# An automated approach for finding spatio-temporal patterns in disease spread

Authors: Prathyush Sambaturu, Parantapa Bhattacharya, Jiangzhuo Chen, Bryan Lewis, Madhav Marathe, Srinivasan Venkatramanan, Anil Vullikanti

## Abstract

**Background**: Agencies such as the Centers for Disease Control and Prevention (CDC) currently release incidence data (e.g., Influenza), along with descriptive summaries of simple spatio-temporal patterns and trends. However, public health researchers, government agencies, as well as the general public, are often interested in deeper patterns and insights into how the disease is spreading, with additional context. Analysis by domain experts is needed for deriving such insights from incidence data.

**Objective:** Our goal is to develop an automated approach for finding interesting spatio-temporal patterns in the spread of a disease over a large region, such as: regions which have specific characteristics, e.g., high incidence in a particular week, those which showed a sudden change in incidence, or regions which have significantly different incidence compared to earlier seasons.

**Methods:** We develop techniques from the area of transactional data mining for characterizing and finding interesting spatio-temporal patterns in disease spread in an automated manner. A key part of our approach involves using the principle of minimum description length (MDL) for representing a given target set in terms of combinations of attributes (referred to as clauses); we consider both positive and negative clauses, and relaxed descriptions, which approximately represent the set, and use integer programming to find such descriptions. Finally, we design an automated approach, which examines a large space of sets corresponding to different spatio-temporal patterns, and ranks them based on the ratio of their size to their description length (referred to as their compression ratio).

**Results:** We apply our methods for finding spatio-temporal patterns in the spread of seasonal Influenza in the United States (US) using state level ILI activity indicator data from the CDC. We observe that the compression ratios are over 2.5 for 50% of the chosen sets, when approximate descriptions and negative clauses are allowed. Sets with high compression ratios (e.g., over 2.5) correspond to interesting patterns in the spatio-temporal dynamics of ILI. Our approach also outperforms other baselines in terms of the compression ratio.

**Conclusions:** Our approach, which is an unsupervised machine learning method, can provide new insights into the patterns and trends in disease spread in an automated manner. Our results show that the description complexity is an effective approach for characterizing sets of interest, which can be easily extended to other diseases and regions, beyond Influenza in the US. Our approach can be easily adapted for automated generation of narratives.

**Keywords:** Epidemic data analysis; Summarization; Spatio-temporal patterns; Transactional data mining.

## Introduction

Large-scale spatio-temporal analyses and forecasts are becoming increasingly common for several diseases, such as, Influenza [1, 2, 3, 4]. There is a lot of public interest in analysis of spatio-temporal trends relating to how these diseases are spreading across the US.–--this includes statements about whether the season has officially started, a listing of regions which have differing levels of activity, the contrast between the current season and earlier seasons, etc. Such analyses have a broad readership, and are popular among news media, the general public, government agencies, as well as public health organizations; this is evidenced by spatio-temporal patterns [5, 6] about the spread of Influenza from news agencies and blogs.

Such patterns are typically identified manually by domain experts, who have significant expertise on specific diseases. Data for such analyses often comes from public health agencies, such as the Centers for Disease Control and Prevention (CDC) [7] and World Health Organization (WHO). Reports generated by CDC contain raw surveillance data on metrics, e.g., activity level from outpatient visits and rates of hospitalization, across states in the US. In addition, summaries of regions with specific characteristics, e.g., those which have high activity levels, are also included in the reports. Such summaries can be found in the CDC reports [7, 8]. For instance, the CDC report in [8] summarizes the states with high Influenza-like Illness (ILI) activity for week ending on Mar 04, 2017 with the number of those states followed by explicit listing of their names.

Such descriptive listings are easy to construct from raw data but are tedious to read and do not provide deeper insights into the disease spread. In contrast, the analysis by Mashable [6] is a *succinct* description of the set of states which have widespread activity, namely, all states in the contiguous U.S., except Oregon. The analysis by the New York Times [5] is also a good and succinct description of the set of states which have reported widespread activity for three consecutive weeks. In addition to descriptions of the set of states with a particular activity level, sets exhibiting specific temporal patterns might also be of interest. An example is the set of states which maintained a stable high activity for three consecutive weeks, ending in the week of January 27, 2018: Most states which had high ILI activity level four weeks back, plus the states of New Jersey, New Mexico, Virginia, Washington, Wyoming. Such descriptions involve identification of features common to these states, which provide additional insights on the outbreak.

The overall objective of our work is to automate the process of identifying "interesting" spatio-temporal patterns from disease surveillance data, and generating succinct descriptions for them. In order to do this, we encode the incidence data as binary matrices (presence or absence of a feature), and use techniques from pattern mining [28, 29] of transactional data to find insights into epidemic spread; we demonstrate its utility using seasonal influenza in the United States as a case study.

