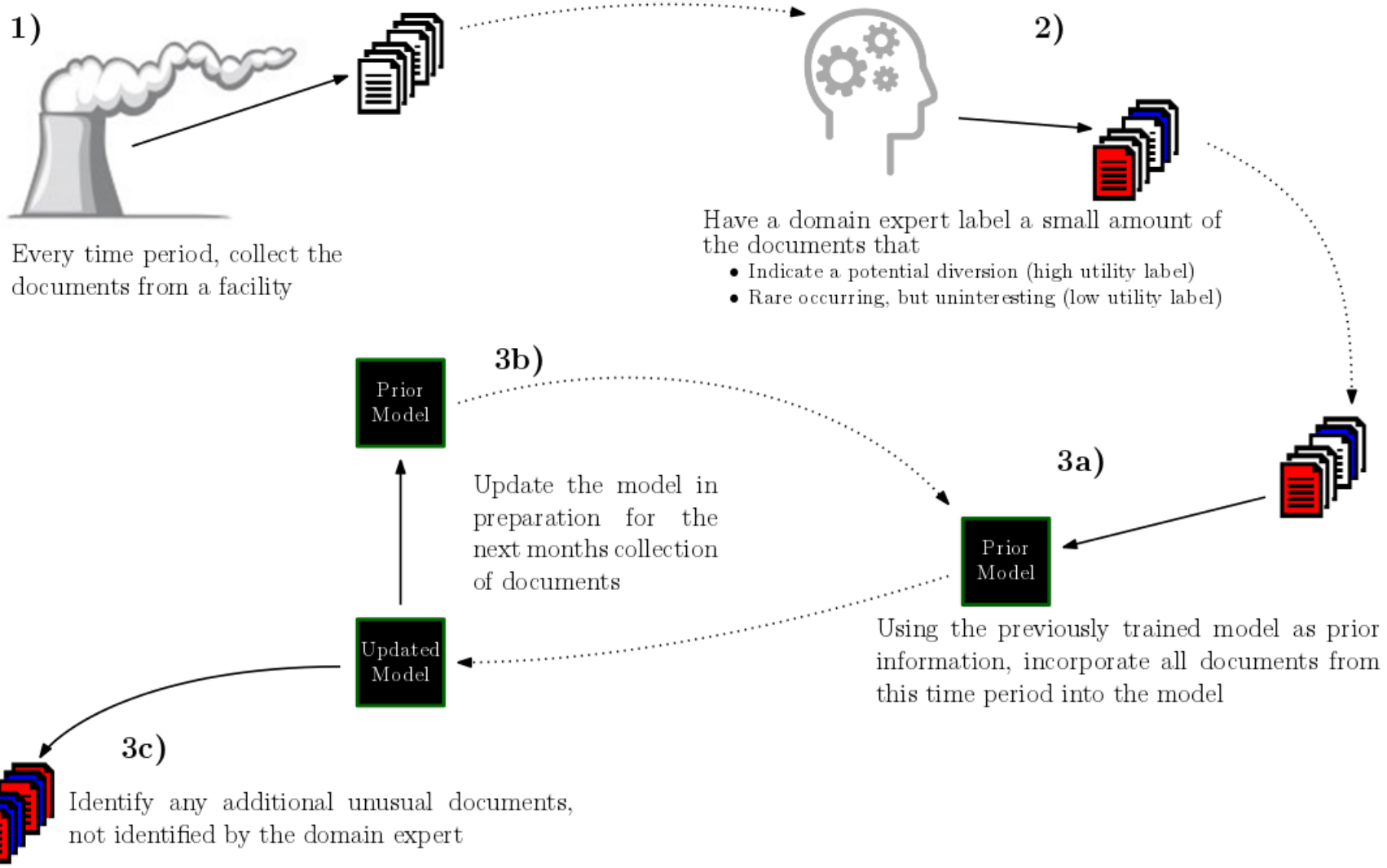

- **Problem:** The number of document features can be very large
- **Problem:** The number of documents increases as time passes, and the entire feature space may not be known at the beginning
- **Solution:** Kernel methods can account for a very large and changing feature space

# Snapshot of Data in Time




We are only interested in certain types of anomalies (red). But we can only observe some of the labels.

