Minería de datos: aprendizaje no supervisado y detección de anomalías.

Dr. Juan Carlos Cubero

Módulo 2: Detección de Anomalías

Data Mining: Anomaly Detection

- Motivation and Introduction
- Supervised Methods
- Semisupervised Methods
- Unsupervised Methods:
  - Graphical and Statistical Approaches
  - Distance-based Approaches
  - Clustering-based
- Evaluation

Anomaly Detection

Data Analysis: Process of inspecting, cleaning, transforming, and modeling data with the goal of discovering useful information, suggesting conclusions, and supporting decision making

→ Common oriented: Find patterns, trends, etc

→ Uncommon oriented: Identify anomalies

Anomaly Detection

What are **outliers/anomalies**?

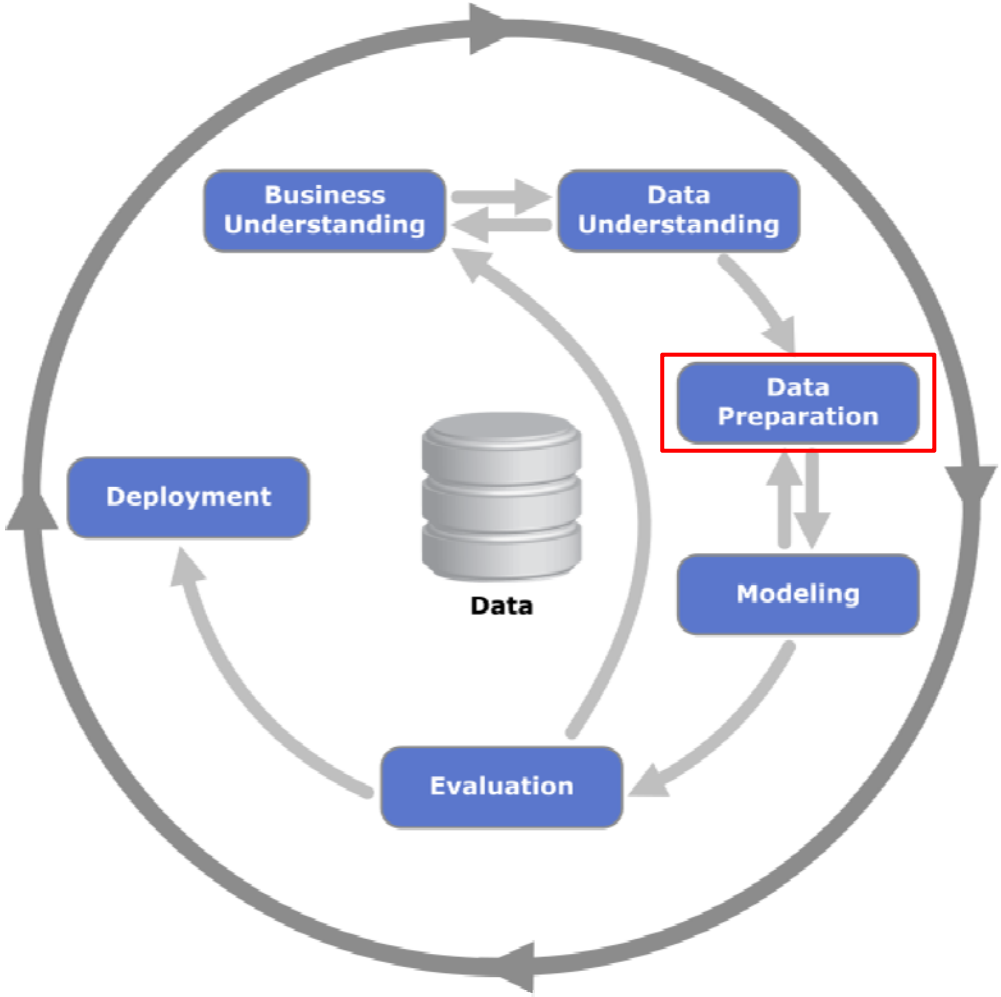
The set of data points that are considerably different than the remainder of the data

Working assumption: There are considerably more "normal" observations than "abnormal" observations (outliers/anomalies) in the data

What to do with an anomaly?

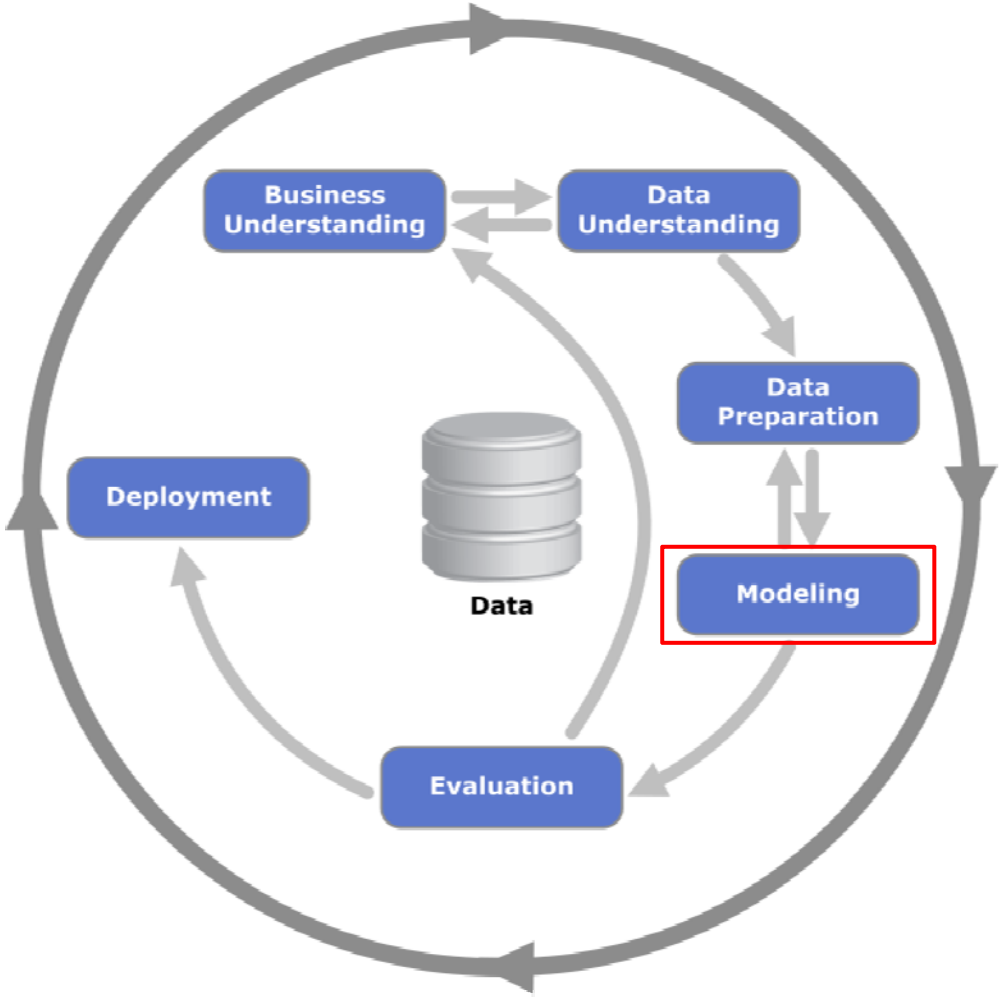
Anomaly Detection What to do with an anomaly?

They may represent errors when capturing data. So, they should be removed from posterior studies



Anomaly Detection What to do with an anomaly?

They may represent real data. So, they should be carefully analyzed



Anomaly Detection What to do with an anomaly?

Some techniques are quite robust to anomaly data, so it's presence is not so important:

- Classification methods, e.g, C4.5 behaves quite well for classifying instances of "majority" classes. They are not affected by the presence of few examples of "minority" classes.

- Association rules methods are based on support and confidence measures, which are robust to "exceptions"

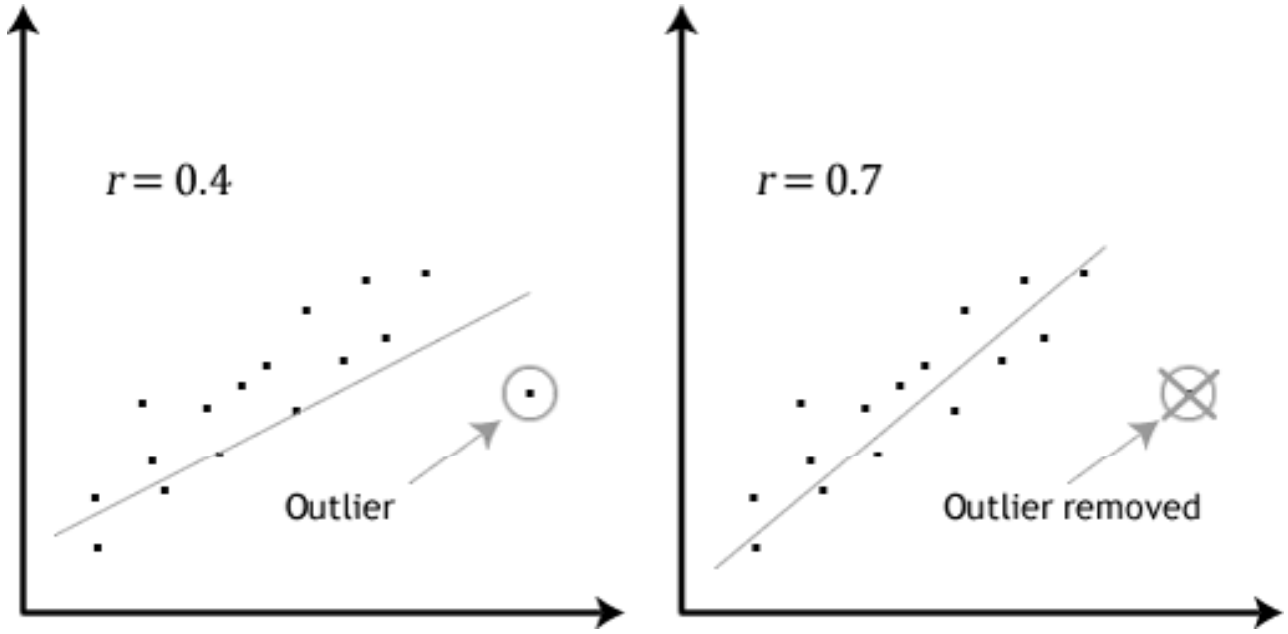
Anomaly Detection What to do with an anomaly?

Other techniques are not robust to anomaly data, so it's presence is quite important:

- Classification methods, e.g, C4.5 are not good for classifying "minority" classes examples because they focus on majority classes.

- Regression, Clustering, hypotheses tests, and any method based on averages are quite sensitive to abnormal values.

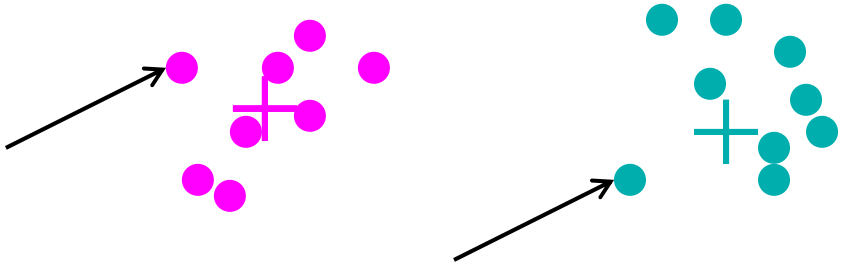
Anomaly Detection What to do with an anomaly?



Anomaly Detection What to do with an anomaly?

Centroids in a clustering with no abnormal data

Centroids in a clustering with abnormal data. In this case, data with abnormal combinations of two variables.



Anomaly Detection What to do with an anomaly?

We should not just remove an anomaly without a detailed analysis.

Bacon, writing in *Novum Organum* about 400 years ago said:

> "Errors of Nature, Sports and Monsters correct the understanding in regard to ordinary things, and reveal general forms. For whoever knows the ways of Nature will more easily notice her deviations; and, on the other hand, whoever knows her deviations will more accurately describe her ways."

Anomaly Detection What's the problem to tackle?

Finding a needle in a haystack is not a correct phrase to refer to the problem of finding anomalies because I know what a needle looks like

Anomaly Detection What's the problem to tackle?

Do I know what I have to find?

I know what I have to find

I have a complete and accurate description of the anomalous entity to be found

I don't know what I have to find

An anomaly is an abnormal entity

Anomaly Detection What's the problem to tackle?

Do I know what I have to find?

Example:

Network Intrusion Detection Systems (NIDS)

- NIDS Signature based:

  I know what I have to find.

- NIDS Anomaly detection based:

  I don't know what I have to find.

Anomaly Detection What's the problem to tackle?

NIDS Signature based:

- I know what I have to find
- The system maintains a collection of known signatures (attacks) and compare them with the network data streams

Blaster virus

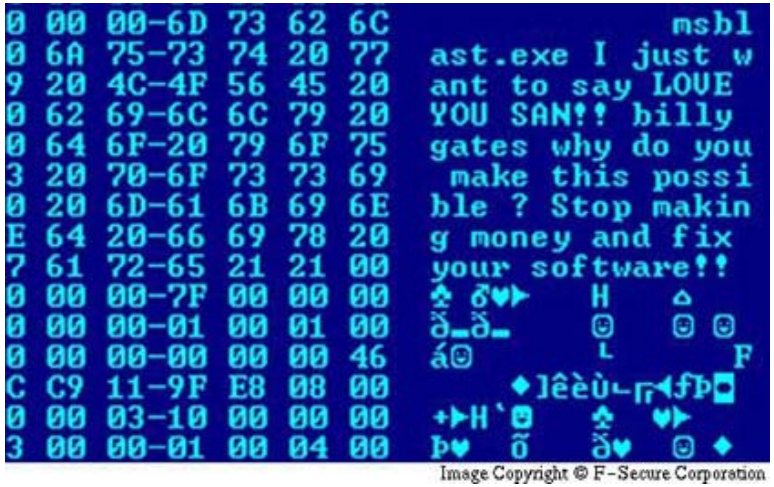


Image Copyright © F-Secure Corporation

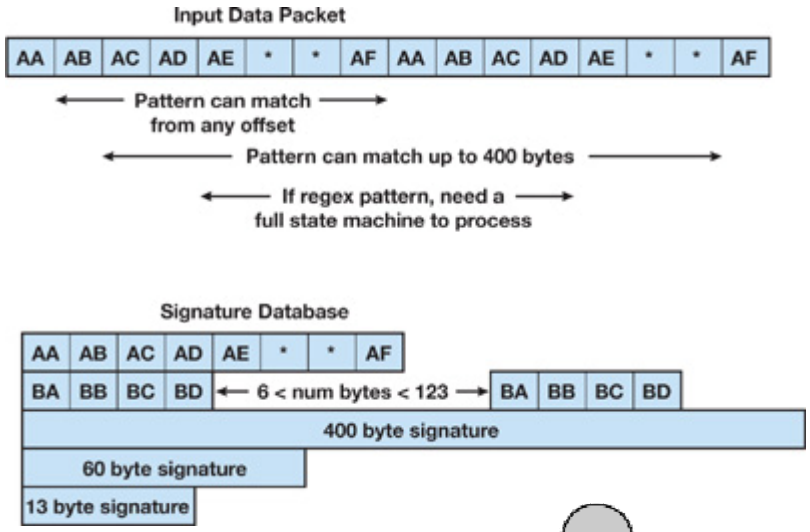
Bagle virus

```
1   lea edi, ptr [ebp+0x4025]   // edi = mem[ebp+...]
2   mov ecx, 0x3ec5             // ecx = 0x3ec5
3   mov edx, 0xef4013a0         // edx = 0xef4013a0
    loop:
4   mov al, byte ptr ds[edi]    // al = mem[ds+edi]
5   sub al, dl                  // al = al - dl
6   sub al, dh                  // al = al - dh
7   xor al, cl                  // al = al ^ cl
8   rol edx, cl                 // rotate edx by cl bits
9   mov byte ptr ds[edi], al    // mem[ds+edi] = al
10  inc edi                     // edi = edi + 1
11  dec ecx                     // ecx = ecx - 1
12  jnz loop                    // jump
13  push edi                    // push args into stack
14  call 0x7c92a950             // call a lib func
```

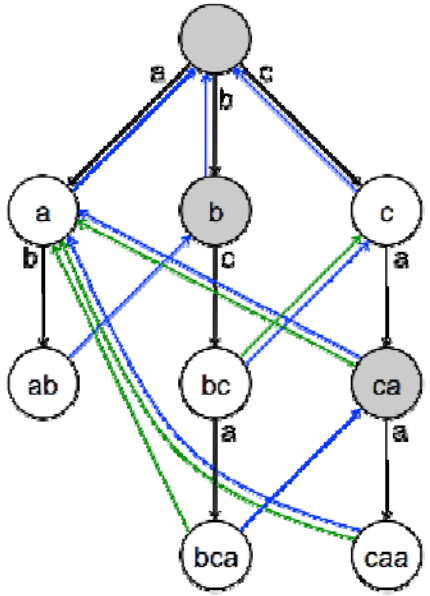
Anomaly Detection What's the problem to tackle?

NIDS Signature based:

- I know what I have to find.
- Main problem: string matching



- Aho-Corasick algorithm (and improvements)
  - A finite state machine is constructed
  - The dictionary is the virus database
  - A new entry is parsed



Anomaly Detection What's the problem to tackle?

NIDS Signature based:

- I know what I have to find.

NIDS Anomaly detection based:

- I don't know what I have to find.

New types of attacks have not a signature.
An attack is an ***abnormal*** entry

We'll focus on this kind of problems

Anomaly Detection: Applications

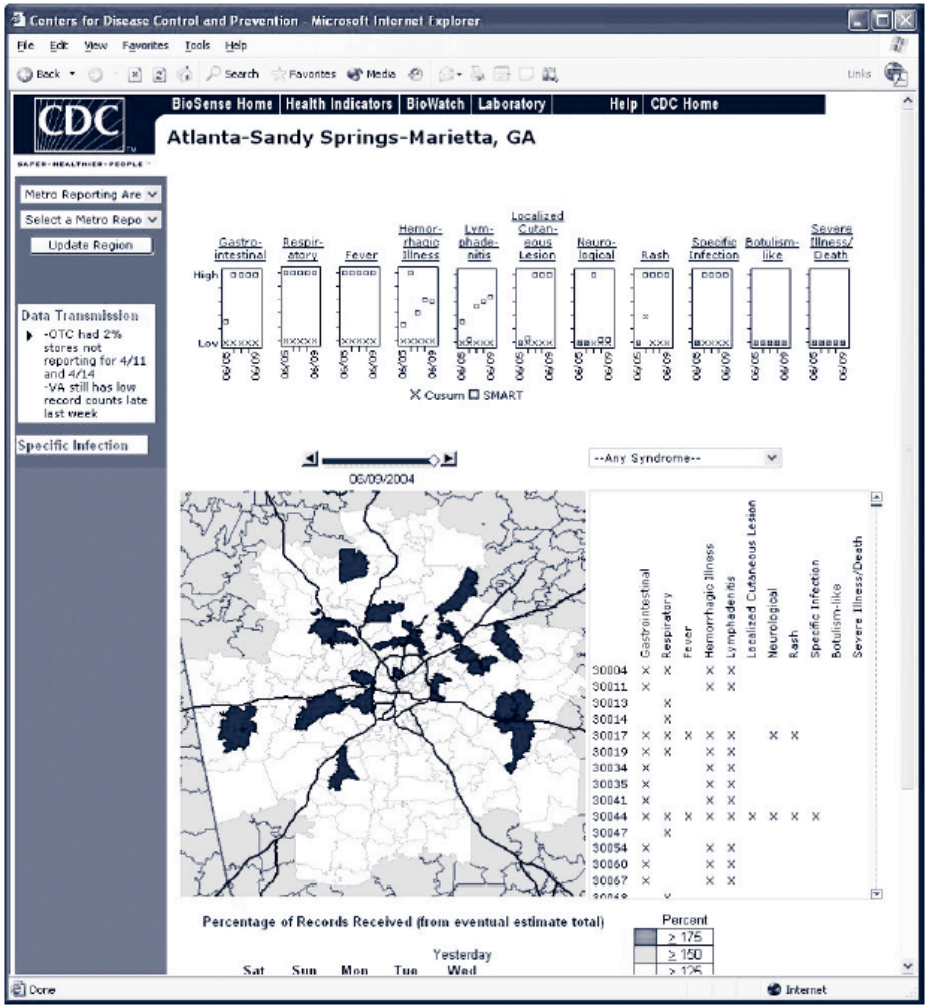
- Network Intrusions
  - A web server involved in *ftp* traffic

- Credit Card Fraud
  - An abnormally high purchase made on a credit card

- Anomalous patient records

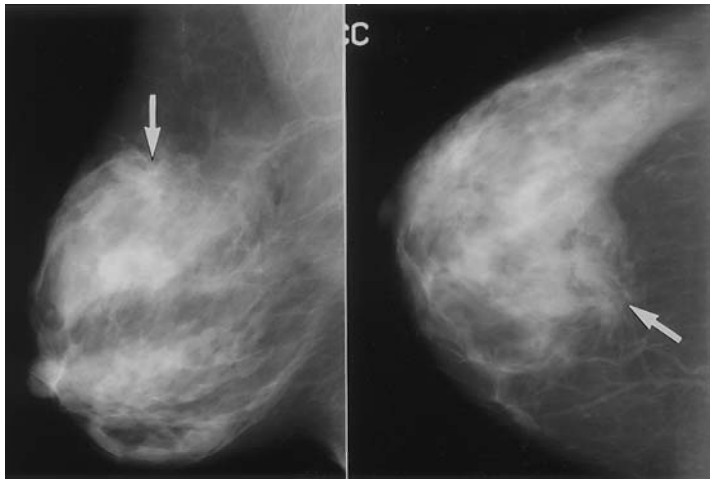
Anomaly Detection: Applications

- Disease Outbreak detection

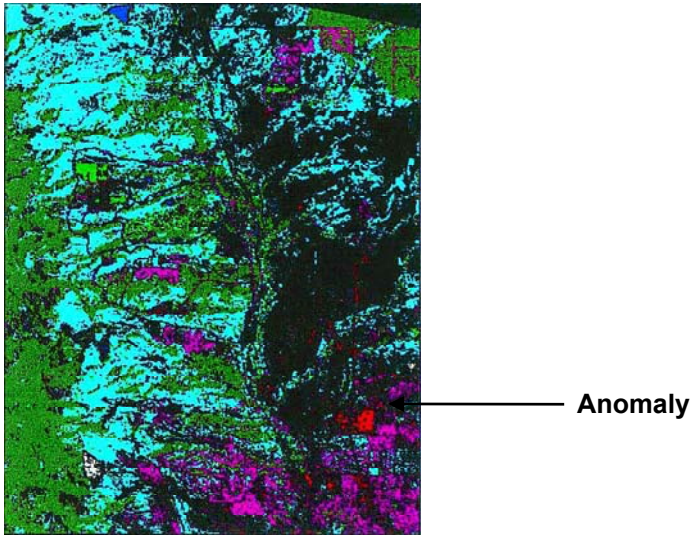


Anomaly Detection: Applications

- Anomalous regions within an image
  - mammography image analysis



  - satellite image analysis



Anomaly Detection: Applications

- Detecting outliers in a image monitored over time

- Suspicious events in video surveillance

Anomaly Detection: Aspects

- Nature of input data
- Availability of supervision
- Type of anomaly: point, contextual, structural
- Output of anomaly detection
- Evaluation of anomaly detection techniques

Anomaly Detection: Nature of Input

- Tabular data.
- More complex data require adaptation or specific techniques:
  - Sequential data
  - Graph data
  - Spatial data
  - Time series

Anomaly Detection: Availability of Supervision

Key questions:

Do I have anomalies in my training set?

Do I know which are the anomalies in my training set?

Anomaly Detection: Availability of Supervision

Supervised Methods →

I have anomalies in my training set and they are labelled

A classification model (including the anomaly class) is built.

Tid	SrcIP	Start time	Dest IP	Dest Port	Number of bytes	Attack
1	206.135.38.95	11:07:20	160.94.179.223	139	192	No
2	206.163.37.95	11:13:56	160.94.179.219	139	195	No
3	206.163.37.95	11:14:29	160.94.179.217	139	180	No
4	206.163.37.95	11:14:30	160.94.179.255	139	199	No
5	206.163.37.95	11:14:32	160.94.179.254	139	19	Yes
6	206.163.37.95	11:14:35	160.94.179.253	139	177	No
7	206.163.37.95	11:14:36	160.94.179.252	139	172	No
8	206.163.37.95	11:14:38	160.94.179.251	139	285	Yes
9	206.163.37.95	11:14:41	160.94.179.250	139	195	No
10	206.163.37.95	11:14:44	160.94.179.249	139	163	Yes

Anomaly Detection: Availability of Supervision

SemiSupervised Methods →

I do not have anomalies in my training set

Tid	SrcIP	Start time	Dest IP	Dest Port	Number of bytes	Attack
1	206.135.38.95	11:07:20	160.94.179.223	139	192	No
2	206.163.37.95	11:13:56	160.94.179.219	139	195	No
3	206.163.37.95	11:14:29	160.94.179.217	139	180	No
4	206.163.37.95	11:14:30	160.94.179.255	139	199	No
5	206.163.37.95	11:14:32	160.94.179.254	139	19	Yes
6	206.163.37.95	11:14:35	160.94.179.253	139	177	No
7	206.163.37.95	11:14:36	160.94.179.252	139	172	No
8	206.163.37.95	11:14:38	160.94.179.251	139	285	Yes
9	206.163.37.95	11:14:41	160.94.179.250	139	195	No
10	206.163.37.95	11:14:44	160.94.179.249	139	163	Yes

Anomaly Detection: Availability of Supervision

UnSupervised Methods →

I have anomalies in my training set but they are not labelled
I don't know if a record is an anomaly or not

Tid	SrcIP	Start time	Dest IP	Dest Port	Number of bytes	Attack
1	206.135.38.95	11:07:20	160.94.179.223	139	192	No
2	206.163.37.95	11:13:56	160.94.179.219	139	195	No
3	206.163.37.95	11:14:29	160.94.179.217	139	180	No
4	206.163.37.95	11:14:30	160.94.179.255	139	199	No
5	206.163.37.95	11:14:32	160.94.179.254	139	19	Yes
6	206.163.37.95	11:14:35	160.94.179.253	139	177	No
7	206.163.37.95	11:14:36	160.94.179.252	139	172	No
8	206.163.37.95	11:14:38	160.94.179.251	139	285	Yes
9	206.163.37.95	11:14:41	160.94.179.250	139	195	No
10	206.163.37.95	11:14:44	160.94.179.249	139	163	Yes

Anomaly Detection: Output?

- The output is a label or a score:

  Label:

  - The procedure assigns to each test instance a label: *normal* or *anomaly*
  - This is especially true in classification-based approaches

  Score:

  - Each test instance is assigned an anomaly score (real number)
    - Allows the output to be ranked
    - Requires an additional threshold parameter

Data Mining: Anomaly Detection

- Motivation and Introduction
- Supervised Methods
- Semisupervised Methods
- Unsupervised Methods:
  - Graphical and Statistical approaches
  - Nearest neighbor based approaches
  - Clustering based approaches
- Evaluation

Supervised Methods

Supervised Methods →

- There are training and test sets available.
- Training cases include a label stating if it is an anomaly or not.
- A classification model (including the anomaly class) is built.

