Detection of Informative Disjunctive Patterns in Support of Clinical Informatics

Artur Dubrawski, Ph.D., M.Eng.
Maheshkumar (Robin) Sabhnani, M.Sc.

The Auton Lab
School of Computer Science
Carnegie Mellon University
www.autonlab.org
awd@cs.cmu.edu

Carnegie Mellon

Auton Lab

Detection of Informative Disjunctive Patterns in Support of Clinical Informatics

1. Brief introduction
2. Patterns in Clinical Data and their representation
3. Example applications:
   A. Discovering patterns of care correlated with increased probability of readmission to an Intensive Care Unit
   B. Microarray analysis in cancer research
   C. Over-densities in outpatient diagnoses
4. Summary

Carnegie Mellon University is in Pittsburgh, PA, USA

CMU Auton Lab: Research and applications

Central topic of our research: scalable, self-adaptive analytic systems with real life impact

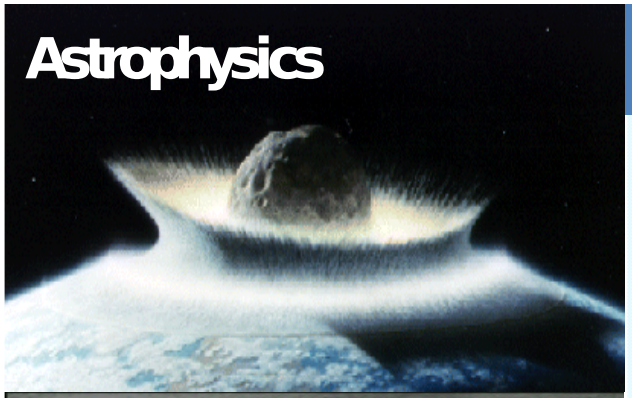
~20 people: 2 regular+3 affiliated faculty, 1 post-doc, 6 analysts and programmers, 6 PhD students; plus a few interns; led by Artur Dubrawski and Jeff Schneider

Working on 10+ sponsored projects.

Current and past funding from NSF, DARPA, DHS, HSARPA, ONR, AFRL, NASA, USDA, FDA, CDC, a few Fortune 100 companies, and a number smaller corporate & academic sponsors and partners

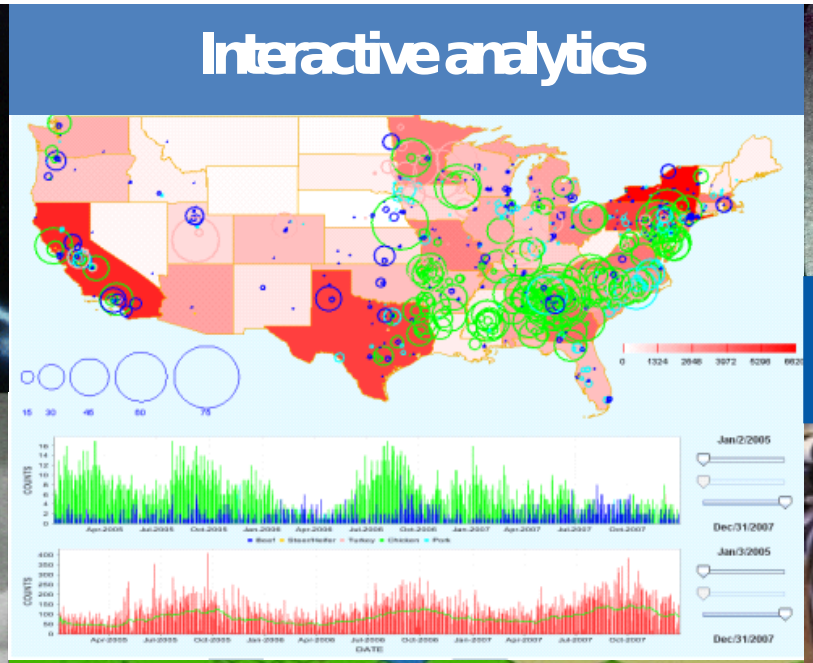
Deliverables

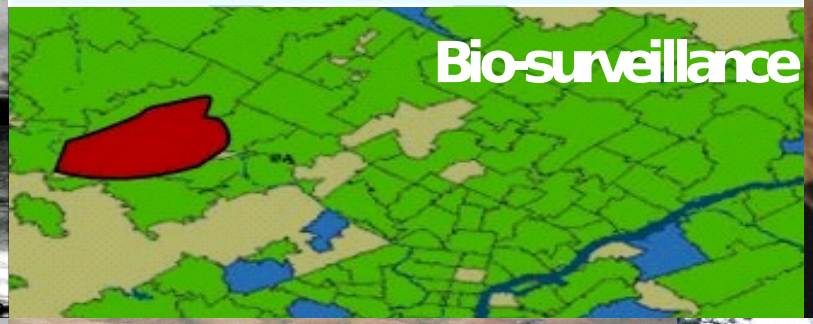
- Algorithms for fast and scalable statistical machine learning and analytics
- Software for embedding in production systems
- Software available for download at www.autonlab.org


Astrophysics

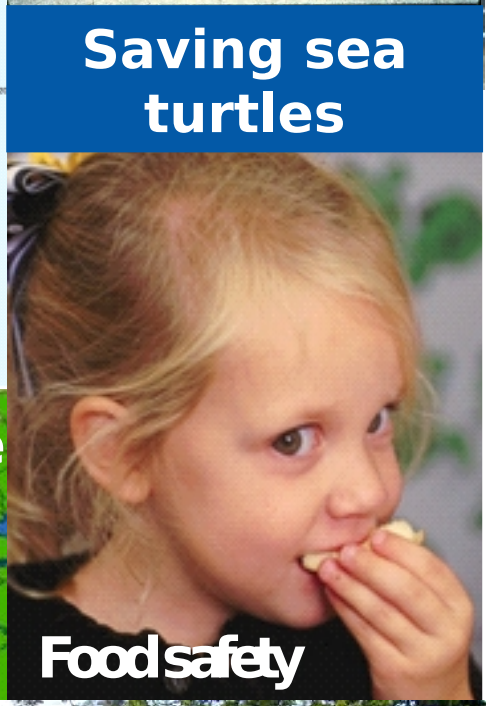

X
United Nations
CTBTO


Nuclear
threat detection


Interactive analytics


Bio-surveillance


Fleet prognostics


Saving sea
turtles
Food safety


Safety
of agriculture

Patterns in Clinical Data

- **A pattern in clinical care may be**
  e.g. predictive of a specific outcome:

If (X is followed) then Prob(Outcome = □) = P

- **One goal of Clinical Informatics is to automate searching through historical records of treatment**
  for such predictive patterns

- **Solutions typically employ any subset of:**
  Data Mining, Machine Learning, Probability, Statistics

- **The key tradeoffs:**
  Accuracy vs. Complexity vs. Tractability

If (X is followed) then Prob(Outcome = ▪) = P

One representation of X:

X : (A = ai AND B = bk) where A, B – attributes of data, and ai , bk – their values

Example: The patient received 12 doses of heparin throughout their stay at ICU (A="10-12 doses") and their stay lasted at most 5 but more than 3 days (B="(3,5]"), and A and B represent dosage of heparin and length of stay, respectively

That statement X was conjunctive

A conjunctive-disjunctive statement would look

Motivation of this talk

Many current approaches to Clinical Data Analysis rely on conjunctive statements to represent relationships between multiple dimensions of data…

This talk argues for conjunctive-disjunctive representation of patterns as a more powerful alternative

Our proposed framework develops models that are conjunctive with respect to features of data and disjunctive with respect to their values

(we assume the features are discrete)

Example application: ICU readmissions

Problem statement:

- **ICU readmissions are undesirable**
- **A readmission occurs if a recently discharged ICU patients returns within 48 hours (common definition)**

Research question:

- **Can the records of treatments help identify who of the current patients are at the elevated risk of readmission?**

Plausible solution:

- **Mine available data for combinations of features and values that correlate with increased readmission rates observed among past patients**

Our data and Approach

- **29,441 admissions to ICUs**

  189 (0.64%) resulted in readmission

- **Available features of data (more than 9,000 overall):**
  - **Demographics: age, gender**
  - **64 Medications: Average volume of medication per day when it was administered (numeric)**
  - **5,034 IOs: Average volume of fluid intake/outlet per hour during ICU stay (numeric)**
  - **4,133 Vitals: was the particular vital measured during the stay (binary)**

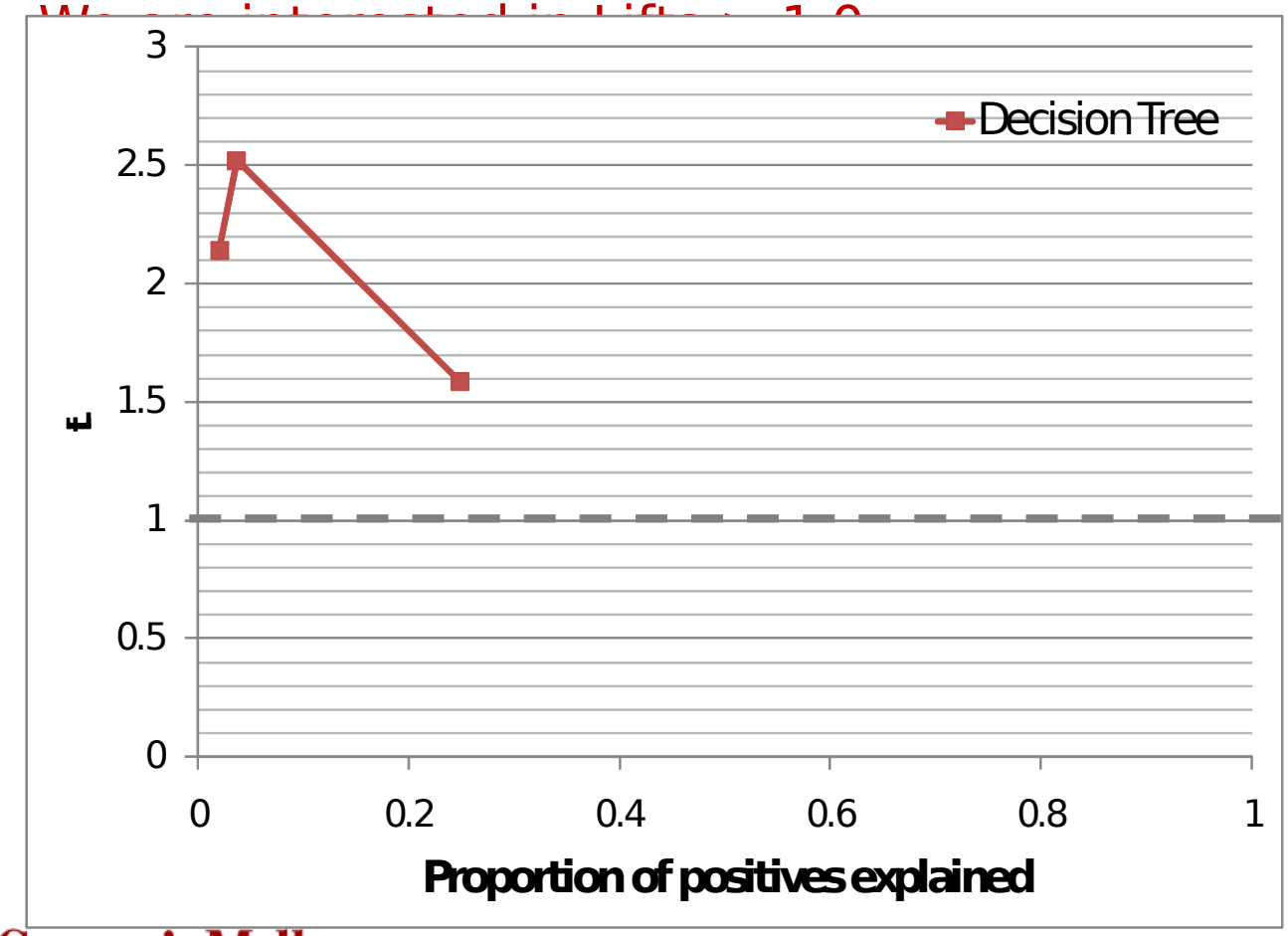
- **Approach:**

  Use this data to train classifiers of the binary outcome: {ReadmissionOccurred, NoReadmission}

Predicting readmissions with Decision Trees

Lift = Rate of readmissions in the selected pattern
/ Prior rate of readmissions (=0.64%)

It reflects predictive utility of a pattern.