## Methods

### Data

We use the Influenza-like Illness (ILI) activity indicator data available at state level from CDC [22]. In the dataset, each state for each week during a given influenza season, is assigned an activity level from 1 to 10 based on the severity of influenza prevalence in that week (measured using the percentage of outpatient visits that show influenza-like symptoms) [23]. These activity levels are also grouped into coarser labels such as Minimal (1-3), Low (4-5), Moderate (6-7), High (8-10). We also incorporate the geographic spread index as published by CDC in [24], which categorizes the states based on the internal spatial spread of influenza. We use a number of features associated with each state, which are defined by the CDC, and can be categorized as follows:

1. Geographical/ Spatial: Features such as Great Lakes, South East, Mid-Atlantic etc.
2. Temporal: Features such as activity level (e.g., high, moderate and low) in the $t$th week before the current one, geographical spread (e.g., widespread, local) in the $t$th week before the current one, whether the number of infections has crossed a threshold, whether the peak has been reached, and similarity with past season. In the description below, these features will be denoted by "was1_high" (states with high ILI activity 1 week ago), "was2_moderate" (states with moderate ILI activity 2 weeks ago), "was52_high" (states with high activity 52 weeks ago), etc. These features capture the spatial, temporal, and severity aspects of the reported cases. Full list of attributes and their description is presented in the Appendix.

In our experiments, we use data corresponding to weeks for the years 2014 to 2017. To generate narratives for a particular week, we use the data from these reports for the current week, the last three weeks, and the data 52 weeks ago to generate the temporal data for each state. This is expressed as a data matrix D with the following characteristics:

1. Number of regions (states) or rows: 51 (50 states and District of Columbia)
2. Number of features or columns: 42 (spatial, temporal, and severity features)

Therefore, the data matrix $D$ for a week has 2142 entries.

### Problem Formulation

Let $D_{n\times m}$ be the data matrix, where each row corresponds to a state and each column to a feature, and $D_{ij} = 1$ if state $i$ has feature $j$. Let $U = \{e_1, ..., e_n\}$ be the universe of elements, in our case, the set of all states. Let $D_j = \{i : D_{ij} = 1\}$ denote the set of elements having feature *j*. Let $S(j_1, \ldots, j_k) = D_{j_1} \cap \ldots \cap D_{j_k}$ denote the set of elements that have features $(j_1, \ldots, j_k)$ (denoted by $\boldsymbol{j}$); referred as a conjunctive *clause*. The clause $S(\boldsymbol{j})$ has length $k$, meaning that it is formed by the intersection of $k$ features.

Given a target set $T \subseteq U$, we consider expressions of $T$ in terms of unions and differences, i.e., $\mathrm{T} = \cup_{\ell=1}^{\mathrm{r}} \mathrm{S}(\mathbf{j}^{\ell}) - \cup_{\ell=\mathrm{r}+1}^{\mathrm{s}} \mathrm{S}(\mathbf{j}^{\ell})$, with an associated cost of $\sum_{\ell=1}^{r} \alpha \cdot \mathrm{NUM}(\boldsymbol{j}^{\ell}) + \sum_{\ell=r+1}^{s} \beta \cdot \mathrm{NUM}(\boldsymbol{j}^{\ell})$, where and $\alpha$ and $\beta$ are the constant parameters associated with positive ($\mathrm{S}(\mathbf{j}^{\ell})\ for\ l\ \in \{1,\ \ldots,\ r\}$) and negative clauses ($\mathrm{S}(\mathbf{j}^{\ell})\ for\ l\ \in \{r+1,\ \ldots,\ s\}$) respectively, and $NUM(\boldsymbol{j}^{\ell}) = k_{\ell}$ denotes the number of features involved in a clause $S(\boldsymbol{j}^{\ell}) = S(j_1^{\ell}, \ldots, j_{k_\ell}^{\ell})$. The negative clauses describe the elements which need to be removed from the set of positive clauses, in order to exactly cover the elements of $T$.

Given a subset $T \subseteq U$ (referred to as a "target" set), and a dataset $D$, the $MinDesc(T, D)$ problem involves finding a set of tuples $\boldsymbol{j}^{\mathbf{1}}, \ldots, \boldsymbol{j}^{\boldsymbol{s}}$, such that $T$ is represented in terms of unions and differences and the associated cost $\sum_{\ell=1}^{r} \alpha \cdot NUM(\boldsymbol{j}^{\ell}) + \sum_{\ell=r+1}^{s} \beta \cdot NUM(\boldsymbol{j}^{\ell})$ is minimized.