Tid	SrcIP	Start time	Dest IP	Dest Port	Number of bytes	Attack
1	206.135.38.95	11:07:20	160.94.179.223	139	192	No
2	206.163.37.95	11:13:56	160.94.179.219	139	195	No
3	206.163.37.95	11:14:29	160.94.179.217	139	180	No
4	206.163.37.95	11:14:30	160.94.179.255	139	199	No
5	206.163.37.95	11:14:32	160.94.179.254	139	19	Yes
6	206.163.37.95	11:14:35	160.94.179.253	139	177	No
7	206.163.37.95	11:14:36	160.94.179.252	139	172	No
8	206.163.37.95	11:14:38	160.94.179.251	139	285	Yes
9	206.163.37.95	11:14:41	160.94.179.250	139	195	No
10	206.163.37.95	11:14:44	160.94.179.249	139	163	Yes

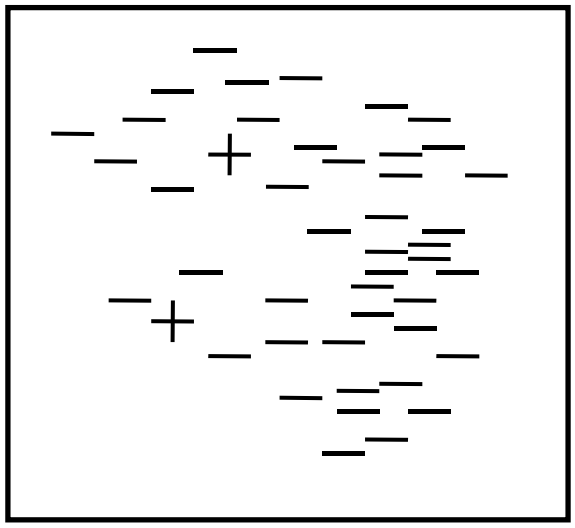
Supervised Methods

- Well known classification models:
  - Decision rules
  - Bayesian nets
  - Support Vector Machines, etc.
- Key issue: → Classification of a class attribute with very rare class values → **Unbalanced classification**

  Why unbalanced data sets require special treatment in a classification model?

  Suppose a intrusion detection problem.
  - Two classes: *normal* (99.9%) and *intrusion* (0.1%)
  - The default classifier, always labeling each new entry as *normal,* would have 99.9% accuracy!

Supervised Methods. Evaluation measures



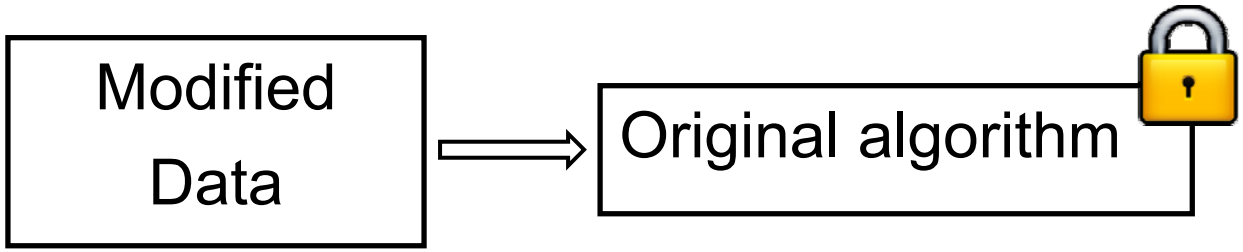
Silly Classification Algorithm Output is Always −

99 % Accuracy !

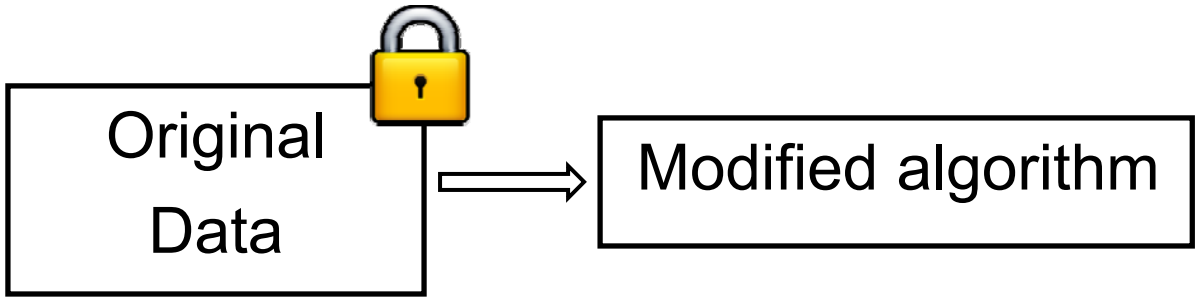
Supervised Methods

There are two approaches to manage the problem of Classification (Supervised Learning) with rare classes.

Instance based methods



Algorithm based methods



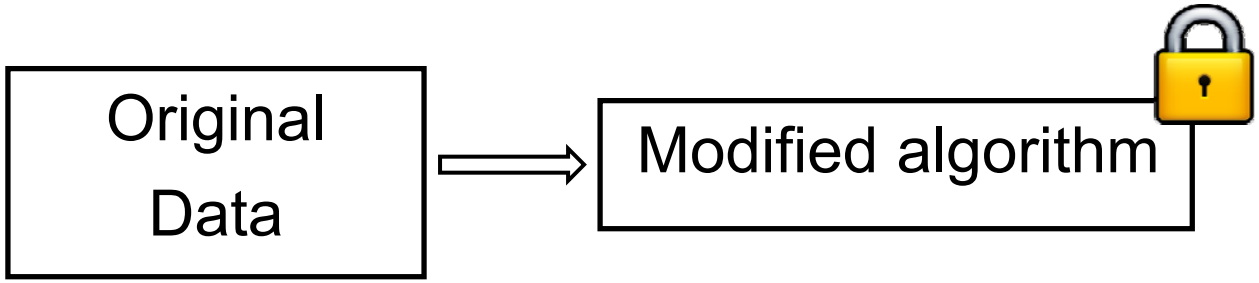
Supervised Methods

- **Instance based methods**. These methods change the learning dataset before applying the classification:
  - Undersampling the majority classes
  - Oversampling the minority classes

- **Algorithm based methods.** These methods do not alter the learning dataset but weight the instances someway:
  - Cost sensitive
  - Bagging
  - Boosting

Supervised Methods: Instance based

Instance based methods



- Undersampling the majority class (Tomek-links, CNN, etc)
- Oversampling the minority class (Smote and variants)

Supervised Methods

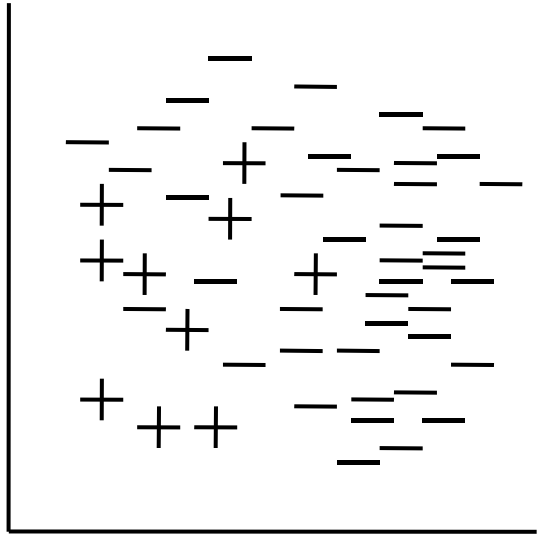
- **Undersampling** by removing instances of the majority classes:
  - Tomek-links
  - CNN: Condensed nearest neighbor. (Hart algoritm)
  - NCL: Neighborhood Cleaning Rule

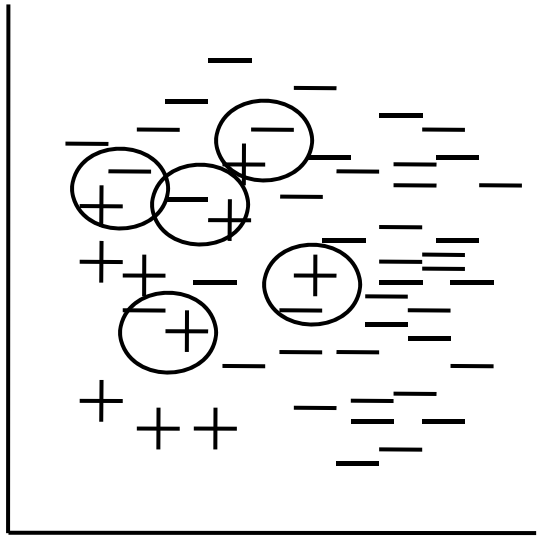
CNN: Wikipedia → k-nearest neighbors algorithm

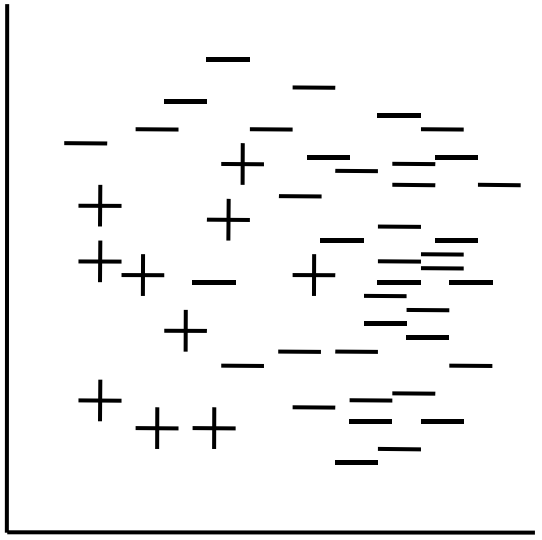
I. Tomek IEEE Transactions on Systems, Man, and Cybernetics - TSMC , vol. 6, no. 11, pp. 769-772, 1976

NCL: Jorma Laurikkala. 2001. Improving Identification of Difficult Small Classes by Balancing Class Distribution. In Proceedings of the 8th Conference on AI in Medicine in Europe: Artificial Intelligence Medicine (AIME '01), Springer-Verlag, London, UK, UK, 63-66.

Supervised Methods: Undersampling. Tomek links



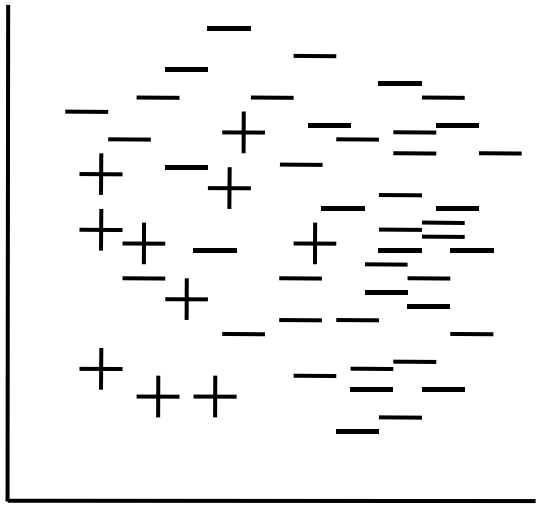




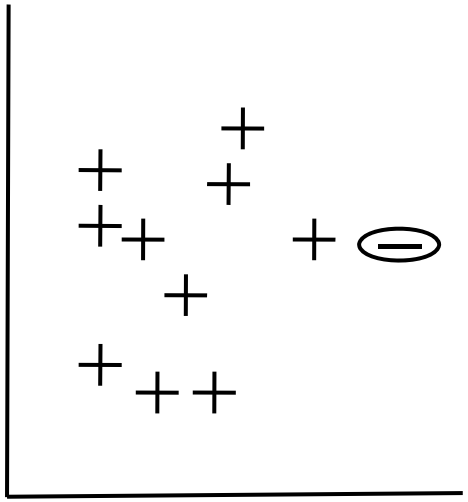
Tomek links: Pair of (+,-) instances with minimum distance. The instances of the majority class are removed.

Final dataset obtained by undersampling with Tomek links

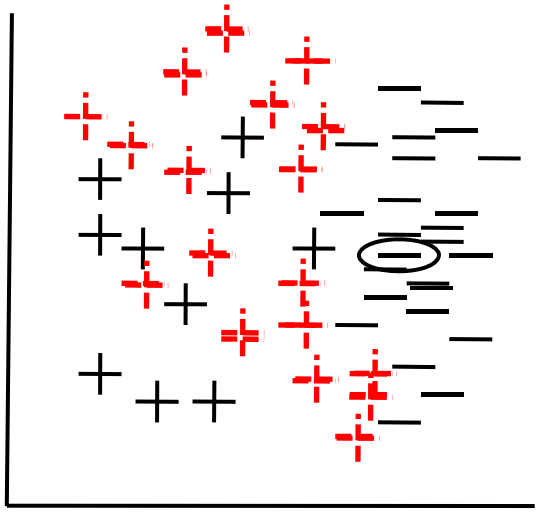
Supervised Methods: Undersampling. CNN



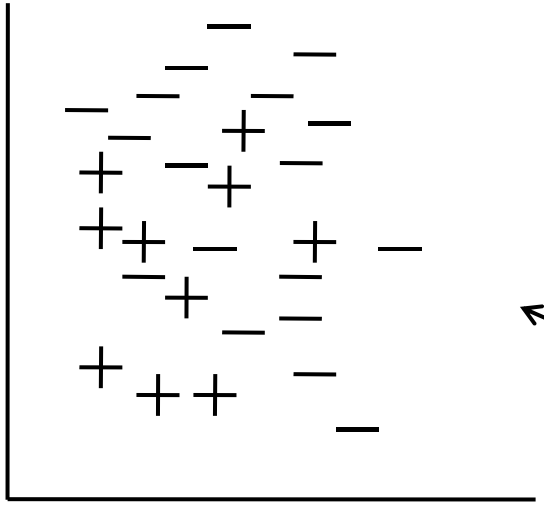
E: Original dataset



E': Pick only positive class instances and a random negative class instance

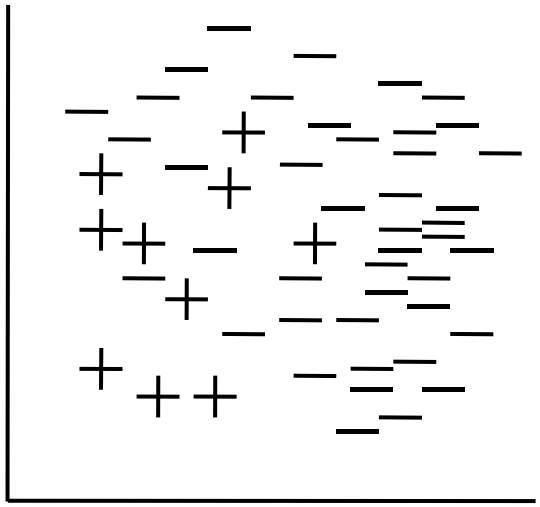


Use E as a test set and E' as learning set. Classify all the instances of E by using 1-NN. Include in E' the instances which have not been correctly classified (in red)

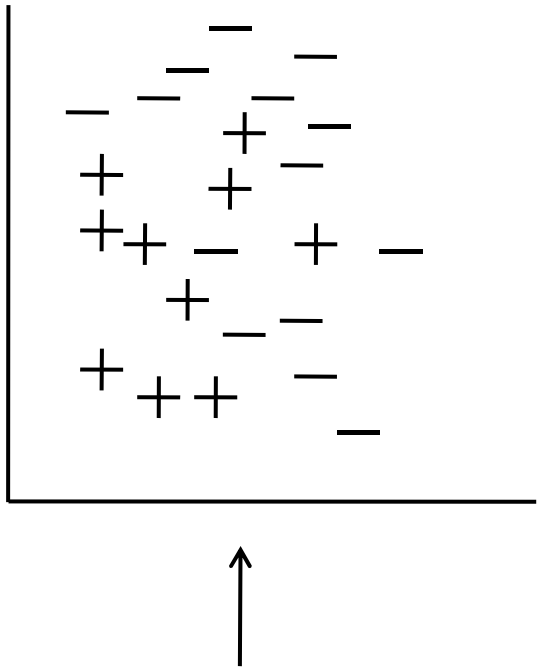


Final dataset obtained by undersampling with CNN

Supervised Methods: Hybrid undersampling



E: Original dataset

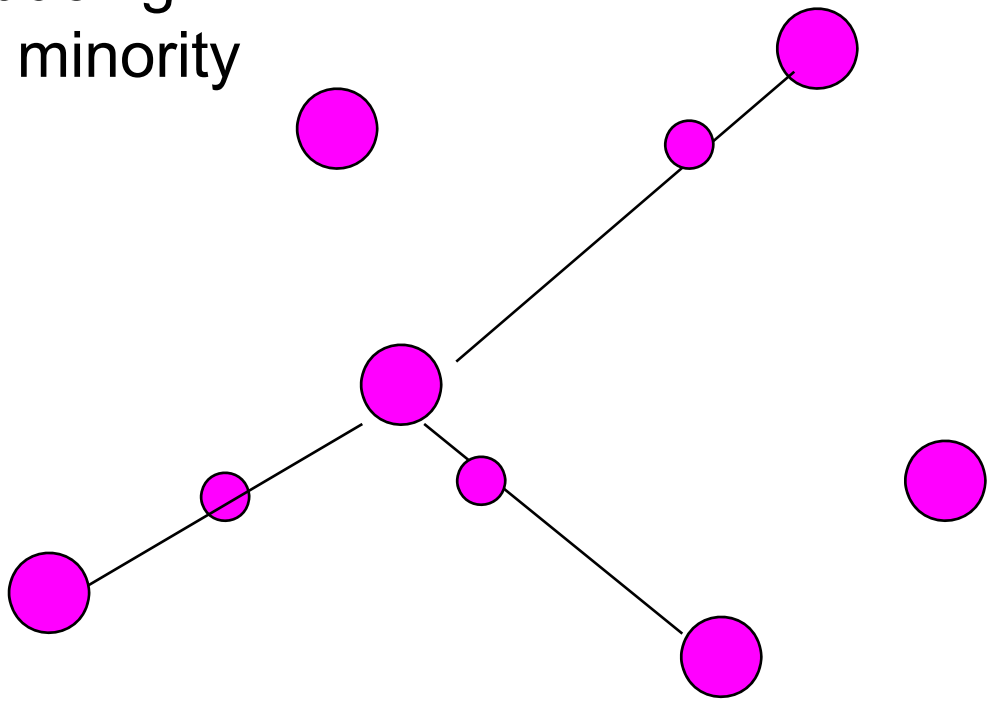


Final dataset obtained by undersampling with Tomek links + CNN

Supervised Methods: Oversampling. Smote

- **Oversampling**: by introducing artificial instances of the minority classes

SMOTE: For a minority class instance, several (no more than five) of it's k minority class nearest neighbors are randomly selected. A new data set is generated in the line between the selected instance and the selected neighbors



: Minority sample

: Synthetic sample

Supervised Methods: Oversampling. Smote

SMOTE also performs an undersampling process over the majority class together with the oversampling over the minority one.

SMOTE: Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. 2002. SMOTE: synthetic minority over-sampling technique. *J. Artif. Int. Res.* 16, 1 (June 2002), 321-357.

Supervised Methods: Oversampling

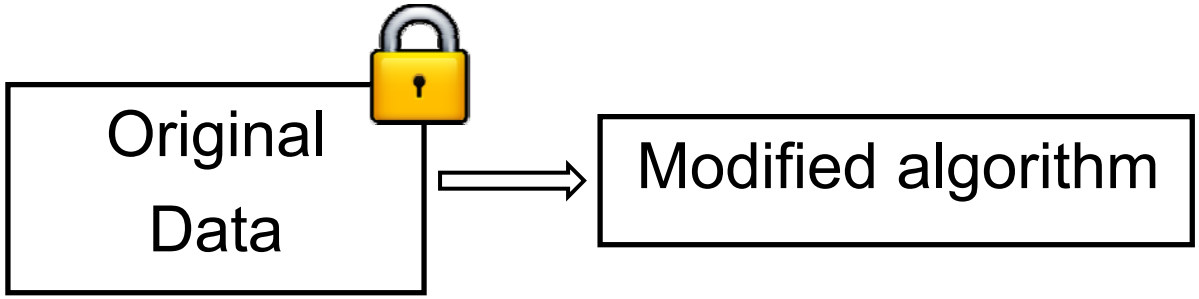
W. Fan, M. Miller, S. Stolfo, W. Lee, and P. Chan. 2004. Using artificial anomalies to detect unknown and known network intrusions. ***Knowledge and Information Systems*** 6, 5 (September 2004), 507-527. DOI=10.1007/s10115-003-0132-7

- Artificial anomalies are injected
- Anomalies are classified using Ripper
- 1999 KDD Cup Dataset is used. Attack types: denial of service (DOS), probing (PRB), remotely gaining illegal remote access to a local account or service (R2L) and local user gaining illegal root access (U2R)

RIPPER Rule	Meaning
guess:- failed_logins $\geq$ 4.	If number of failed logins is at least 4, then this telnet connection is "guess", a guessing password attack.
overflow:- hot $\geq$ 3, compromised $\geq$ 2, root_shell = 1.	If the number of hot indicators is at least 3, the number of compromised conditions is at least 2, and a root shell is obtained, then this telnet connection is a buffer overflow attack.
. . .	. . .
normal:- true.	If none of the above, then this connection is "normal".

Supervised Methods: Algorithm based

Algorithm based methods



- Cost-sensitive
- Boosting
- Bagging
- Otros

Supervised Methods: Algorithm based

- **Algorithm based methods**. These methods do not change the learning dataset but they modify the learning algorithm to give more weight to instances of minority classes.
  - Cost-sensitive methods (assigning high cost to the minority class value)
  - Bagging → by including more minority class instances in each step of the bagging algorithm.
  - Boosting → by giving more weight to the minority class instances in each step of the boosting algorithm
  - Specific adaptations of rule based methods, neural networks, SVM's. etc.
  - Hybrid methods: SmoteBoosting, SmoteBagging, etc

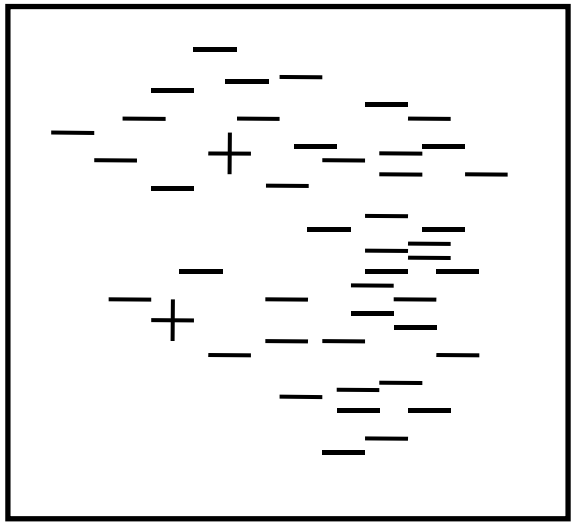
Supervised Methods

- In summary, classifications models should be properly modified to deal with unbalanced sets:

  - By modifying the learning dataset
  - By giving more “weight” to the minority class during the learning process.

Haibo He and Edwardo A. Garcia. 2009. Learning from Imbalanced Data. IEEE Trans. on Knowl. and Data Eng. 21, 9 (September 2009), 1263-1284. DOI=10.1109/TKDE.2008.239 http://dx.doi.org/10.1109/TKDE.2008.239

Supervised Methods. Evaluation measures



Silly Classification Algorithm Output is Always −

99 % Accuracy !

Supervised Methods. Evaluation measures

- **Evaluation measures**.
  - Remember that accuracy is not a good evaluation measure when dealing with rare class values: The default classifier, always labeling each new entry as normal, would have 99.9% accuracy when the anomaly has a presence of 0.1% .
  - We need other evaluation measures as alternatives to accuracy: Recall, Precision, F-measure, ROC-curves, etc
  - BUT there's a specific problem for anomalies when they are extremely unbalanced → Base rate fallacy
  - More at the end of these slides.

Data Mining: Anomaly Detection

- Motivation and Introduction
- Supervised Methods
- Semisupervised Methods
- Unsupervised Methods:
  - Graphical and Statistical approaches
  - Nearest neighbor based approaches
  - Clustering based approaches
- Evaluation

Semi-Supervised Methods

- More than often, labelled training sets (regarding the anomaly) are not available.

  → Supervised methods are not applicable.

Semi-Supervised Methods

Semi-supervised Methods →

Training cases do not include anomalies.

A **profile/model** for *normal behavior* is constructed. A new case is an anomaly if it doesn't match the profile.

How to model *normal behavior*?

Tid	SrcIP	Start time	Dest IP	Dest Port	Number of bytes	Attack
1	206.135.38.95	11:07:20	160.94.179.223	139	192	No
2	206.163.37.95	11:13:56	160.94.179.219	139	195	No
3	206.163.37.95	11:14:29	160.94.179.217	139	180	No
4	206.163.37.95	11:14:30	160.94.179.255	139	199	No
5	206.163.37.95	11:14:32	160.94.179.254	139	19	Yes
6	206.163.37.95	11:14:35	160.94.179.253	139	177	No
7	206.163.37.95	11:14:36	160.94.179.252	139	172	No
8	206.163.37.95	11:14:38	160.94.179.251	139	285	Yes
9	206.163.37.95	11:14:41	160.94.179.250	139	195	No
10	206.163.37.95	11:14:44	160.94.179.249	139	163	Yes

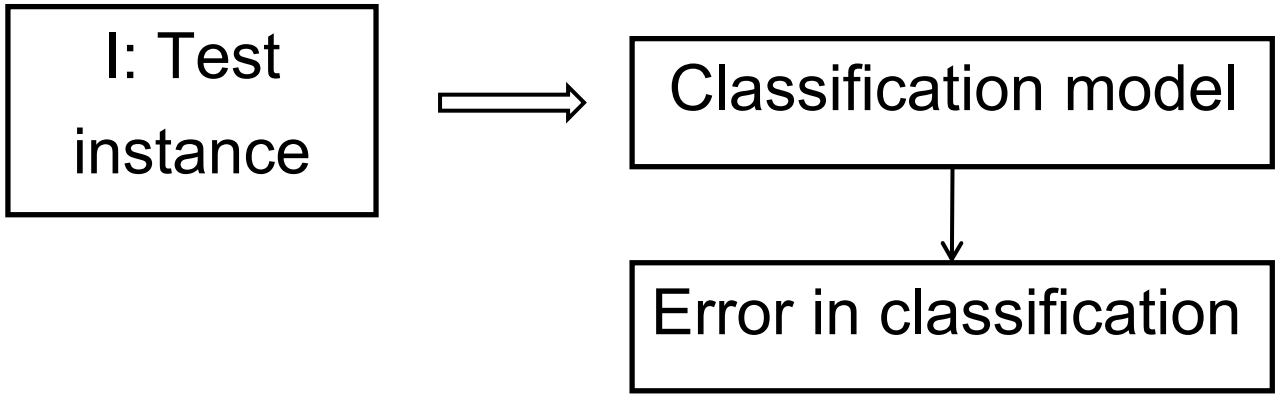
Semi-Supervised Methods

How to model *normal behavior*?

- Classification based,
- Association rules,
- Support vector description,
- Other methods

Semi-Supervised Methods : Classification based

- A classification model for an arbitrary (not unbalanced) class attribute is available.



I is declared to be an anomaly

Problem: Too many false positives

Semi-Supervised Methods: Classification based

Classification Based Methods

- A classification model is available, but there's no an *Anomalous* class attribute as in supervised methods)
- A new case is passed through the model
- A mismatch of the classifier is considered an anomaly.