# Process Overview Part 1



# Sequential LatLapMED

Latent Laplacian Maximum Entropy Discrimination (LatLapMED) gives a posterior distribution whose mode is a maximum margin classifier

**Sequential Estimation:**

- Use the posterior at time t as the prior for data at time t + 1
- Get a posterior at time t + 1 and use the new discriminant function to classify the data
- Because the posterior from t is used as prior at t + 1, every posterior will contain all information from the previous time points

Time: t $\quad \mathrm{P}_0(\boldsymbol{\Theta}|\mathcal{D}^1,\hat{\boldsymbol{\eta}}^1,\ldots,\mathcal{D}^{t-1},\hat{\boldsymbol{\eta}}^{t-1})\mathrm{P}(\mathcal{D}^t|\boldsymbol{\Theta}) \rightarrow \mathrm{P}(\boldsymbol{\Theta}|\mathcal{D}^1,\hat{\boldsymbol{\eta}}^1,\ldots,\mathcal{D}^t,\hat{\boldsymbol{\eta}}^t)$

Time: t+1 $\quad \mathrm{P}_0(\boldsymbol{\Theta}|\mathcal{D}^1,\hat{\boldsymbol{\eta}}^1,\ldots,\mathcal{D}^t,\hat{\boldsymbol{\eta}}^t)\mathrm{P}(\mathcal{D}^{t+1}|\boldsymbol{\Theta}) \rightarrow \mathrm{P}(\boldsymbol{\Theta}|\mathcal{D}^1,\hat{\boldsymbol{\eta}}^1,\ldots,\mathcal{D}^{t+1},\hat{\boldsymbol{\eta}}^{t+1})$

# LatLapMED

$$\min_{\mathrm{P}(\boldsymbol{\Theta},\boldsymbol{\gamma},\lambda|\mathcal{D},\boldsymbol{\eta})} \mathrm{KL}\left(\mathrm{P}(\boldsymbol{\Theta},\boldsymbol{\gamma},\lambda|\mathcal{D},\boldsymbol{\eta})||\mathrm{P}_0(\boldsymbol{\Theta},\boldsymbol{\gamma},\lambda)\right)$$

subject to

$$\iiint \mathrm{P}(\boldsymbol{\Theta},\boldsymbol{\gamma},\lambda|\mathcal{D},\boldsymbol{\eta})\left(y_i\eta_1\left(\boldsymbol{X}_i\boldsymbol{\theta}+b\right)-\gamma_1\right)d\boldsymbol{\theta}dbd\boldsymbol{\gamma} \geq 0$$

