Introduction to Data Mining

Christopher Leckie

Department of Computing and Information Systems
The University of Melbourne

Styles of Decision Making

Overview

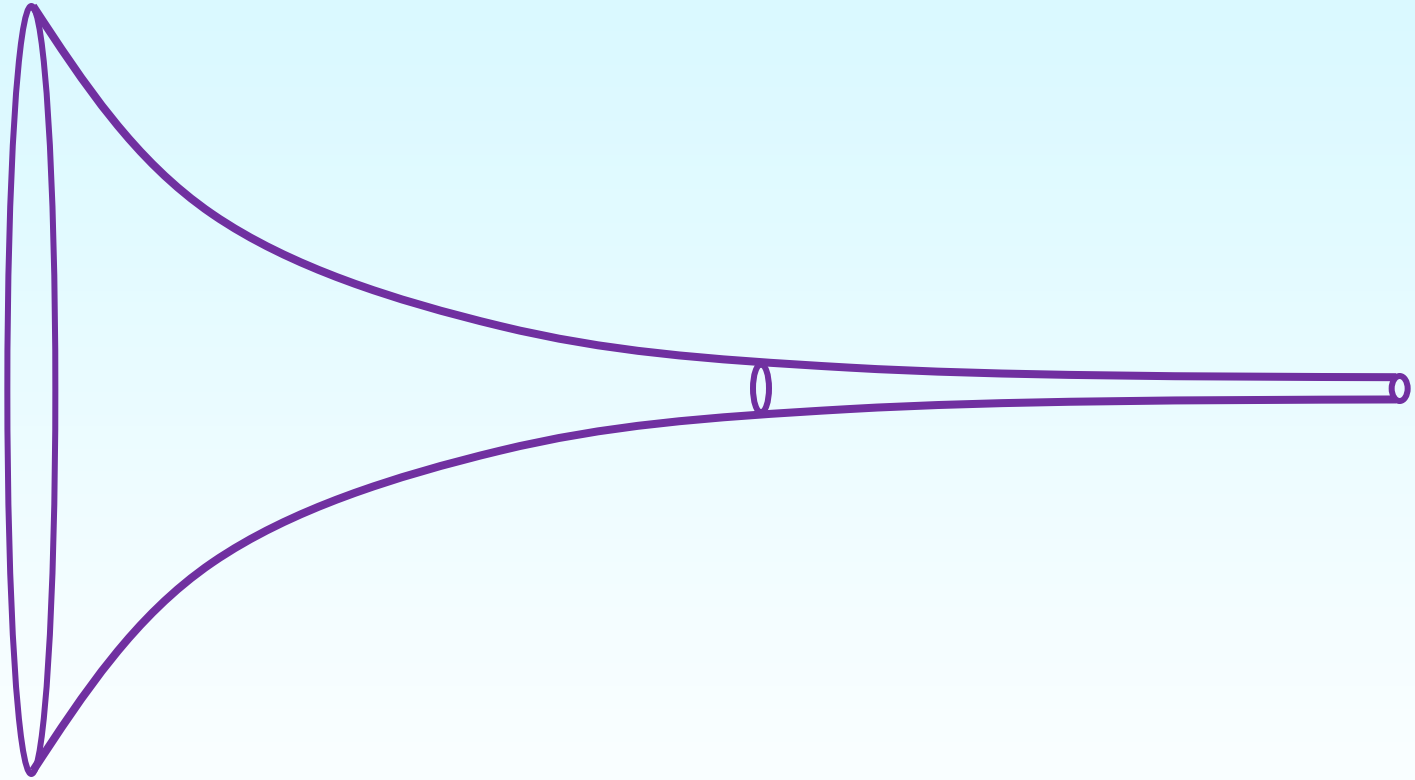
Data mining aims to find useful patterns in large databases

For example:

- **Market segmentation studies**
  - **Find categories of customers with similar buying behaviour**
  - **Example of “unsupervised learning”**

- **Predictive modelling**
  - **Find customers who are likely to commit fraud based on their transaction history**
  - **Example of “supervised learning”**