For instance, a rule based classification model:

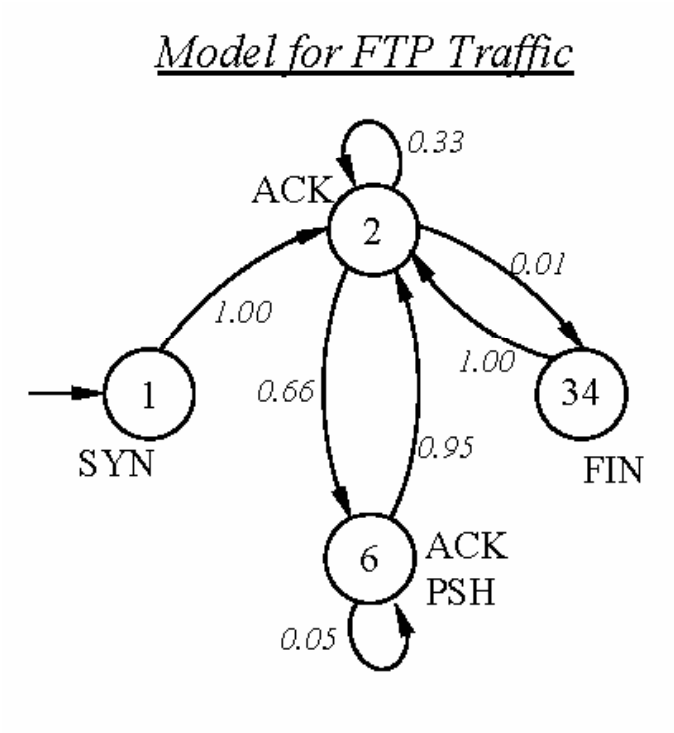
A → C (supp=65, conf=0.9)
else
B → D (supp=110, conf=0.8)
….

An anomaly score factor is reported, depending on the degree of quality of the covering rule(s) → Depending on the rule based method (RIPPER for instance) only one rule or several rules would be fired.

The same method applies to other classifiers such as decision trees

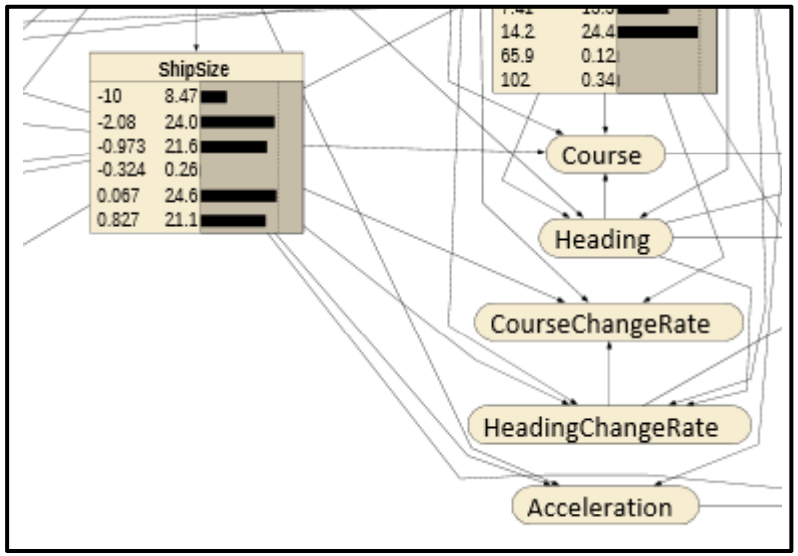
Semi-Supervised Methods : Classification based

Classification Based : Bayesian Methods / Markov Models



S1 S2 S6 S6 S2 S34

is an anomaly



Semi-Supervised Methods: Classification based

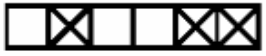
Classification Based : Bayesian Methods / Markov Models

Juan M. Estevez-Tapiador, Pedro Garcia-Teodoro, and Jesus E. Diaz-Verdejo. **2003**. Stochastic Protocol Modeling for Anomaly Based Network Intrusion Detection. In *Proceedings of the First IEEE International Workshop on Information Assurance (IWIA'03)* (IEEE-IWIA '03). IEEE Computer Society

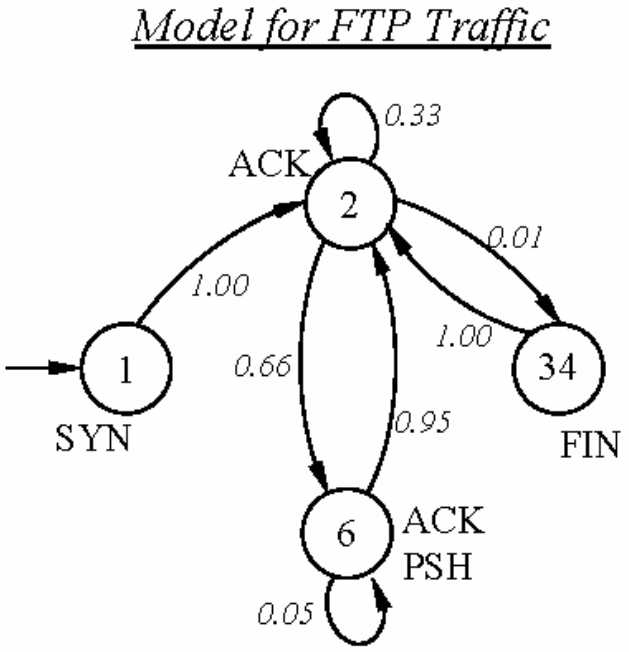
Analysis of ftp, http, ssh traffic through PDU (Protocol Data Unit) headers.

FIN	URG	RST	PSH	ACK	SYN
32	16	8	4	2	1

Examples:

Flags configuration	*Symbol*
	S_6
	S_{34}
	S_{29}
	S_{19}

Markov chain



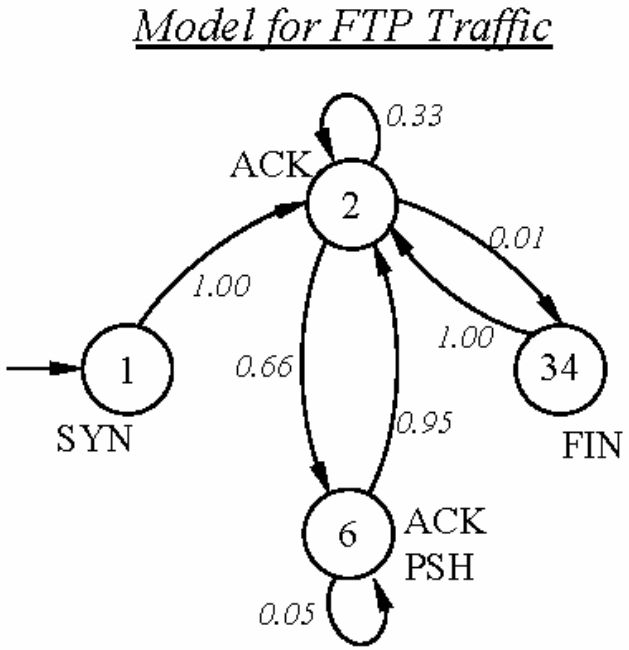
Semi-Supervised Methods : Classification based

Classification Based : Bayesian Methods / Markov Models

Evaluation: Extract headers from the packets of the network flow
→ If the sequence is not probable in the learned model, it is an anomaly

S1 S2 S6 S6 S2 S34

is an anomaly



Network intrusion detection systems (NIDS) review, **2009**. P. García-Teodoro, J. Díaz-Verdejo, G. Maciá-Fernández, E. Vázquez. *Computers & Security.* Volume 28, Issues 1–2, 18–28

Semi-Supervised Methods : Classification based

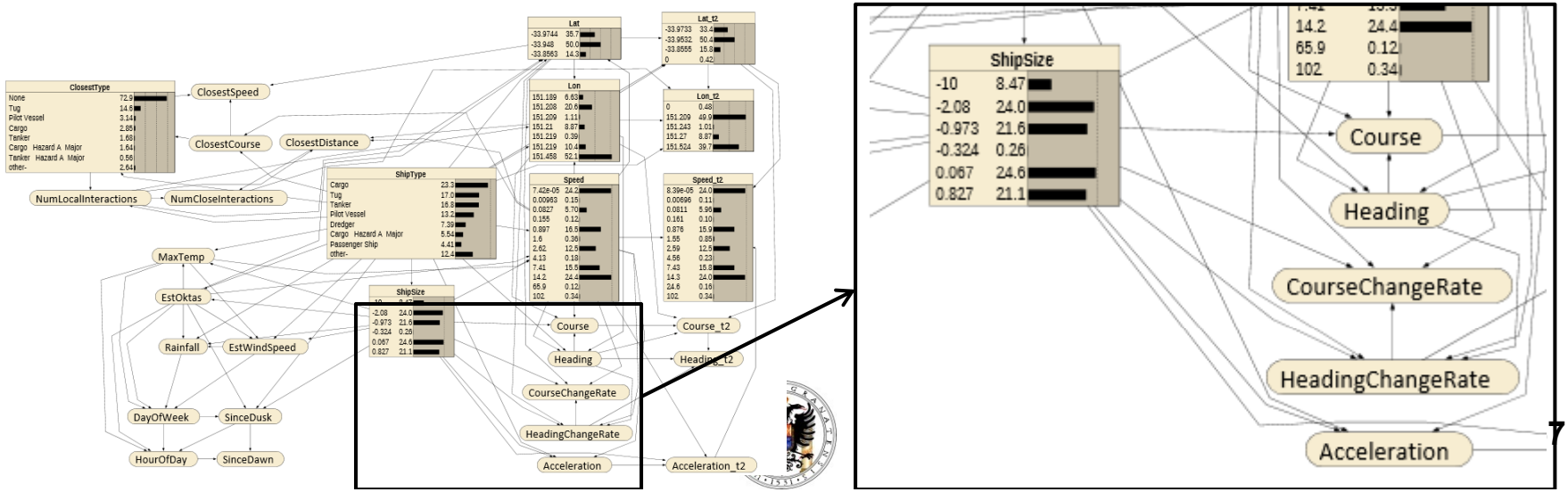
Classification Based : Bayesian Methods / Markov Models

Steven Mascaroa, Ann E. Nicholsob, Kevin B. Korbb. **2013**. Anomaly detection in vessel tracks using Bayesian networks. International Journal of Approximate Reasoning, online in 2013.

Data contains information about vessel type, weather, position, time, etc. Each one has a corresponding node in the BN.

The structure of the BN is learned (CaMML).

Records with low probabilities given the learned BN are anomalies



Semi-Supervised Methods: Rules Based

Rules Based Methods: Frequent Patterns and Association Rules

LERAD: Instead of generating rules with high support and confidence, the objective is to generate rules U → W with low P(¬W|U)
How to estimate P(¬W|U)? ¬W is not observed!

Training set A: U → W_1

Training set B: U → W_1 or W_2 or W_3 ← More likely to be violated
Smaller predictive power

A new case violating the rule will have a higher outlier factor in case A.

$$SrcIp = 128.1.2.3 \bigwedge DestIp = 128.4.5.6$$
$$\Rightarrow DestPort \in \{21, 25, 80\} \quad [p = r/n = 3/100]$$

n is the support
r is the number of consequents
p=r/n estimates P(¬W|U)

Gaurav Tandon and Philip K. Chan. **2007**. Weighting versus pruning in rule validation for detecting network and host anomalies. In *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining* (**KDD '07**). ACM, New York, NY, USA, 697-706. DOI=10.1145/1281192.1281267

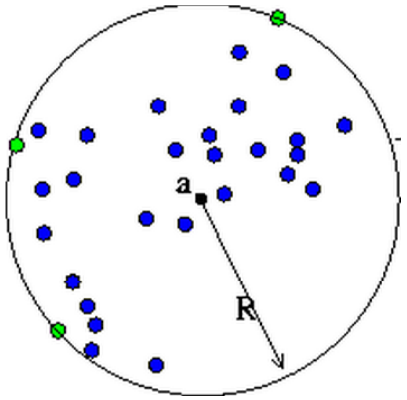
Semi-Supervised Methods: kernel based

Support Vector Machines Based Methods

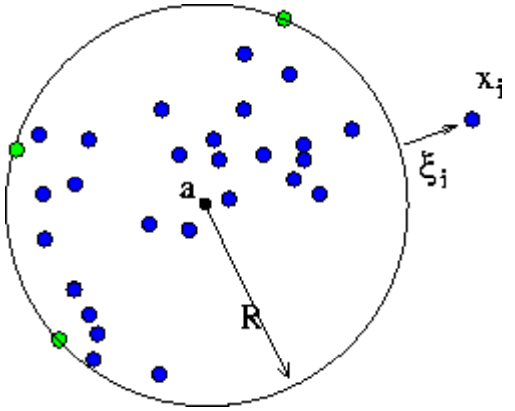
$$\min_{a,R} R^2 \text{ such that } (x_i - a)^2 \leq R^2 \; \forall i$$

Assumption:
There's only one *normal* class.

The *normal* profile is built by determining a boundary region.

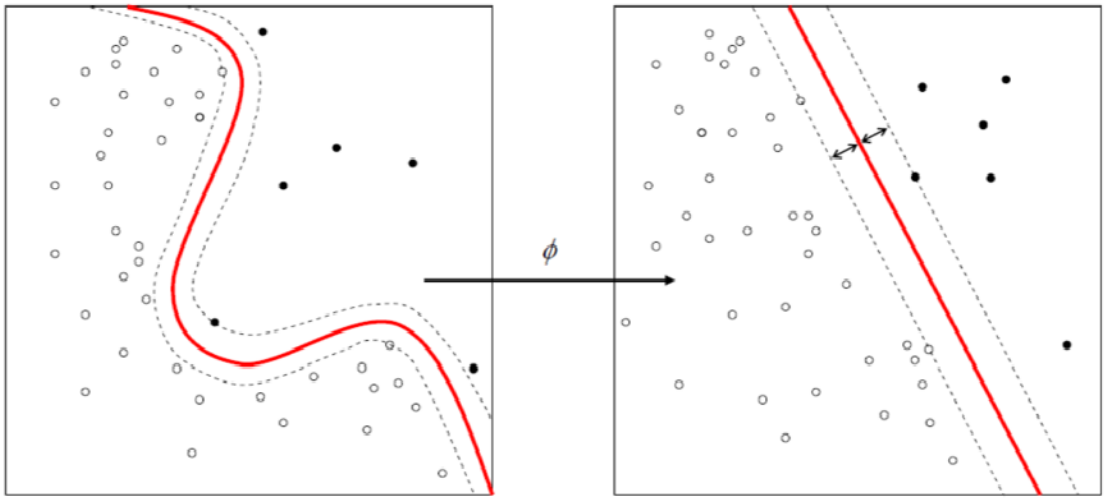


A new case is classified as anomaly if it lies outside the boundary region



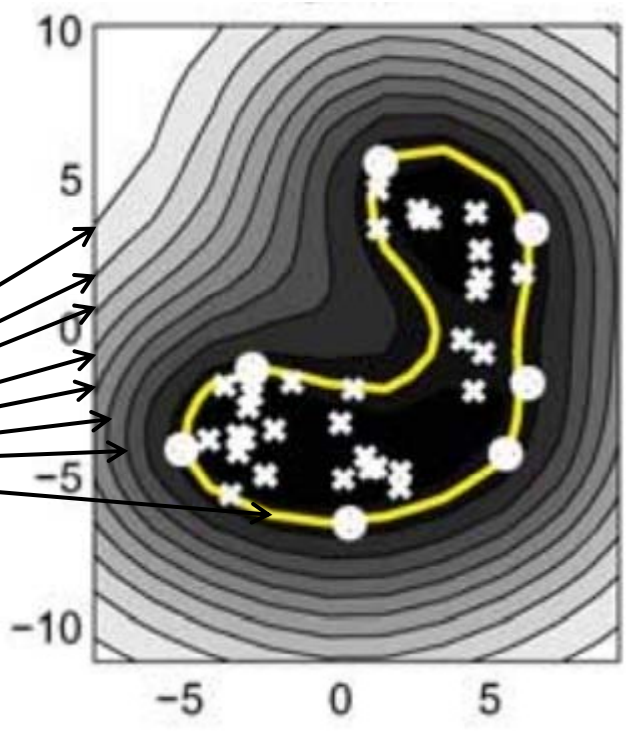
Semi-Supervised Methods : kernel based

By using other kernels functions, we get non-spheric shapes



Support vector data description

Boundaries (outlier) regions



Semi-Supervised Methods : kernel based

Support vector data description.

David M. J. Tax and Robert P. W. Duin. **2004**. Support Vector Data Description. *Mach. Learn.* 54, 1 (January 2004), 45-66. DOI=10.1023/B:MACH.0000008084.60811.49

Phuoc Nguyen, Dat Tran, Xu Huang, and Wanli Ma. **2013**. Parallel support vector data description. (IWANN'13 Vol. Part I. 280-290. DOI=10.1007/978-3-642-38679-4_27

Semi-Supervised Methods : kernel based

One class support vector machines for detecting **anomalous windows registry accesses** (**2003**) by K Heller, K Svore, A Keromytis, Stolfo. Proceedings of the workshop on Data Mining for Computer Security

Process: EXPLORER.EXE
Query: OpenKey
Key: HKCR\CLSID\B41DB860-8EE4-11D2-9906-E49FADC173CA\shellex\MayChange DefaultMenu
Response: SUCCESS
ResultValue: NOTFOUND

Each 5-tuple record is transformed into a higher dimensional space, considering all the possible values each attribute can take → (0,0,1,0,0,0,0,1,0,0,0, …. , 1,0)

Semi-Supervised Methods : Historical records

Other methods: Counting historical records

Assumption:

Anomalies are current events which have not been observed in the past

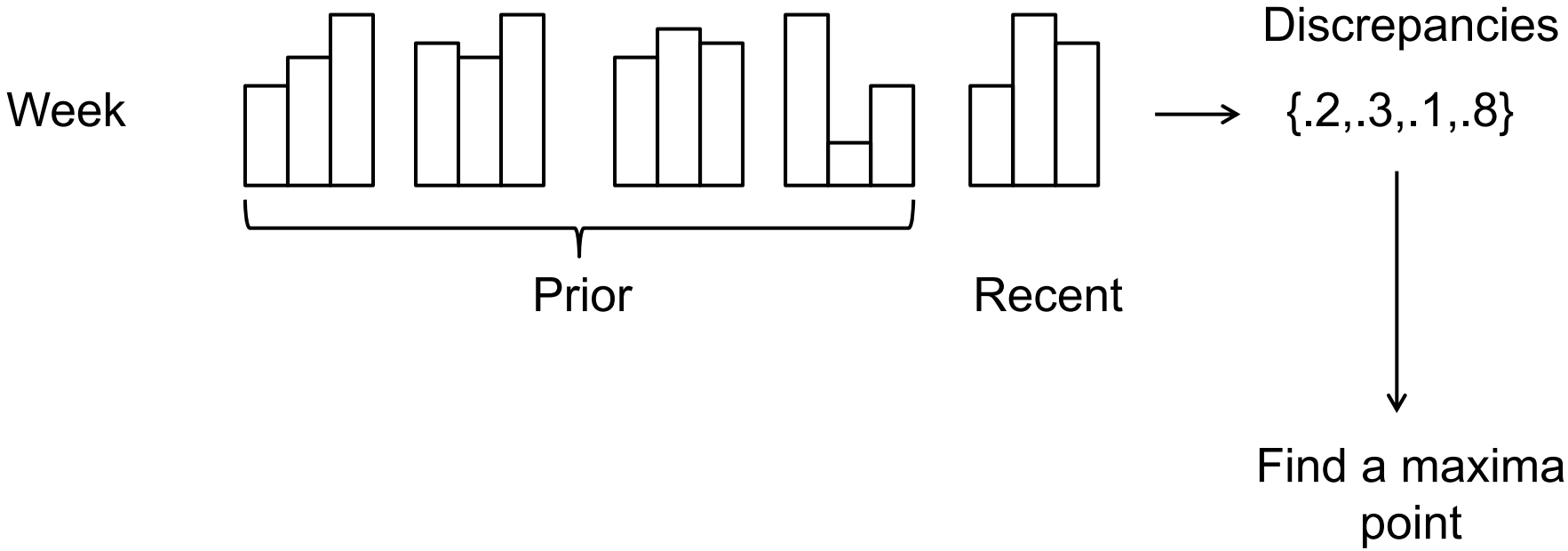
Approach: Compare counting of current events with counting of historical events in order to detect anomalies.

- Choose some attributes or characteristics of interest
- Count the number of observations of each attribute in a particular time
- Compare this count with historical data

M. S. Beigi, S.-F. Chang, S. Ebadollahi, and D. C. Verma. **2011**. Anomaly detection in information streams without prior domain knowledge. IBM J. Res. Dev. 55, 5 (September 2011), 550-560. DOI=10.1147/JRD.2011.2163280 http://dx.doi.org/10.1147/JRD.2011.2163280

Semi-Supervised Methods : Historical records

Other methods: Counting historical records



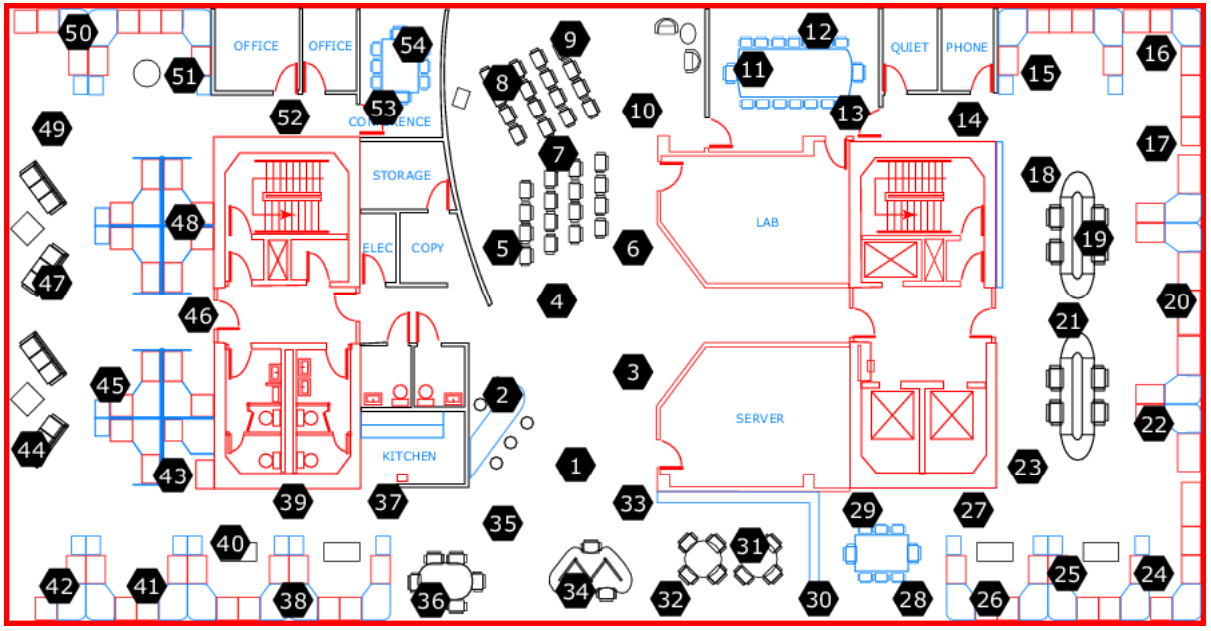
Apply the same procedure with other time scales: Month, Year, etc.

Semi-Supervised Methods : Historical records

Other methods: Counting historical records

Applications:

- Detecting anomalies in ambient conditions in a Laboratory



Data Mining: Anomaly Detection

- Motivation and Introduction
- Supervised Methods
- Semisupervised Methods
- Unsupervised Methods:
  - Graphical and Statistical approaches
  - Nearest neighbor based approaches
  - Clustering based approaches
- Evaluation

Unsupervised Methods

Unsupervised Methods →

Training cases include anomalies and they are not labelled.

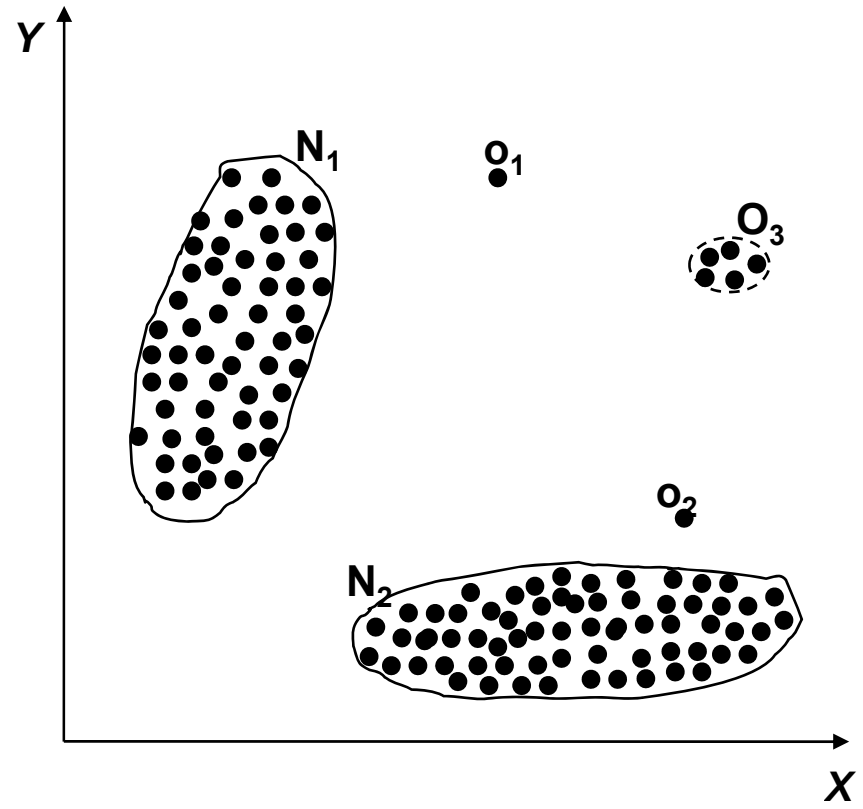
Tid	SrcIP	Start time	Dest IP	Dest Port	Number of bytes	Attack
1	206.135.38.95	11:07:20	160.94.179.223	139	192	No
2	206.163.37.95	11:13:56	160.94.179.219	139	195	No
3	206.163.37.95	11:14:29	160.94.179.217	139	180	No
4	206.163.37.95	11:14:30	160.94.179.255	139	199	No
5	206.163.37.95	11:14:32	160.94.179.254	139	19	Yes
6	206.163.37.95	11:14:35	160.94.179.253	139	177	No
7	206.163.37.95	11:14:36	160.94.179.252	139	172	No
8	206.163.37.95	11:14:38	160.94.179.251	139	285	Yes
9	206.163.37.95	11:14:41	160.94.179.250	139	195	No
10	206.163.37.95	11:14:44	160.94.179.249	139	163	Yes

Unsupervised Methods

Unsupervised Methods →

Training cases may include anomalies but there are no classification labels.

A new case is considered an anomaly by taken into account its relation with the rest of data.



Unsupervised Methods

- **Graphical approaches**:
  Given a database D, inspects it visually and determines which points are anomalies

- **Statistical-based:**
  Given a database D, and a data point **x** $\in$ D, a statistical test determines whether **x** is an anomaly or not, at a significance level p. These tests assume a latent distribution.

- **Distance-based:**
  There's available a distance measure which can be applied to any pair of data instances and is able to discriminate between the anomalies and normal instances well enough.