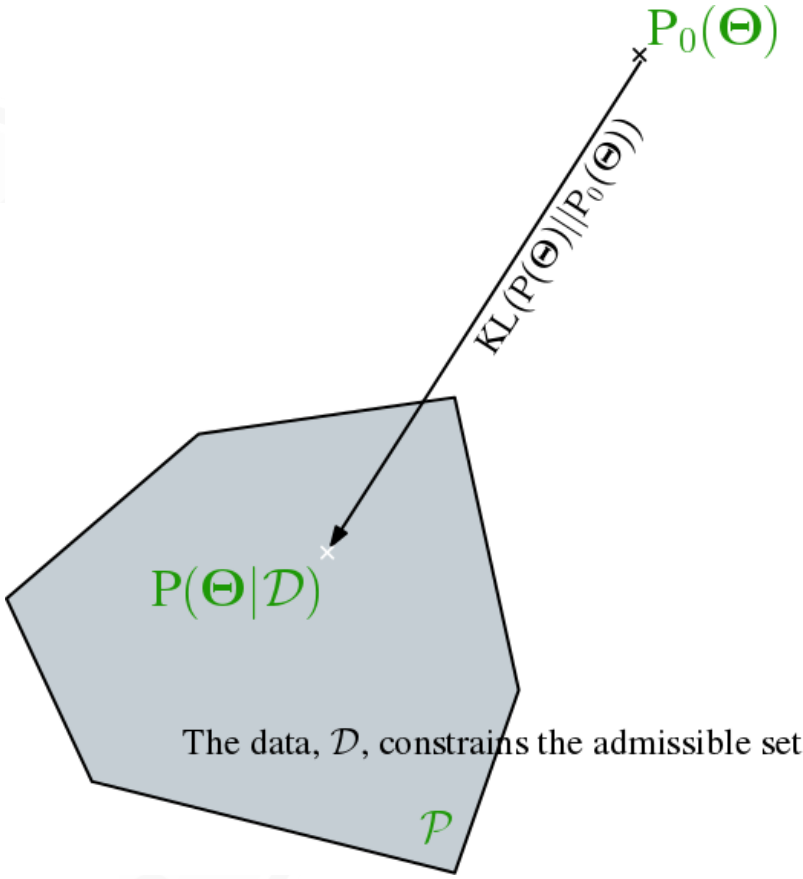
$$\vdots$$

$$\iiint \mathrm{P}(\boldsymbol{\Theta},\boldsymbol{\gamma},\lambda|\mathcal{D},\boldsymbol{\eta})\left(y_l\eta_l\left(\boldsymbol{X}_l\boldsymbol{\theta}+b\right)-\gamma_l\right)d\boldsymbol{\theta}dbd\boldsymbol{\gamma} \geq 0$$

Hinge Loss Constraints

$$\iint \mathrm{P}(\boldsymbol{\Theta},\boldsymbol{\gamma},\lambda|\mathcal{D},\boldsymbol{\eta})\left(\int_{x\in\mathcal{M}_{\boldsymbol{\eta}}} ||\nabla_{\mathcal{M}_\eta}\boldsymbol{X}\boldsymbol{\theta}||^2\, d\mathcal{P}_{X_\eta}\right) d\boldsymbol{\theta}d\lambda \leq 0$$

Constraint for Partially Labeled Points



# Experiment Settings

**Scenario:**

- A website with articles about sports, science, technology, and guns
- The domain expert is interested in identifying articles specifically about nuclear weapons
- Each month, the website produces 2,000 documents, the domain expert will label 15 articles about guns and 15 about nuclear weapons
  - Popular dataset of newsgroup documents used often in text classification and text clustering

**Goal:**

- Update the previous month's model with the current month's labeled and unlabeled documents
- Identify more articles about nuclear weapons from among the unlabeled documents