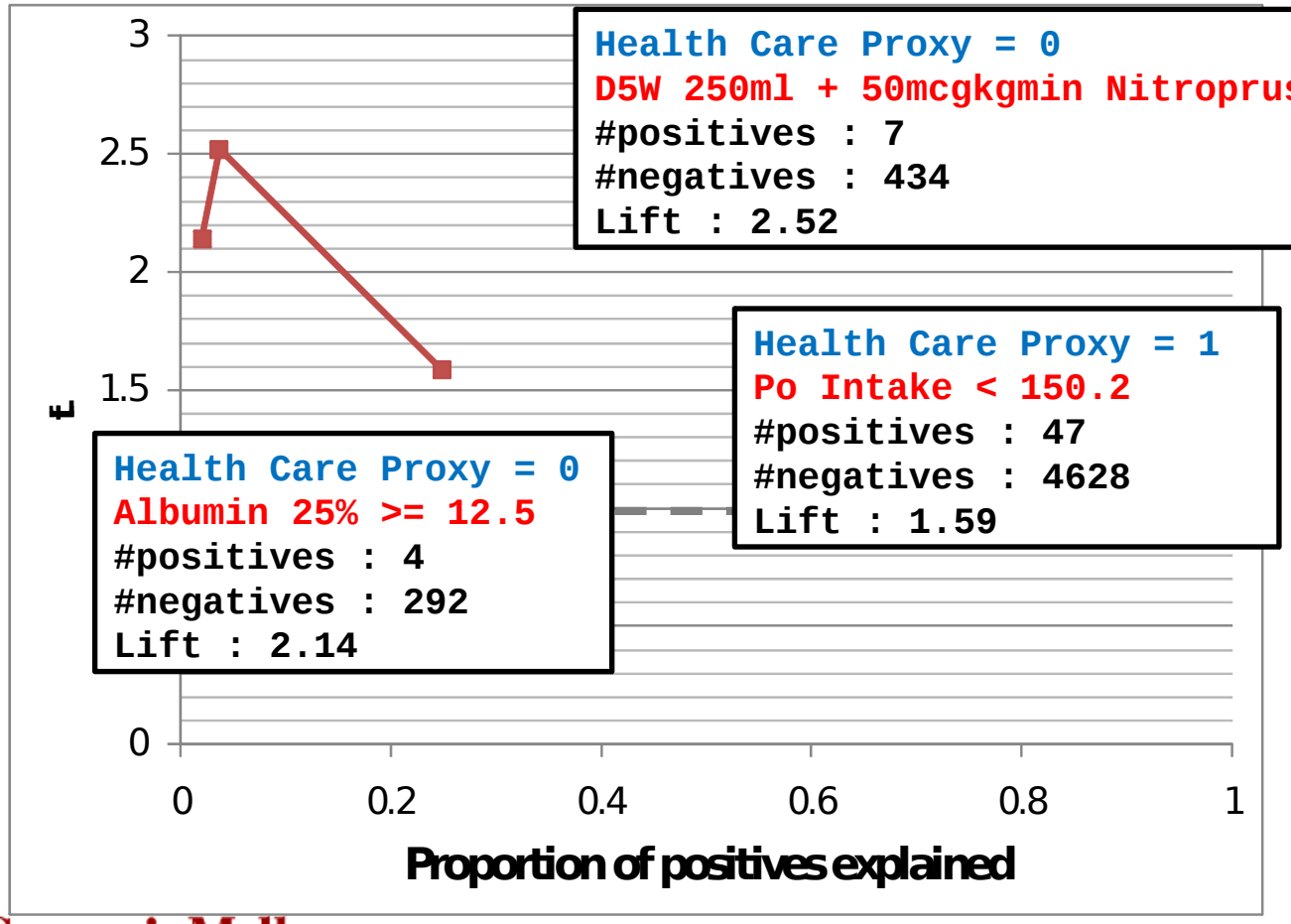
The decision tree was able to reliably identify 3 conjunctive patterns of varying utilities (lifts) and data coverages (recalls)

Predicting readmissions with Decision Trees

Lift = Rate of readmissions in the selected pattern
/ Prior rate of readmissions (=0.64%)

It reflects predictive utility of a pattern.

We are interested in Lifts > 1.0



The decision tree was able to reliably identify 3 conjunctive patterns of varying utilities (lifts) and data coverages (recalls)

A relatively simple disjunctive algorithm

The “Subset Search” algorithm:

Iterate through all features in data, picking one at a time:

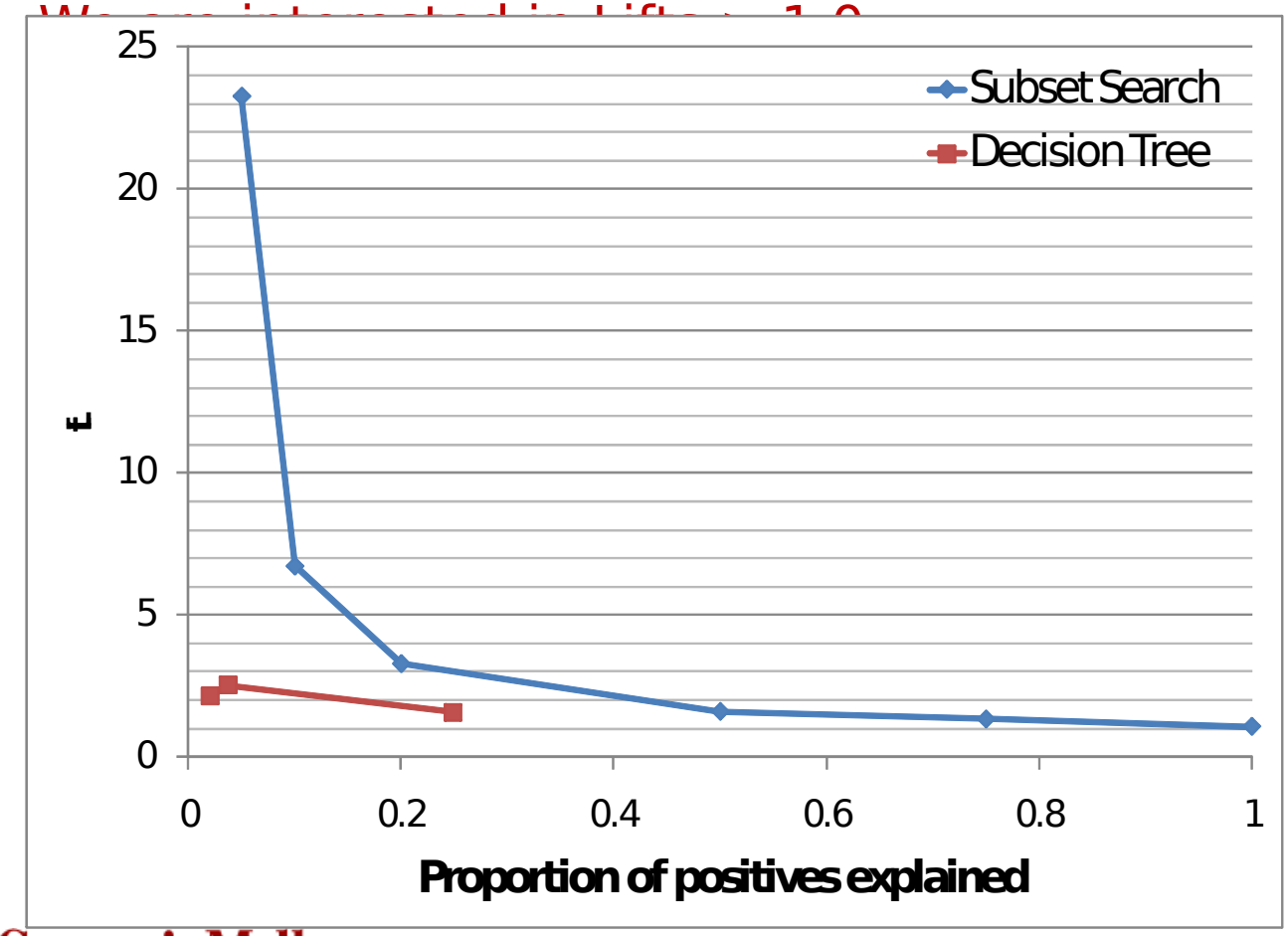
1. If the feature is numeric, discretize it, keeping distribution of positive examples in data uniform
2. Find a subset of values of the feature that maximizes the lift and selects the minimum count of positive examples in data
3. Check the usefulness of this set of values:
4. It must increase the combined lift of the developed conjunctive-disjunctive pattern, if it were added to it, by the amount greater than some threshold
5. If it is useful, add it to the learned pattern
6. If the resultant lift is sufficiently high, terminate algorithm
7. Remove the current feature from consideration and recursively execute the algorithm picking another feature
8. Return the current feature to the set of selectable features and continue

The “Subset Search” algorithm relatively quickly learns from data predictive rules that have the conjunctive-disjunctive form

Predicting readmissions with Subset Search

Lift = Rate of readmissions in the selected pattern
/ Prior rate of readmissions (=0.64%)

It reflects predictive utility of a pattern.



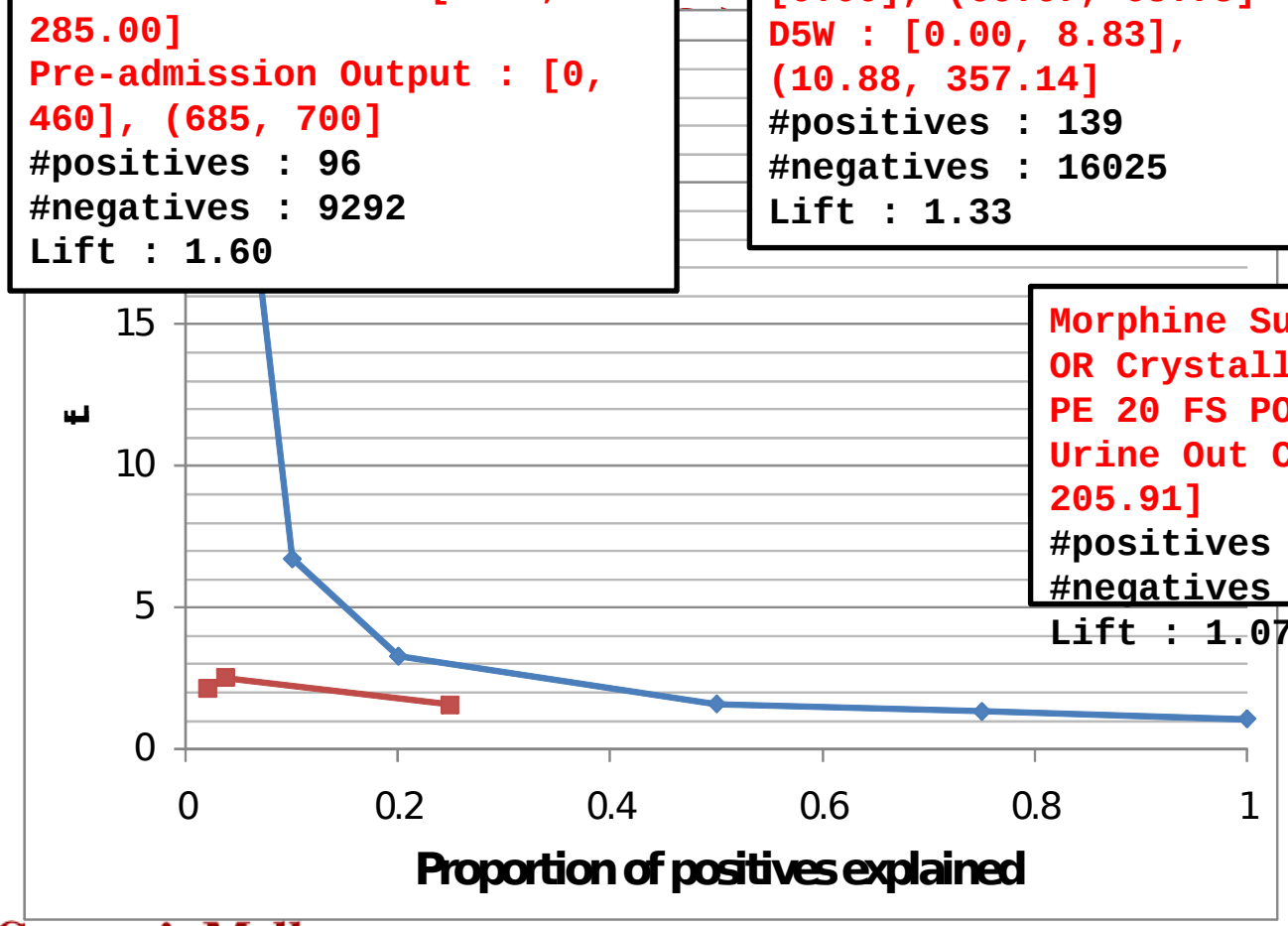
Using varying sets of parameters, the Subset Search algorithm found a few patterns of superior utility to those identified by the decision tree