The Common Theme – Big Data

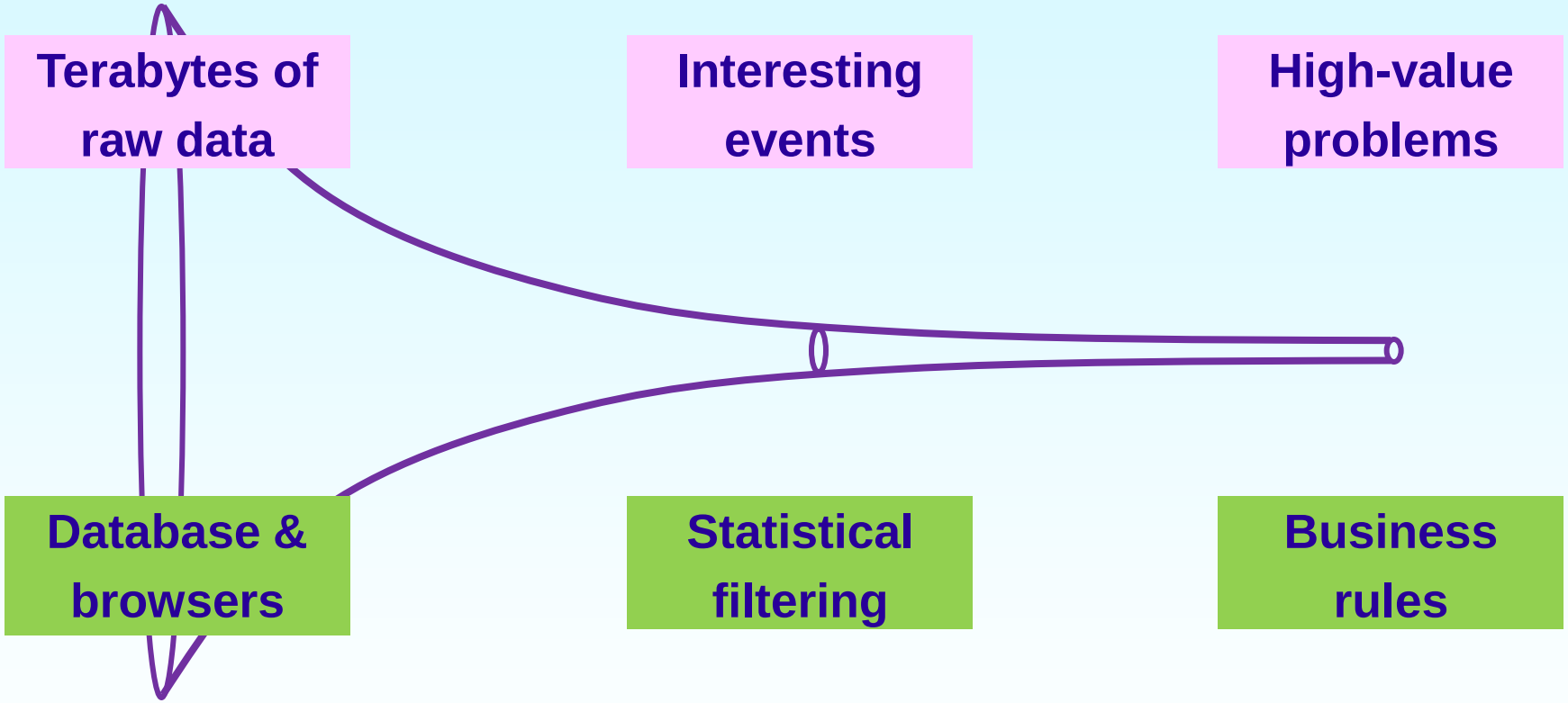


Terabytes of raw data

Interesting events

High-value problems

Automating the Data Analysis Pipeline



Part of the field of **data analytics / machine learning**

Clustering to Learn Categories (Unsupervised Learning)

What are the natural categories in a database?

Consider a database of animals.

How many different types of animals are there here?

Learning a Classifier (Supervised Learning)

Training a classifier

cat cat dog dog cat

Classifier

Classifying new examples

Classifier

dog

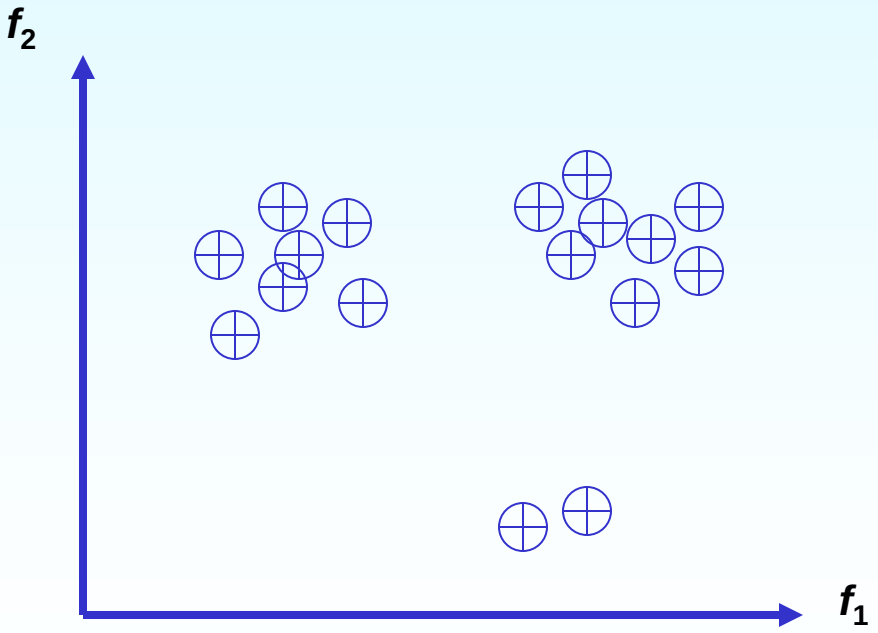
Learning Unusual Patterns (Anomaly Detection)

- Learn a model of “normal” database records
- Use this model to test new records for anomalies
- Any anomalies can be either interesting or errors

Unsupervised Anomaly Detection

[Eskin et al. 2002]

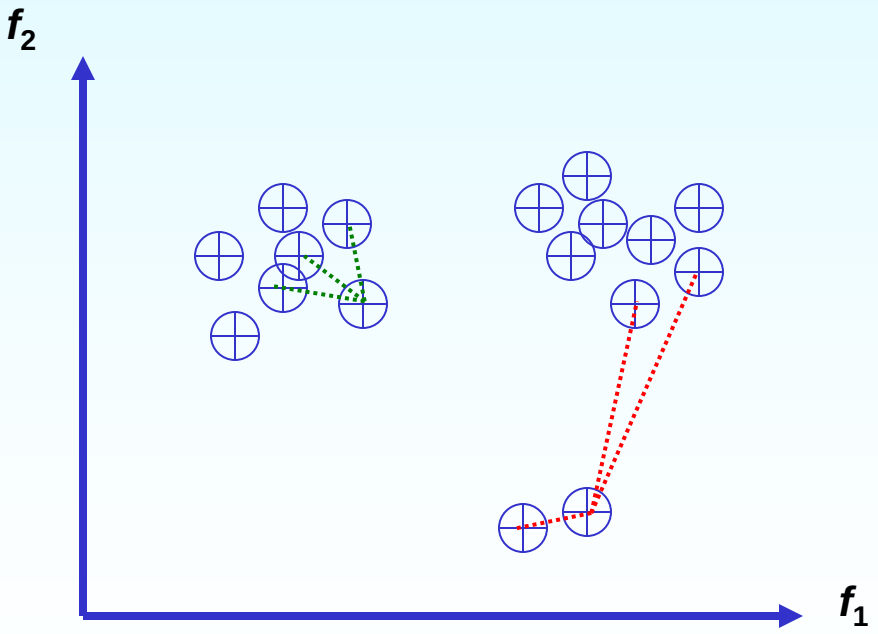
- **Map record fields into a feature space {f_1… f_k}**
- **Cluster similar records**
- **Use large clusters to represent normal records**



Unsupervised Anomaly Detection

K-nearest neighbours:

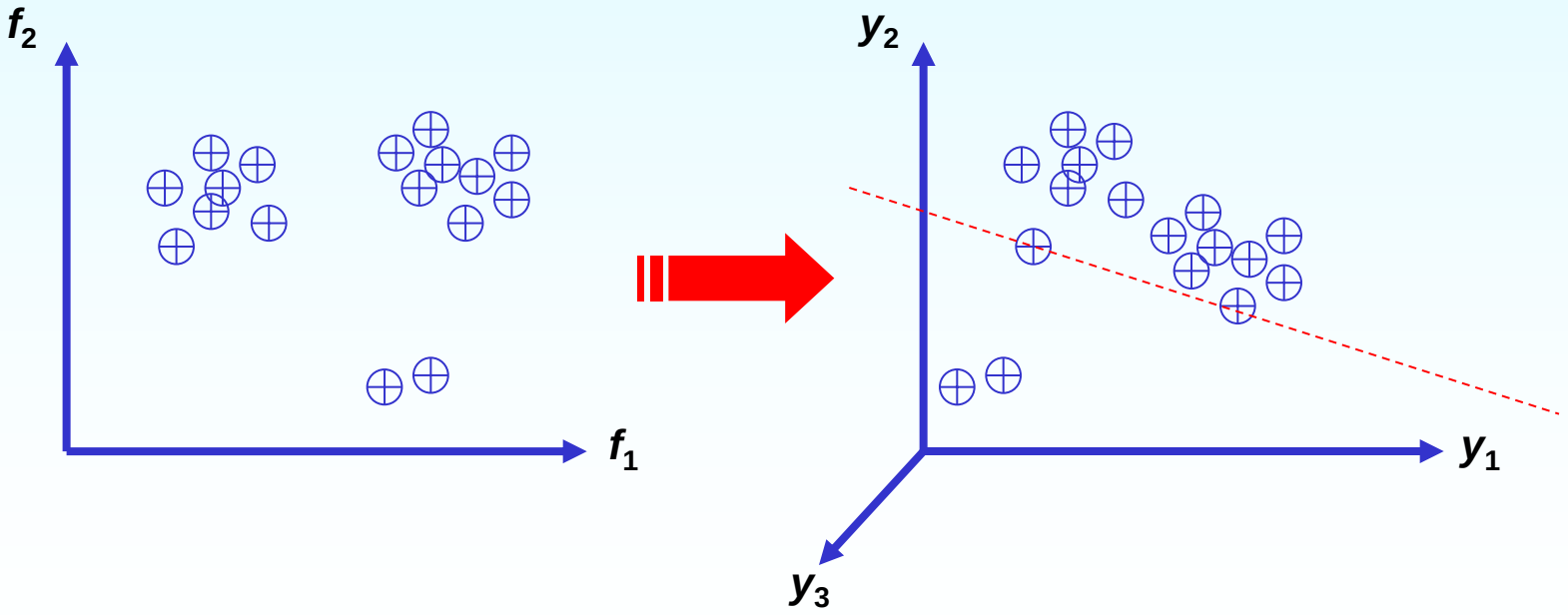
- **Find *k* nearest neighbours of each point**
- **Data points with high kNN distance are in sparse regions of space**



Unsupervised Anomaly Detection

One-class Support Vector Machine:

- **Map data points into a higher dimensional space**
- **Find a hyperplane that is *maximally distant* from origin while separating *most points* from origin**



Unsupervised Anomaly Detection

Fixed-width clustering:

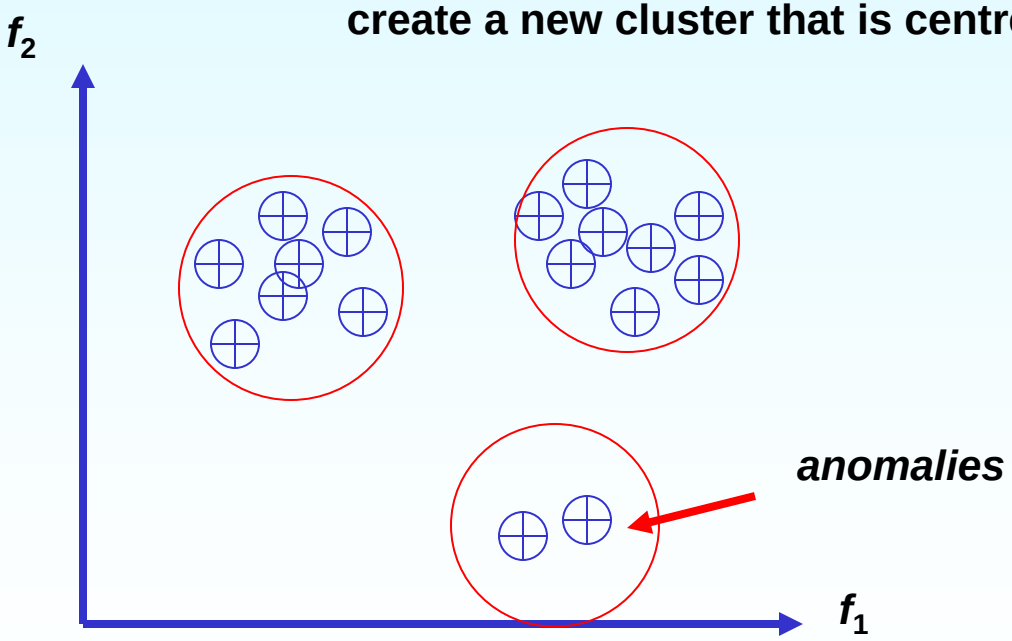
For each data point x

If distance(x, centroid of nearest cluster c) < w

add x to cluster c

Else

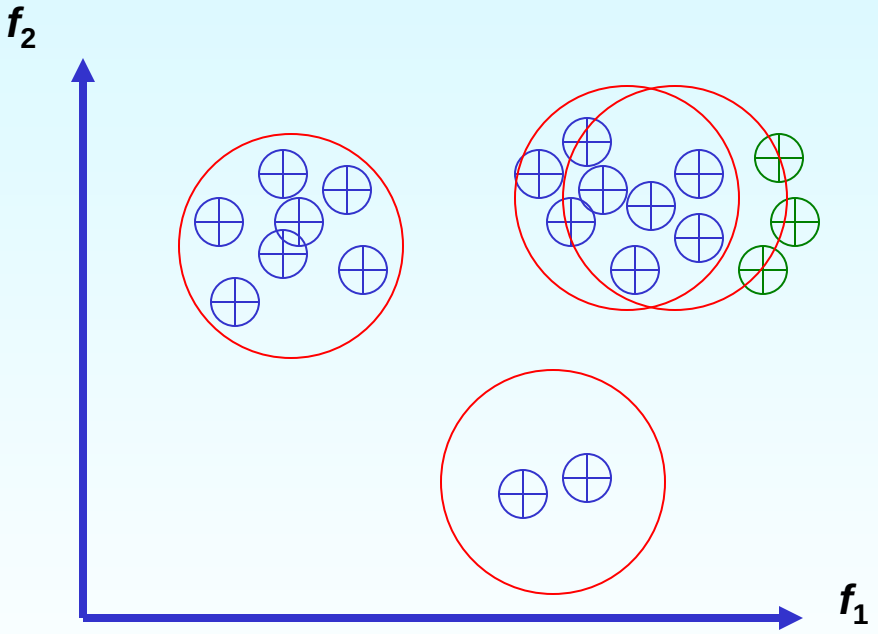
create a new cluster that is centred on x



Challenge: changing data patterns cause false positives

Time-Varying Clustering

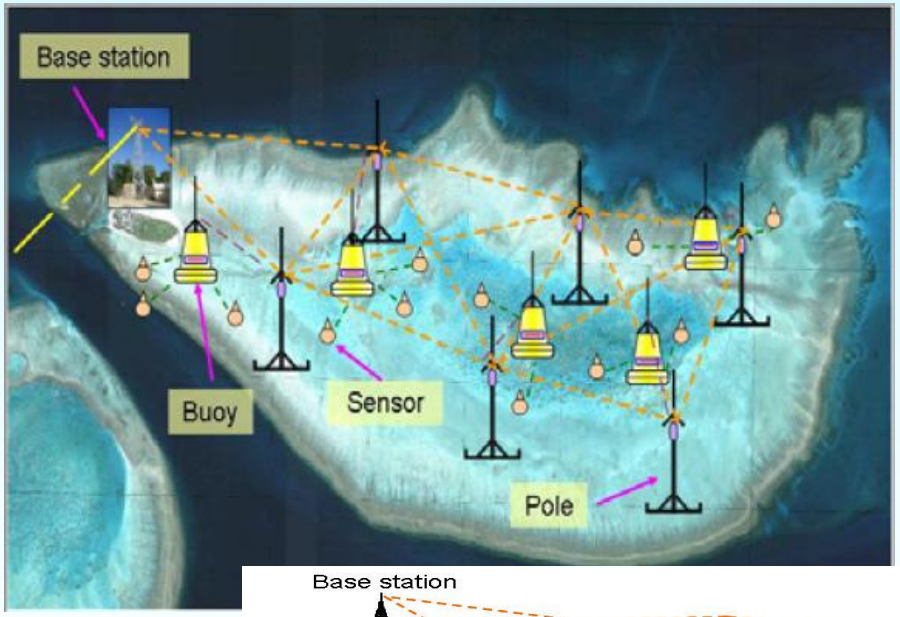
Need to adapt to changing data patterns

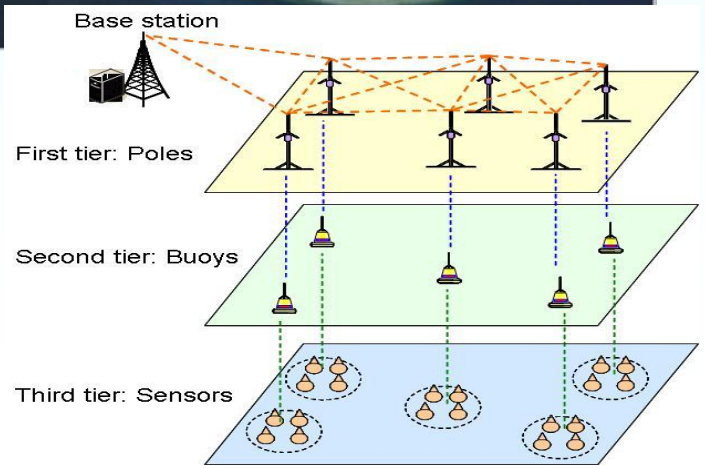


$$mean(cluster) = \frac{\gamma \times mean(cluster) + example}{\gamma + 1}$$

Scenario 1 – Environmental Management

What is the impact of global warming on the Great Barrier Reef?





http://www.coralreefeon.org/sensor networking the Great Bar-rier Reefa.pdf. http://wallpaper.digiocto.com/O,water/,R. Szewczyk, E. Osterweil, J. Polastre, M. Hamilton, A. Mainwaring, and D. Estrin. Habitat monitoring with sensor networks. In CACM, vol 47, pg 34–40, June 2004, Courtesy: Stuart Kininmonth, AIMS

Wireless Sensor Networks

- Wireless nodes for remote monitoring and control
- Self-configuring multi-hop network
- Limited
  - Power (Battery)
  - Bandwidth
  - Memory
  - Computation capability
- Heterogeneous nodes with varying capabilities

http://www.sensortaiwan.com/asptest/Prd_Wireless.asp

Unusual events in sensor measurements

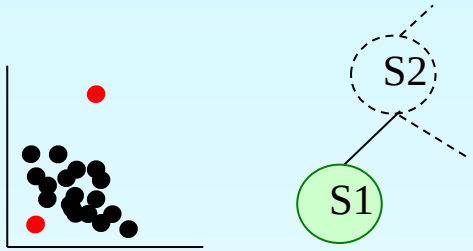
Observations that are inconsistent with the remainder of the data set (anomalies)

- Causes of anomalies
  - Sudden change in the environment
  - Faulty nodes (loss of calibration)
  - Malicious attacks (data injection)
  - Noise
- Identifying anomalies
  - Analyse measurement or traffic data in the network
  - Build model of normal behavior to classify anomalies

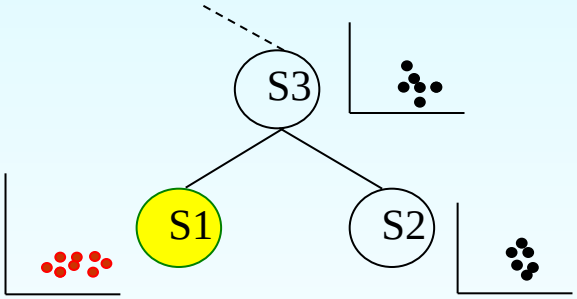
S. Rajasegarar, C. Leckie and M. Palaniswami. Anomaly Detection in Wireless Sensor Networks. In IEEE Wireless Communications, vol. 15, no. 4, pp. 34-40, August 2008.

