UNIVERSIDAD
DE LA RIOJA

APPLIED DATAMINING IN INDUSTRIAL ENVIRONMENTS

J. Ordieres
University of la Rioja. SPAIN

UNIVERSIDAD
DE LA RIOJA

App. Data Mining. Table of Contents

- Introduction
- Interest for using DM in Ind. Env.
- Opportunity regarding EU market
- Main concerns
- Dealing with massive data
- Dealing with outliers
- A way for mixed strategies
- Conclusion

App. Data Mining. Introduction

J. Ordieres:

Professor at the Univ. of la Rioja.

Currently involved in 3 EU funded projects related to DM.

Also involved in 2 contracts with private companies in order to improve their processes by using DM.

So … I’m just a ***practitioner*** in these fields. Pls. be patient.

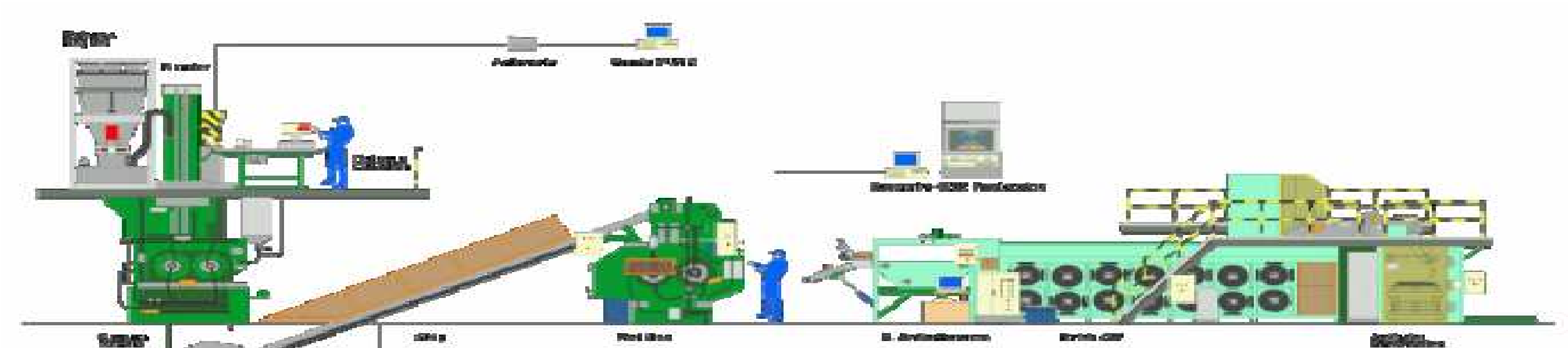
App. Data Mining. Table of Contents

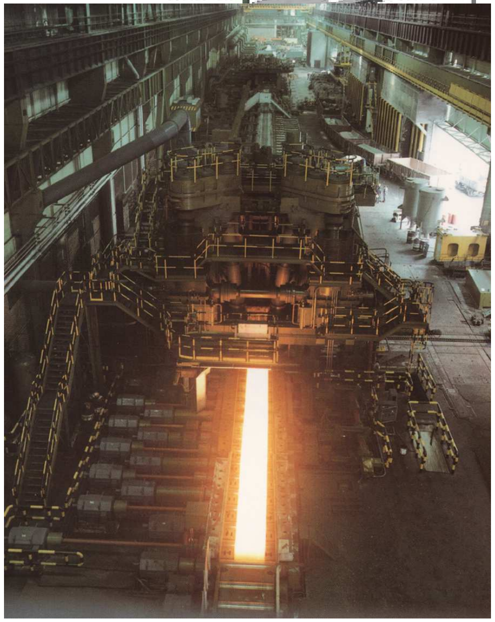
- Introduction
- Interest for using DM in Ind. Env.
- Opportunity regarding EU market
- Main concerns
- Dealing with massive data
- Dealing with outliers
- A way for mixed strategies
- Conclusion

App. Data Mining. Interest in IE

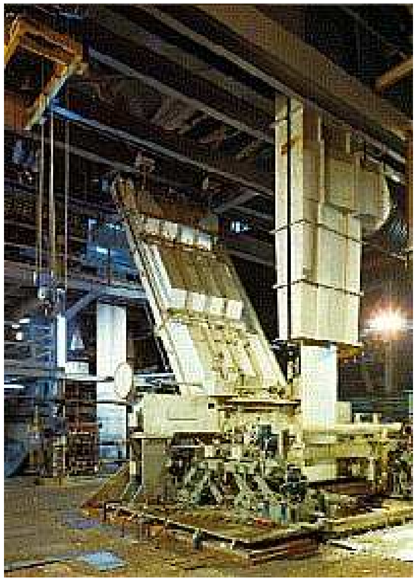
In those industrial processes operating under open loop control strategies, a model is needed for estimation of:

a) control commands.
b) the magnitude of control actions.
c) the most appropriate control command values.
d) those input variables that command the process.
e) the levels of driving parameters.