Predicting readmissions with Subset Search

```
Propofol : [0.0], (265.00, 348.75], (535.25, 595.00], (895.00, 2267.50]
0.9% Normal Saline 1000ml : (0.00, 1000.00]
Cath Lab Intake : [0.00, 285.00]
Pre-admission Output : [0, 460], (685, 700]
#positives : 96
#negatives : 9292
Lift : 1.60
```

```
Age : (0.00, 86.00]
Pre-Admission Intake : [0, 665], (1400, 3000], [5000, 5500]
Sterile Water 100ml : [0.00], (66.67, 68.75]
D5W : [0.00, 8.83], (10.88, 357.14]
#positives : 139
#negatives : 16025
Lift : 1.33
```

```
Morphine Sulfate : [0.00, 30.00]
OR Crystalloid : [0.00, 3600.00]
PE 20 FS PO : [0.00, 10.00]
Urine Out Condom Cath : [0.0, 205.91]
#positives : 189
#negatives : 27327
Lift : 1.07
```



algorithm found a few patterns of superior utility to those identified by the decision tree

Predicting readmissions with Subset Search

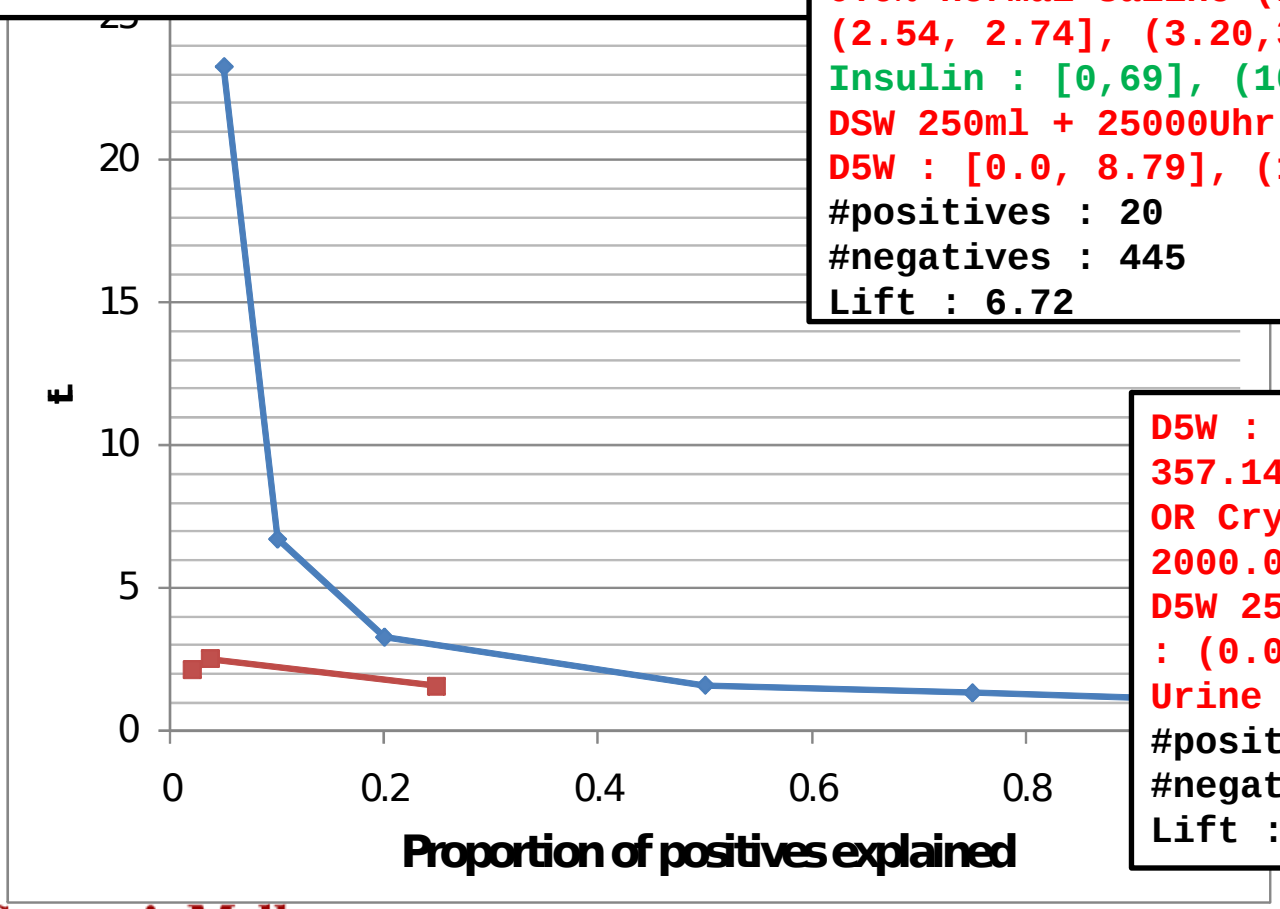
```
TF residual : (19.50, 26.39], (39.65, 50.72]
Pre-admission intake :  0.0
Free water bolus :  (0.0, 237.86]
Urine out forey : [0.0, 129.69]
#positives : 11
#negatives : 57
Lift : 23
```

ected pattern

4%)

rn.

```
0.9% normal saline (100ml) + 100Uhr Insulin :
(2.54, 2.74], (3.20,3.40], (5.10, 9.35]
Insulin : [0,69], (102.67, 222.20]
DSW 250ml + 25000Uhr Heparin :  [0.0, 7.68]
D5W : [0.0, 8.79], (10.00, 10.59]
#positives : 20
#negatives : 445
Lift : 6.72
```