Roadmap of research

- Local anomalies
  - Detecting anomalies that occur with respect to data at a single node



- Global anomalies
  - Detecting nodes whose data is anomalous with respect to other nodes

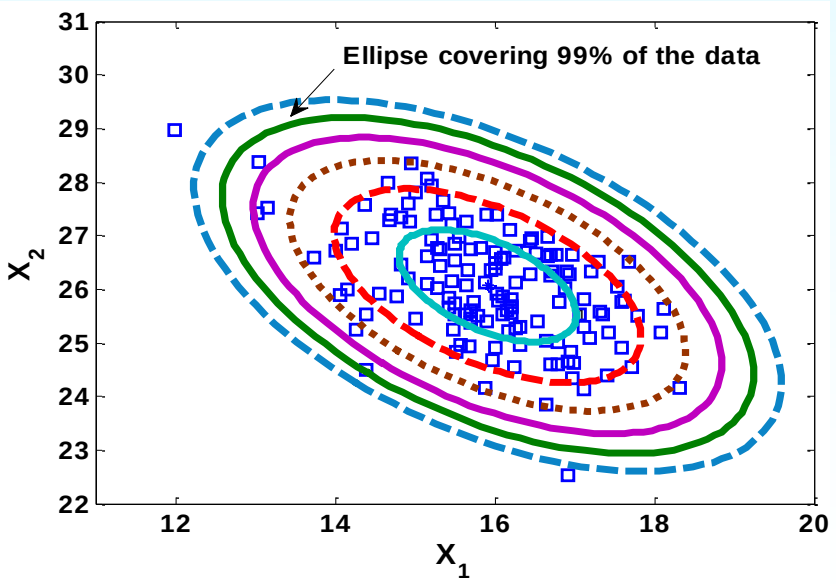


- Modelling complex events
  - Detecting unusual events that span different time scales and spatial scales

Building hyper-ellipsoidal models

- Computationally efficient representation of raw data
- Batch learning
  - Random vector $\mathbf{X} = (X_1, \ldots, X_d)^T$ with sample mean and covariance (μ, Σ)
  - Construct level set of all vectors that have same Mahalanobis distance to the mean:

$$Q(\mathbf{x}-\boldsymbol{\mu}) = (\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}} \Sigma^{-1} (\mathbf{x}-\boldsymbol{\mu}) = \left\| (\mathbf{x}-\boldsymbol{\mu}) \right\|_{\Sigma^{-1}}^2$$



Rajasegarar, S., Bezdek, J. C., Leckie, C. and Palaniswami, M. (2009). “Elliptical Anomalies in Wireless Sensor Networks,” ACM Transactions on Sensor Networks, 6(1), 1550-1579.

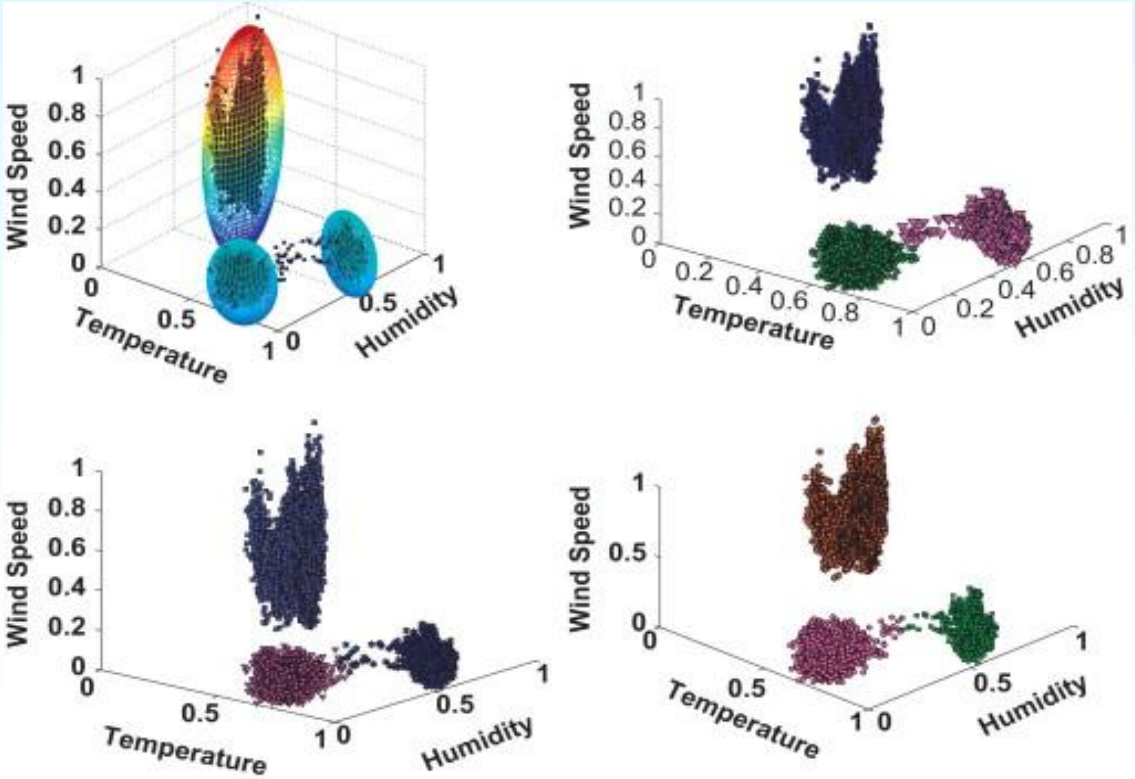
Hyper-ellipsoidal clustering algorithm

Require an efficient clustering algorithm that can run on a sensor node:

(1) automatic selection of the number of clusters

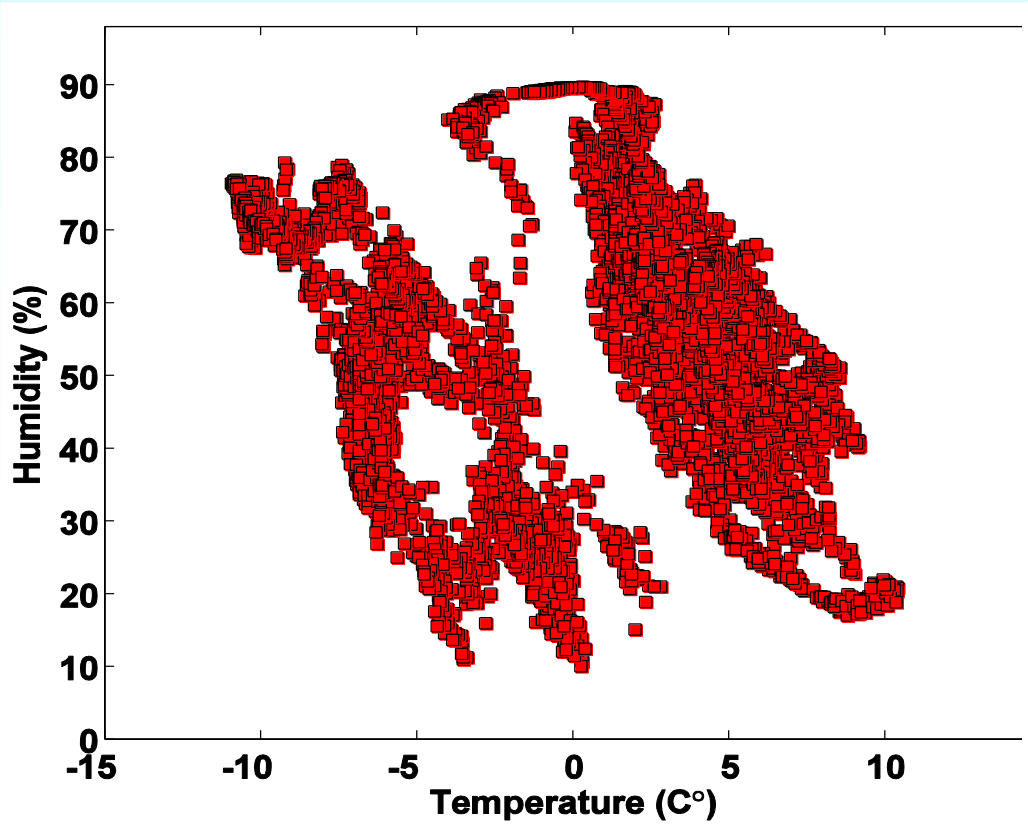
(2) low computational cost (O(N))

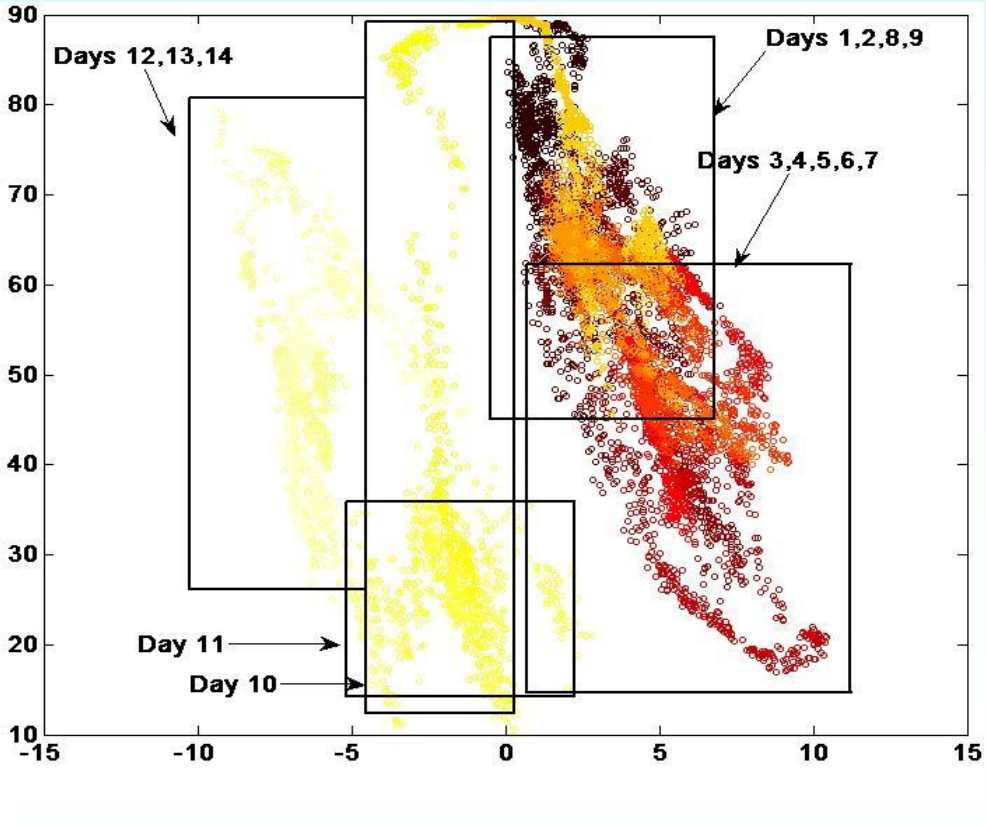
(3) explicit cluster boundary detection



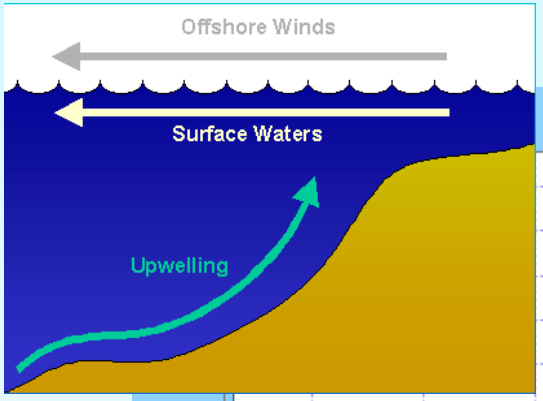
M. Moshtaghi, S.Rajasegarar, C. Leckie, S. Karunasekera, "An Efficient Hyperellipsoidal Clustering Algorithm for Resource-Constrained Environments", Pattern Recognition, Volume 44, Issue 9, Sept. 2011