  > Nearest neighbor based approaches

  > Cluster based approaches

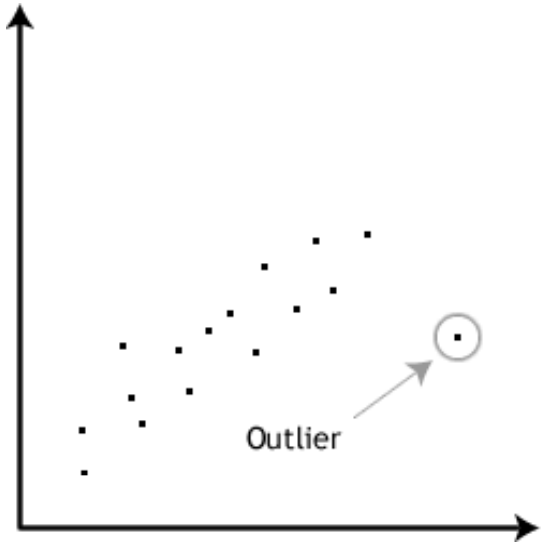
Data Mining: Anomaly Detection

- Motivation and Introduction
- Supervised Methods
- Semisupervised Methods
- Unsupervised Methods:
  - Graphical and Statistical approaches
  - Nearest neighbor based approaches
  - Clustering based approaches
- Evaluation

Graphical Approaches

- **Graphical approaches**:
  Given a database D, inspects it visually and determines which points are anomalies

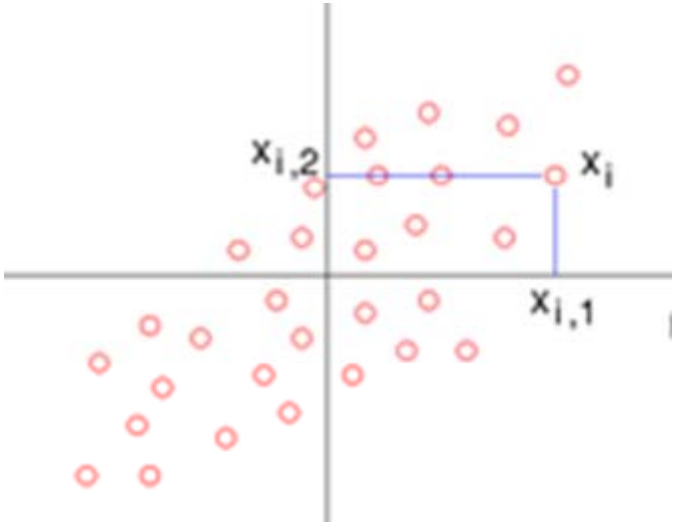
- **Limitations**
  - Time consuming
  - Subjective



Graphical Approaches

How to resume the information given by several attributes into a couple of dimensions? One possibility: by projecting. Obviously, we loose some information. PCA and other techniques tries to loose as less as possible.

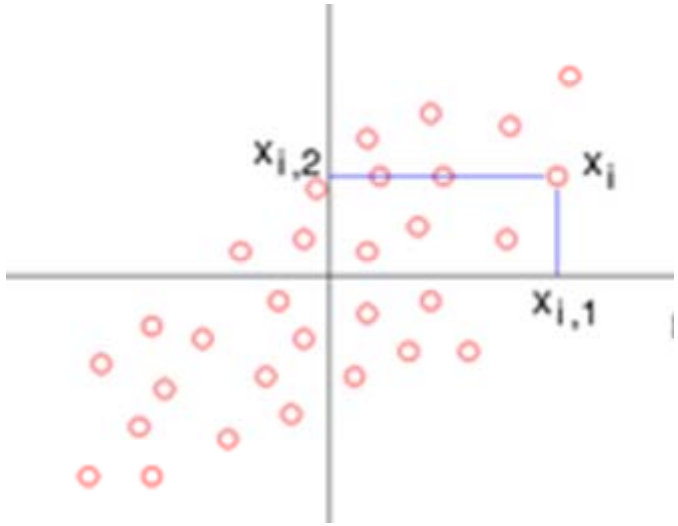
The “new” data is constructed by considering the coordinates of the “old” data points in the new set of dimensions

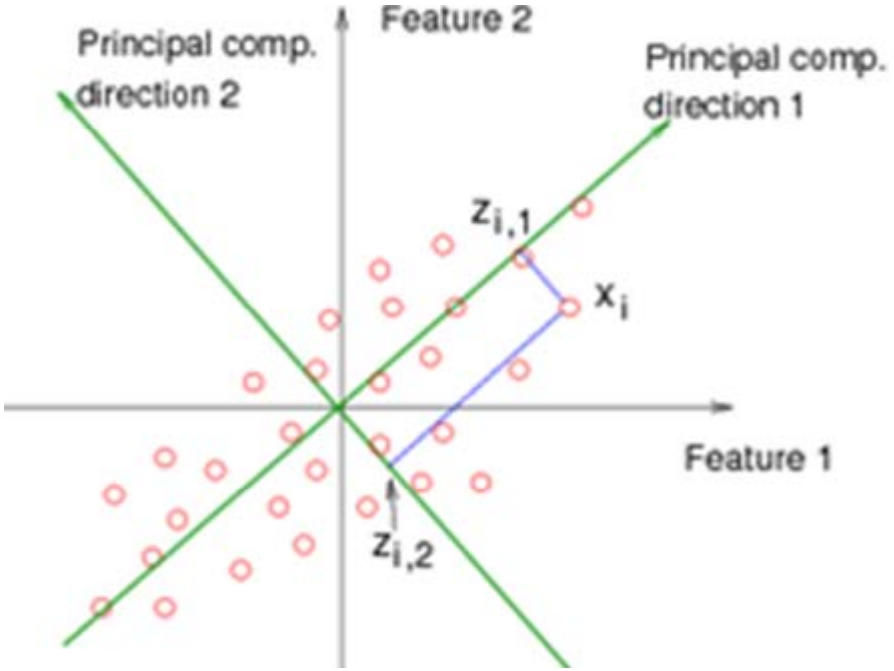


1-dimension

Graphical Approaches

2-dimensions





The original data was two-dimensional.

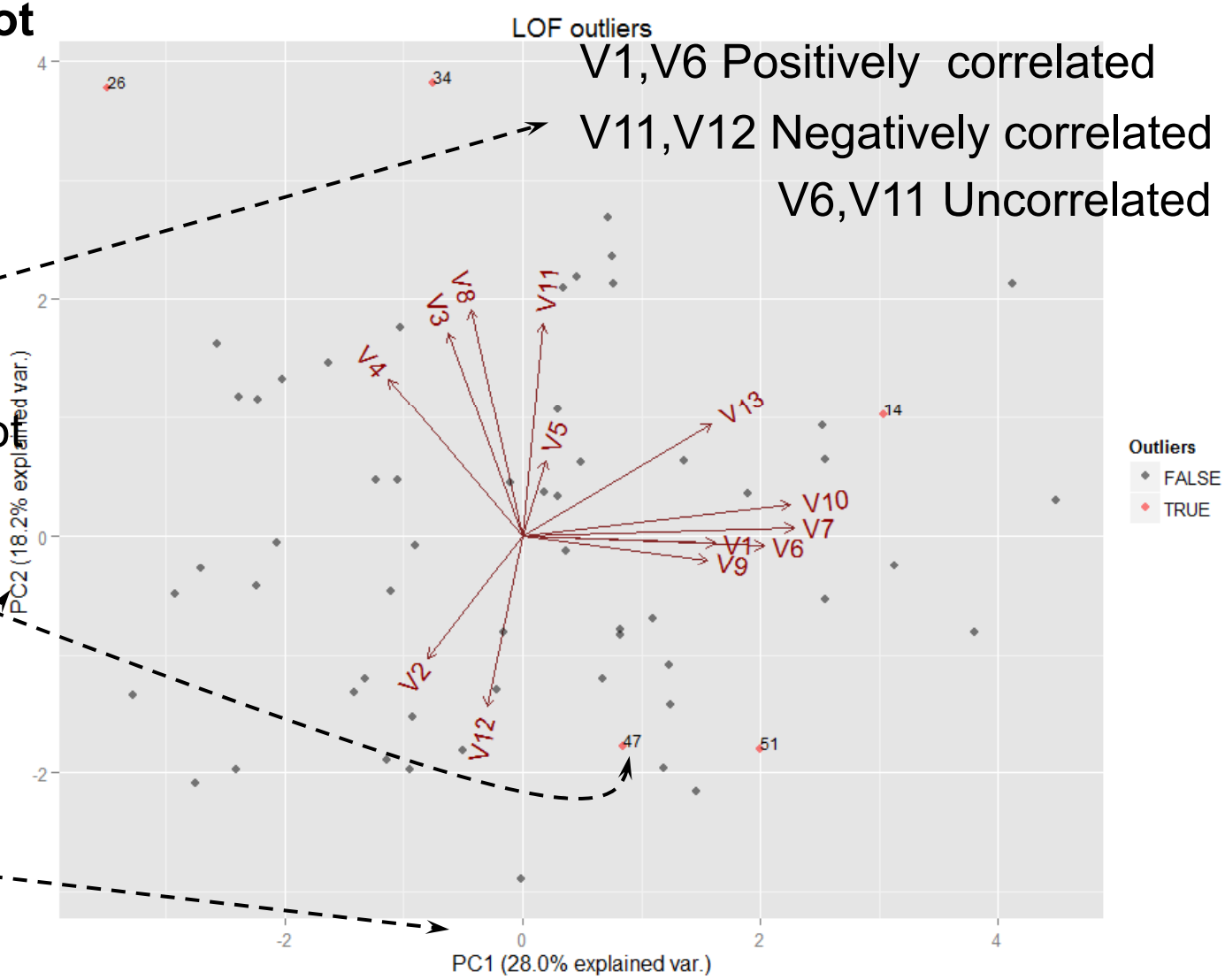
So, in the new two-dimensional space there's no information lost.

Graphical Approaches

In general, a **biplot** performs a PCA projection into 2 dimensions and add arrows for each original attribute.

Anyway, it may not reveal outliers because of the projections

Only 28+18.2=46.2% variance explained



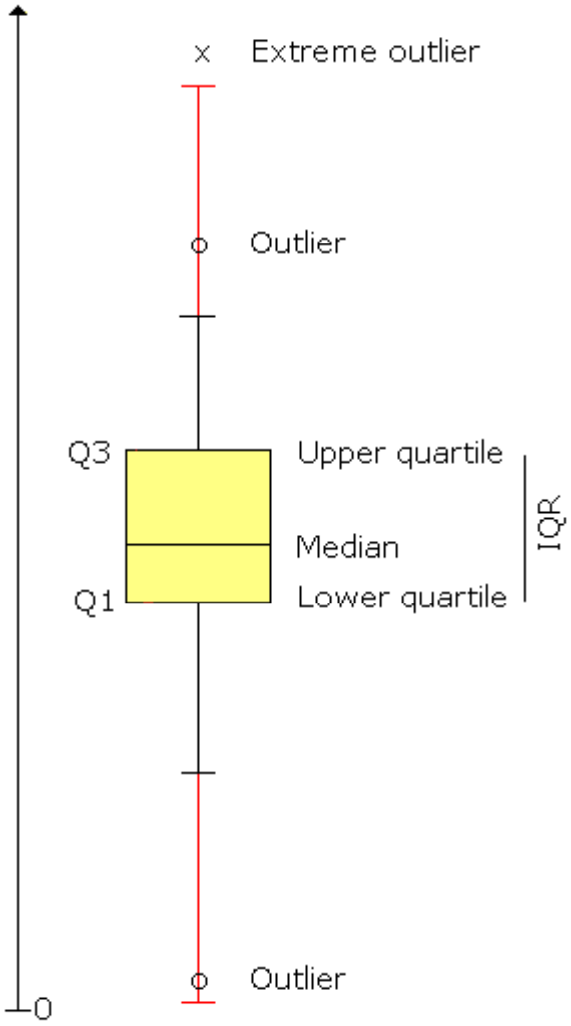
Statistical Approaches: Non Parametric

- This approach apply properties of the Normal distribution to non-Normal distributions.

  Only 1 dimension

IQR = Q3 - Q1

P is an Outlier if P > Q3 + 1.5 IQR
P is an Outlier if P < Q1 - 1.5 IQR
P is an Extreme Outlier if P > Q3 + 3 IQR
P is an Extreme Outlier if P < Q1 - 3 IQR



Statistical Approaches: Parametric

- Assume a parametric model describing the distribution of the data (e.g., normal distribution)

- Given a database D, and a data point $x \in D$, a statistical test determines whether x is an outlier or not, at a significance level p. The test depends on:

  - Data distribution
  - Parameter of distribution (e.g., mean, variance)
  - Number of expected outliers

Statistical Approaches: Parametric

- 1-variate Normal distribution:

  One outlier: Grubb's test

  k-Outliers: Tietjen and Moore's test

  Less than k-Outliers: Rosner's test

- Multi-variate Normal distribution:

  Non Robust Methods: Mahalanobis distances

  Robust Methods: Mahalanobis distances with MCD and median instead of Covariance and mean.

Statistical Approaches: Parametric

Univariate data:

- Instead of asking if a particular x is an outlier, we wonder whether the random variable has outliers or not.
- Grubb's test for **normal distribution** (R package: "outliers")
  - H_0: There are no outliers in data
  - H_A: There is **exactly one** outlier

Grubbs' test statistic: $$G = \frac{\max_{i=1..N} \left| X_i - \overline{X} \right|}{S}$$

The outlier itself participates in computing the mean and the standard deviation *S*

The Grubbs test statistic is the largest absolute deviation from the sample mean in units of the sample standard deviation. This is the two sided version. There are also one sided versions for the maximum (minimum).

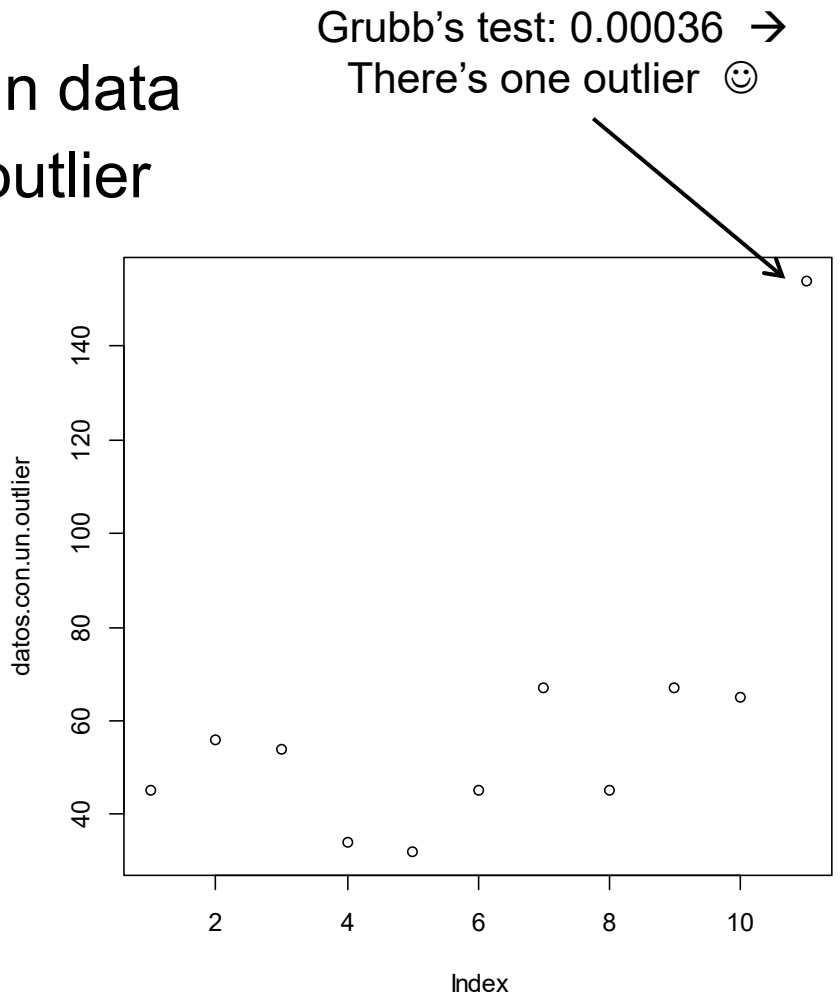
Reject H_0 if: $$G > \frac{(N-1)}{\sqrt{N}} \sqrt{\frac{t^2_{(\alpha/(2N),N-2)}}{N-2+t^2_{(\alpha/(2N),N-2)}}}$$

http://www.graphpad.com/quickcalcs/Grubbs1.cfm

Statistical Approaches: Parametric

Grubb's test for

- H_0: There are no outliers in data
- H_A: There is exactly one outlier



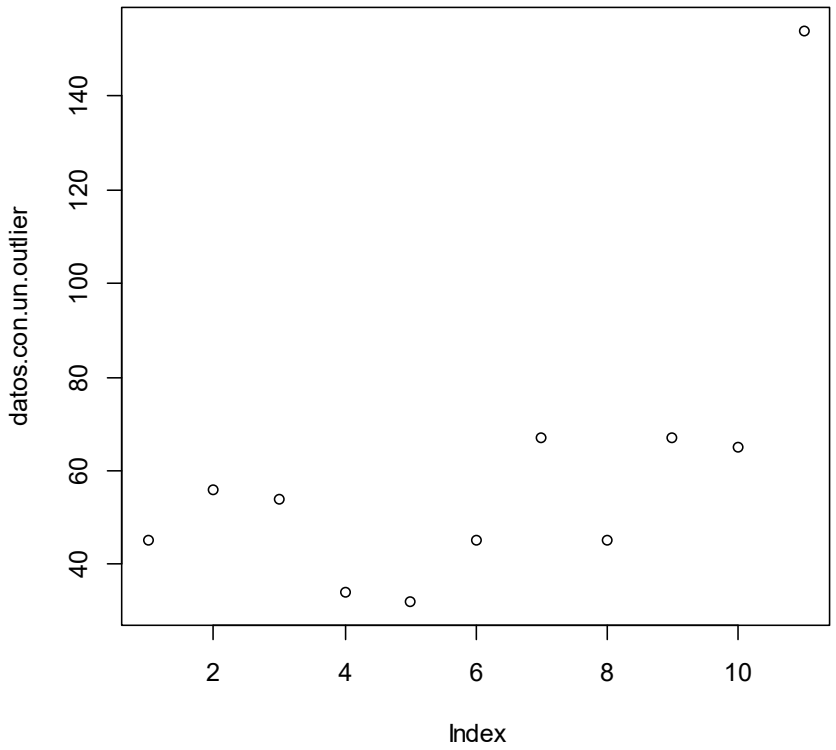
Statistical Approaches: Parametric

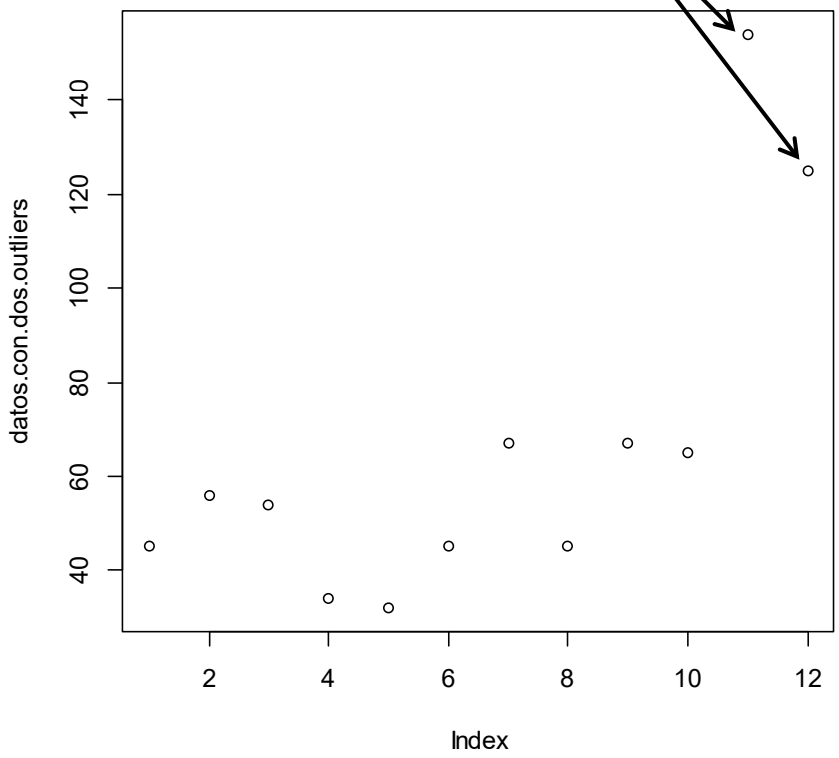
- Grubb's test is for a ***single*** outlier in 1-variate Normal distribution (there are other proposals as Dixon's test)
- What to do if there are ***multiple outliers***, i.e, when there's more than a single outlier?
- Detect one outlier at a time by applying a statistical test, remove the outlier, and repeat the test. Problem: ***Masking***
- Masking appears when a test for a single outlier fails to detect it because there is another outlier which *hides* the first one.

Statistical Approaches: Parametric

Masking

These points skews the mean and S toward them, i.e, they become too large: So, these points are not too far from the mean.





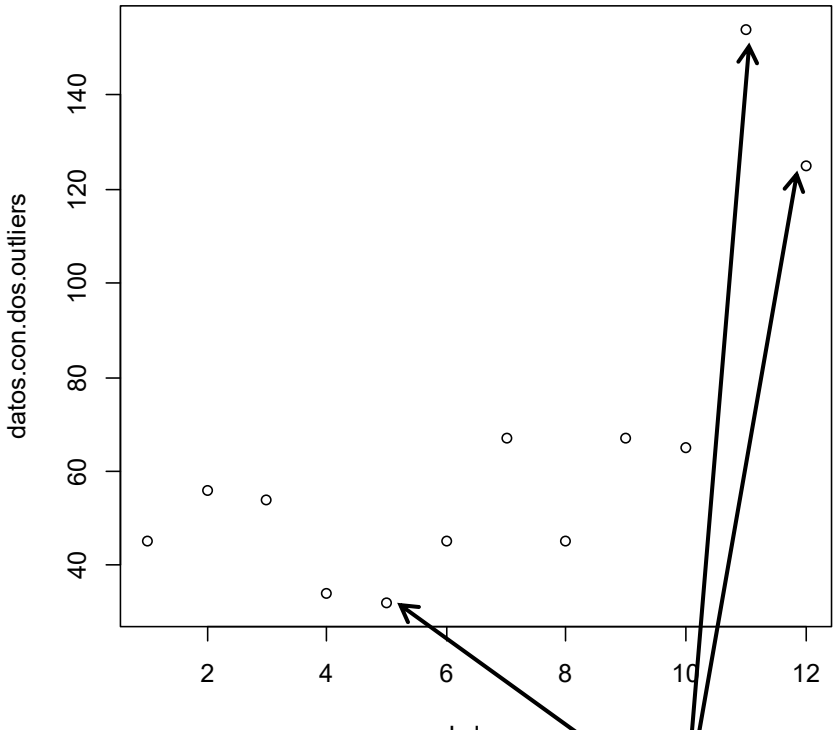
Grubb's test: 0.00036 → There's one outlier ☺

Grubb's test: 0.056 → Grubb's test fail to detect any outlier, when in fact there are two ☹

Statistical Approaches: Parametric

- Another possibility: Apply a test to check if there are **k** outliers.

- Tietjen and Moore's test:
  - H_0: There are no outliers in the data
  - H_A: There are **exactly k** outliers
- Problem: ***Swamping***.
- Swamping appears when a test for k outliers declares there are k outliers when in fact the number of outliers is lower.



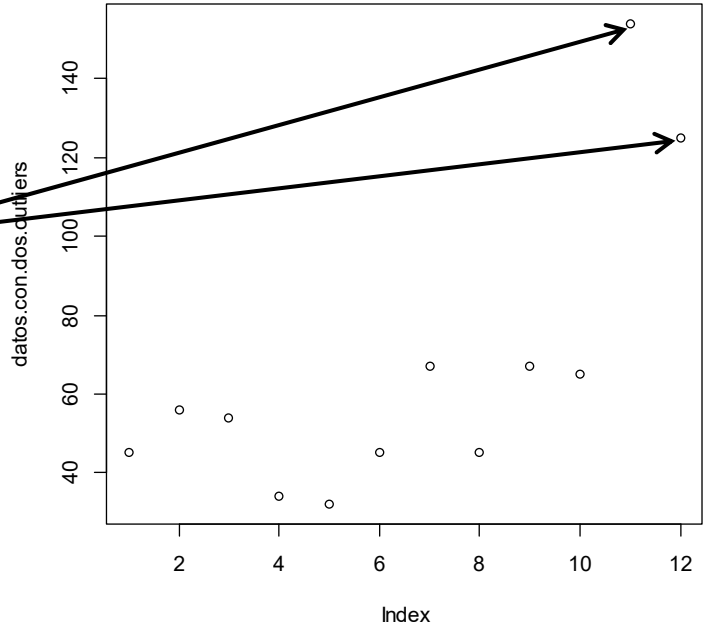
If we apply the test for k=3, it is significative because the block of the three most extreme values has an extreme statistic value, due to the two real outliers ☹

Statistical Approaches: Parametric

- Another possibility: Apply a test to check if there are **less than k** outliers.