Using varying

sets of

```
D5W : (0.00, 8.29], (16.37,
357.14]
OR Crystalloid :  [0.00,
2000.00]
D5W 250ml + 4mcgkgmin Levophed-k
: (0.0, 4.75]
Urine Out Void : [0.00, 203.00]
#positives : 39
#negatives : 1747
Lift : 3.28
```

tree

So, what has just happened?

- We have seen how much benefit can be attained by relaxing constraints about the form of the patterns in Clinical Data

Our "Subset Search" algorithm was able to find patterns of substantially greater utility than the simpler conjunctive scheme used by the decision tree

- But what are the down-sides?
  - Greater complexity of the resulting patterns carries the greater risk of over-fitting (we need to carefully test for that)
  - Computation costs to learn models from data may be higher than for the alternative conjunctive models

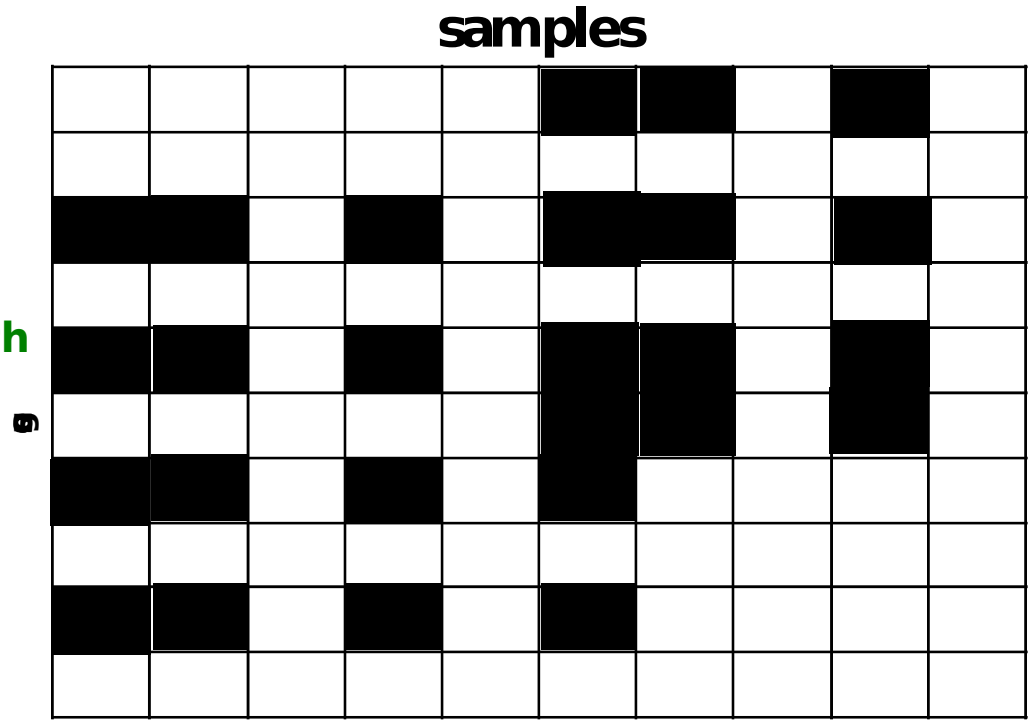
- However, if we are brave enough, we may consider aiming at an even more formidable task:

Learning OVERLAPPING conjunctive-disjunctive patterns

Example application: Microarray analysis

Problem statement:

- In cancer research we often want to identify bi-clusters of genes and samples with elevated activation levels
- They can help design effective therapies



Research challenge:

- The bi-clusters may overlap, with additive effects on activation

Standard solution:

- Iteratively find the best bi-cluster, remove its effect from data, repeat
- It is often sub-optimal

Our data and Approach

- **Two real-world tumor data sets:**
  1. **Breast cancer data (13,666 genes, 117 samples)**
  2. **Lung cancer data (12,625 genes, 125 samples)**

- **Competing algorithms:**
  - **Large Average Submatrix algorithm (LAS)**

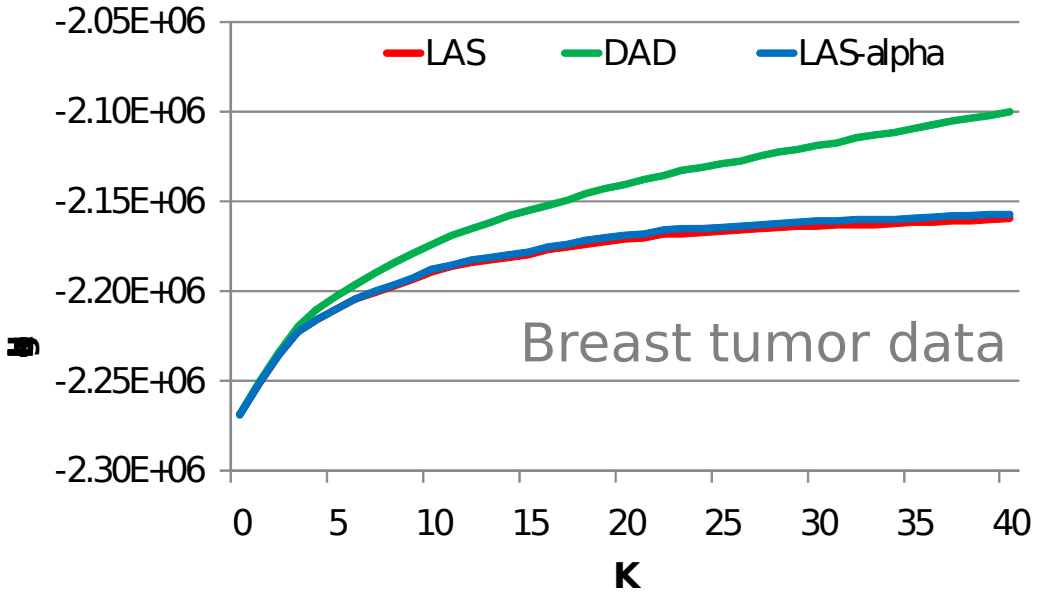
[Shabalin et al. Finding large average submatrices in high dimensional data. *Annals of Statistics*, 3(3):985–1012, 2009]