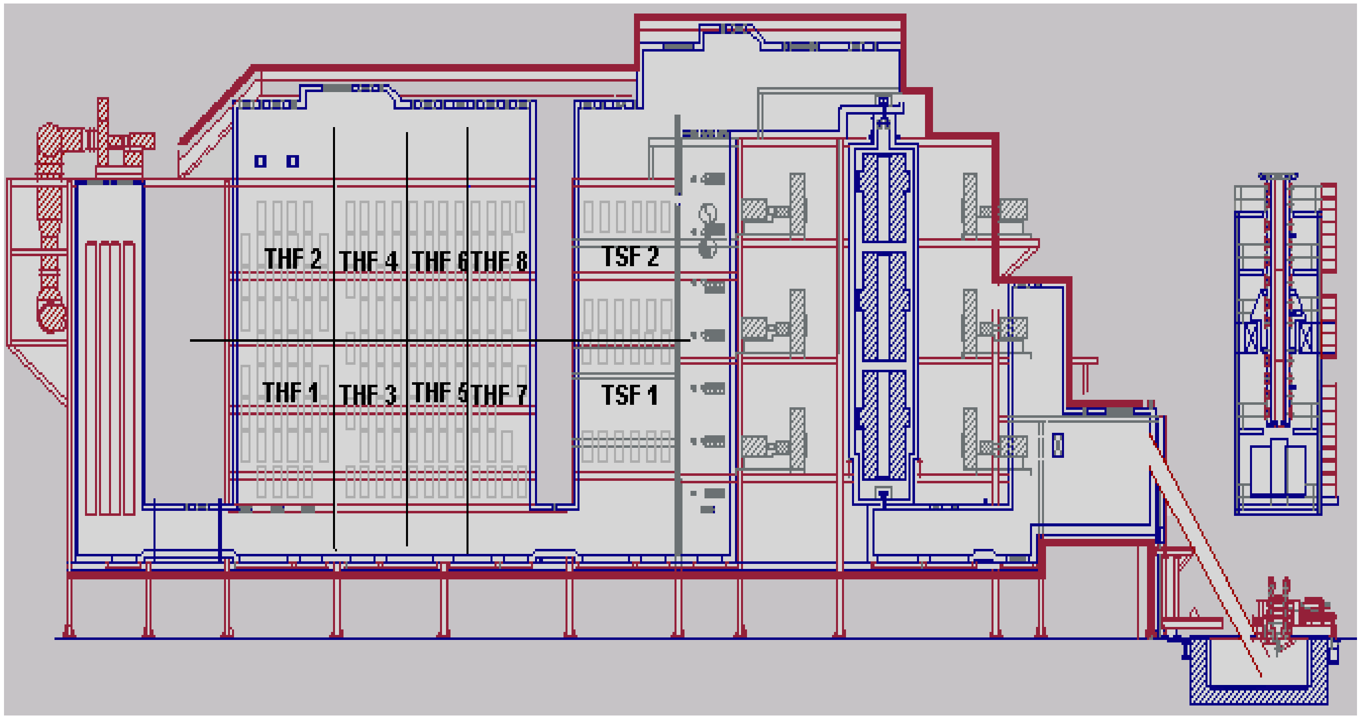


App. Data Mining. Interest in IE



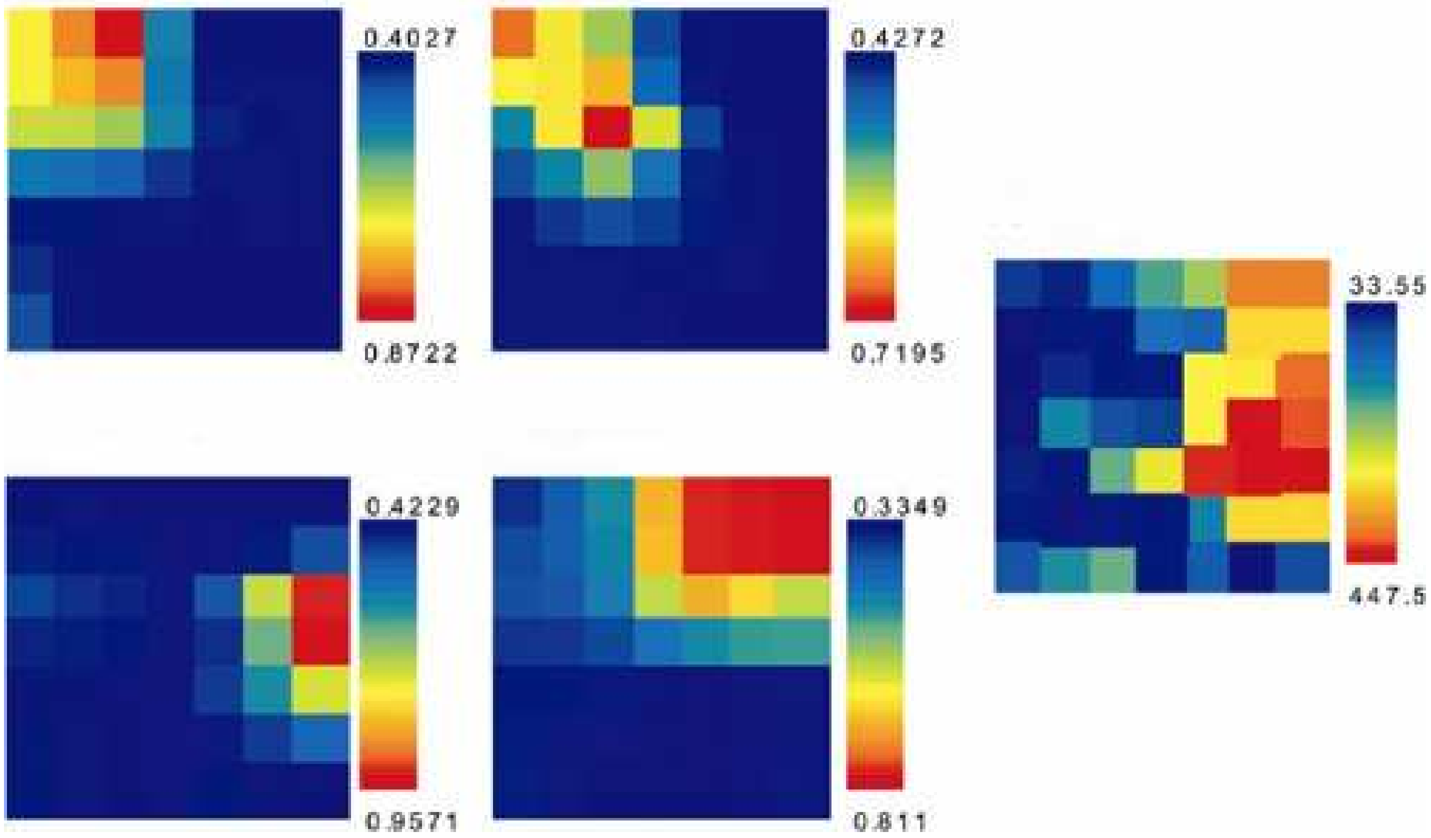
A constant topic is related to be able to improve the final product.