In order to make the descriptions interpretable, we will restrict the sizes of these clauses, i.e., the number $k_{\ell}$ of columns whose intersection is allowed; here, we will focus on $k_{\ell} \leq 2$, though our approach extends to any $k$.

Our main idea for finding patterns of interest is to explore the space of all target sets and identify those which have low cost descriptions. This is motivated by the *Minimum Description Length* (MDL) Principle, that forms the basis of many machine learning methods to find such descriptions; we refer to [15, 30] for details on this topic.

In some cases, the target set $T$ does not have a small description, but we can find a set $T^{'}$ which is *close* to $T$, and has a smaller description than $T$. We model this as finding a representation for a subset $T^{'}$ such that $T^{'} \approx T$, which is formalized as the $MinApproxDesc$ problem: Given a target set $T \subseteq U$, a dataset $D$, and constant parameters $\alpha, \beta, \gamma$, the $MinApproxDesc(T, D)$ problem involves finding a set of tuples $\boldsymbol{j}^{\mathbf{1}}, \ldots, \boldsymbol{j}^{\boldsymbol{s}}$, for representation of $T^{'}$as unions and differences, such that the symmetric difference of $T\ and\ T^{'}$is of size at most $\gamma|T|$, and the associated cost is minimized. Since *MinApproxDesc* is a generalization of *MinDesc*, we only consider the $MinApproxDesc$ problem in the rest of the paper. The $MinDesc$ and $MinApproxDesc$ problems are both NP-complete, even when $k_{\ell} = 1$, which corresponds to the *set cover* problem (we refer to [17] for discussion on this topic).

## Approach and Implementation

We use an integer programming approach described in the Appendix, which is able to scale well for the problems of interest in epidemic analysis. We use the Gurobi optimization software [18] to solve the resulting Integer program. The size of the instances encountered results in programs that can be solved very efficiently. So, we expect our method will scale to much larger datasets easily.

### Generate Set Descriptions.

We consider the set of states with a high activity level in the current week, as a target set $T$. We prepare the data matrix $D$ for the current week. These states have value 1 in the column 'high' of the matrix. Then, we use our method to compute the succinct descriptions for the target set $T$ for the parameters $(\alpha, \beta, \gamma)$ = (2, 2, 0). From the MDL principle, a set $T$ is likely to be an interesting pattern if it has a high compression ratio.

We also study the impact of the parameter $\gamma$ on the description length. Recall that the parameter $\gamma$ controls how accurately we attempt to describe the target set. A larger $\gamma$ would mean greater error, but should lead to a more succinct description. The target set $T$ is the set of states with high activity in the current week. We run our method for a given week with target set $T$ and, for each value of $\gamma \in \{0.1,\ 0.2,\ 0.3\}$.

### Ranking Set Descriptions.

It is not known a priori which target sets would give interesting patterns. We search from a large space of possible target sets corresponding to all clauses with up to $k$ terms (i.e., sets formed by intersections of up to k columns), compute their MDL scores, and rank them based on their compression ratio, and other characteristics.

### Baselines and Evaluation Measures

The work of Xiang et al. [11] is directly related to our approach and can be considered as a special case of *MinDesc*, where only positive clauses are allowed. We refer to this as DBS. We use the number of clauses used by DBS and *MinApproxDesc* for comparison.

We use the compression ratio as a metric for evaluating the performance of our method. Let the number of clauses used in description by *MinApproxDesc* for a target set $T$ be $s$. The compression ratio provided by *MinApproxDesc* is defined as the ratio of the target set size $|T|$ to the number of clauses used in the description by solution to *MinApproxDesc*,

$$compression\ ratio = \frac{|T|}{s}$$

We also provide a scoring system to determine the "interestingness" of a target set. Sets consisting of states with high activity level are likely to be more interesting than those with moderate, low or minimal activity levels; therefore, these are assigned scores 4, 3, 2, 1 respectively (i.e., 4 for sets with high activity level, and so on). Next, states exhibiting a sudden change in activity level (e.g., from low to high, or vice versa) are more interesting than those having no change in activity levels we assign a score of 5 for the former type, and 2 for the latter. Then, “a set of states with high this week and minimal 1 week ago” has a score of 9, while “a set of states with minimal this week and minimal 1 week ago” has a score of 3. This process is described in detail in the Appendix. The score assigned to each target set/ description measures its “interestingness”.

## Results

### Generate Set Descriptions.

The text description, in Table 1, is hand generated, corresponding to the solution computed using our method. The average Compression Ratio over all the rows in the table is over 2.6. This shows that our method can easily find succinct descriptions for different kinds of target sets. Using additional attributes for the regions might allow for more succinct descriptions.