Incremental learning of hyper-ellipsoidal models





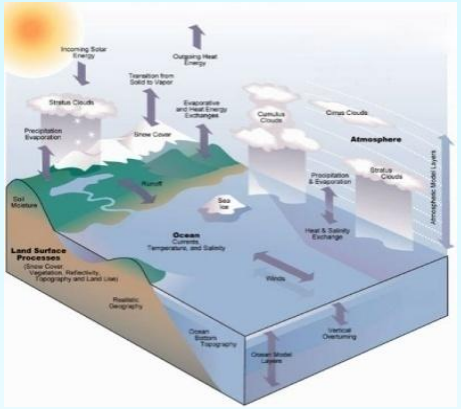
Detecting Interesting Events

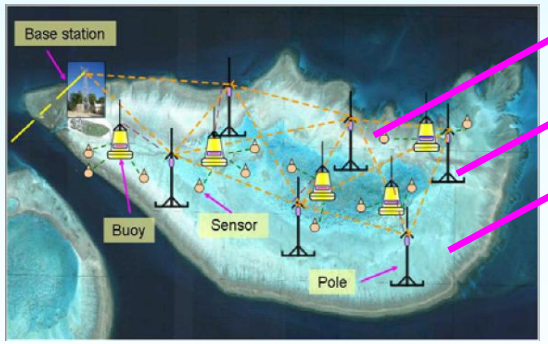


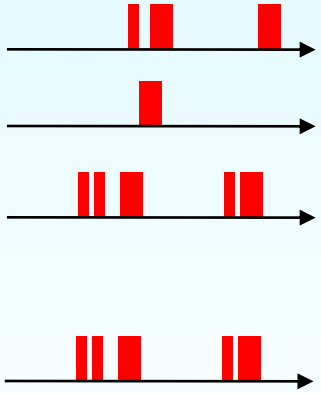
Developed by Andika Widjaja

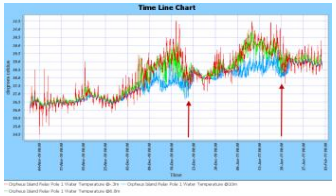
Future work – learning complex events

Aim: model and detect elaborate activities in complex sensing environments









Complex activities and trends

Localised activities

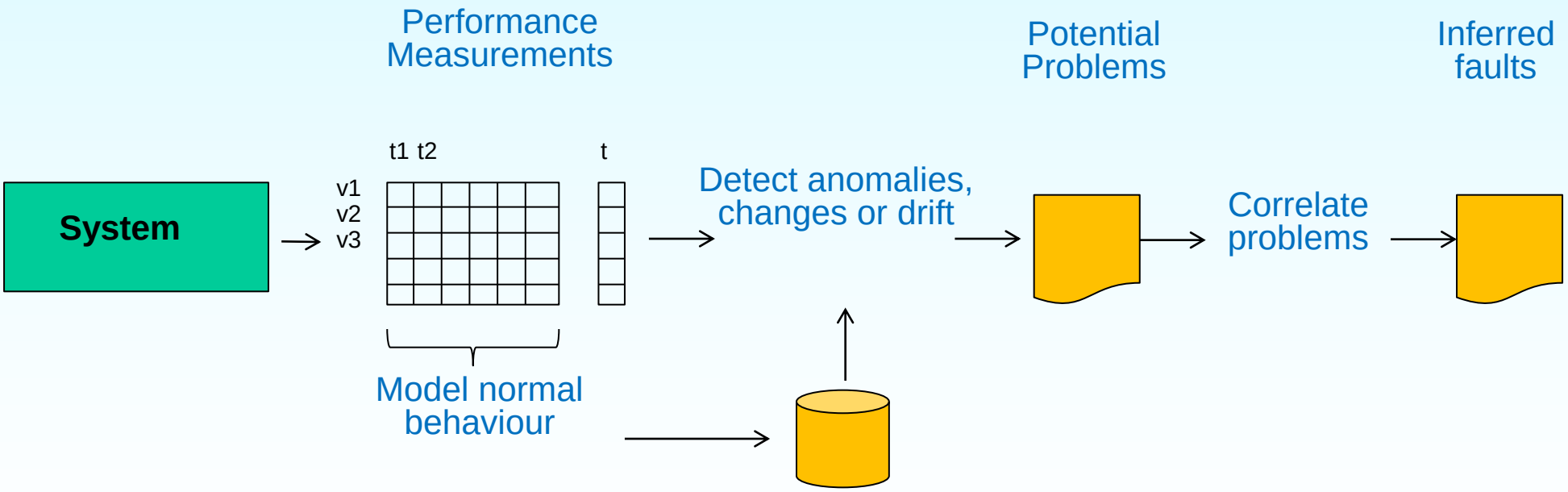
Inferred events

Sensor data streams

Scenario 2 – Fault Diagnosis and Preventive Maintenance

- **Failure driven maintenance**
  - Customer complains -> Fix fault
  - Quality of service: low
  - Cost: low-high
- **Periodic maintenance**
  - Regular downtime -> Replace / retune (even if not needed)
  - Quality of service: high
  - Cost: high
- **Predictive maintenance**
  - Detect incipient problem -> Replace / retune (before customer impact)
  - Quality of service: high
  - Cost: low-medium