Also, predicting some properties is relevant.



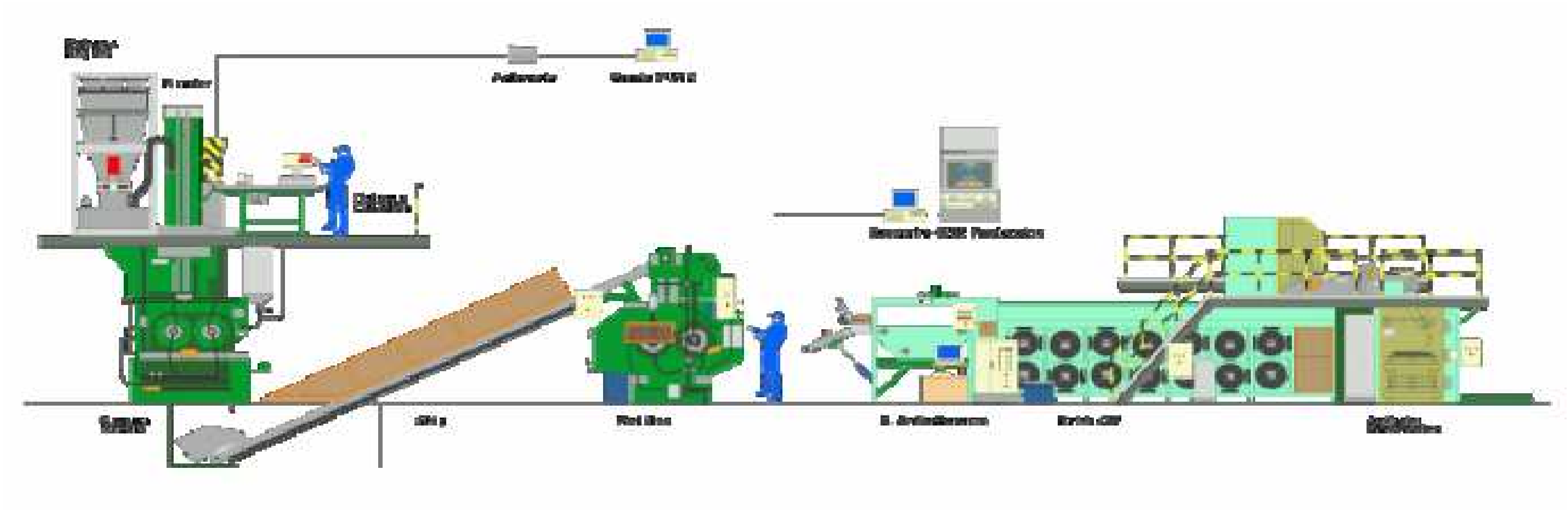
App. Data Mining. Interest in IE

Because these models are expensive, the first idea was to be able to replace them by trying to set up a ‘black box’ model, taking advantage of data from the process itself…



App. Data Mining. Interest in IE

In those cases where the ‘product quality’ depends on factors that can’t be measured on-line …



It is necessary to perform a reliable estimation of the commanding action levels and to carry out laboratory measures later.

App. Data Mining. Interest in IE

Some help or ‘knowledge’ is supposed to be available from the process signals themselves.

This is one of our issues …

Just an example:

When a coil, by error, is processed using a harder material than normal and ends up with a client, significant damage may be produced in the installation process.

An artificial lock needs to be provided in order to ‘predict’ the elongation regarding the tension and pressure used in the skin-pass. If the predicted elongation is a lot different than the measured elongation, the coil is then removed from the queue for further analyses …