- Rosner's test (R package: "EnvStats"):
  - H_0: There are no outliers in the data
  - H_A: There are **less than k** outliers

Rosner's test with k=3 only declares two outliers ☺



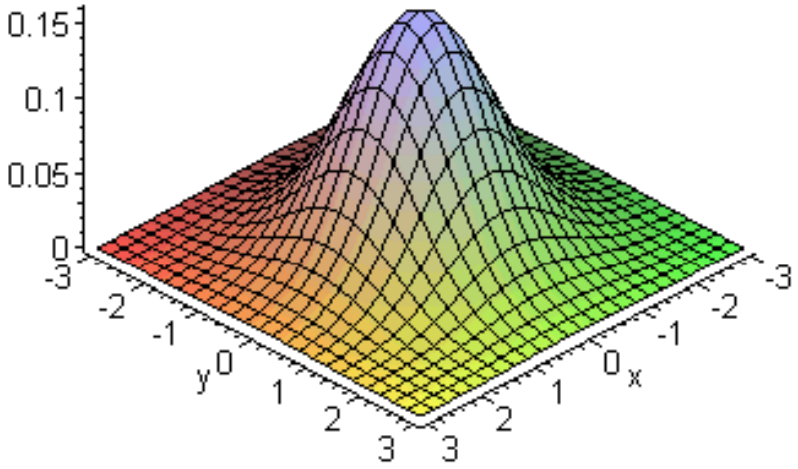
This test performs multiple comparisons, and therefore applies a correction to control FWER

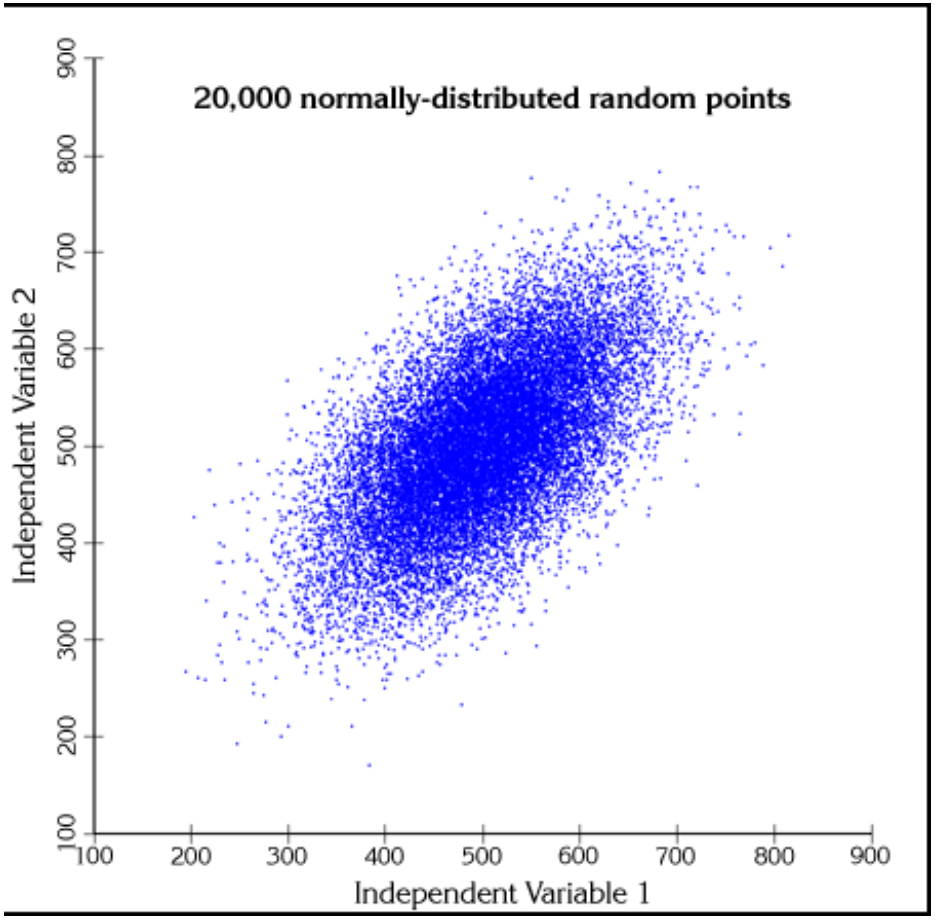
Statistical Approaches: Parametric

- Working with several *(p)* dimensions

$$N(x) = \frac{1}{\sqrt{(2\pi)^p \, |\Sigma|}} e^{-\frac{(x-\mu)^{\mathbf{T}} \boldsymbol{\Sigma}^{-1} (x-\mu)}{2}}$$

Bivariate Normal



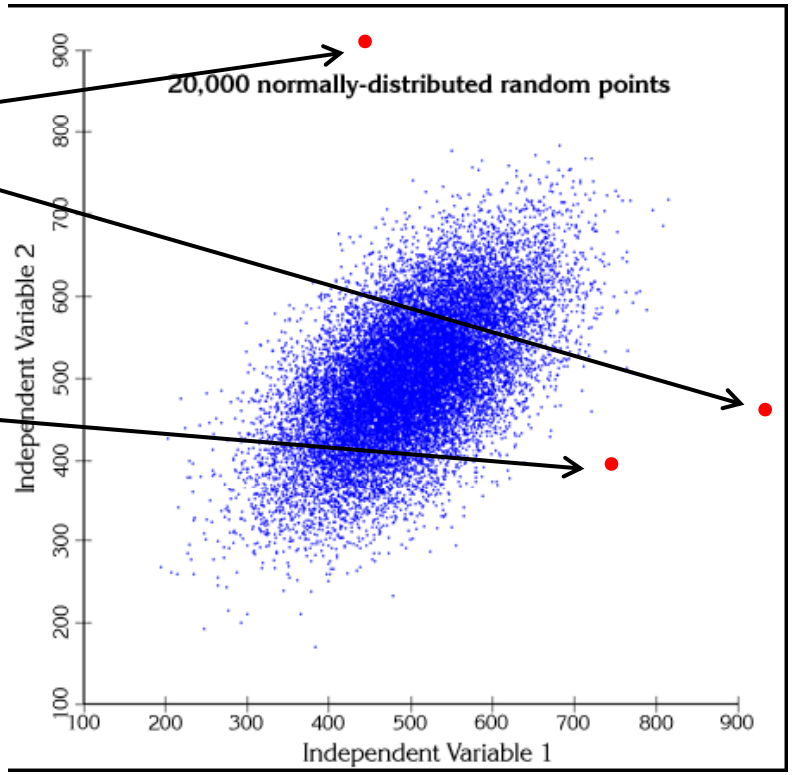


Statistical Approaches: Parametric

What's an outlier in *p*-dimensions?

- A data with an extreme value in some attribute(s)

- A data with an abnormal combination of (common) attribute values



Statistical Approaches: Parametric

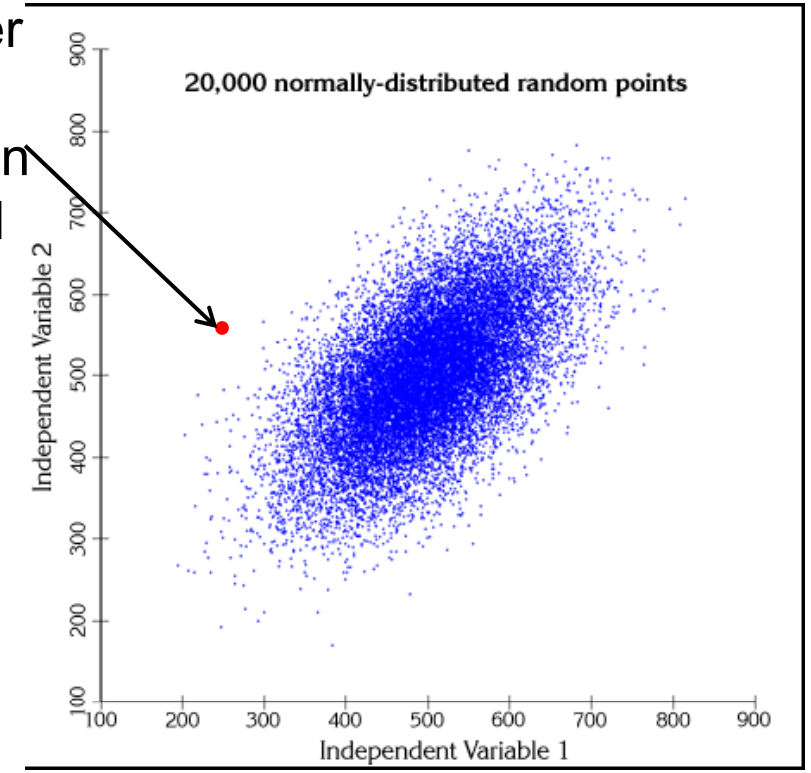
In the Normal distribution we can detect abnormal values in one attribute and abnormal combinations of attribute values by taking into account distance to the mean, variance and covariance

$$s_{xy} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{n-1}$$

$$\mathbf{S} = (s_{jk}) = \begin{pmatrix} s_{11} & s_{12} & \cdots & s_{1p} \\ s_{21} & s_{22} & \cdots & s_{2p} \\ \vdots & \vdots & & \vdots \\ s_{p1} & s_{p2} & \cdots & s_{pp} \end{pmatrix}$$

p attributes

The outlier makes *s* higher than expected



Statistical Approaches: Parametric

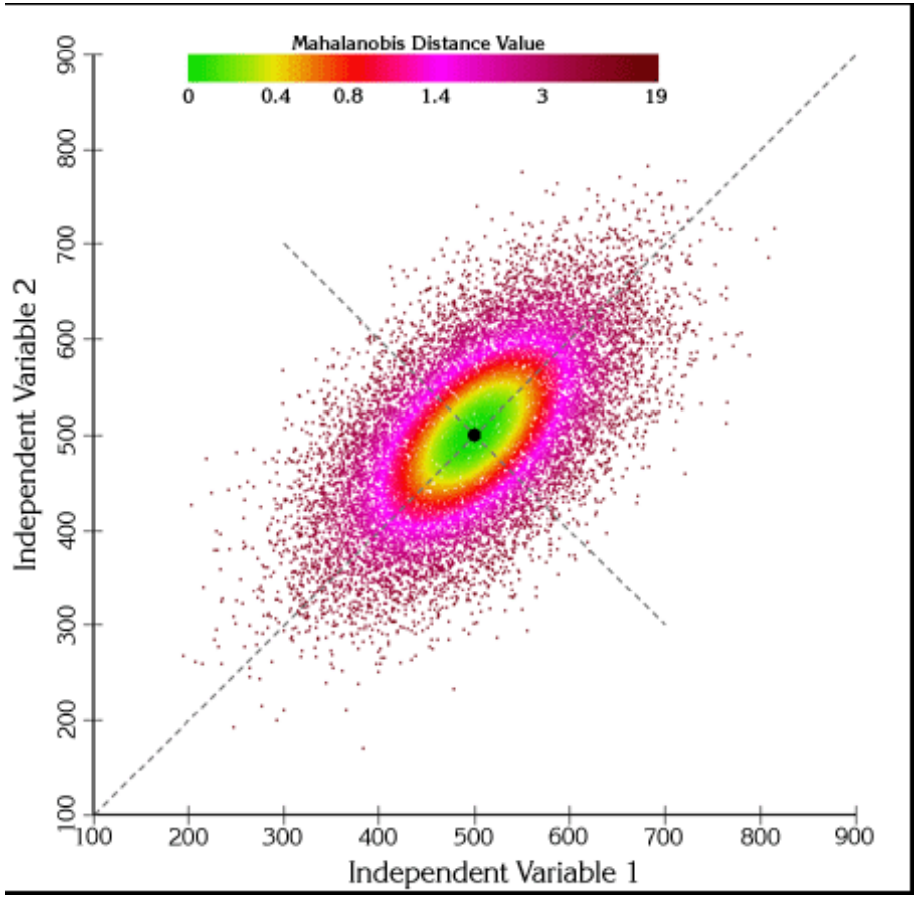
Mahalanobis distance of point *x* **to a Normal sample distribution**

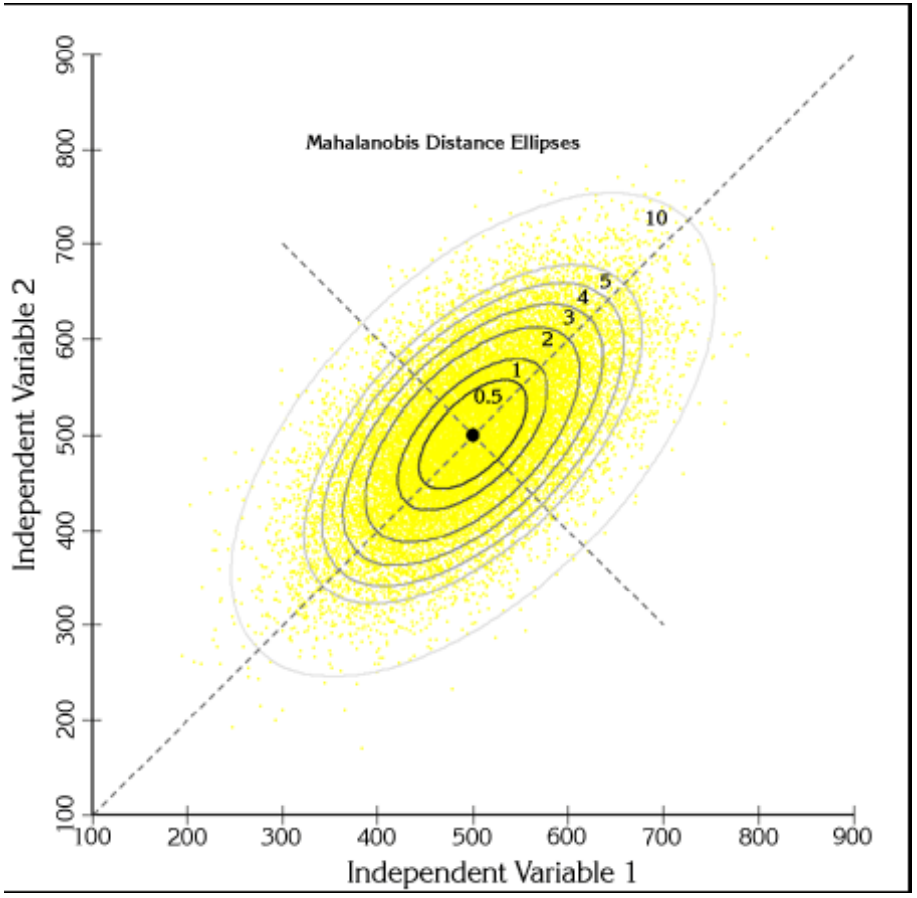
$$d_{S,\bar{x}}(x_i) = \sqrt{(x_i - \bar{x})^{\mathbf{T}} S^{-1} (x_i - \bar{x})}$$

$$d^2{}_{S,\bar{x}}(x_i) = (x_i - \bar{x})^{\mathbf{T}} S^{-1} (x_i - \bar{x})$$

Mahalanobis distance is a multidimensional version of a z-score (Mah.dist. is applied to the data without previously been normalized). It measures the distance of a case to the centroid of a normal distribution.

Statistical Approaches: Parametric





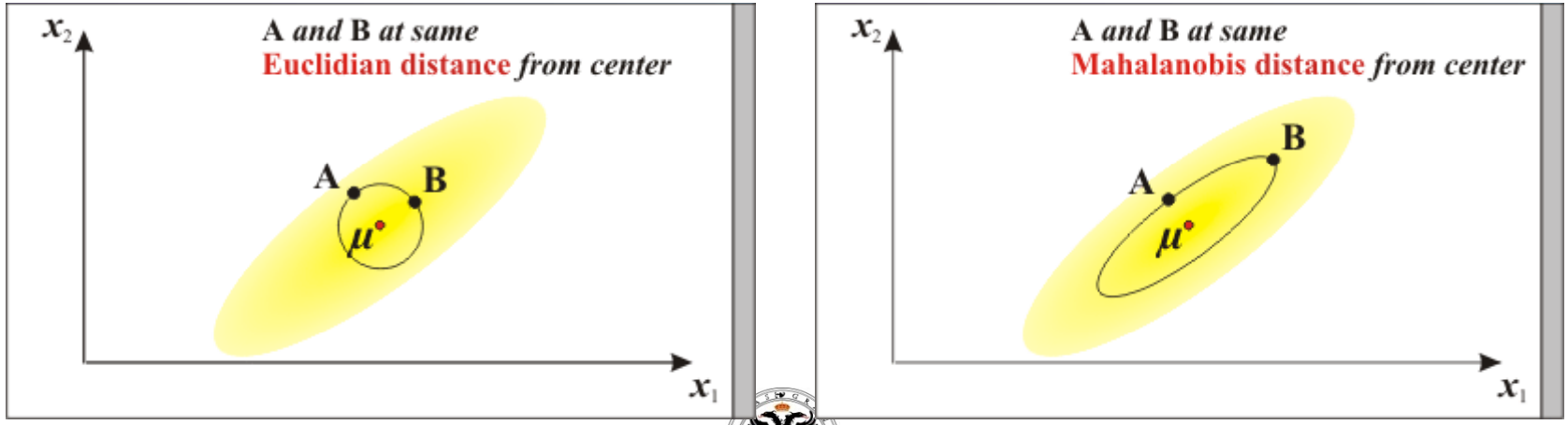
Statistical Approaches: Parametric

Formally, the Mahalanobis distance measures the distance of a case to the multidimensional mean μ (centroid) of a distribution, given the covariance Σ (multidimensional variance) of the distribution

$$N(x) = \frac{1}{\sqrt{(2\pi)^p \,|\Sigma|}} e^{-\frac{(x-\mu)^{\mathbf{T}}\boldsymbol{\Sigma}^{-1}(x-\mu)}{2}}$$

$$d_{\boldsymbol{\Sigma},\mu}(x) = \sqrt{(x-\mu)^{\mathbf{T}}\boldsymbol{\Sigma}^{-1}(x-\mu)}$$

$$d^2{}_{\boldsymbol{\Sigma},\mu}(x) = (x-\mu)^{\mathbf{T}}\boldsymbol{\Sigma}^{-1}(x-\mu)$$



Statistical Approaches: Parametric

Theoretical Mahalanobis distances distribution $d^2_{\Sigma,\mu}(x_i) \sim \chi^2_p$

The parameters are unknown, so we work with their estimators.

$d^2_{S,\bar{x}}(x_i) = (x_i - \bar{x})^{\mathbf{T}} S^{-1} (x_i - \bar{x})$ is an estimator of $d^2_{\Sigma,\mu}(x_i)$

Exact Sample Mahalanobis distances distribution

$$\left(\frac{(N-1)^2}{N}\right) d^2_{S,\bar{x}}(x_i) \sim Beta\left(\frac{p}{2}, \frac{(N-p-1)}{2}\right)$$

Approximate Sample Mahalanobis distances distribution

$$d^2_{S,\bar{x}}(x_i) \approx \chi^2_p \qquad \left(\frac{Np}{N-p}\right) d^2_{S,\bar{x}}(x_i) \approx F_{p,N-p}$$

Statistical Approaches: Parametric

Which test should be considered?

If we are interested on testing if there is exactly one outlier, a single test should be considered:

$$H_0 = x_{highest} \sim N(\mu, \Sigma)$$

Where $x_{highest}$ is the data value with highest Mah. distance

If we are interested on testing whether there are outliers, a multiple comparison should be performed on the dataset with size *N*

$$H_{0i} = x_i \sim N(\mu, \Sigma) \quad i = 1 \ldots N$$

A correction is used to control the FWER error

Statistical Approaches: Parametric

Which α should be considered?

Traditionally α = 0.025 was considered.
But we must take into account FWER error.

$$H_{0i} = x_i \sim N(\mu, \Sigma) \quad i = 1...N$$

Remember: When testing multiple hypotheses (*N* in this case) in the same experiment:
Prob(Making one or more type I error) = $1-(1-\alpha)^{1/N}$

So, any method to control FWER error rate like Bonferroni correction or Holm's procedure should be applied.
Traditionally, Sidak correction is used:

$$\alpha_N = 1-(1-\alpha)^{1/N}\,(Sidak) \cong (\geq)\alpha / N\,(Bonferroni)$$

Statistical Approaches: Parametric

In summary. To check which values are outliers in a sample of size *N* of a Multivariate Normal Distribution, the following tests have to be considered:

$$H_{0i} = x_i \sim N(\mu, \Sigma) \quad i = 1...N$$

Fixing a typical type I error $\alpha = 0.05$ for instance, the following correction is used to control the FWER error

$$\alpha_N = 1 - (1-\alpha)^{1/N} \qquad \alpha = 0.05 \rightarrow \alpha_{1000} \approx 0.00005!$$

The following approximations can be used (*F* being better)

$$Label\ x_i\ as\ outlier\ if\ d^2_{S,\bar{x}}(x_i) \geq \chi^2_p(1-\alpha_N)$$

$$Label\ x_i\ as\ outlier\ if\ \left(\frac{Np}{N-p}\right) d^2_{S,\bar{x}}(x_i) \geq F_{p,N-p}(1-\alpha_N)$$

Statistical Approaches: Parametric

All the previous tests (1 or k)-variate require Normality
They are based on the sample mean and variance/covariances , which are not ***robust*** estimators:
Mean(1,2,3,4,5) = 3 Mean(1,2,3,4,200) = 42

The ***breakdown point*** of an estimator is the proportion of incorrect observations (e.g. arbitrarily large observations) an estimator can handle before giving an incorrect (e.g., arbitrarily large) result. It is a value between 0 and 0.5

Breakdown point of the mean: 0 (just one point may alter the result) → It's not a robust estimator.

Breakdown point of the median: 0.5 → It's a robust estimator.

Breakdown point of the X% trimmed mean (X% of the greatest and lowest values are discarded to compute the mean): 0.X

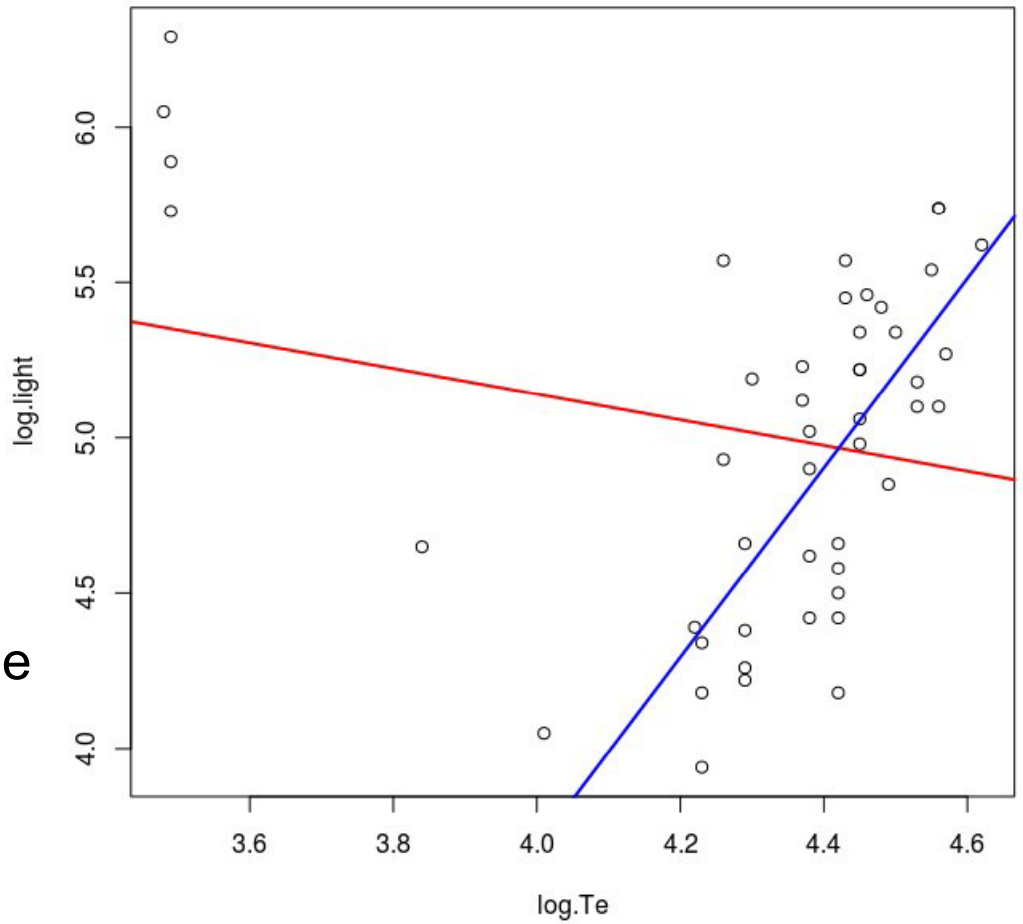
Statistical Approaches: Parametric

Example of how outliers could affect a Linear Regression (using non robust estimators -red- and robust ones -blue-)

DataSet: starsCYG

log.Te: Logarithm of the effective temperature at the surface of the star (Te).

log.light: Logarithm of its light intensity (*L/L_0*)

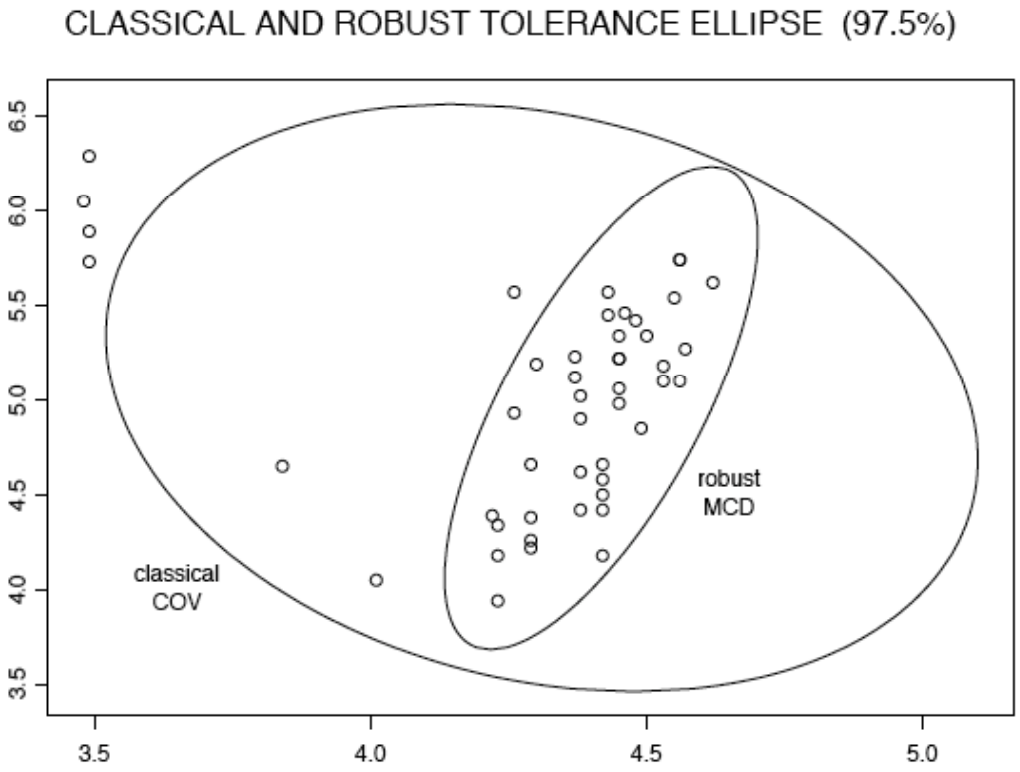


Statistical Approaches: Parametric

In the multivariate case robust estimators S^* of the covariance matrix are used → MCD (minimum covariance determinant) is an iterative (similar to k-means) method introduced by Rousseeuw, (84,99)

The robust estimator MCD of the Cov.Matrix is the sample covariance of the *good* points

The robust estimator of the mean is the sample mean of those points included in the computation of MCD



Statistical Approaches: Parametric

Squared Mahalanobis distance estimator:

$$d^2_{S,\bar{x}}(x_i) = (x_i - \bar{x})^{\mathbf{T}} S^{-1} (x_i - \bar{x})$$

Robust (MCD) Squared Mahalanobis distance estimator:

$$d^2_{S^*,\bar{x}^*}(x_i) = (x_i - \bar{x}^*)^{\mathbf{T}} S^{*-1} (x_i - \bar{x}^*)$$

The exact distribution of the robust distance estimator is unknown. Hardin and Rocked (2005) showed that:

- The ChiSq approximation is well suited for the points used in the computation of *S**
- The *F* approximation (properly adjusted to take into account the size of the sample used to estimate Σ) is better suited for points which are independent of those used to compute *S** (this is the case of outliers)

Statistical Approaches: Parametric

Peter J. Rousseeuw and Mia Hubert. Robust statistics for outlier detection. 2011. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, Volume 1, Issue 1, pages 73–79, January/February 2011

Hubert, M. and Debruyne, M. (2010), Minimum covariance determinant. WIREs Comp Stat, 2: 36–43

Hardin, J., Rocke, D.M., 2005. The distribution of robust distances. Journal of Computational and Graphical Statistics 14

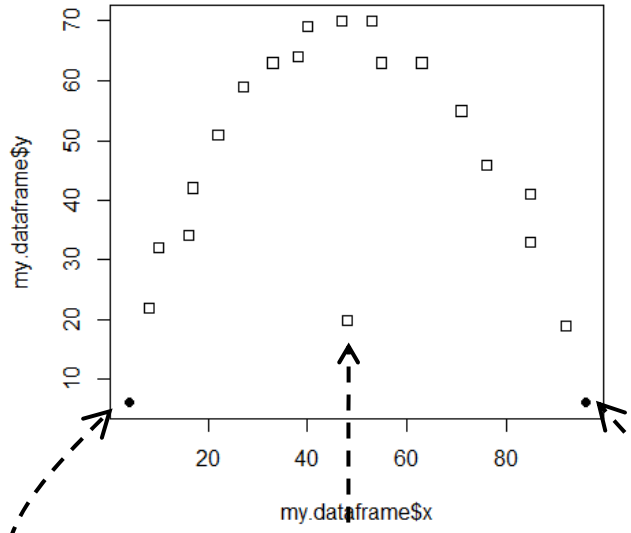
Cerioli, Andrea, (2010), Multivariate Outlier Detection With High-Breakdown Estimators, Journal of the American Statistical Association, 105, issue 489, p. 147-156.

Andrea Cerioli, Alessio Farcomeni, Error rates for multivariate outlier detection, Computational Statistics & Data Analysis, Volume 55, Issue 1, 1 January 2011, Pages 544-553

Statistical Approaches: Parametric

Limitations when using Mahalanobis distance estimators.