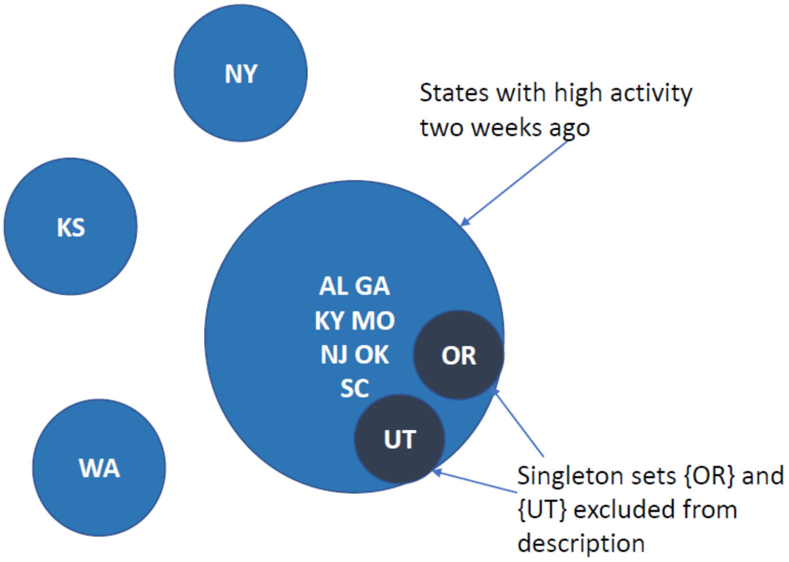
Table 1: Description for the set of states with high activity levels. The abbreviations are used for state names [20].

| S.No. | Week | Descriptions of states with high ILI activity in the week | No. of clauses | Target set T | \|T\| | Compression Ratio |
|---|---|---|---|---|---|---|
| 1 | 2017-01-21 | KS, NY, WA, and states with high activity two weeks back excluding OR and UT | 6 | AL, GA, KS, KY, MO, NJ, NY, OK, SC, WA | 10 | 1.67 |
| 2 | 2017-02-18 | Ak, IL, MD, MN, states with high activity a week ago, states with low activity two weeks ago, and states with minimal activity three weeks ago excluding WY | 7 | Al, AK, AR, CT, GA, IL, IN, KS, KY, LA, MD, MI, MN, MS, MO, NJ, NM, NY, NC, OK, PA, RI, SC, SD, TN, TX, VA | 27 | 3.86 |
| 3 | 2017-03-25 | States with high activity for last two weeks, excluding LA, MS and TX | 4 | AI, AR, GA, KS, KY, NC, OK, SC, TN, VA | 10 | 2.5 |
| 4 | 2017-04-08 | KY, SC | 2 | KY, SC | 2 | 1 |
| 5 | 2015-01-03 | CA, NV, NY, and states with high or moderate activity levels a week ago excluding FL and GA | 7 | Al, AR, CA, CO, HI, ID, IL, IN, Ks, KY, LA, MD, MN, MS, MO, NV, NM, NY, NC, OH, OK, PA, SC, TN, TX, UT, VA, WV, WI | 29 | 4.14 |

We now qualitatively evaluate the descriptions shown in Table 1 and examine the insights about epidemic outbreaks this gives. Some of the rows involve large target sets, e.g., rows 2 and 5 correspond to 27 and 29 states, respectively. The CDC descriptions for these weeks would be very long lists (as in Column 5), which are unlikely to give useful insights or identify any patterns. The description in row 5 (week 2015-01-03) is succinct and gives the following insights: Almost all the states with high or moderate activity level in the previous week are high in the current week. Three new states that were not experiencing high/moderate activity are now at the high activity level. Florida and Georgia have experienced a sharp decline in activity levels within a week.

We also note that some of the descriptions may not be insightful. For instance, the one for the week of 2017-04-08 (row 4) is simply a list. It is possible that there were no common characteristics of these two states, so that the most succinct description is just a list. The description for the week of 2017-02-18 (row 2) is quite complicated: it combines three sets of states with different activity levels in different times in the past. Figure 1 shows that a set of 10 states with high ILI for week 2017-01-21 is represented by our algorithm using 6 sets/clauses. The Compression ratio achieved here is 1.67 as we only use 6 clauses instead of listing 10 state names. However, automated generation of such descriptions will allow a human expert to filter and select appropriate descriptions, instead of creating them from scratch.

Figure 1: The set representation of the description for week 2017-01-21 (row 1). Each circle is a set and the states in the set are listed with their respective abbreviations. The states in the blue region correspond to the target set T. OR and UT are the singleton subsets (in dark blue) with high ILI activity two weeks ago but not in the current week.