App. Data Mining. Introduction

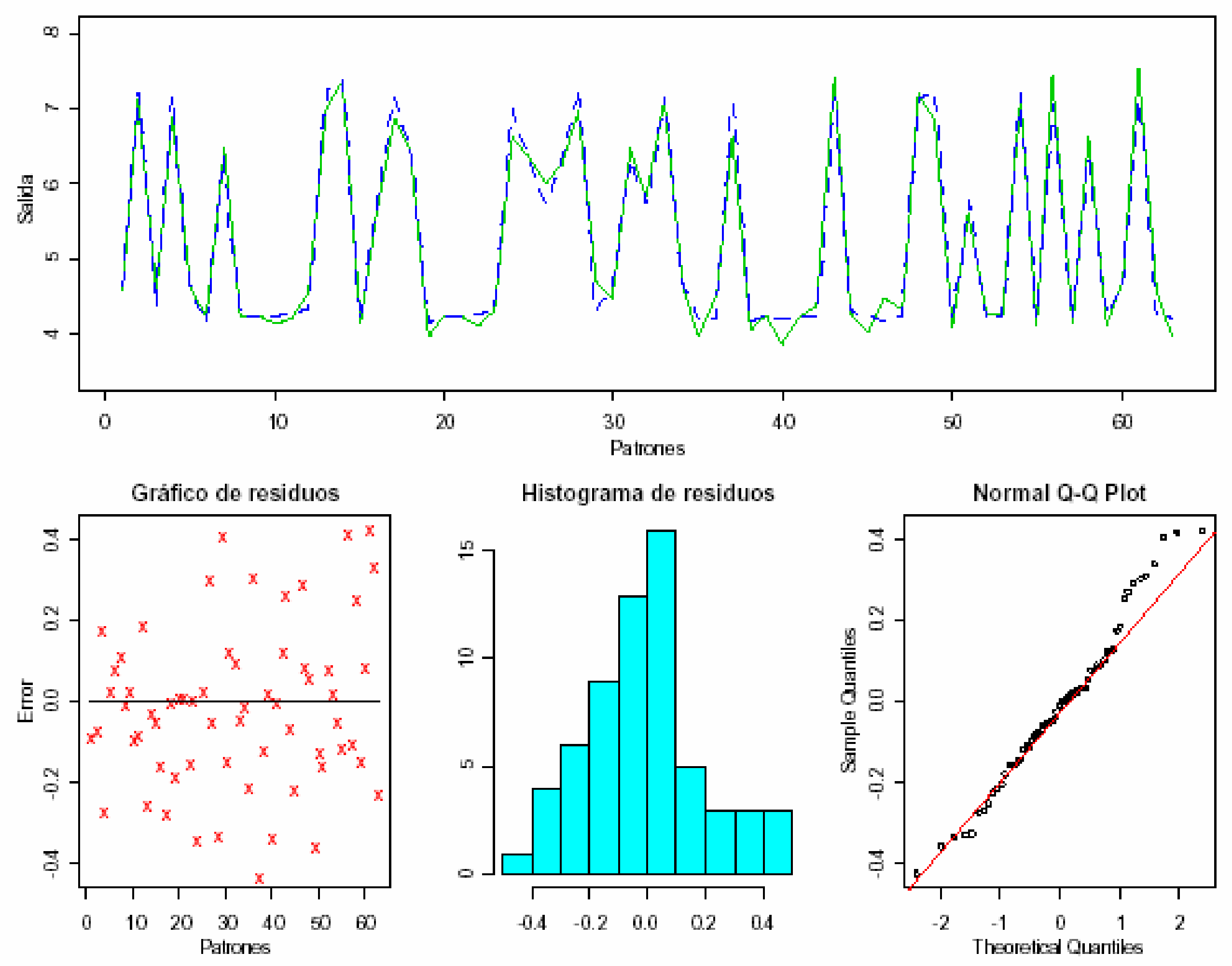
Normally, classical models are based on physical laws and they are improved with some ‘local corrections’ …

Moreover, improvements are very difficult as far as local adjustments conflict with the global approach and, often, many improvements are required for material or parameter changes.

In this cases, specific models derived from the data relationships themselves can help.

App. Data Mining. Introduction

Then, when the new model exists … it is quite easy to simulate the effect of set-up variations …



From here, new models arising make it easier to develop the control process …

App. Data Mining. Table of Contents

- Introduction
- Interest for using DM in Ind. Env.
- Opportunity regarding EU market
- Main concerns
- Dealing with massive data
- Dealing with outliers
- A way for mixed strategies
- Conclusion

App. Data Mining. Opportunity in EU

Applied research is quite well funded, especially if new APPLIED knowledge is envisaged and a EU dimension is required.

After a critical point, improvements are very difficult as far as local adjustments conflict with the global approach and, often, regular improvements are required for material or parameter changes.

Specific tools need to be developed.

App. Data Mining. Opportunity in EU

Specific tools need to be developed combining others well known:

- Decision Tree-based Methods
- Rule-based Methods
- Memory-based reasoning
- Neural Networks
- Genetic Algorithms
- Naïve Bayes and Bayesian Belief Networks
- Support Vector Machines
- …

UNIVERSIDAD DE LA RIOJA

App. Data Mining. Table of Contents

- Introduction
- Interest for using DM in Ind. Env.
- Opportunity regarding EU market
- Main concerns
- Dealing with massive data
- Dealing with outliers
- A way for mixed strategies
- Conclusion

Data Mining. Main concerns

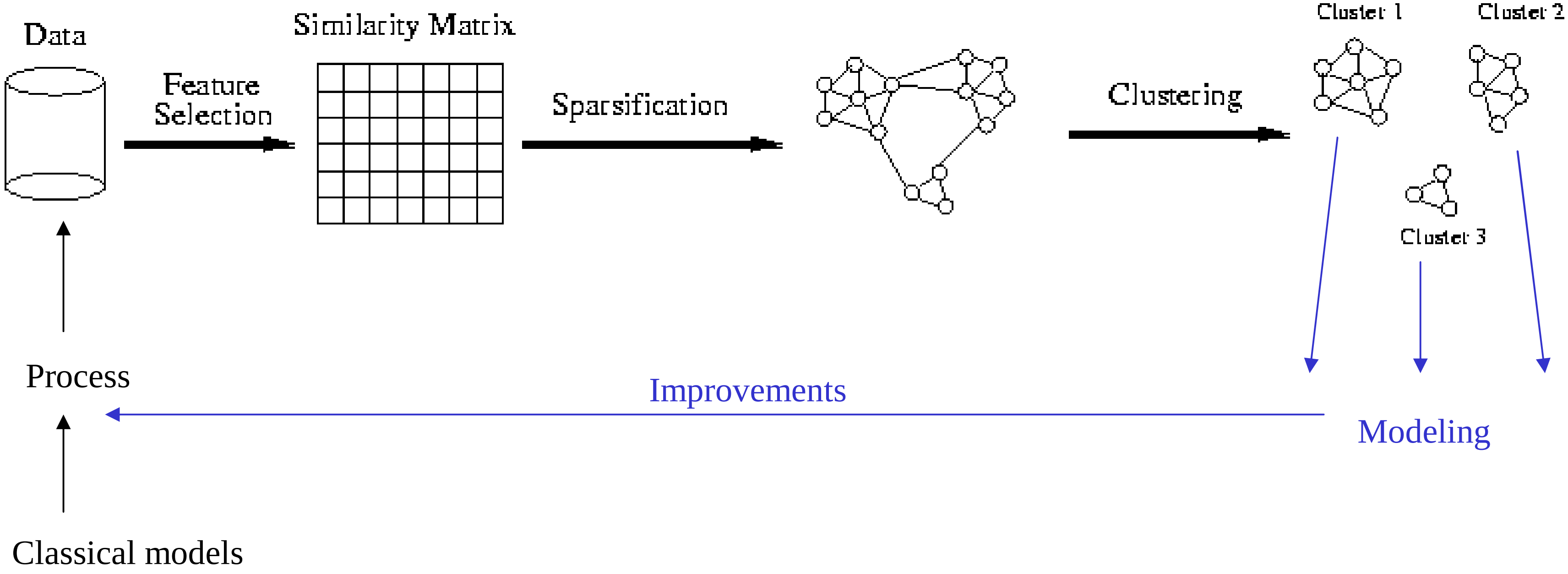
We need to know what kinds of problems are possible, i.e., what sorts of situations correspond to poor data quality.