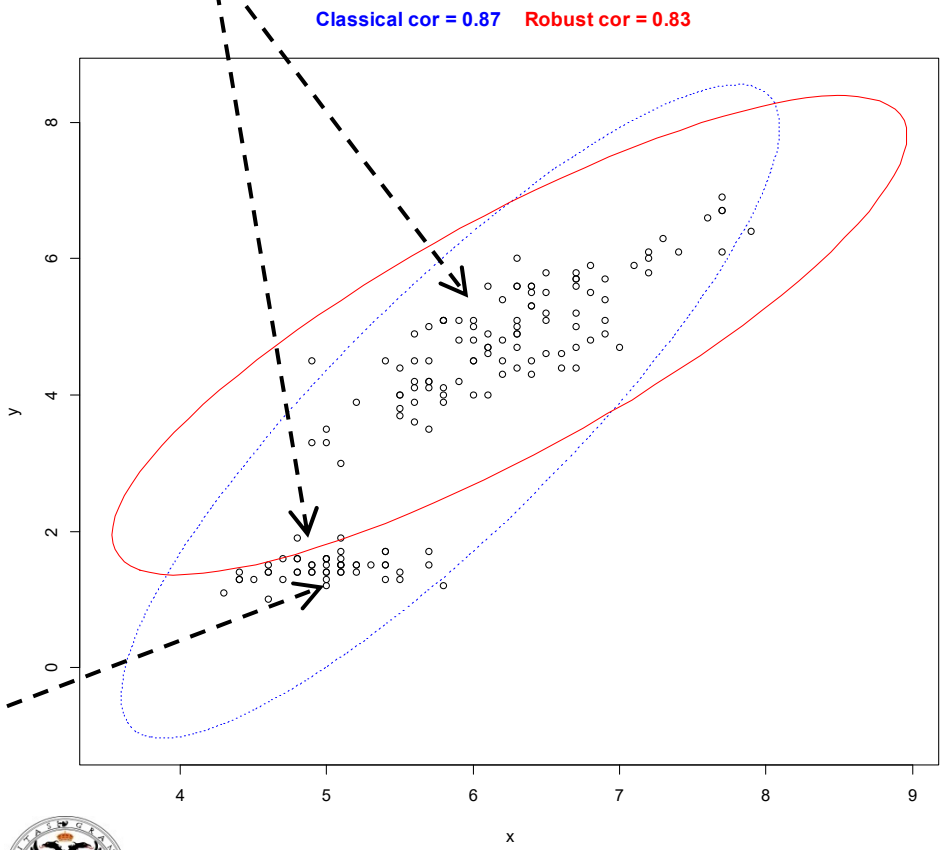
Non linear relations



Labelled as not outlier ☹

Labelled as outliers ☹

Mixture of Normal Distributions



Data Mining: Anomaly Detection

- Motivation and Introduction
- Supervised Methods
- Semisupervised Methods
- Unsupervised Methods:
  - Graphical and Statistical approaches
  - Nearest neighbor based approaches
  - Clustering based approaches
- Evaluation

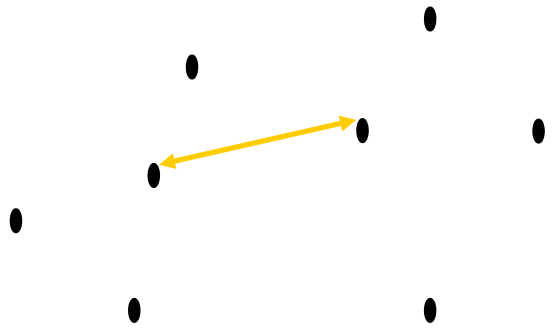
Unsupervised: Nearest neighbor

Limitations of statistical approaches:

- In many cases, data distribution is not normal or it may not be known
- High dimensional data does not usually follow a specific multivariate distribution

Unsupervised: Nearest neighbor

- **Data is represented as a vector of features.**
  We have a distance measure to evaluate nearness between two points
- **Distance-based methods:**
  Given a database D, and a data point **x** $\in$ D, the method assigns an anomaly score to **x,** based on the distance of **x** to the other points



Unsupervised: Nearest neighbor

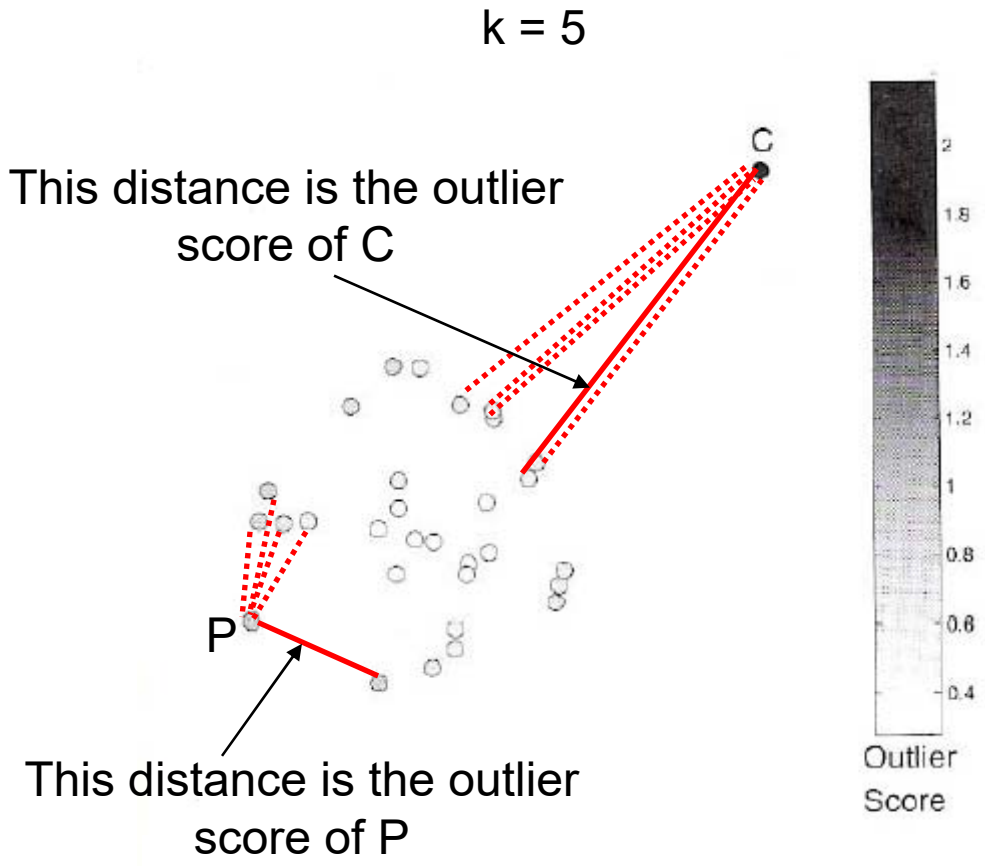
- Nearest neighbor approaches are score-based:
  Given a database D, and a data point **x** $\in$ D, the method assigns an anomaly score to **x**
  - Given a database D, find all the data points **x** $\in$ D with anomaly scores **greater than** some threshold t
  - Given a database D, find all the data points **x** $\in$ D having the **top-n** largest anomaly scores f(**x**)
  - Given a database D, containing mostly normal (but unlabeled) data points, and a **test point x**, compute the anomaly score of **x** with respect to D

- Two major approaches
  - Nearest-neighbor based
  - Nearest-neighbor density based

Unsupervised: Nearest neighbor

- Approach:
  - Compute the distance (proximity) between every pair of data points
  - Fix a magic number k representing the k-th nearest point to another point
  - For a given point P, compute its *outlier score* as the distance of P to its k-nearest neighbor.
    There are no clusters. Neighbor refers to a point
  - Consider as outliers those points with *high* outlier score.

Unsupervised: Nearest neighbor



Unsupervised: Nearest neighbor

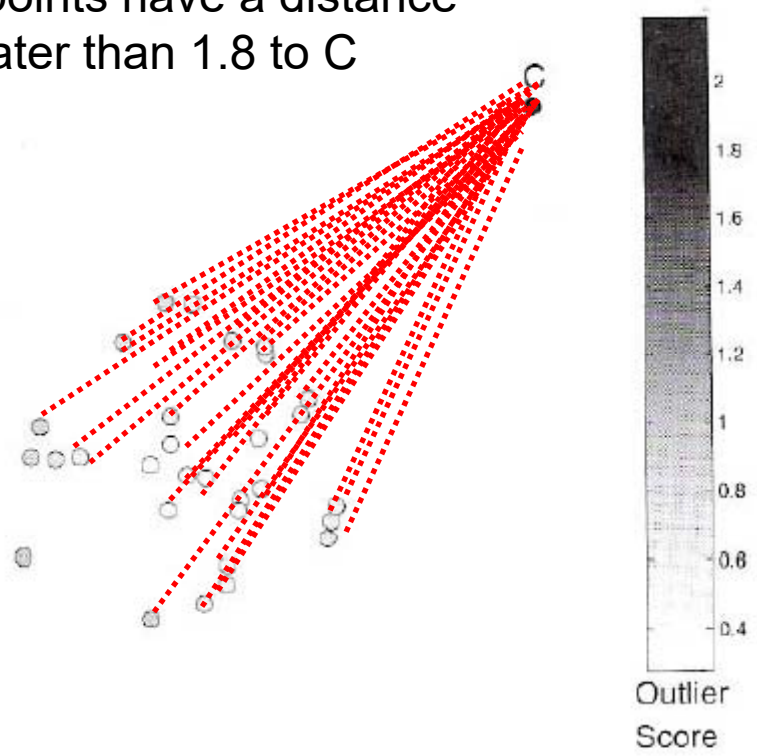
Edwin M. Knorr, Raymond T. Ng, and Vladimir Tucakov. **2000**. Distance-based outliers: algorithms and applications. *The VLDB Journal* 8, 3-4 (February 2000), 237-253. DOI=10.1007/s007780050006 http://dx.doi.org/10.1007/s007780050006

- Similar Approach:
  - Instead of fixing k, a distance D is fixed. Then, the method consider the percentage of points which are far away from the outlier.
  - An object O in a dataset T is a distance based DB(p,D) outlier if at least fraction p of the objects in T is located greater than distance D from O.

Unsupervised: Nearest neighbor

Knorr et al.

97% of points have a distance greater than 1.8 to C



Unsupervised: Nearest neighbor

- Computation requires comparing many distances.

  → $O(N^2)$

- Some improvements can be considered: Divide the data space in cells and use this spatial information to prune the search

  → O(dN) where d is the dimensionality and N is the size of the data

Unsupervised: Nearest neighbor

Knorr et al.

Application: Video Trajectory Surveillance

Trajectories are NOT represented in a 2D-position space.

Trajectories are summarized by the following features:

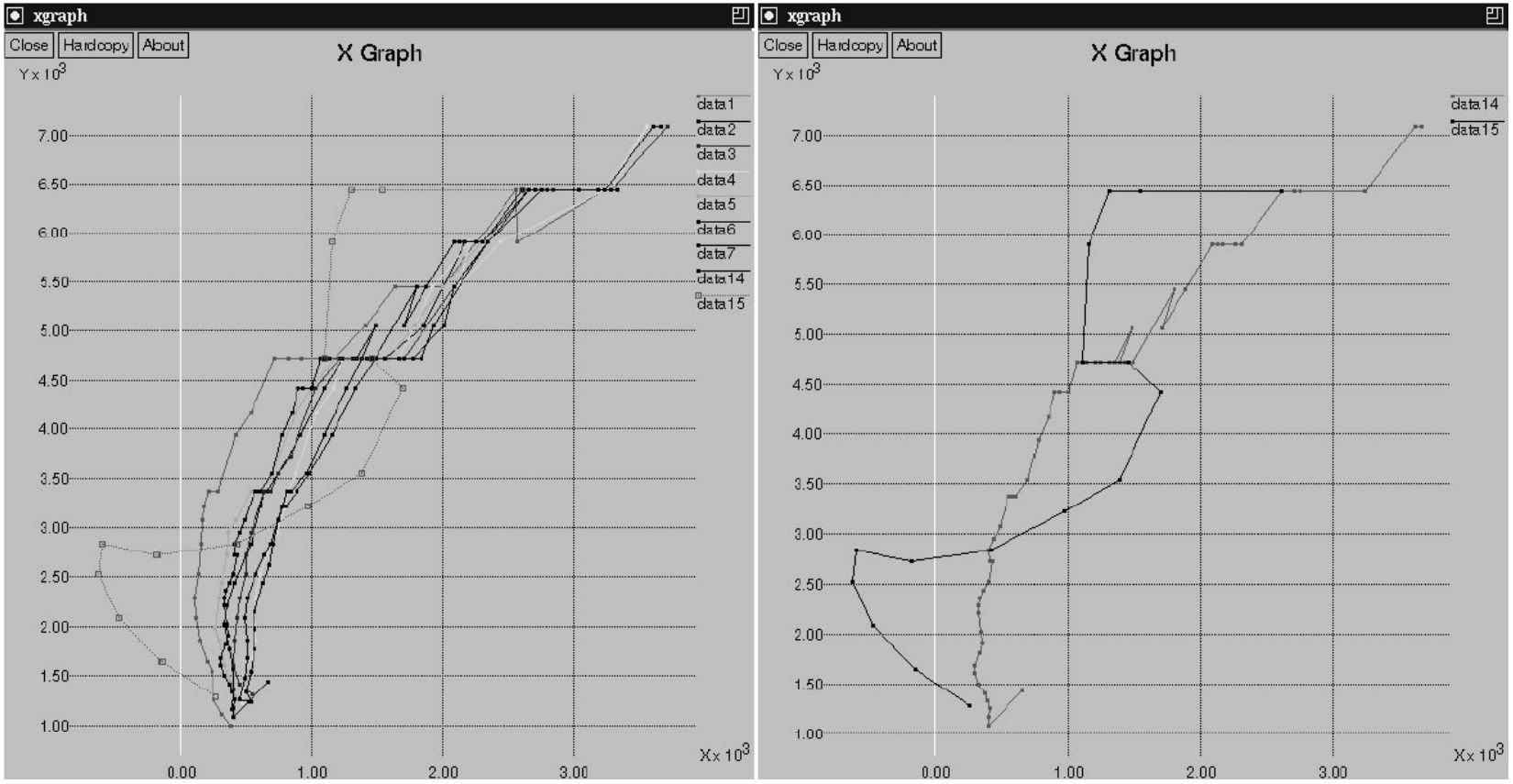
- Start and end points.
- Number of points: the length of the trajectory.
- Heading: the average, minimum, and maximum values of the directional vector of the tangent of the trajectory at each point.
- Velocity: average, minimum, and maximum velocity of the person during the trajectory.

An ad hoc distance measure is defined in this space

Unsupervised: Nearest neighbor

Knorr et al.

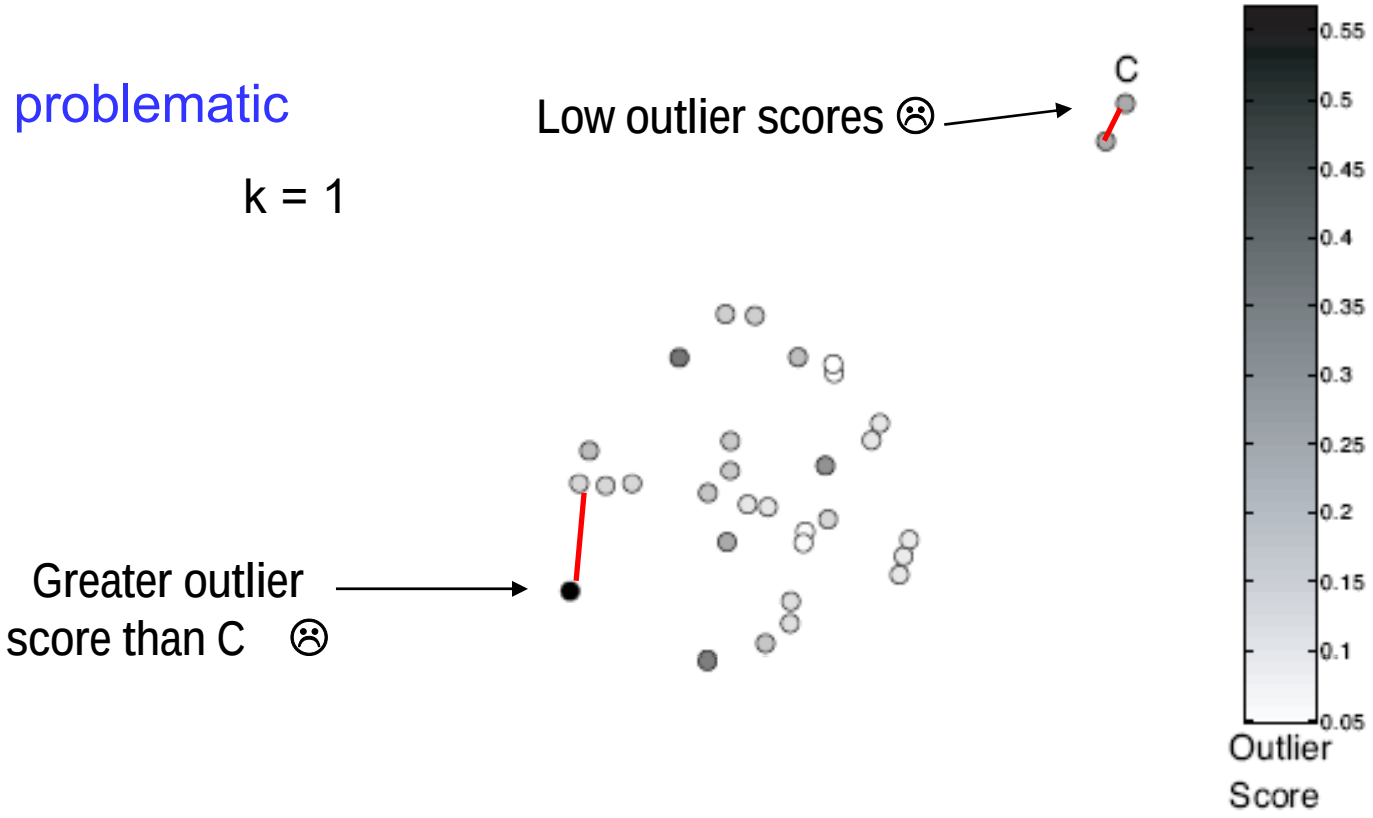
Video Trajectory Surveillance



Unsupervised: Nearest neighbor

Choice of k is problematic

k = 1

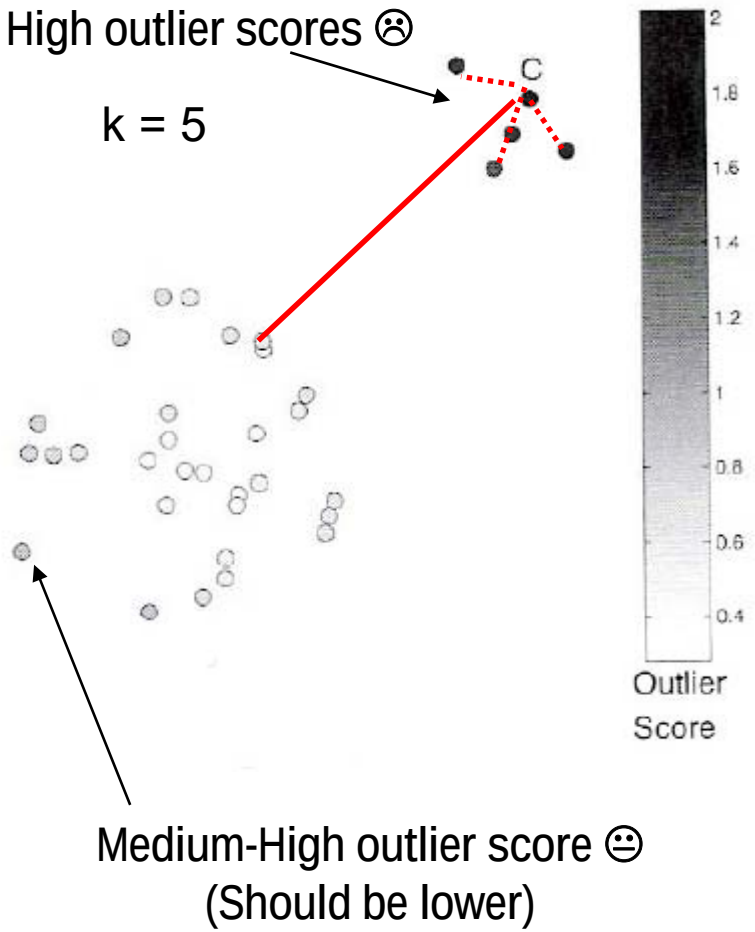


Unsupervised: Nearest neighbor

Choice of k is problematic

All the points in any isolated natural cluster with fewer points than k, have high outlier score

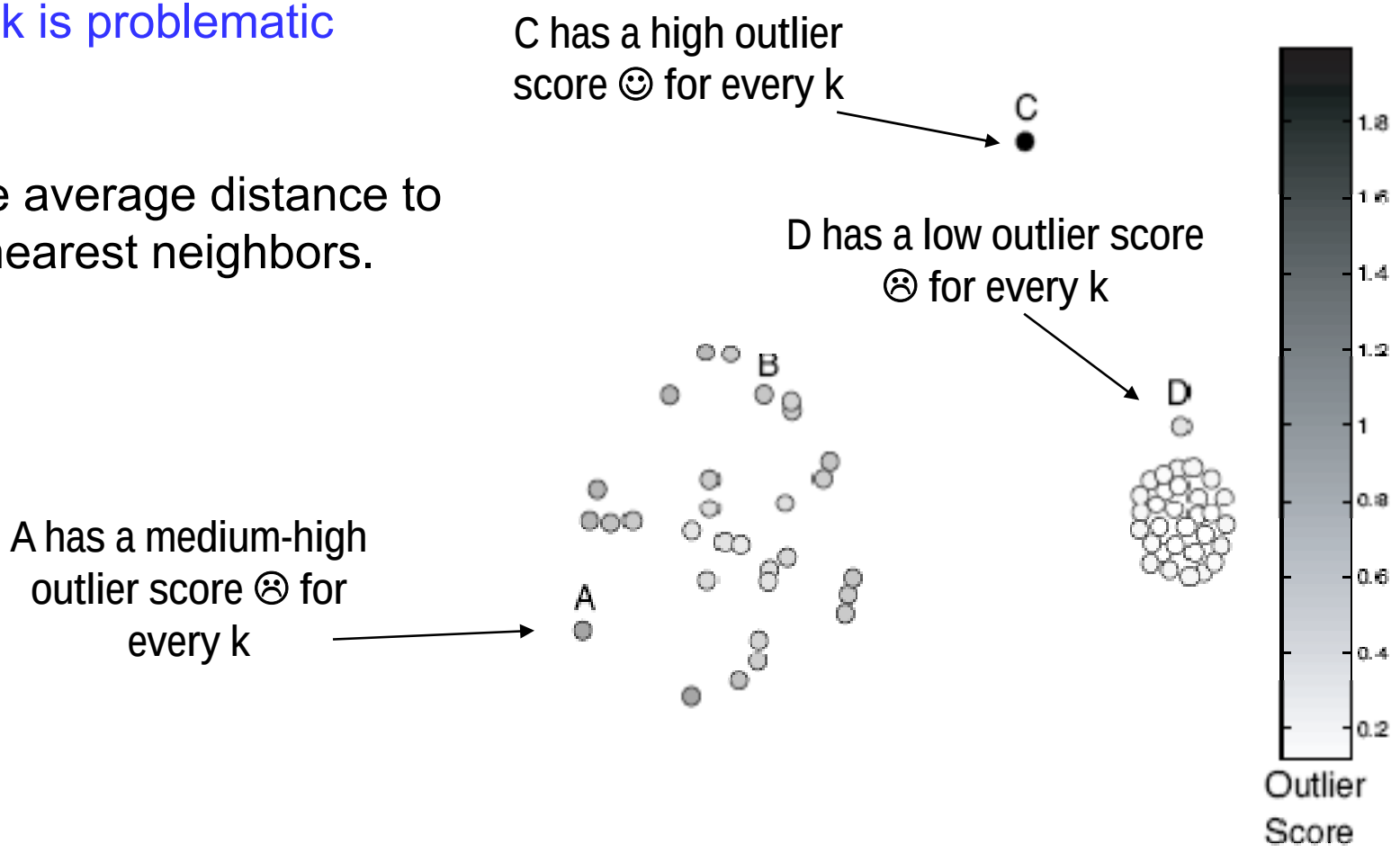
We could mitigate the problem by taking the average distance to the k-nearest neighbors but is still poor.



Unsupervised: Nearest neighbor

Choice of k is problematic

Taking the average distance to the k-nearest neighbors.



Unsupervised: Nearest neighbor

Markus M. Breunig, Hans-Peter Kriegel, Raymond T. Ng, and Jörg Sander. **2000**. **LOF**: identifying density-based local outliers. ***SIGMOD Rec.*** 29, 2 (May 2000), 93-104. DOI=10.1145/335191.335388

- Define the *k-density of a point as the inverse of the average sum of the distances to its k-nearest neighbors.*

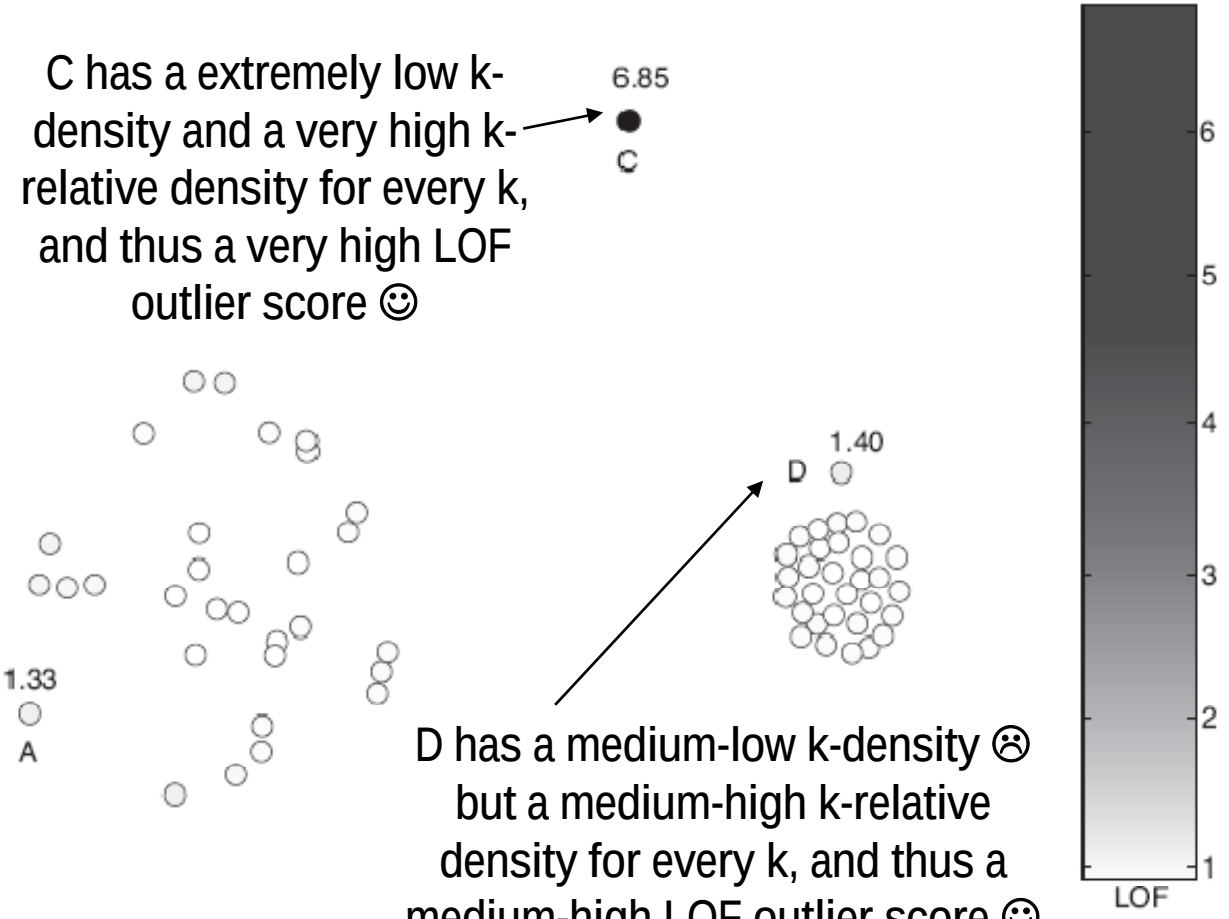
- Define the *k-relative density* of a point P as the ratio between its *k-density* and the average *k-densities* of its *k*-nearest neigbhors

- The outlier score of a point P (called LOF for this method) is its *k-relative density.* LOF is implemented in R.