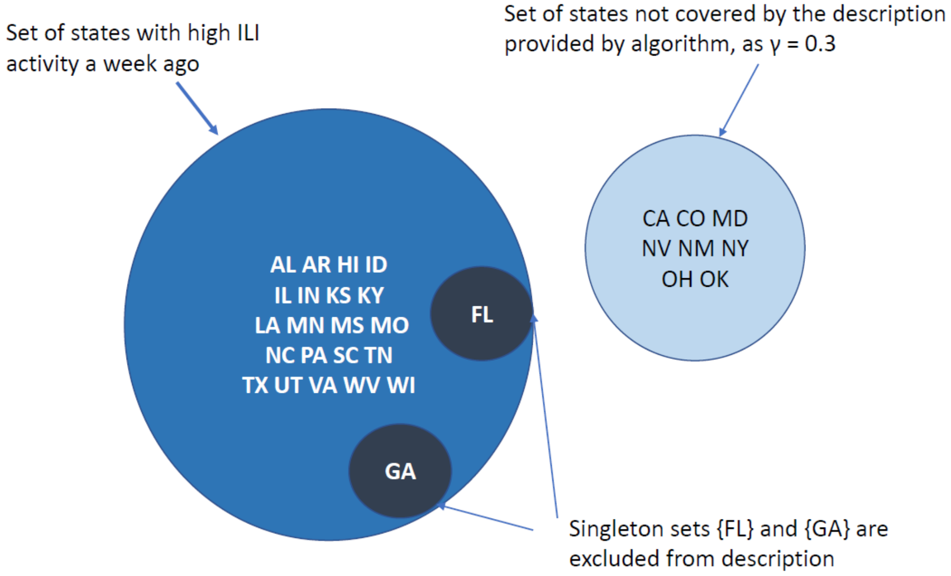

NY
KS
WA
AL GA
KY MO
NJ OK
SC
OR
UT
States with high activity two weeks ago
Singleton sets {OR} and {UT} excluded from description

The Compression Ratio increases as we increase the relaxation factor γ. Figure 2 shows that a set of 29 states with high ILI for week 2015-01-03 can be represented using only 3 sets/clauses. Although, 8 out of the 29 states are omitted from the description (shown in light blue region), as relaxation parameter is set to 0.3.

Table 2: Impact of varying relaxation factor γ on the description and compression ratio. Rows 1.1-1.3 correspond to row 1 in Table 1, and rows 5.1-5.3 correspond to row 5 in Table 1.

| **S.No.** | **γ** | **Description** | **# Clauses** | **Compression Ratio** |
|---|---|---|---|---|
| 1.1 | 0.1 | KS, WA, and states with high activity two weeks ago, excluding OR and UT | 5 | 2 |
| 1.2 | 0.2 | NY and states with high activity two weeks back, excluding OR and UT | 4 | 2.5 |
| 1.3 | 0.3 | States with high activity two weeks back excluding OR and UT | 3 | 3.33 |
| 5.1 | 0.1 | NY, and states with high or moderate activity levels a week ago excluding FL and GA | 5 | 5.8 |
| 5.2 | 0.2 | States with high or moderate activity levels a week ago excluding FL and GA | 4 | 7.25 |
| 5.3 | 0.3 | States with high activity level a week ago excluding FL and GA | 3 | 9.67 |

Figure 2: The set representation of description of set of states with high ILI activity on 2015-01-03 (row 8). The blue set corresponds to the states with high activity a week ago. The dark blue colored singletons FL and GA are subsets of the blue set, but do not have high activity in the current week. The faded blue colored set consists of the states omitted from the description due to relaxation.


Set of states with high ILI activity a week ago
Set of states not covered by the description provided by algorithm, as γ = 0.3
AL AR HI ID
IL IN KS KY
LA MN MS MO
NC PA SC TN
TX UT VA WV WI
FL
GA
CA CO MD
NV NM NY
OH OK
Singleton sets {FL} and {GA} are excluded from description

## Ranking Set Descriptions

Table 4: “Interestingness” scores.

| S.No. | Week | $\gamma, \alpha, \beta$ | Target set/ Pattern | Description | Score |
|---|---|---|---|---|---|
| 1 | 2018-01-27 | (0, 2, 2) | States with high activity this week, low activity two weeks ago, and moderate three weeks ago | HI, MD, NC, OH | 14 |
| | | | States with moderate activity a week ago, minimal activity two weeks ago, and low three weeks ago | ND | 13 |
| | | | States with low activity two weeks ago, moderate three weeks ago, and minimal four weeks ago | MD, NC, OH | 7 |
| 2 | 2017-02-25 | (0.3, 2, 4) | States with high activity one week ago, low activity two weeks ago, and moderate three weeks ago | IA | 14 |
| | | | States that had moderate activity levels one week ago, minimal activity levels three weeks ago and minimal activity levels four weeks ago | MA, OH, WI | 8 |