The following are well known problems:

noise and outliers
missing values
duplicate data
inconsistent values

Data Mining. Main concerns

If the strategy can be pictured as below, it is practically necessary at each step to deal with some threats …



Process

Improvements

Modeling

Classical models

App. Data Mining. Table of Contents

- Introduction
- Interest for using DM in Ind. Env.
- Opportunity regarding EU market
- Main concerns
- Dealing with massive data
- Dealing with outliers
- A way for mixed strategies
- Conclusion

Data Mining. Dealing with massive data

Just as an example (1 process line):

Hot rolling:

2010 process variables,

sampled each 100m of coil

900 m/min of medium speed

24/24 7/7

1 year: 4,7 million patterns in R^{2010}

9508 million of variables

It is necessary to find one sampling policy.

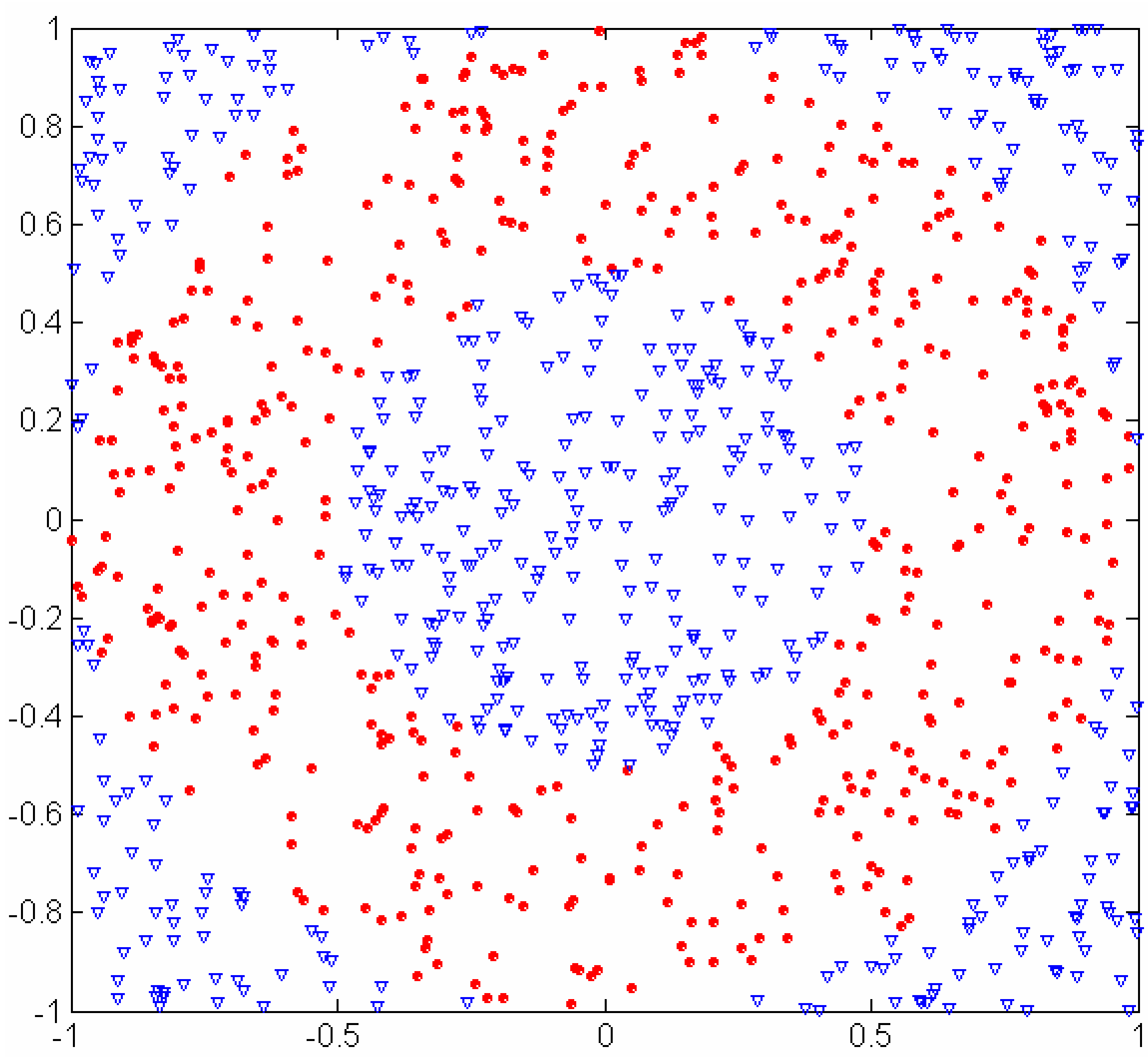
Data Mining. Dealing with massive data

- Sampling is the main technique employed for data selection.
  - It is often used for both the preliminary investigation of the data and the final data analysis.