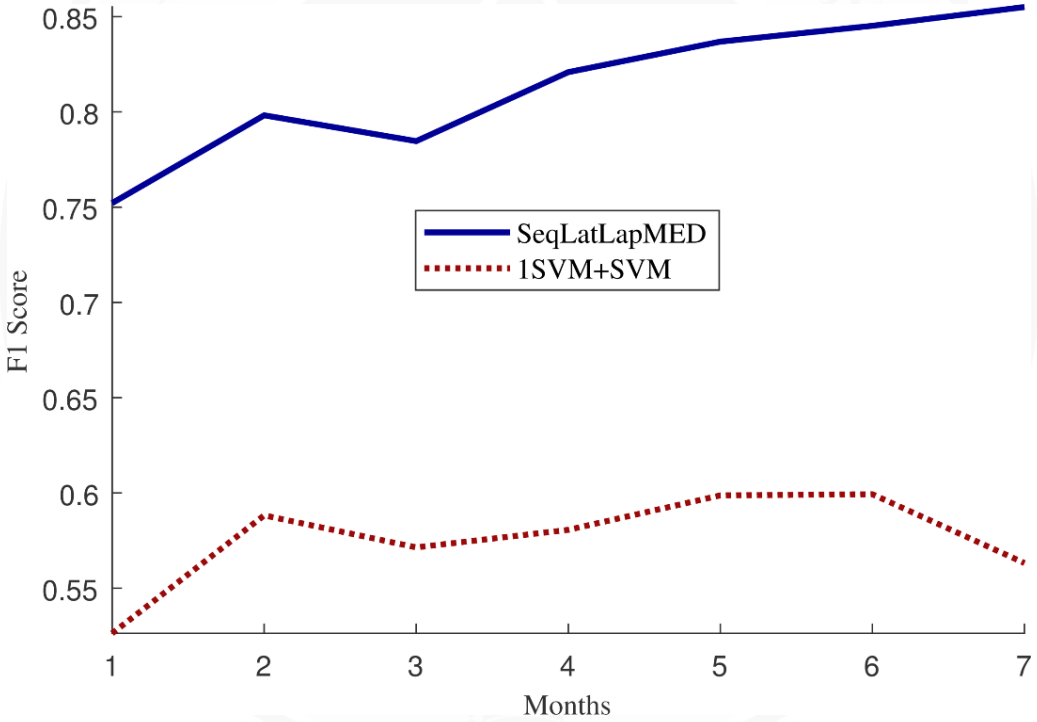
# Dataset

- Text data from the 20 Newsgroups data set, which has 60,648 distinct words
  - Popular dataset of newsgroup documents used often in text classification and text clustering
- Total dataset has 13,674 articles, broken into 7 pieces (months)
- **Nominal:** combine the articles with labels comp, rec, and sci
- **Anomalous:** articles with talk.politics.guns label
  - **High Utility:** anomalous articles injected with passages from NYT articles about nuclear weapons*
  - Each piece has articles only from a specific month (April – Oct 2017)
- Roughly 7% of the dataset is anomalous and 1.5% of the dataset is labeled

*Mengle, S., & Goharian, N. (2008, April). Detecting hidden passages from documents. In proceedings of SIAM Conference on Data Mining (SDM 2008) Workshop.