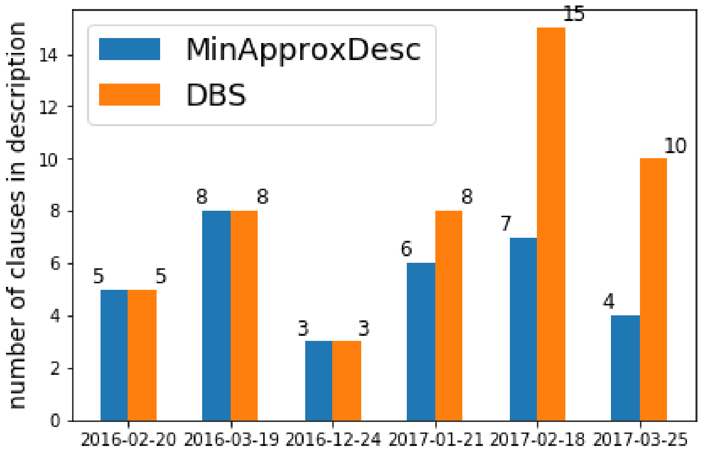
We find that the top scoring narratives generally are trends. Some examples of trend type of descriptions found by our method are the following: *Gradual increase in the activity levels over consecutive weeks*: The states AL, GA, MS, and TN had high activity in the week of 2016-03-12, moderate the previous week, and minimal two weeks ago. *Stable high activity for consecutive weeks*: In the week ending 2018-01-27, the states NJ, NM, VA, WA, WY, and the states with high activity four weeks earlier, excluding NE and TN, had high activity levels for three consecutive weeks. *Gradual decrease in ILI activity over consecutive weeks*: For the week of 2014-02-01, the activity levels in NC decreased from high to moderate to low in three consecutive weeks.

Examples of surprise events identified by our methods are as follows: The activity level in NC, NM, SD, and WY jumped from low to high within a week, for the week ending 2017-02-04. The activity level in NH and TN changed from high to low within a week, for the week ending 2013-02-02.

## Comparison with baselines

*MinApproxDesc* (see Figure 3) clearly provides summaries of smaller cost compared to that of DBS for the weeks 2017-01-21, 2017-02-18, and 2017-03-25. For the remaining weeks, it provides summaries of same cost as that of DBS.

Figure 3: Solution comparison: MinApproxDesc vs. DBS.



## Discussion

There is a lot of work on finding spatio-temporal patterns in different datasets. These are typically unsupervised machine learning methods, and we refer the readers to [25, 26] for surveys on different algorithms and their applications to various datasets. As is the case with other unsupervised methods, the specific technique depends on the application. The approach of finding patterns based on compression and small description have been found to be useful in many settings, e.g., [27] and Xiang et al. [11]. As we show in our results, we find that our description length based approach gives useful insights into spatio-temporal patterns in incidence of ILI, especially when negative clauses are allowed. However, no prior methods handle negative clauses, to the best of our knowledge. In addition to negative clauses, we also find that the relaxed versions can also significantly reduce the complexity of descriptions in many cases.

Our ranking method also provides a systematic approach to identify trends and surprises in the spread of ILI. However, the descriptions of high score are not always intuitive or interesting, which is often the case with unsupervised machine learning methods. Instead, our ranking based approach (or other variations of it) could help provide new insights to a domain expert, who might be able to find interesting spatio-temporal patterns more easily. Thus, such an approach could be a first step in processing epidemic incidence data. We believe that including more characteristics for the data (i.e., more columns in the data matrix $D$) can help find more succinct descriptions. Further, the integer programming based approach is quite powerful, and more constraints can be easily added to generate descriptions with specific kinds of properties. Though the descriptions reported here were generated by hand, these are all very well structured, and could conceivably be generated using natural language processing techniques easily.

We compare the performance of our method with two other pattern detection methods in the literature as baselines, though, as mentioned earlier, they do not consider negative clauses. The first method, *Apriori* [9] is a very popular approach for association rule mining and pattern detection in a database containing transactions. Each transaction is seen as a set of items called itemset. The Apriori algorithm finds the frequent item sets in the database, the item sets that appear frequently among the transactions of the database. We observe that the rules generated by Apriori are trivial in nature and are not very informative.