- Statisticians sample because obtaining the entire set of data of interest is too expensive and/or time consuming.

- Sampling is used in data mining because it is too expensive and/or time consuming to process all the data

Data Mining. Dealing with massive data



Sometimes it is not easy to identify the underlying structure

500 circular and 500 triangular data points.

Circular points:

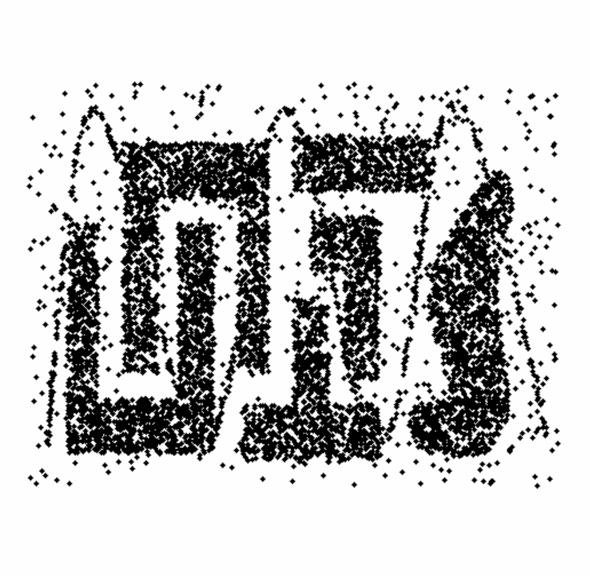
$0.5 \leq \mathrm{sqrt}(x_1^2+x_2^2) \leq 1$

Triangular points:

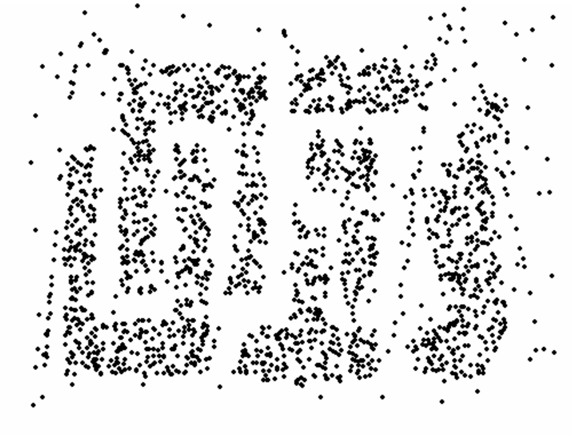
$\mathrm{sqrt}(x_1^2+x_2^2) > 0.5$ or

$\mathrm{sqrt}(x_1^2+x_2^2) < 1$

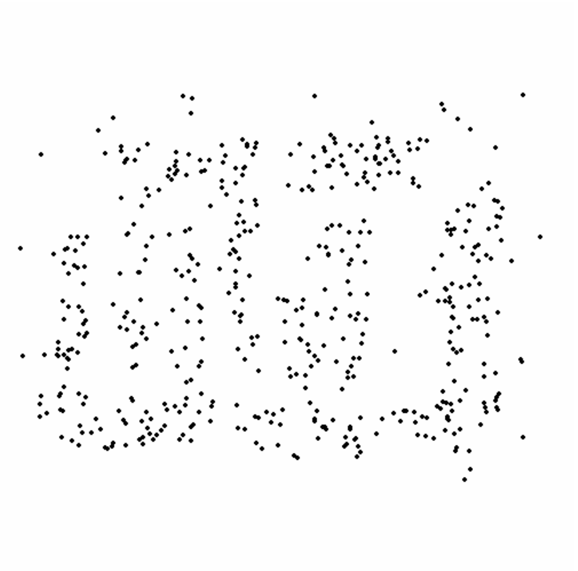
Data Mining. Dealing with massive data



8000 points



2000 Points



500 Points

It must be managed carefully, because you can run into problems.

Data Mining. Dealing with massive data

- Redundant features
  - duplicate much or all of the information contained in one or more of the attributes, e.g., the coil format and coil code (where the format code is included).

- Irrelevant features
  - contain no information that is useful for the data mining task at hand, e.g., students' ID numbers are irrelevant to the task of predicting their grade point averages.

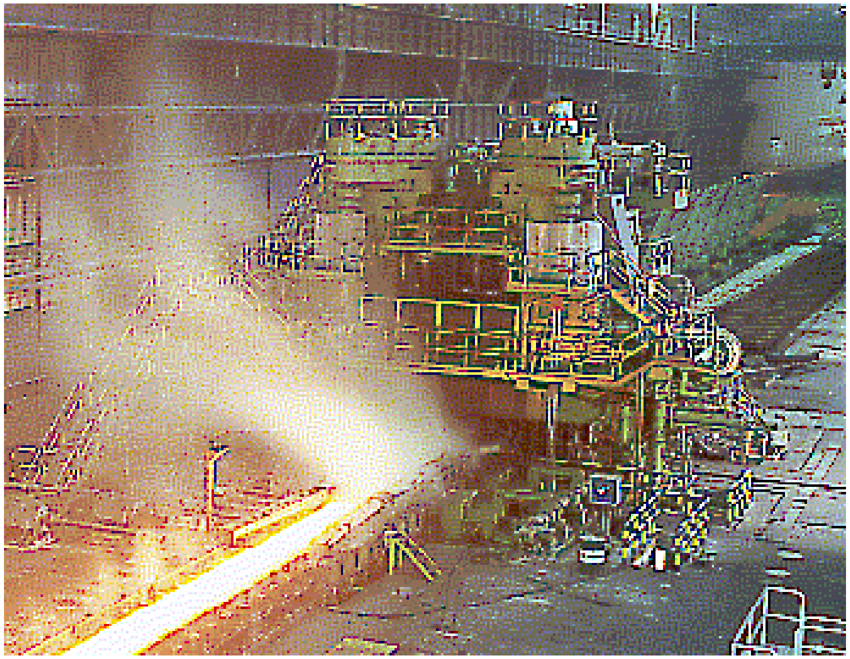
App. Data Mining. Table of Contents

- Introduction
- Interest for using DM in Ind. Env.
- Opportunity regarding EU market
- Main concerns
- Dealing with massive data
- Dealing with outliers
- A way for mixed strategies
- Conclusion