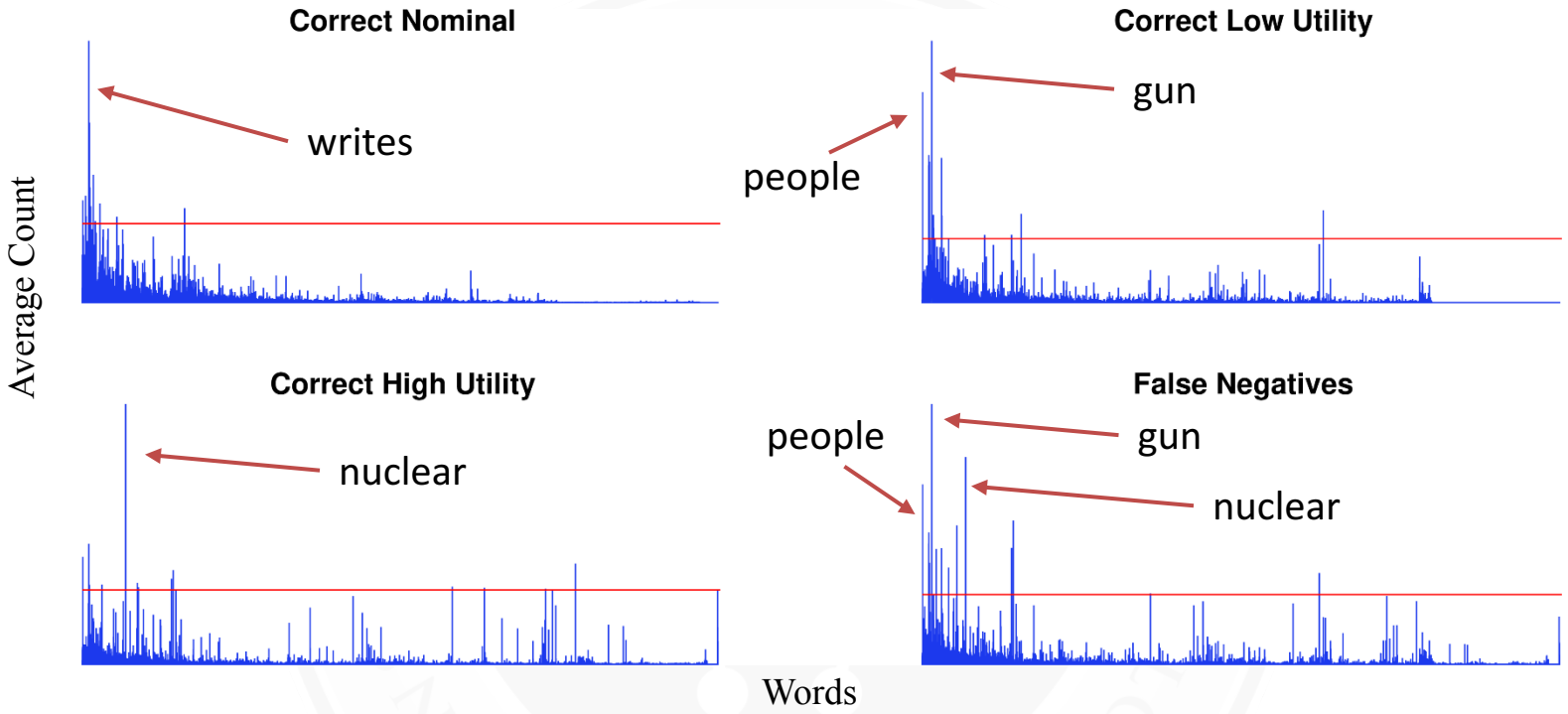
# Experiment Results

**F1 Score:** $F1 = 2\,\frac{Precision * Recall}{Precsion + Recall}$, the harmonic mean of the precision and recall. It measures accuracy: the ability to correctly identify only the nuclear weapons articles without missing any of them.




The sequential estimator is able to improve its F1 score. The other estimator cannot because it doesn't not learn from previous data.

# Experiment Results




- The word "nuclear" is the most frequently occurring in articles that SeqLatLapMED was able to correctly identify as about nuclear weapons
- The words "gun" and "people" are the 1st and 3rd most frequently occurring in articles that SeqLatLapMED failed to correctly identify as about nuclear weapons and they are the 1st and 2nd most frequently occurring in articles about guns