The work of Xiang et al. [11] (DBS) can be considered as a special case of *MinDesc*, where only positive clauses are allowed. Xiang et al. give a logarithmic approximation for the DBS problem for such instances. We implement an Integer Linear Program to solve this problem exactly. By comparing the solutions provided by *MinApproxDesc* with that of DBS, we demonstrate the benefit of allowing differences in generating compact descriptions. This analysis shows that these two baselines do not give very useful insights for the type of dataset considered here.

Our methodology can be easily extended to other diseases and applications involving spatio-temporal data. The method can handle very general kinds of features and clauses formed by them. The ranking method has to be designed based on the specific domain.

## Limitations

The feature values are real numbers, e.g., the similarity with a past season can be a correlation metric, not binary. One way to handle this issue would be to map the non-binary values to binary using discretization of the weights. Since we limited our focus to only meaningful features, our current approach explores target sets with temporal properties over small time intervals. In case of an increase in number of features by a few orders of magnitude than we considered, the ILP may not be able to scale well. One way to address this problem is to design scalable heuristics that give some theoretical/ experimental guarantees.

## Conclusion

Automated generation of interesting spatio-temporal patterns and trends is an important problem, and can be very useful to public health experts, as well as the general public. Our approach, based on techniques from pattern mining, provides a short-list of patterns in ILI data from the CDC. We find that sets with high compression ratio tend have common characteristics, which are often interesting. This is, however, an unsupervised machine learning method, and needs to be verified manually. Our ranking method is one way to select interesting patterns in an automated manner. The techniques developed in this paper could potentially be applied for other diseases, and other public health domains.

## Acknowledgements

The work of the authors has been partially supported by the following grants: NSF grant IIS-1633028, NSF grant ACI-1443054, Defense Threat Reduction Agency (DTRA) grants HDTRA1-11-1-0016 and HDTRA1-17-D-0023.

## Authors' Contribution

PS, PB, BL and AV designed the study. PS, PB and AV developed the methods. All the authors helped in the evaluation and writing.

## Conflicts of Interest

None

## References

1. Chakraborty P, Khadivi P, Lewis B, Mahendiran A, Chen J, Butler P, Nsoesie E, Mekaru S, Brownstein J, Marathe M, Ramakrishnan N. Forecasting a moving target: Ensemble models for ILI case count predictions. SIAM International Conference on Data Mining; 2014. p. 262-270. [doi:10.1137/1.9781611973440.30]

2. Tizzoni M, Bajardi P, Poletto C, Ramasco J, Balcan D, Goncalves B, Perra N, Colizza V, Vespignani A. Real-time numerical forecast of global epidemic spreading: case study of 2009 A/H1N1pdm; 2012: 23(1):169-214. [doi:10.1186/1741-7015-10-165]

3. Wang Z, Chakraborty P, Mekaru S, Brownstein J, Ye J, Ramakrishnan N. Dynamic poisson autoregression for influenza-like-illness case count prediction. ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; 2015; p. 1285-1294. [doi:10.1145/2783258.2783291]

4. Brooks LC, Farrow DC, Hyun S, Tibshirani RJ, Rosenfeld R. Flexible modeling of epidemics with an empirical framework; PLoS Comput Biol.; 2015. [doi:10.1371/journal.pcbi.1004382]

5. "This flu season is the worst in nearly a decade, new york times, 2018. URL: https://www.nytimes.com/2018/01/26/health/flu-rates-deaths.html. [accessed 2018-11-15]. [WebCite Cache ID 73xIRUdhv] .

6. Mashable, cdc reports flu season is worsening, as 17 more children die, 2018. URL: https://mashable.com/2018/02/02/cdc-says-2018-flu-season-worse-children-deaths/#6KaneYhQEmqf. [accessed 2018-11-08]. [WebCite Cache ID 73mqMIFTH]

7. 2017-18 influenza season week 6 ending feb 10, 2018. URL: https://www.cdc.gov/flu/weekly/weeklyarchives2017-2018/Week06.htm. [accessed 2018-11-08] [WebCite Cache ID 73mqpvW7z]

8. 2016-17 influenza season week 9 ending mar 04, 2017. URL: https://www.cdc.gov/flu/weekly/weeklyarchives2016-2017/Week09.htm. [accessed 2018-11-15] [WebCite Cache ID 73xI7qtXo]

9. Agrawal R, Srikant R. Fast algorithms for mining association rules in large databases. International Conference on Very Large Data Bases (VLDB); 1994: p.487-99. [URL: http://dl.acm.org/citation.cfm?id=645920.672836]

10. Madeira SC, Oliveira AL. Biclustering algorithms for biological data analysis: a survey. IEEE/ACM Transactions on Computational Biology and Bioinformatics; 2004: p.24-45. [doi: 10.1109/TCBB.2004.2]