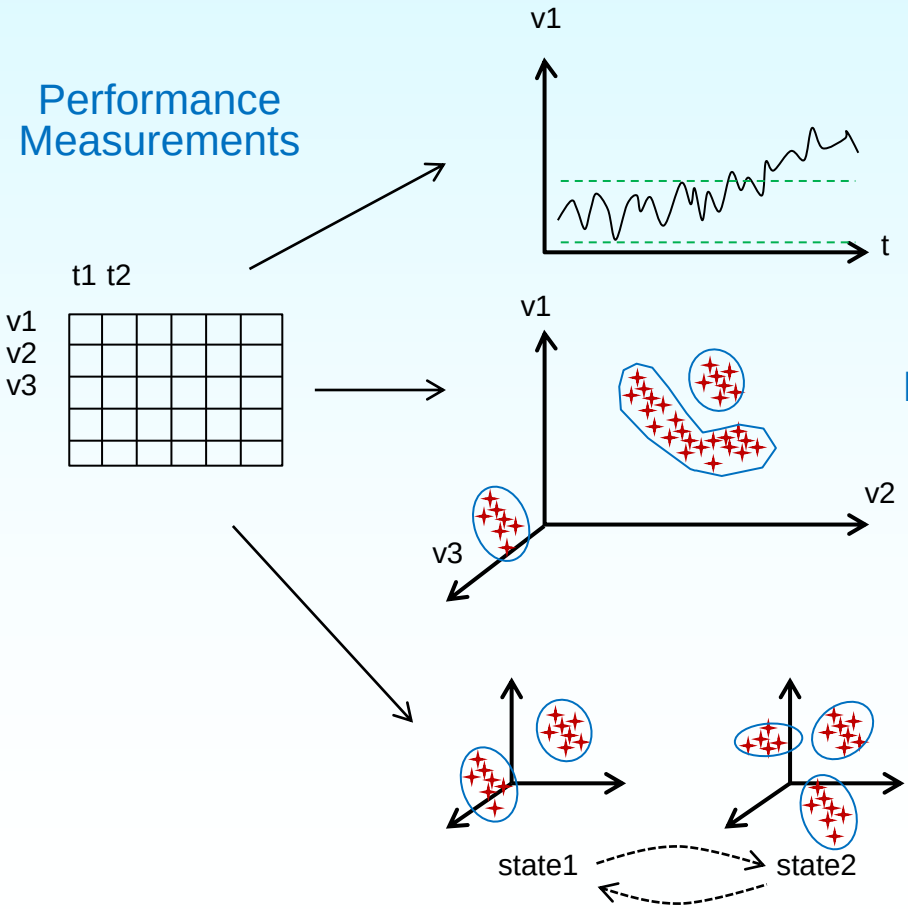
Predictive Maintenance



Modelling Normal Behaviour

System

Performance Measurements

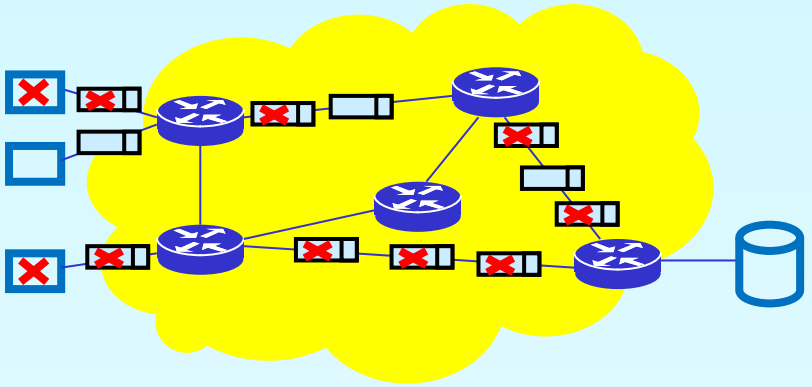


Change point detection in time series

Multivariate cluster models, kernel-based methods (SVM)

Switching data streams

Applications focus



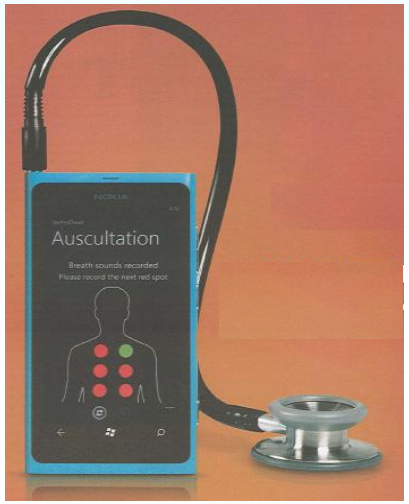
cyber-security, telecommunications

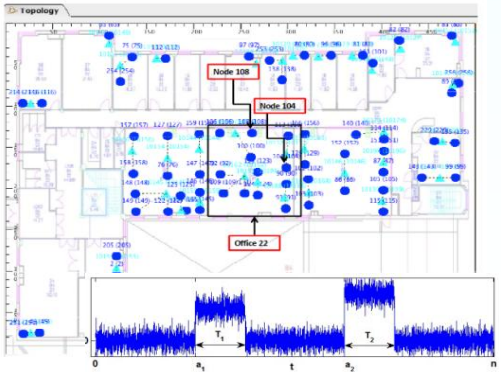
transport

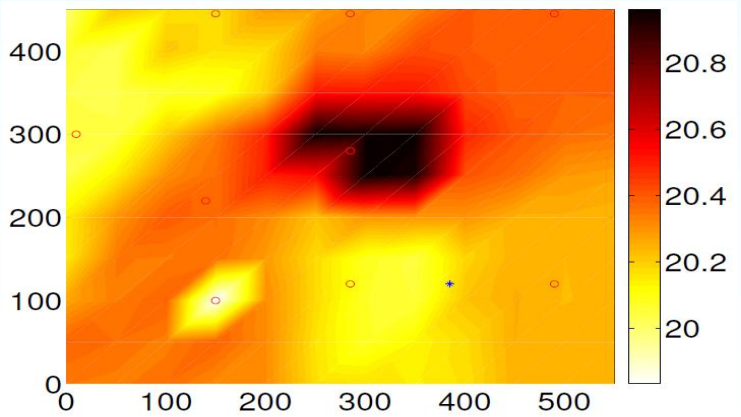
environmental monitoring

smart cities

participatory sensing







Conclusion

Data mining aims to find useful patterns in large databases

Useful in environmental monitoring, operations, security …

Many patterns discovered using data mining are interesting, but which ones are useful?

Curious for more?

COMP90049 Knowledge Technologies

Topics include: data encoding and markup, web crawling, clustering, pattern mining, Bayesian learning, instance-based learning, document indexing, database storage and indexing, and text retrieval

COMP90042 Web Search and Text Analysis

Topics include: search engines, cross-language information retrieval, machine translation, text mining, question answering, summarisation

COMP90051 Statistical and Evolutionary Learning

Topics include: statistical learning, evolutionary algorithms, swarm intelligence, neural networks, numeric prediction, weakly supervised classification, discretisation, feature selection