Unsupervised: Nearest neighbor

Breunig et al (LOF)

C has a extremely low k-density and a very high k-relative density for every k, and thus a very high LOF outlier score ☺

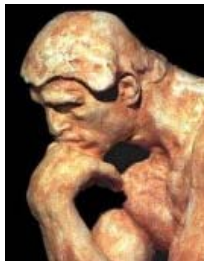


A has a very low k-density 😐 but a medium-low k-relative density for every k, and thus a medium-low LOF outlier score ☺

D has a medium-low k-density ☹ but a medium-high k-relative density for every k, and thus a medium-high LOF outlier score ☺

Unsupervised: Nearest neighbor

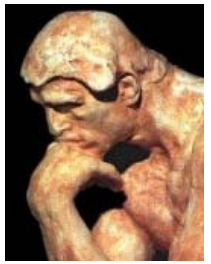
MINDS – MINnesota INtrusion Detection System (LOF based)



- Basic features of individual TCP connections
  - source & destination IP/port, protocol, number of bytes, duration, number of packets
- Time based features: detect fast scans -e.g: DoS attacks-
  - For the same source (destination) IP address, number of unique destination (source) IP addresses inside the network *in last T seconds*
  - Number of connections from source (destination) IP to the same destination (source) port *in last T seconds*
- Connection based features: detect slow scans
  - For the same source (destination) IP address, number of unique destination (source) IP addresses inside the network *in last N connections*
  - Number of connections from source (destination) IP to the same destination (source) port *in last N connections*

Unsupervised: Nearest neighbor

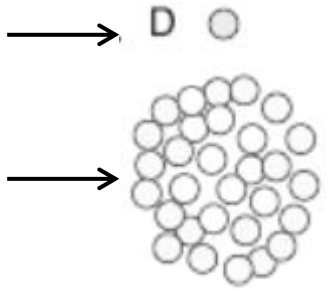
MINDS – MINnesota INtrusion Detection System (LOF based)



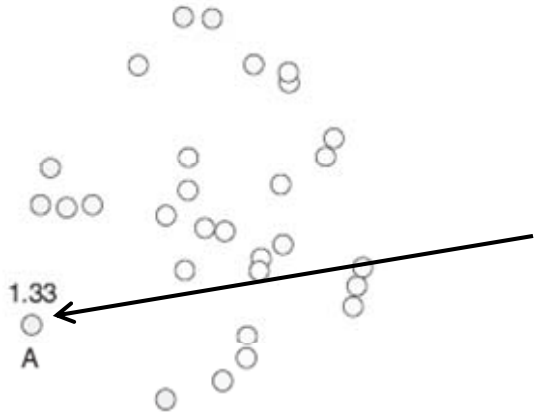
In order to avoid computation time, MINDS uses a sample of non-anomalous data entries and compare new entries with this sample (in a ”semisupervised way”)

Example: Slapper worn → Not detected with a simple distance based approach but detected with LOF:

Combination of source-destination port very rare. Detected as anomaly ☺



Similar scans →



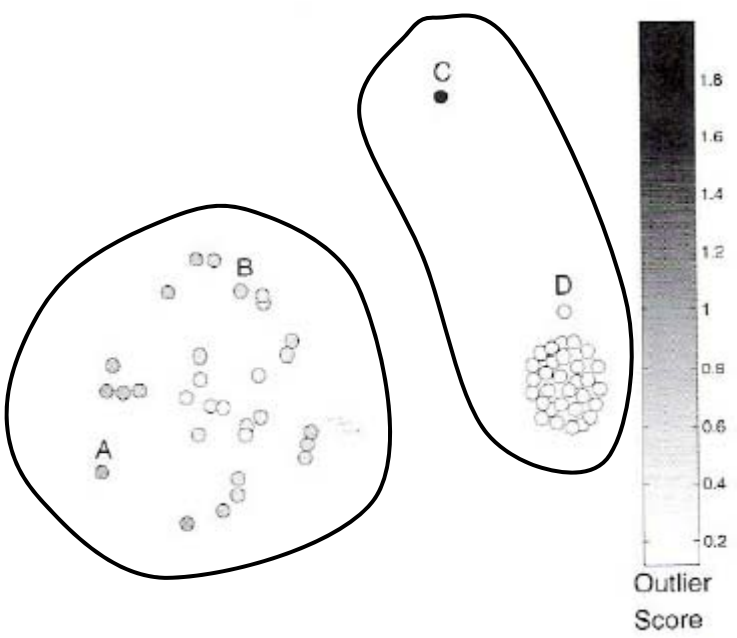
Some worms not detected ☹ : Portsweeps scans which are located in the sparse region of normal data.

Data Mining: Anomaly Detection

- Motivation and Introduction
- Supervised Methods
- Semisupervised Methods
- Unsupervised Methods:
  - Graphical and Statistical approaches
  - Nearest neighbor based approaches
  - Clustering based approaches
- Evaluation

Unsupervised: Clustering Based

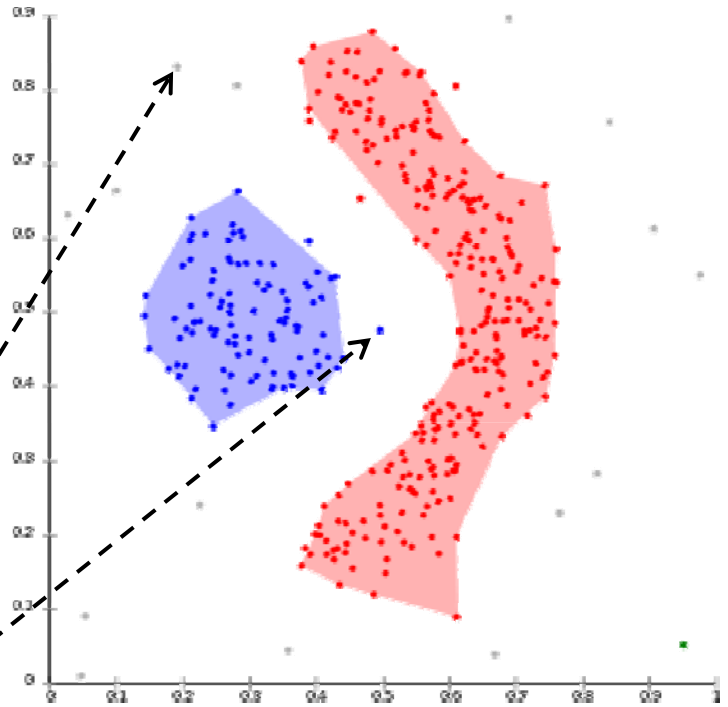
- Basic idea:
  - A set of clusters has already been constructed by any clustering method.
  - An existing object or a new one is compared to these clusters in order to determine if it is an anomaly



Unsupervised: Clustering Based

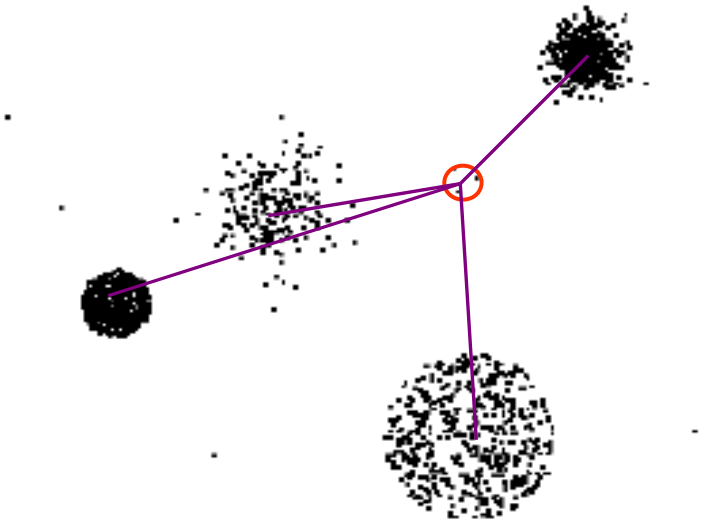
- Approaches
  - One possibility: An object is an anomaly if it doesn't belong to any cluster
  - Only for clustering methods which may not assign objects to clusters.Typically, density based algorithms as DBSCAN:

    → *Noise Points*



Unsupervised: Clustering Based

- Approaches
  - Another possibility: The anomaly score is given by the distance to its nearest centroid.

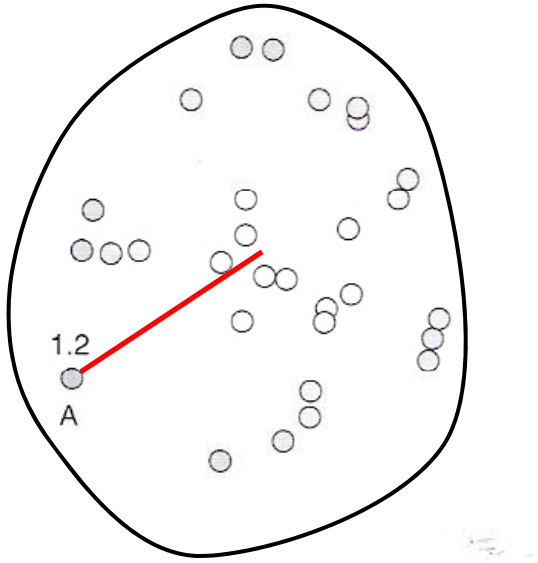


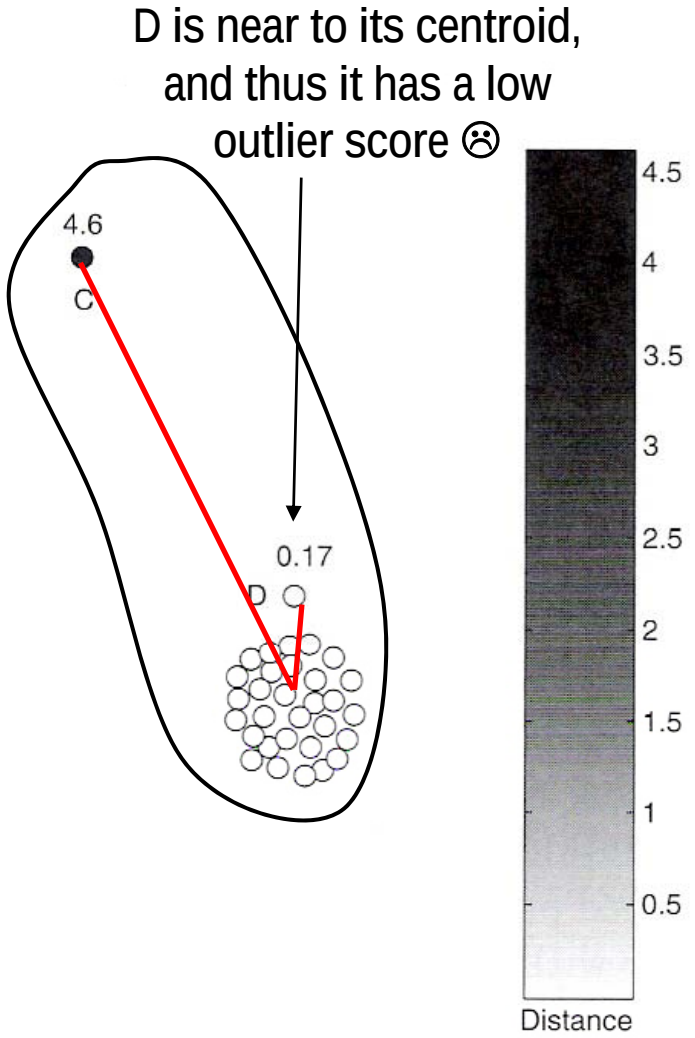
How do we measure it?

Unsupervised: Clustering Based

- Alternative a)

  By measuring the Euclidean distance to its closest centroid



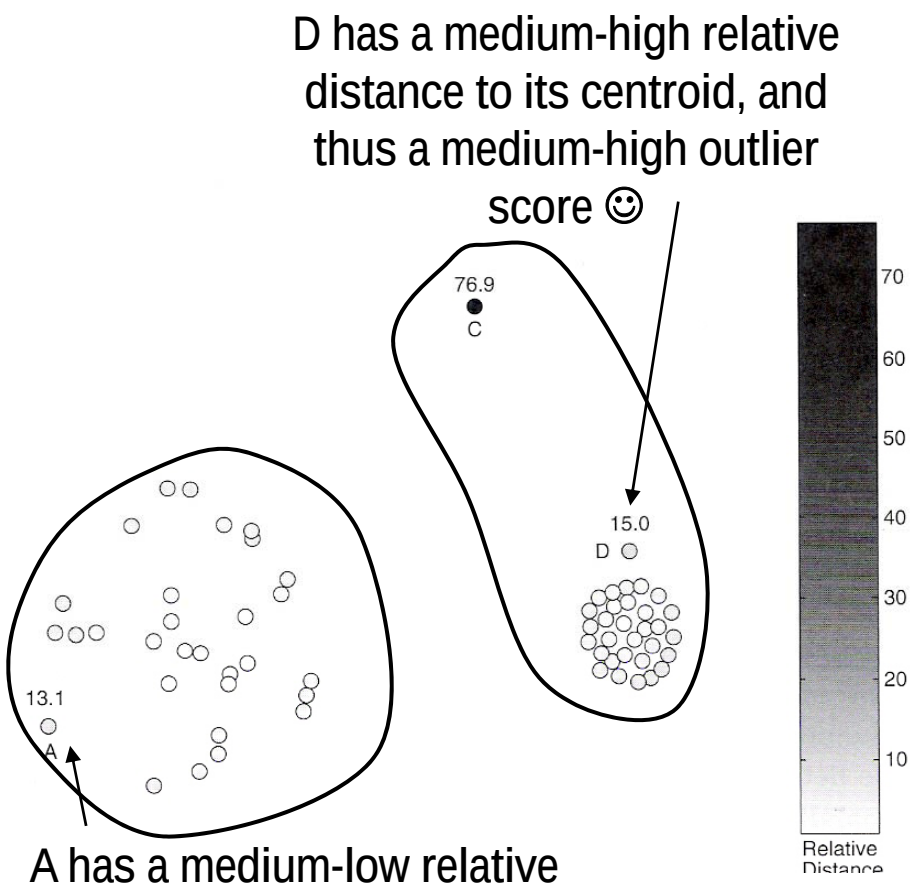


Unsupervised: Clustering Based

- Alternative b)

  By measuring the relative distance to its closest centroid.

  Relative distance is the ratio of the point's distance from the centroid to the median distance of all the points in the cluster from the centroid.

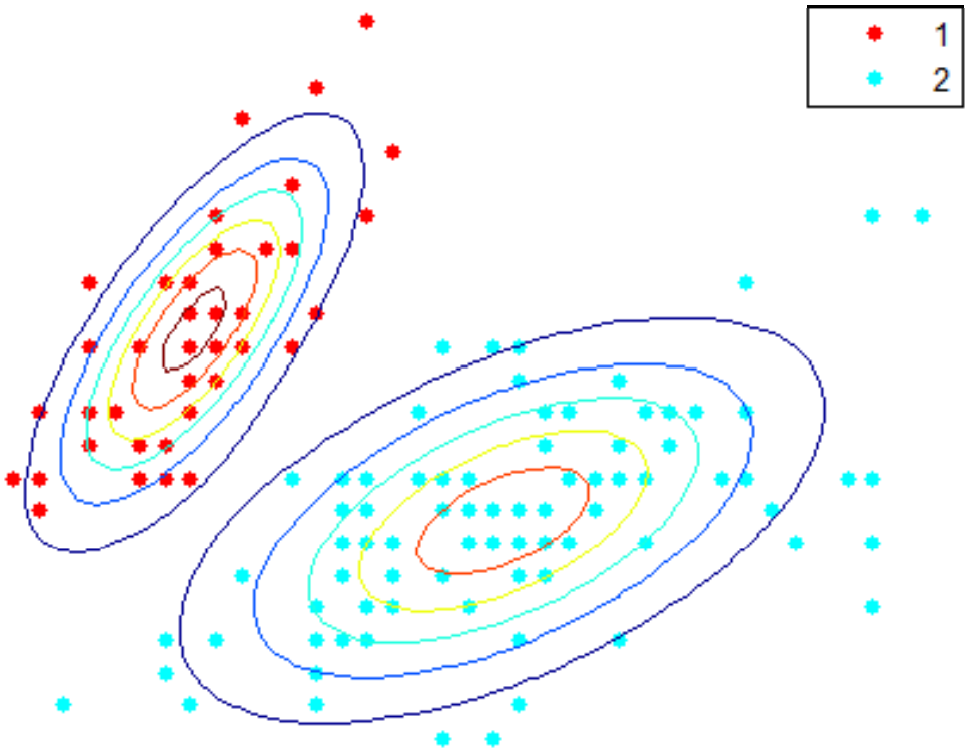


Unsupervised: Clustering Based

- Alternative c)

  By measuring the Mahalanobis distance of each point to the cluster distribution

  Mahalanobis distance involves computation of the covariance matrix of each cluster, which is quite expensive



Unsupervised: Clustering Based

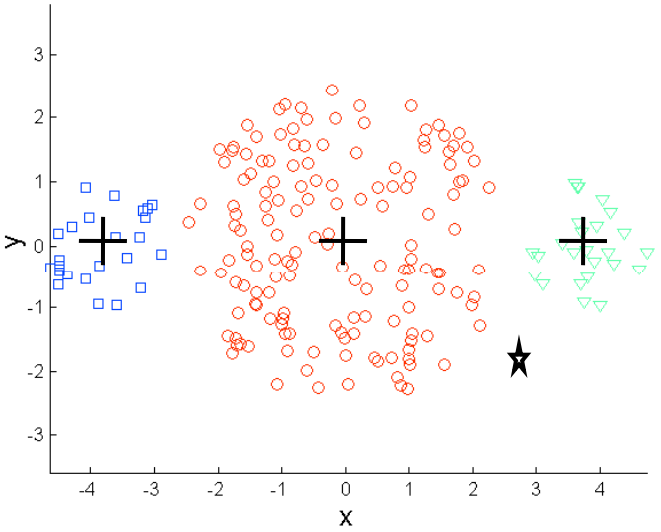
Choice of k is problematic

(k is now the number of clusters)

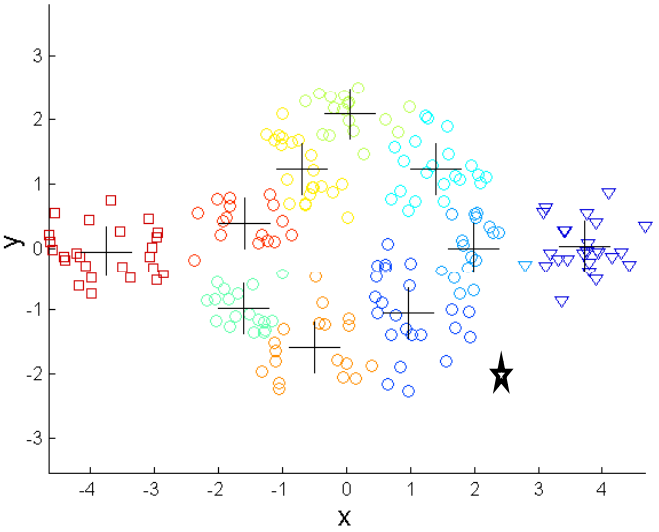
Usually, it's better to work with a large number of small clusters.

An object identified as outlier when there is a large number of small clusters, it's likely to be a true outlier.

k = 3



k = 10

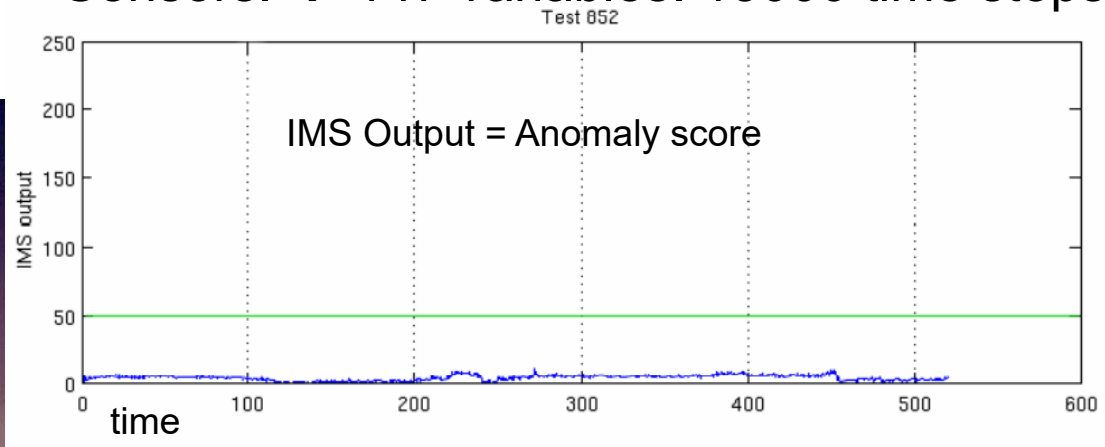


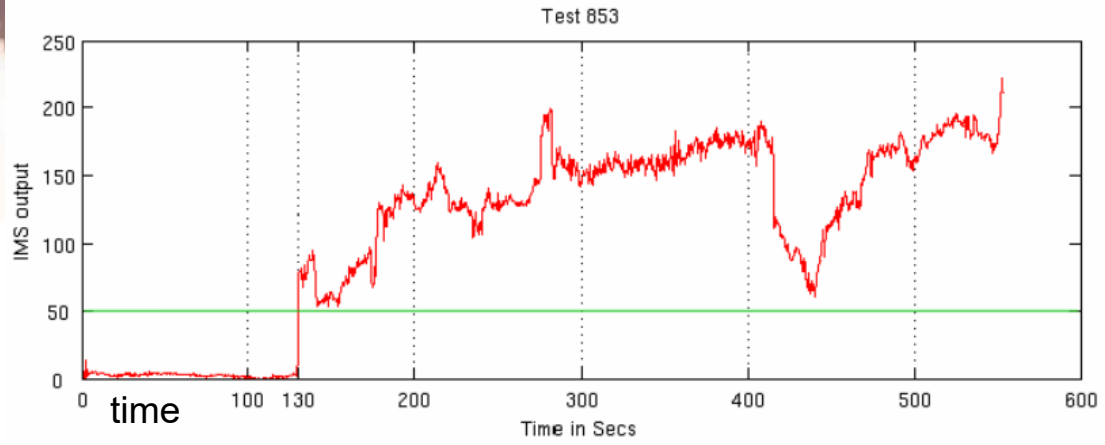
Unsupervised: Clustering Based

Mark Schwabacher , Nikunj Oza, Bryan Matthews, **2007**. Unsupervised Anomaly Detection for Liquid-Fueled Rocket. Propulsion Health Monitoring," AIAA Infotech@Aerospace Conference, American Institute of Aeronautics and Astronautics

SSME:
Space Shuttle Main Engines (reusable)

Inductive Monitoring System(IMS)
Pressure, temperature, vibration, etc.
Sensors. → 147 variables. 13000 time steps





Juan-Carlos Cubero. University of Granada

Hybrid Approaches

In complex scenarios, several techniques work together.

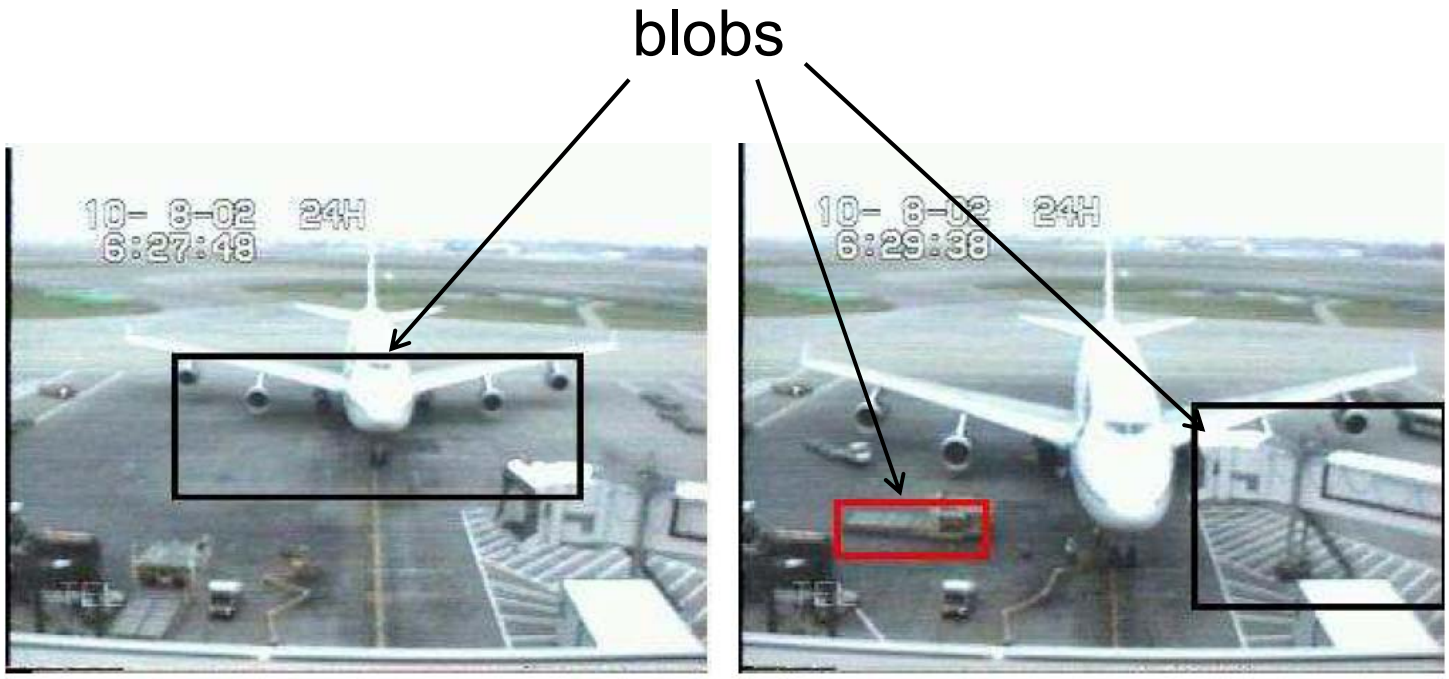
Anomaly detection sequences in video

Tao Xiang and Shaogang Gong. **2008**. Video Behavior Profiling for Anomaly Detection. *IEEE Trans. Pattern Anal. Mach. Intell.* 30, 5 (May 2008), 893-908. DOI=10.1109/TPAMI.2007.70731 http://dx.doi.org/10.1109/TPAMI.2007.70731

Frame 200

Frame 370

Hybrid Approaches



Frame 200 Frame 370

Hybrid Approaches

- A foreground detection method is applied to detect changes in image sequences

- Foreground pixels in a vicinity are grouped into a blob

- Instead of working with pixels, the method uses:

$$f = (\overline{x}, \overline{y}, w, h, Rf, M_p x, M_p y)$$

$(\overline{x}, \overline{y})$ are location features.