11. Xiang Y, Jin R, Fuhry D, Dragan FF. Summarizing transactional databases with overlapped hyperrectangles. Data Min. Knowl. Discov.; 2011; 23(2): p.215–251. [doi: 10.1007/s10618-010-0203-9]

12. Wu ST, Li Y, Xu Y, Pham B, Chen P. Automatic pattern taxonomy extraction for web mining. International Conference on Web Intelligence; 2004: p.242-48. [doi: 10.1109/WI.2004.10132]

13. Chandola V, Kumar V. Summarization - compressing data into an informative representation. IEEE International Conference on Data Mining (ICDM'05); 2005. [doi: 10.1007/s10115-006-0039-1]

14. Miettinen P, Vreeken J. Model order selection for boolean matrix factorization. ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD); 2011: p. 51-59. [doi: 10.1145/2020408.2020424]

15. Grünwald P. The Minimum Description Length Principle. MIT Press; 2007. URL: https://mitpress.mit.edu/books/minimum-description-length-principle. [accessed 2018-11-16]. [WebCite Cache ID 73yYgrtrL]

16. Vreeken J, Leeuwen MV, Siebes A. Krimp: mining itemsets that compress. Data Mining and Knowledge Discovery; 2011; 23(1) p.169-214. [doi: 10.1007/s10618-010-0202-x]

17. Garey MR, Johnson DS. Computers and Intractability: A Guide to the Theory of NP-Completeness. W.H. Freeman and Co.; 1979. URL: https://en.wikipedia.org/wiki/Computers_and_Intractability. [accessed 2018-11-16]. [WebCite Cache ID: 73yZKCXj2]

18. Gurobi. URL: http://www.gurobi.com/. [accessed 2018-11-08]. [WebCite Cache ID 73mAgeuFX]

19. Eibe Frank, Mark A. Hall, and Ian H. Witten. The WEKA Workbench. Online Appendix for "Data Mining: Practical Machine Learning Tools and Techniques". Morgan Kaufmann, Fourth Edition, 2016 [https://dl.acm.org/citation.cfm?id=3086818]

20. List of us state abbreviations. URL: https://en.wikipedia.org/wiki/List_of_U.S._state_abbreviations. [accessed 2018-11-15]. [WebCite Cache ID 73xIzdXzb]

21. CDC. Past Weekly Surveillance Reports. URL: https://www.cdc.gov/flu/weekly/pastreports.htm. [accessed 2019-06-16]. [WebCite Cache ID y5uqe3wr]

22. CDC. FluView. URL: https://gis.cdc.gov/grasp/fluview/main.html [accessed 2019-12-02].

23. CDC FluView Dashboard. URL: https://gis.cdc.gov/grasp/fluview/fluportaldashboard.html. [accessed 2019-12-06]

24. CDC ILI Geographic Spread Index. URL: https://gis.cdc.gov/grasp/fluview/FluView8.html. [accessed 2019-12-06]

25. Gowtham Atluri, Anuj Karpatne, and Vipin Kumar. 2018. Spatio-Temporal Data Mining: A Survey of Problems and Methods. *ACM Comput. Surv.* 51, 4, Article 83 (August 2018), 41 pages. URL: DOI: https://doi.org/10.1145/3161602.

26. Zhenhui Li. 2014. Spatiotemporal pattern mining: Algorithms and applications. In Frequent Pattern Mining. Springer, 283--306. URL: doi="10.1007/978-3-319-07821-2_12".

27. Jae-Gil Lee, Jiawei Han, and Kyu-Young Whang. 2007. Trajectory clustering: a partition-and group framework. In *Proceedings of the 2007 ACM SIGMOD international conference on Management of data* (SIGMOD '07). ACM, New York, NY, USA, 593-604. URL: https://doi.org/10.1145/1247480.1247546.

28. J. Han, H. Cheng, D. Xin and X. Yan. 2007. Frequent Pattern Mining: Current Status and Future Directions. Data Mining and Knowledge Discovery archive, Vol. 15 Issue 1, pp. 55 – 86. URL: https://link.springer.com/article/10.1007/s10618-006-0059-1.

29. Jiawei Han, Jian Pei, Behzad Mortazavi-Asl, Qiming Chen, Umeshwar Dayal, and Mei-Chun Hsu. 2000. FreeSpan: frequent pattern-projected sequential pattern mining. In Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining (KDD '00). ACM, New York, NY, USA, 355-359. DOI=http://dx.doi.org/10.1145/347090.347167.

30. Peter D. Grnwald, In Jae Myung, and Mark A. Pitt. 2005. Advances in Minimum Description Length: Theory and Applications (Neural Information Processing). The MIT Press. URL: https://dl.acm.org/citation.cfm?id=1051706.