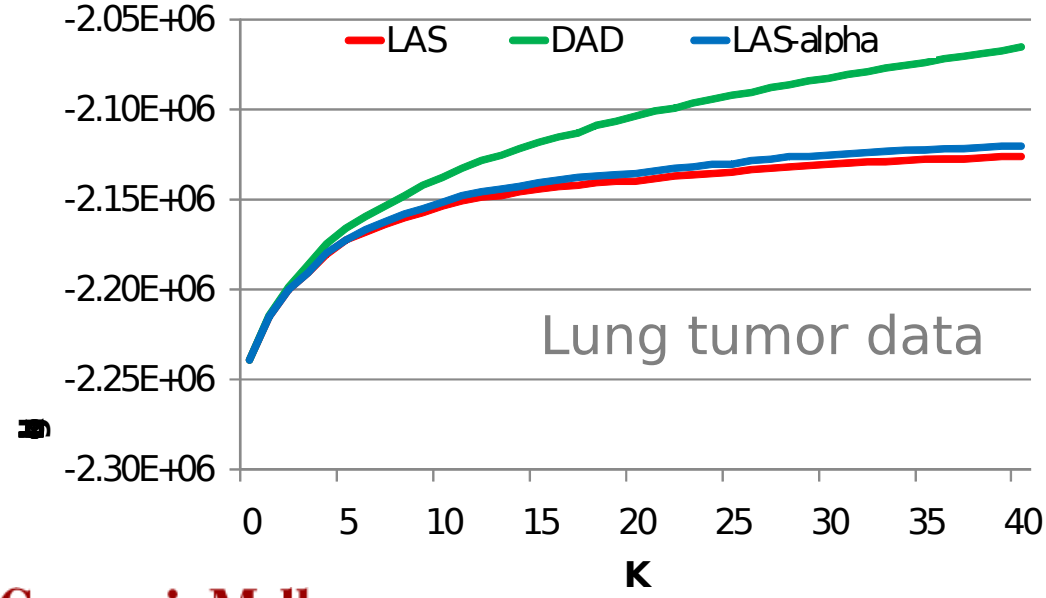
  - **Find the best fitting cluster in data via hill climbing optimization in feature-value space, using random restarts to avoid local maxima**
  - **Remove the effect of the best fitting cluster from data and repeat**
  - **Disjunctive Anomaly Detection algorithm (DAD)**

[Sabhnani et al. Detection of disjunctive anomalous patterns in multidimensional data. *Advances in Disease Surveillance*, 2009.]

  - **Iteratively find the next best fitting cluster in data using round-robin hill climbing optimization with random restarts**
  - **Each time the new cluster is found, refit all clusters added so far using non-negative least-square regression, then continue**

Detecting bi-clusters in cancer microarrays





As we hoped for, the algorithm designed to handle overlapping patterns did better

Much better

The level of accuracy of disjunctive representation that

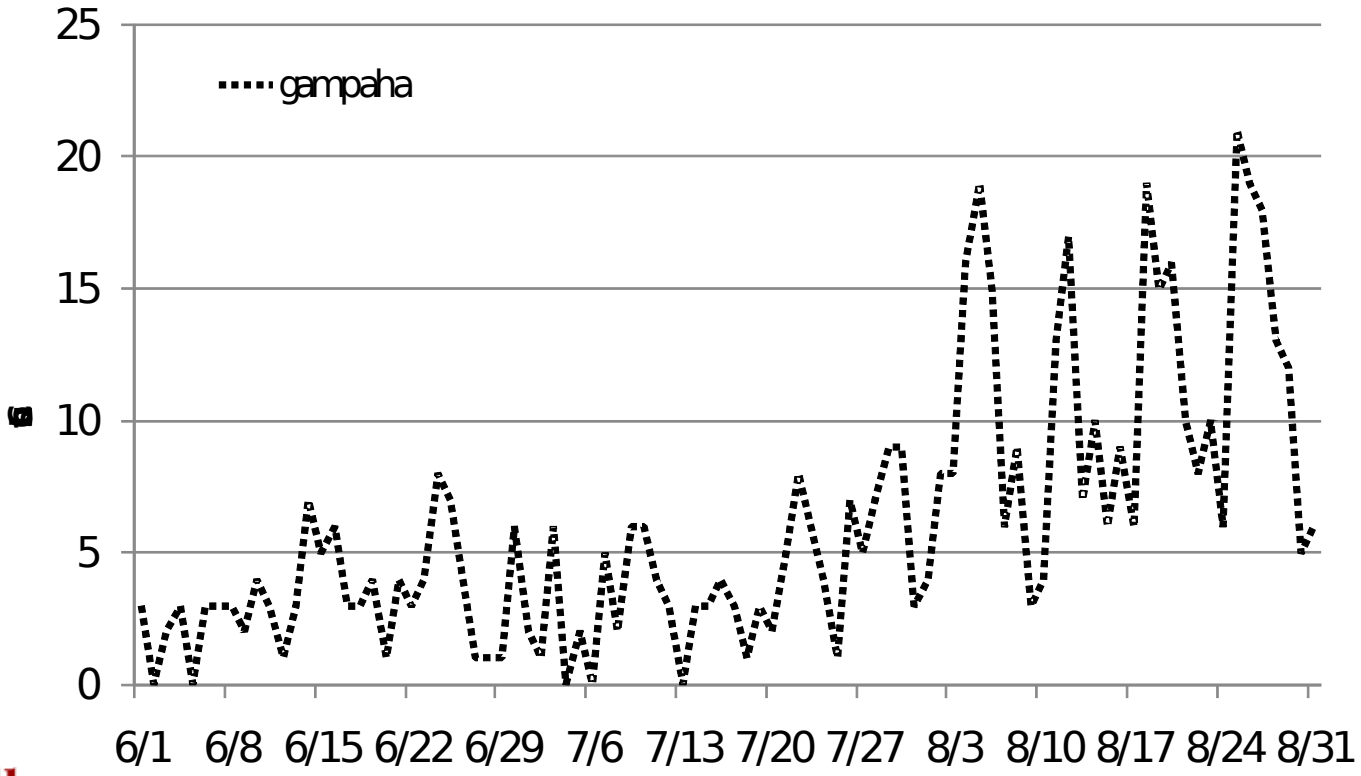
LAS algorithm achieves using 40 clusters,

DAD accomplishes using fewer than 15 clusters

Where else do we see the need for models capable of representing multiple disjunctive overlapping patterns?