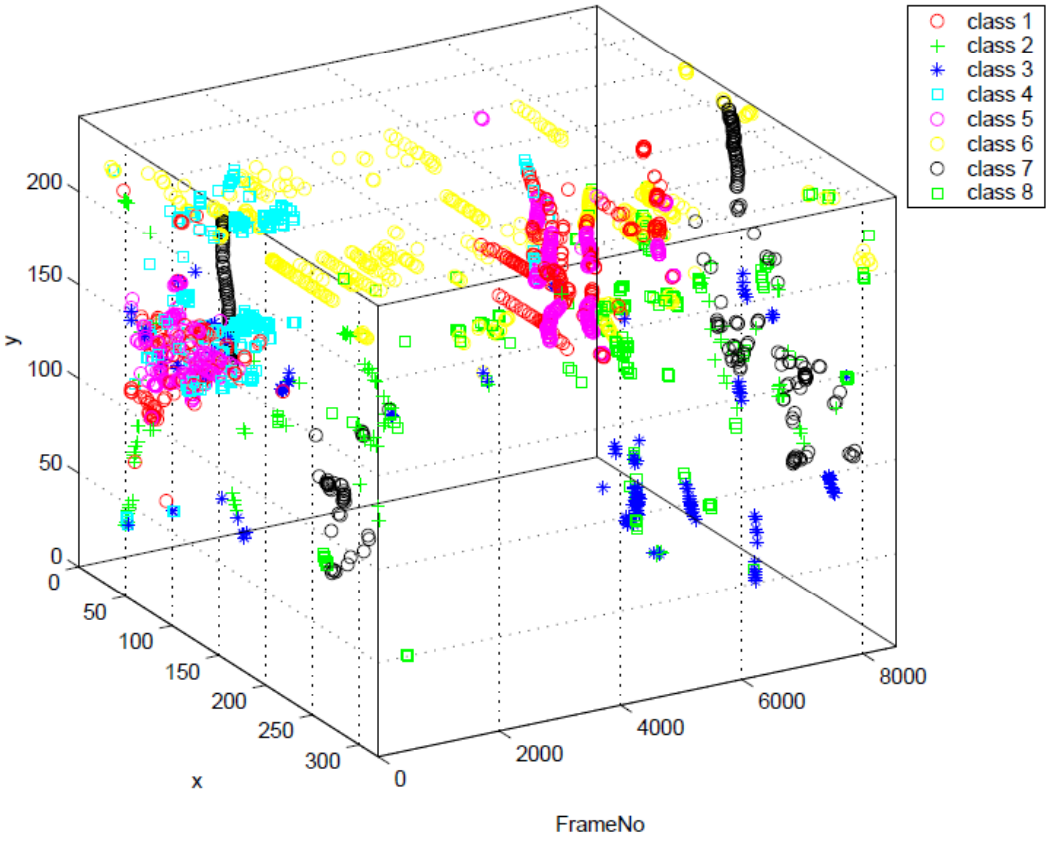
(w, h) and Rf are mainly shape features but also contain some indirect motion information

$(M_p x, M_p y)$ are motion features capturing the direction of object motion.

Hybrid Approaches

A clustering is performed in the 7-dimension feature space

$$f = (\overline{x}, \overline{y}, w, h, Rf, M_p x, M_p y)$$



Each cluster (class) represents an scene-event:

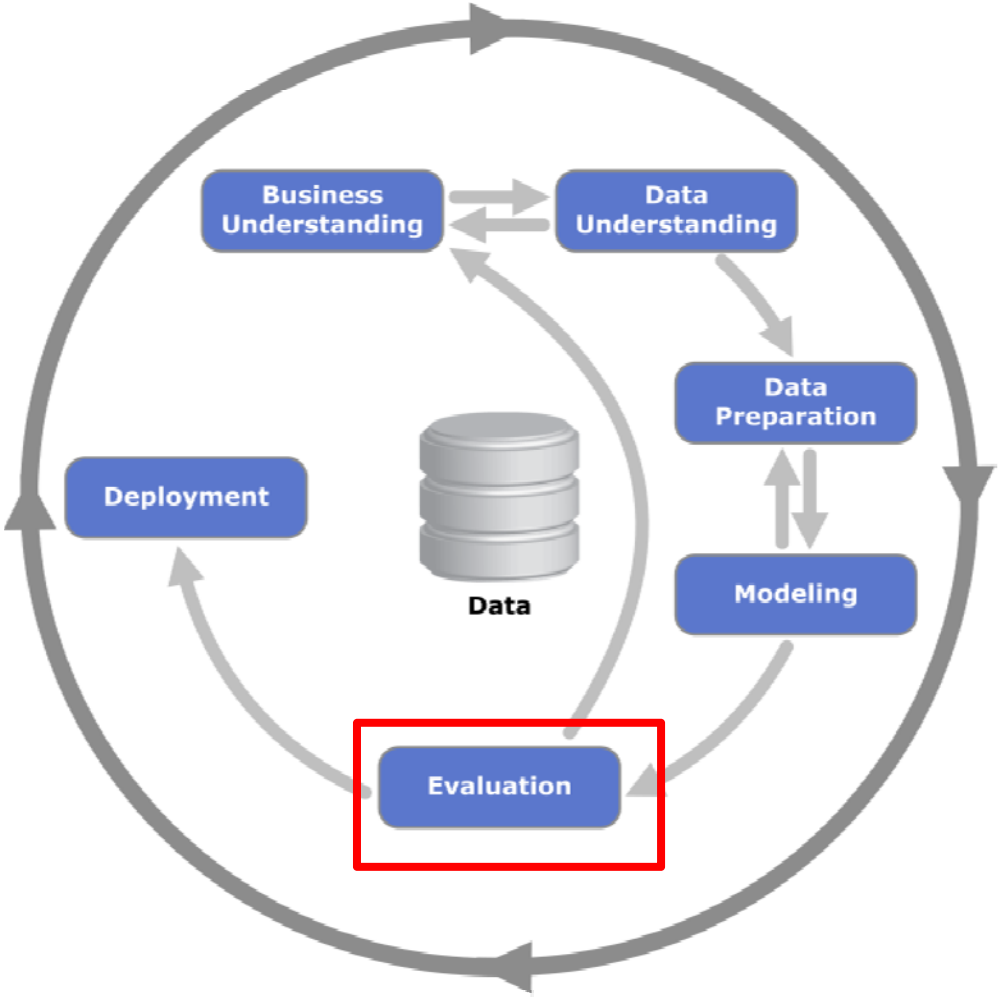
- aircraft moving and stopping
- airbridge moving
- rear catering vehicles
-….

A Bayesian net to model the scene-events interaction is constructed

Data Mining: Anomaly Detection

- Motivation and Introduction
- Supervised Methods
- Semisupervised Methods
- Unsupervised Methods:
  - Graphical and Statistical approaches
  - Nearest neighbor based approaches
  - Clustering based approaches
- Evaluation
- Abnormal regularities

Data Mining: Phases



Evaluation : Evaluation measures

Assumptions:

- The anomaly detection method output is yes / no (it's an anomaly / it isn't an anomaly)
- There is a procedure (maybe an expert) capable of saying if the prediction for new entries is correct, i.e, a test set is available.

Confusion matrix		Predicted class	
		NC	A
Actual class	NC	TN	FP
	A	FN	TP

anomaly class – A
normal class – NC

P: Positive (Prediction is A)
N: Negative (Prediction is NC)
T: True (Prediction is correct)
F: False (Prediction is incorrect)

Evaluation: Evaluation measures

Confusion matrix		Predicted class	
		NC	A
Actual class	NC	TN	FP
	A	FN	TP

- **(Global) Accuracy = (TN+TP) /(TN+TP+FN+FP)**

  What percentage of your predictions (for A or NC) were correct?

  Suppose the following.
  - Two classes: *normal* (99.9%) and *anomaly* (0.1%)
  - The default classifier, always labeling each new entry as *normal,* would have 99.9% accuracy!

We need other evaluation measures as alternatives to accuracy (Recall, Precision, F-measure, ROC-curves)

Evaluation: Evaluation measures

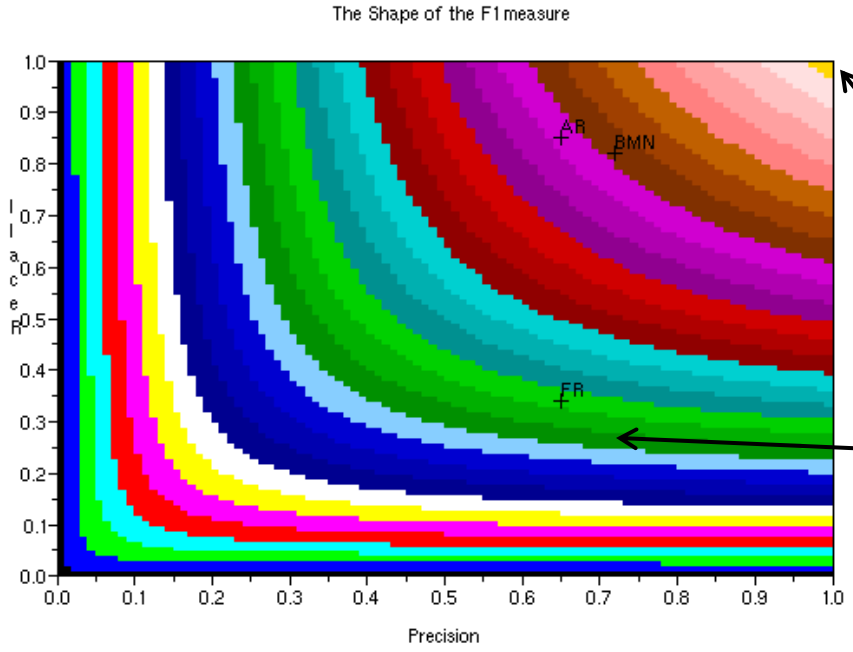
Confusion matrix		Predicted class	
		NC	A
Actual class	NC	TN	FP
	A	FN	TP

All the following measures are focused to an unique value (A in this case)

- **Precision for the anomaly A = TP/(TP + FP) = Prob(Real=A|Predict=A)**
  What percentage of anomaly predictions (A) were correct?
- **Recall (Sensitivity, True Positive Rate) = TP/(TP + FN) =**
  Prob(Predict=A|Real=A)
  What percentage of the anomaly cases (A) did you catch?
- **False Positive Rate FPR = FP/(FP + TN) = Prob(Predict=A|Real=NC)**
  What percentage of normal cases were erroneously predicted as anomalies.
- **Specifity (True Negative Rate) =**
  = TN/(TN+ FP) = Prob(Predict=NC|Real=NC) = 1-FPR
  What percentage of the normal cases did you catch?

Evaluation: Evaluation measures

It's not possible to improve all these measures at the same time.
If precision rises, recall decreases, for instance.
In order to balance both, a popular measure is the F_1 -score



$$F_1 = 2\frac{precision \cdot recall}{precision + recall}$$

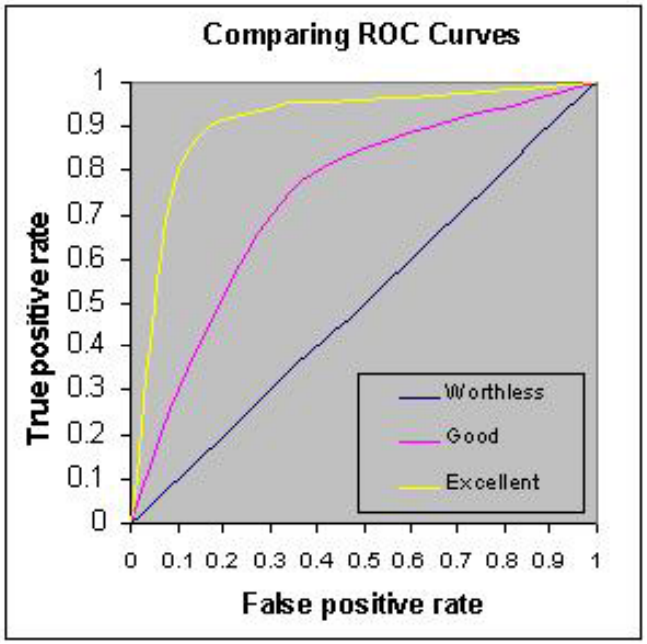
Best case

F_1 score decreases quite fast when recall and precision are not balanced

Evaluation: Evaluation measures

Let us consider an anomaly detection method giving a score as output. How to choose a **threshold** to label new entries as anomalies or non anomalies?

- Calculate the outlier scores of the instances in a test set.
- Sort the instances according to the scores.
- Apply threshold at each unique value of the score, getting the predictions label “yes” (it’s an anomaly) or “no”
- Compare these labels with the real labels and compute the desired measures



Now, we can apply, for instance, ROC curve analysis to find the threshold which provides the best balance between TPR and FPR

Evaluation: Evaluation measures

In Anomaly detection, Precision/Recall curves are also very common.

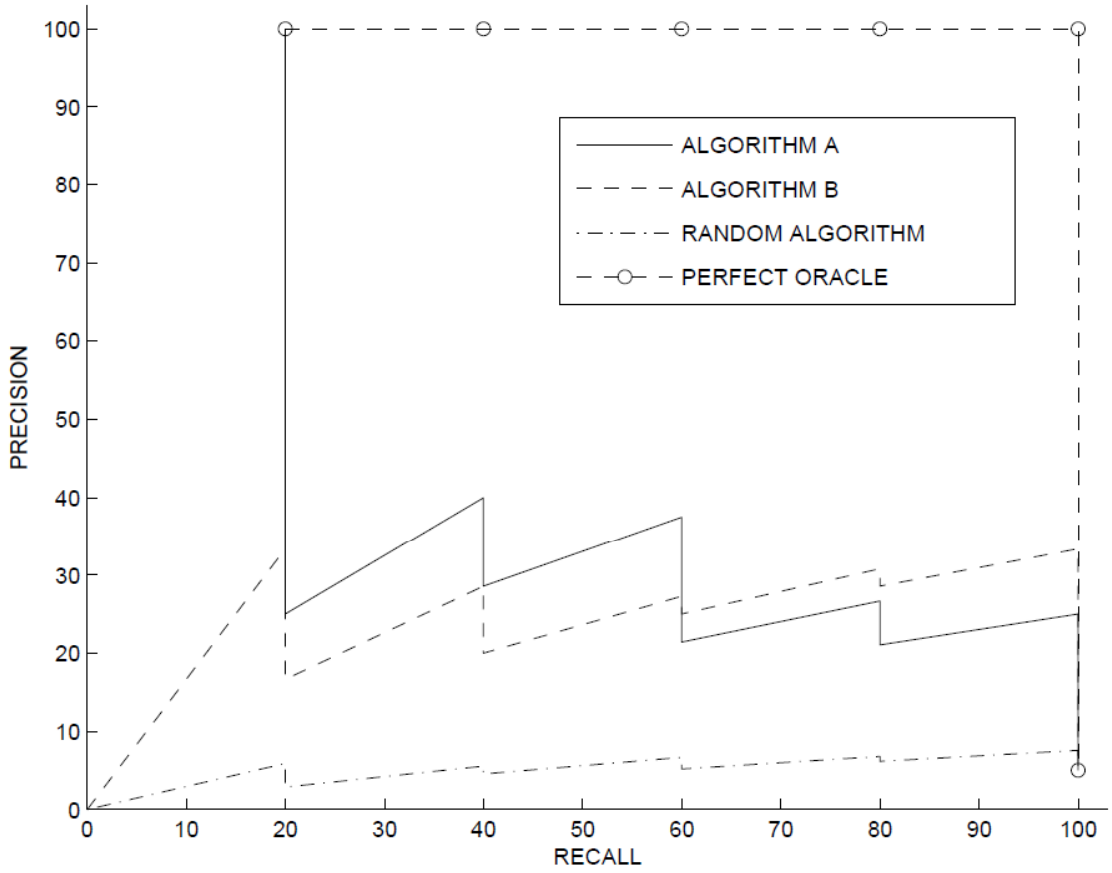
Let us suppose there are 5 true outliers and we want to compare two algorithms A and B. Each algorithm is executed, the scores are obtained and a rank is constructed.

Algorithm	Rank of Ground-truth Outliers
Algorithm A	1, 5, 8, 15, 20
Algorithm B	3, 7, 11, 13, 15
Random Algorithm	17, 36, 45, 59, 66
Perfect Oracle	1, 2, 3, 4, 5

Evaluation: Evaluation measures

P/R curve:

Algorithm	Rank of Ground-truth Outliers
Algorithm A	1, 5, 8, 15, 20
Algorithm B	3, 7, 11, 13, 15
Random Algorithm	17, 36, 45, 59, 66
Perfect Oracle	1, 2, 3, 4, 5



alg. A performs better than alg. B to detect the first three outliers. But alg. B is better to detect the five outliers.

Taken from Aggarwal's book

Evaluation: Base Rate Fallacy

- Precision, Recall and F_1 measures are useful evaluation measures in unbalanced data sets, but new problems arise when working with extremely unbalanced data sets.

Stefan Axelsson. **2000**. The base-rate fallacy and the difficulty of intrusion detection. ***ACM Trans. Inf. Syst. Secur.*** 3, 3 (August 2000), 186-205. DOI=10.1145/357830.357849

https://www.**cia**.gov/library/center-for-the-study-of-intelligence/csi-publications/books-and-monographs/psychology-of-intelligence-analysis/art15.html

Evaluation: Base Rate Fallacy

- Suppose a very rare disease (prob=10^{-5}) and a good diagnostic method. What's good? In Medicine, two measures are usually used: Sensitivity (Recall or TPR) and Specifity (TNR = 1-FPR)

- Suppose 99% Sensitivity (when the test was administered to a test population all of which had the disease, 99% of the tests indicated disease) and 99% Specifity (when the test population was known to be 100% free of the disease, 99% of the test results were negative)

- If the test is positive, do you have a 99% chance of having the disease?

Evaluation: Base Rate Fallacy

- Bayes theorem tells us that we have to take into account the PRIOR probability of the disease. For a formal proof see the next slides. Intuitively:
- For each 10^2 positive predictions (the test prediction is "you have the disease"), one of them is incorrect. But it's not probable to find the disease in 10^2 people. We should pick 10^5 people to find an occurrence of such disease. So, the test fails 10^{5-2} times (except one where it succeeds).
- Chances of really having the disease = 1 in 10^{5-2} = 1 in 1000
  This is the Precision for the disease =
  Prob(real is disease | prediction is disease) =
  Prob(real is anomaly | prediction is anomaly) = 0.001
- These are good news for you but not for the prediction

Evaluation: Base Rate Fallacy

- A typical problem: Intrusion Detection. The anomaly is the intrusion

I	: It is an intrusive behavior	*True*
¬*I*	: It is not a non-intrusive behavior	*False*
Alarm:	alarm (labeled as intrusion)	*Positive*
¬*Alarm*:	no alarm (labeled as no intrusion)	*Negative*

- The objective is to maximize (as usually, in a classification problem):
  - The precision for *I*, $P(I|Alarm)$
    If the alarm fires, it is an intrusion
    (also called *Bayesian detection rate*)
  - The precision for ¬*I*, $P(\neg I|\neg Alarm)$
    (if the alarm does not fire, it is not an intrusion)

- A good classification system will have:

  - A high **detection rate** (true positive rate or **Recall** for *I*) $P(Alarm|I)$

  - A low **false alarm rate**: $P(Alarm|\neg I)$

Evaluation: Base Rate Fallacy

$$P(A \mid B) = \frac{P(A) \cdot P(B \mid A)}{P(B)} \quad \longrightarrow \quad P(A \mid B) = \frac{P(A) \cdot P(B \mid A)}{\sum_i P(A_i) P(B \mid A_i)}$$

$$\{A_i\} = \{A, \neg A\} \longrightarrow \quad P(A \mid B) = \frac{P(A) \cdot P(B \mid A)}{P(A) \cdot P(B \mid A) + P(\neg A) \cdot P(B \mid \neg A)}$$

Let us consider $A = I, B = Alarm, \{A_i\} = \{I, \neg I\}$

$$P(I \mid Alarm) = \frac{P(I) \cdot P(Alarm \mid I)}{P(I) \cdot P(Alarm \mid I) + P(\neg I) \cdot P(Alarm \mid \neg I)}$$

Evaluation: Base Rate Fallacy

$$P(I \mid Alarm) = \frac{P(I) \cdot P(Alarm \mid I)}{P(I) \cdot P(Alarm \mid I) + P(\neg I) \cdot P(Alarm \mid \neg I)}$$

In an anomaly detection system, we have very low *P(I)* values (for instance, $2*10^{-5}$). So, *P(¬I)* is very high

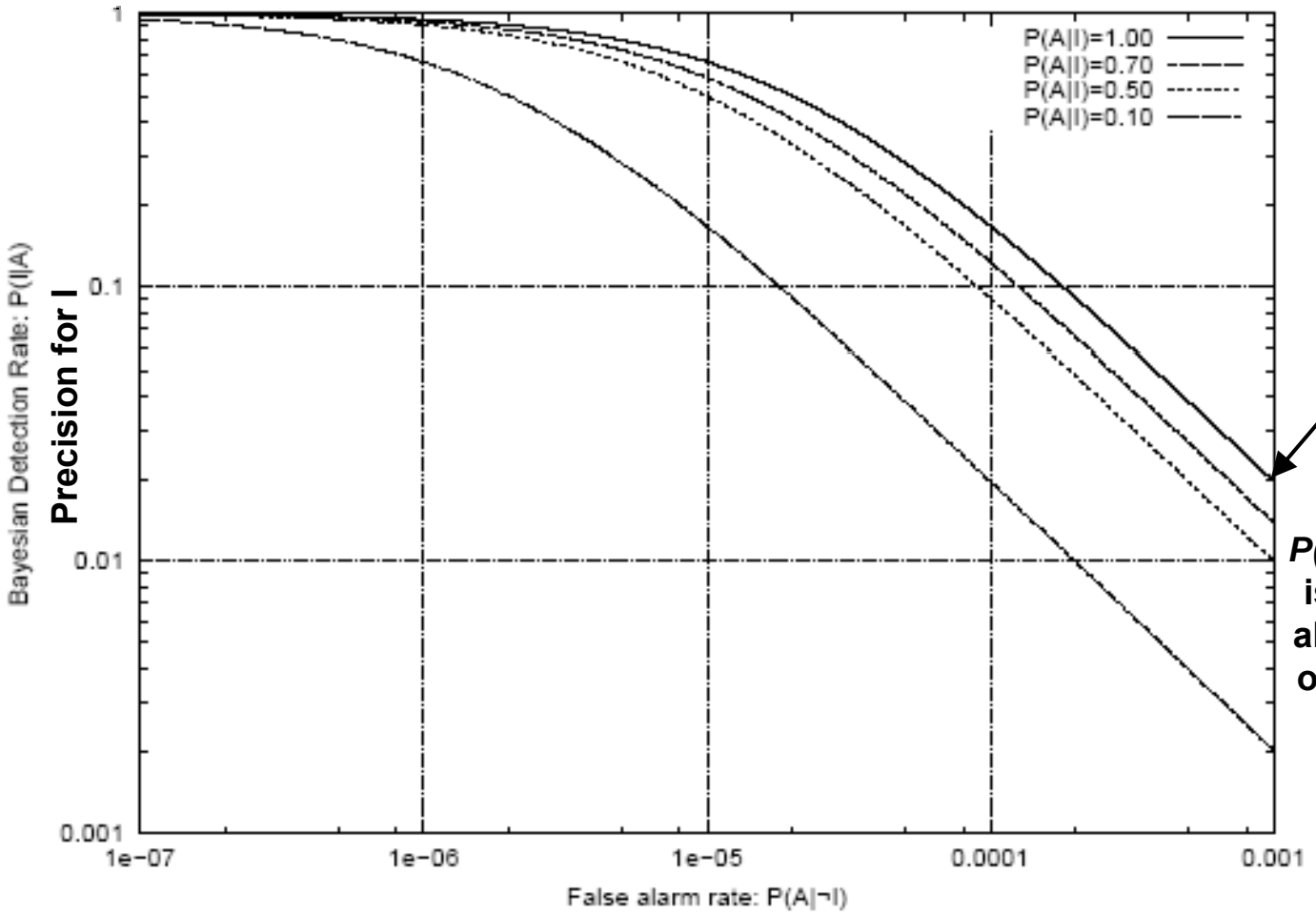
$$P(I \mid Alarm) = \frac{2 \cdot 10^{-5} \cdot P(Alarm \mid I)}{2 \cdot 10^{-5} \cdot P(Alarm \mid I) + 0.99998 \cdot P(Alarm \mid \neg I)}$$

The final value of *P*(*I*|*Alarm*) (precision for I) is dominated by the false alarm rate *P*(*Alarm*|¬*I*).
P(*Alarm*|¬*I*) should have a very low value to compensate 0.99998, and thus get a high value of *P*(*I*|*Alarm*) (precision for I).

BUT even a very good classification system, does not have such a false alarm rate. ☹

Evaluation: Base Rate Fallacy

P(I)=1e-05



Consider a classification with the best possible Recall *P(Alarm|I)*=1 and an extremely good false alarm rate *P(Alarm|¬I)* of 0.001 = 1e-03

In this "ideal" case, *P(I|Alarm)*=0.02 (the scale is logarithmic) So, If the alarm fires 50 times, only one is a real intrusion. ☹

Evaluation: Base Rate Fallacy

Conclusion:

Classification systems for anomalies, should always report the precision for the anomaly and the false alarm rate.

The lower the false alarm rate is, the higher the precision is

How much?

The false alarm rate should be less than one order of magnitude (*10) of anomaly's prior probability.

When the anomaly is extremely rare (10^{-5}), this is really difficult to attain for a classification system ☹

Evaluation: Base Rate Fallacy

Some methods try to lower the false alarm rate, usually by incorporating ad hoc information.

Georgios P. Spathoulas, Sokratis K. Katsikas, Reducing false positives in intrusion detection systems, **Computers &; Security**, Volume 29, Issue 1, February **2010**, Pages 35-44, ISSN 0167-4048, 10.1016/j.cose.2009.07.008.

Tadeusz Pietraszek. On the use of ROC analysis for the optimization of abstaining classifiers. **Machine Learning**, Volume 68, Number 2, 137-169, **2007**. DOI: 10.1007/s10994-007-5013-y

Fu Xiao, Xie Li, Using Outlier Detection to Reduce False Positives in Intrusion Detection, Network and Parallel Computing Workshops, IFIP International Conference on, pp. 26-33, 2008 IFIP International Conference on Network and Parallel Computing, **2008**

Bibliografía básica

Varun Chandola, Arindam Banerjee, and Vipin Kumar. 2009. Anomaly detection: A survey. ACM Comput. Surv. 41, 3, Article 15 (July 2009), 58 pages. DOI=http://dx.doi.org/10.1145/1541880.1541882

Pang-Ning Tan, Michael Steinbach, and Vipin Kumar. 2005. Introduction to Data Mining, (First Edition). Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA.

Charu C. Aggarwal. 2013. Outlier Analysis. Springer Publishing Company, Incorporated. http://charuaggarwal.net/outlierbook.pdf

Jiawei Han. 2012. Data Mining: Concepts and Techniques. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA; Third edition.

Vic Barnett , Toby Lewis , 1994. Outliers in Statistical Data. Wiley; Third edition.