UNIVERSIDAD
DE LA RIOJA

Data Mining. Dealing with outliers

Some variables are measured by physical sensors very quickly and in harsh environments (high temperatures, wet conditions, high pressures, etc.)

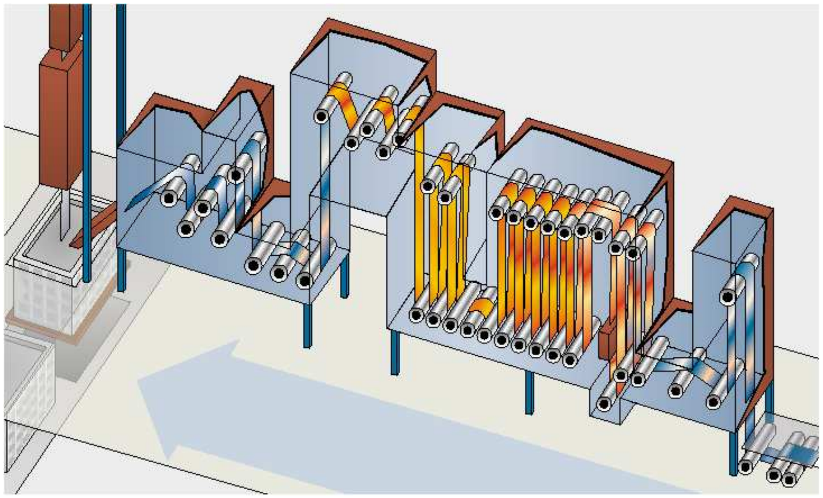


It is common to find ‘measurement errors’, but there are other sources of samples to be managed carefully.

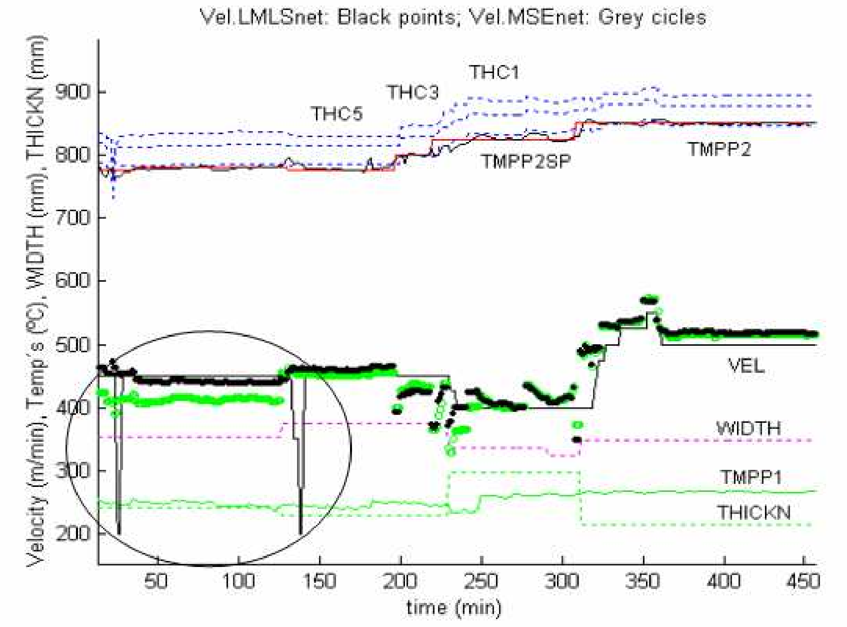
It is necessary to identify them

Data Mining. Dealing with outliers

In other cases the process is stopped by other problems, *e.g. welding problems during coil extension as shown below*, and also in these cases, even when there are no measurement errors, it would be necessary to identify these points in a sample set.