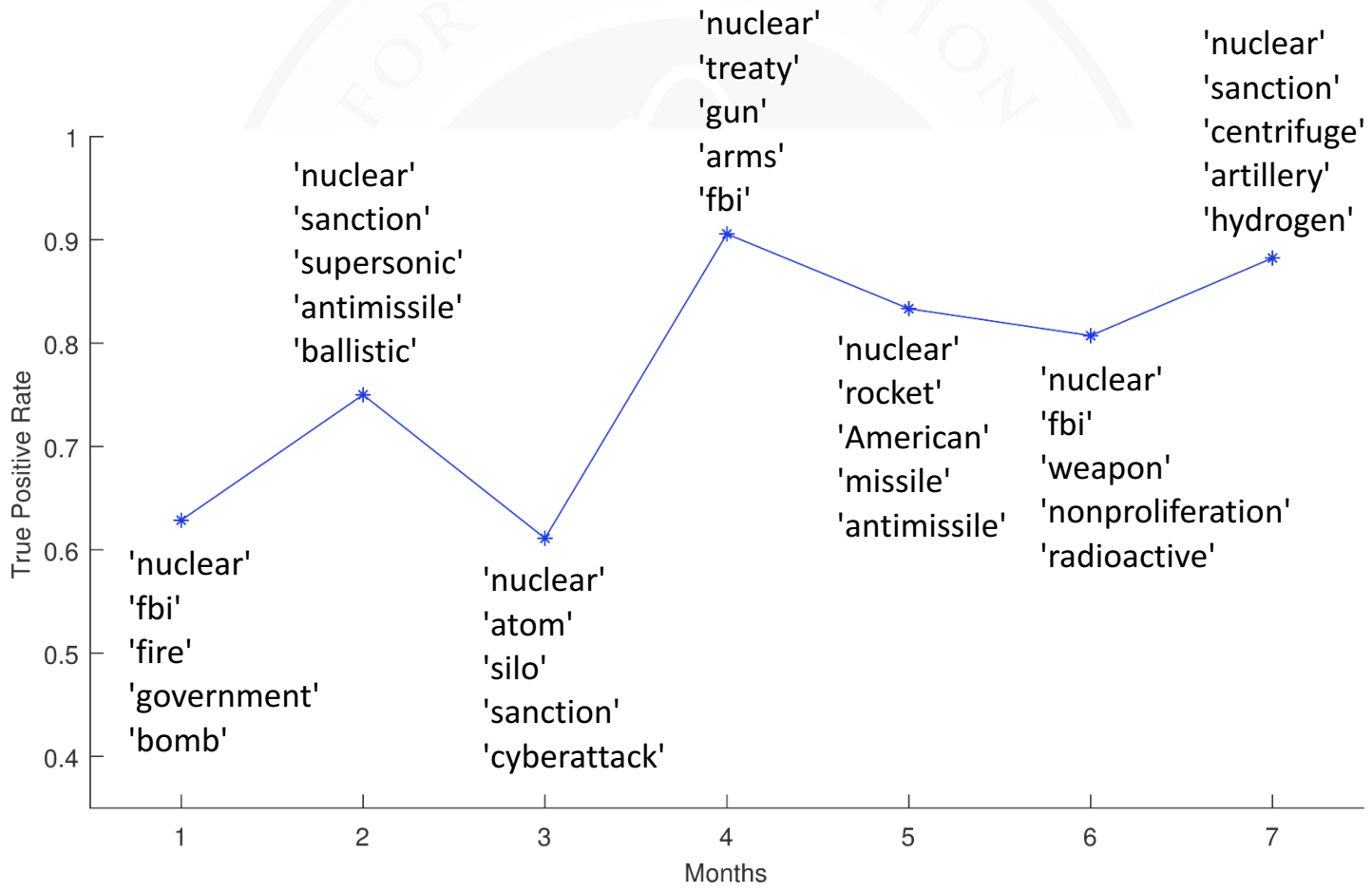
# Experiment Results

The top 15 most frequently occurring words for each category

| TN (nominal) | TN (low utility) | TP (high utility) | FN (high utility) |
|---|---|---|---|
| 'writes' | 'gun' | 'nuclear' | 'gun' |
| 'article' | 'people' | 'writes' | 'nuclear' |
| 'don' | 'writes' | 'article' | 'people' |
| 'time' | 'guns' | 'people' | 'fbi' |
| 'good' | 'don' | 'sanction' | 'file' |
| 'system' | 'article' | 'fbi' | 'writes' |
| 'people' | 'firearms' | 'fire' | 'article' |
| 'apr' | 'weapons' | 'bomb' | 'fire' |
| 've' | 'time' | 'government' | 'guns' |
| 'windows' | 'government' | 'don' | 'children' |
| 'ca' | 'law' | 'arsenal' | 'don' |
| 'work' | 'state' | 'atom' | 'control' |
| 'file' | 'fire' | 'ballistic' | 'militia' |
| 'problem' | 'police' | 'warhead' | 'batf' |
| 'mail' | 'control' | 'intercontinental' | 'state' |

# Experiment Results

The top 5 most frequently occurring unique words used by SeqLatLapMED to correctly identify articles about nuclear weapons



# Summary

- We propose a method to detect anomalous documents that combines rare event detection with semi-supervised diversion classification where only a small subset of the anomalous documents have been labeled as diversions by domain experts
- We provide an efficient algorithm to continuously update the detector and identify new diversions without re-fitting the entire model
- We demonstrate that significantly better diversion detection performance (low number of false positives and negatives) can be achieved by using the proposed algorithm