- **In Sri Lanka!**

Weekly Epidemiological Reports show an increase in disease rates in Gampaha in summer 2008

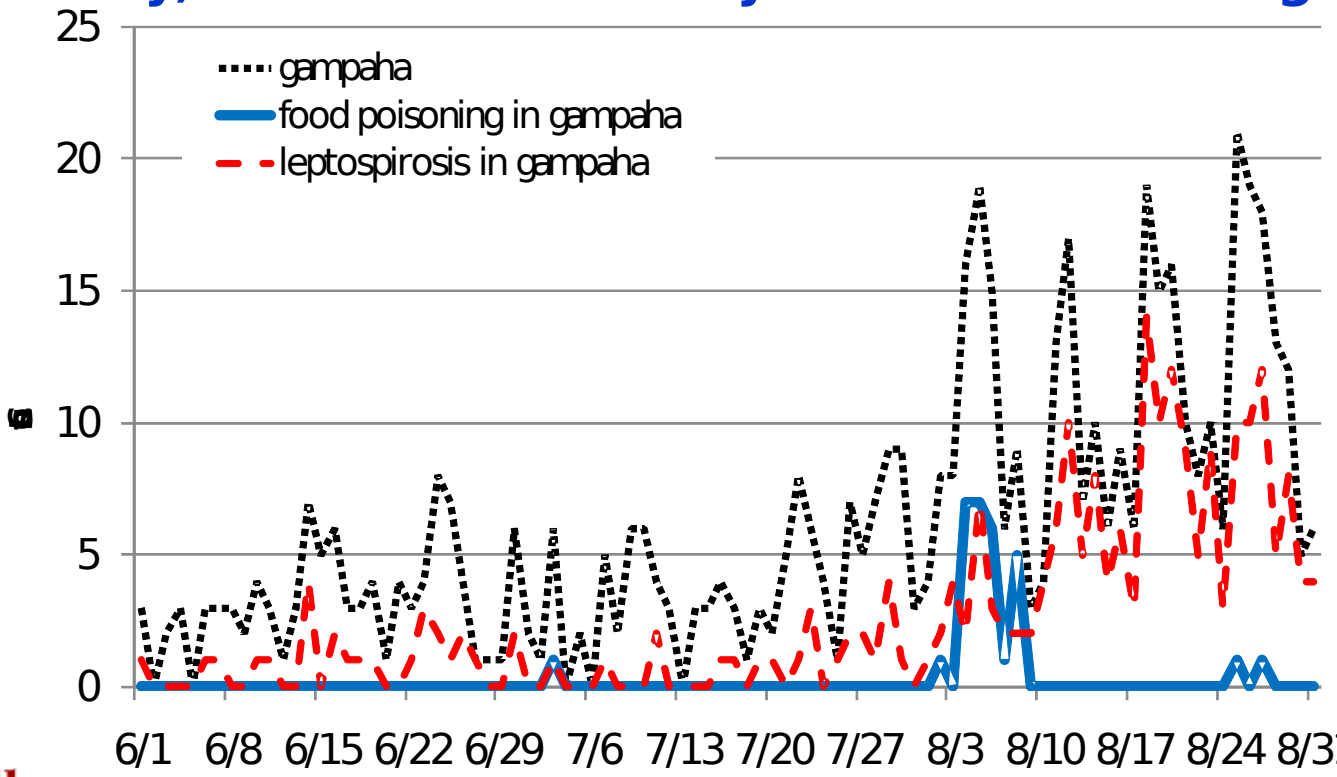


Where else do we see the need for models capable of representing multiple disjunctive overlapping patterns?

- **In Sri Lanka!**

Weekly Epidemiological Reports show an increase in disease rates in Gampaha in summer 2008

Apparently, it did not involve just one escalating disease pattern:



More details in an upcoming Auton Lab paper by Sabhnani et al.

Summary

- **Disjunctive patterns are present in many types of data encountered in Clinical Informatics**
- **Ability to discover them can be quite useful in a range of applications**
  - We have examined in-patient and out-patient data, as well as clinical study data
  - There is a need and room for new computational approaches to handle such patterns
  - Current, popular algorithms are predominantly conjunctive, and/or use a greedy search-then-eliminate strategy
  - In our preliminary comparative studies, we have observed improved accuracies and/or lower complexities of the disjunctive models when compared against well performing competitors
  - Challenges to overcome before moving fast-forward remain
  - Computational scalability is a formidable one

This material is based upon work supported by the National Science Foundation, grant NSF IIS-091103

Summary

- **Disjunctive patterns are present in many types of data encountered in Clinical Informatics**
- **Ability to discover them can be quite useful in a range of applications**
  - We have examined in-patient and out-patient data, as well as clinical study data
  - There is a need and room for new computational approaches to handle such patterns
  - Current, popular algorithms are predominantly conjunctive, and/or a greedy search-then-eliminate strategy
  - In our preliminary comparative studies, we have observed improved accuracies and/or lower complexities of the disjunctive models when compared against well performing competitors
  - Challenges to overcome before moving fast-forward remain
  - Computational scalability is a formidable one

By the way, if you are interested in real-time bio-surveillance taking place in Sri Lanka, take a look at this movie:

http://www.autonlab.org/autonweb/library/videos.html

This material is based upon work supported by the National Science Foundation, grant NSF IIS-091103