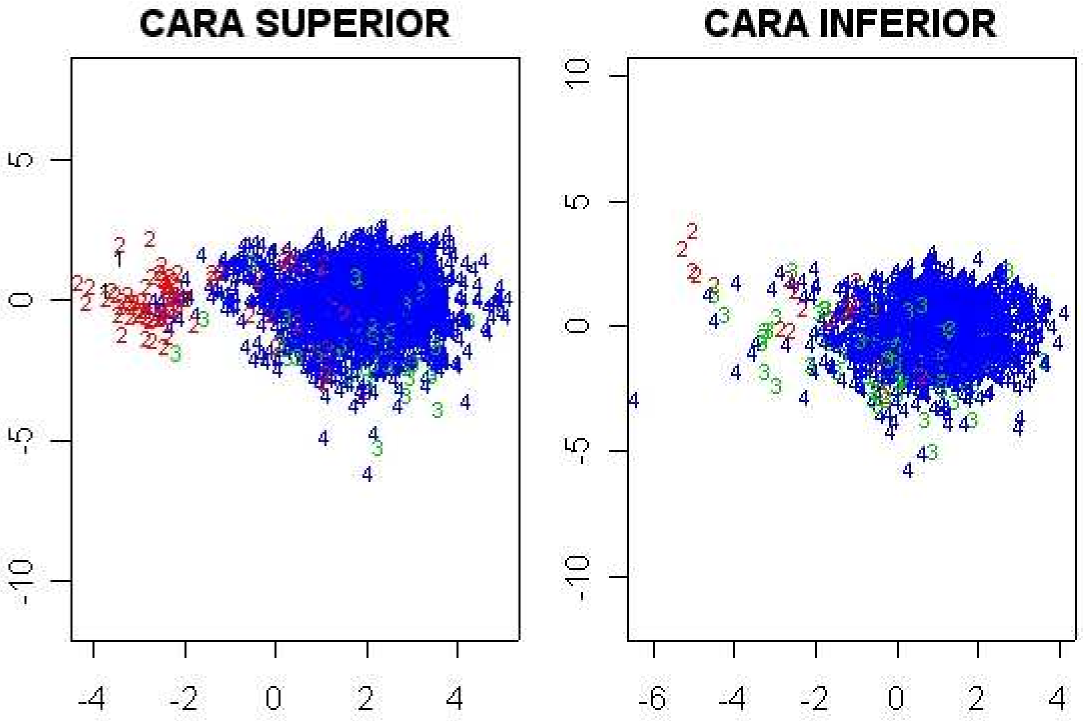


Especially if they are interfering with data used to build a model, e.g. to estimate line speed when the material format is changing and the temperature needs to be under control.



Data Mining. Dealing with outliers

In those cases where an indirect, automatic control system tries to keep the process under control, the errors are usually non-normal, so outlier identification must be managed carefully.



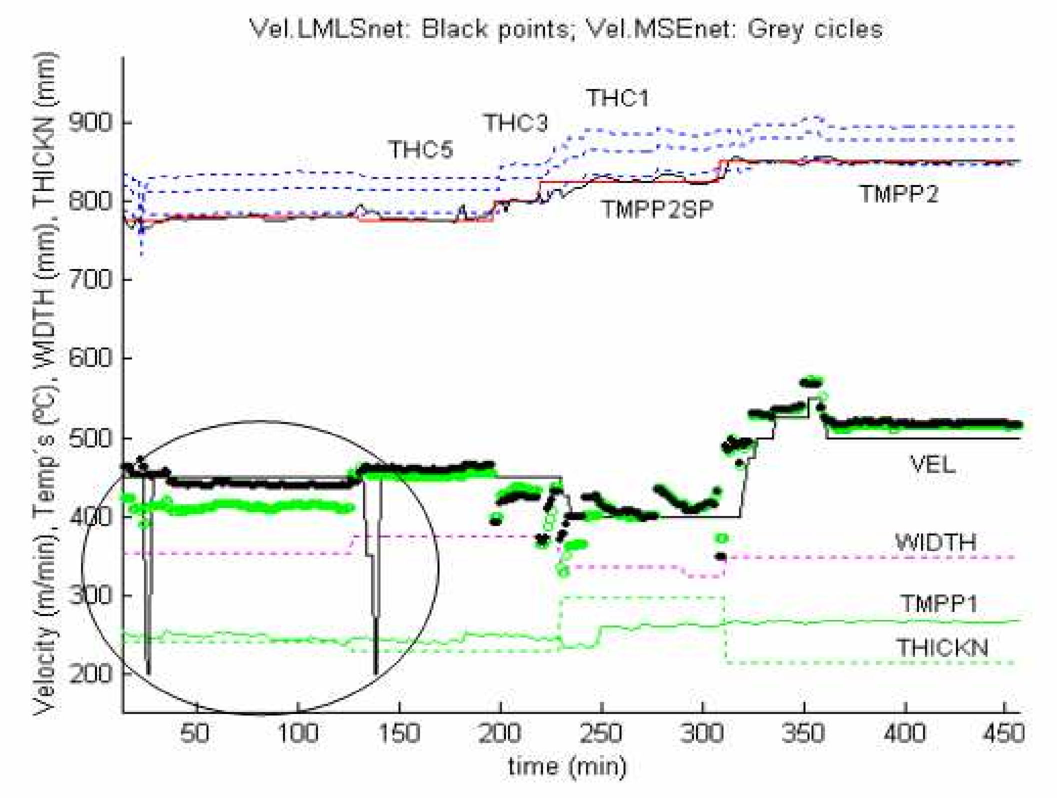
A new algorithm were developed (PAELLA), based on local metrics concepts and probability to be outlier as it becomes respect to local models.

CASTEJON, ORDIERES, PISÓN & VERGARA.
"Outlier Detection and Data Cleaning in Multivariate Non-Normal Samples: The PAELLA Algorithm."
Data Mining and Knowledge Discovery, 9, 171-187, 2004.

Data Mining. Dealing with outliers

In other cases it’s interesting to try working with potential outliers in order to build a model, e.g. if there is not a large amount of data.

But robust properties are required



An integration of neural networks concepts and robustness is produced in order to help us.

A. Espinoza, J. Ordieres, F.J. M. de Pisón, A. G. Marcos.
“TAO-ROBUST BACKPROPAGATION LEARNING ALGORITM”
*Sent for publishing to “**Neural Networks** “ journal.*

UNIVERSIDAD
DE LA RIOJA

App. Data Mining. Table of Contents

- Introduction
- Interest for using DM in Ind. Env.
- Opportunity regarding EU market
- Main concerns
- Dealing with massive data
- Dealing with outliers
- A way for mixed strategies
- Conclusion

UNIVERSIDAD
DE LA RIOJA

Data Mining. A way for mix strategies

Real, every-day problems demand powerful algorithms at each step of a selected methodology (CRISP-DM or whatever you want)

Cluster identification needs to manage large sample sizes:

Some authors (*like Prof. Ciampi*) propose that SOM technology be mixed before clustering. This is just an example of mixed strategies.

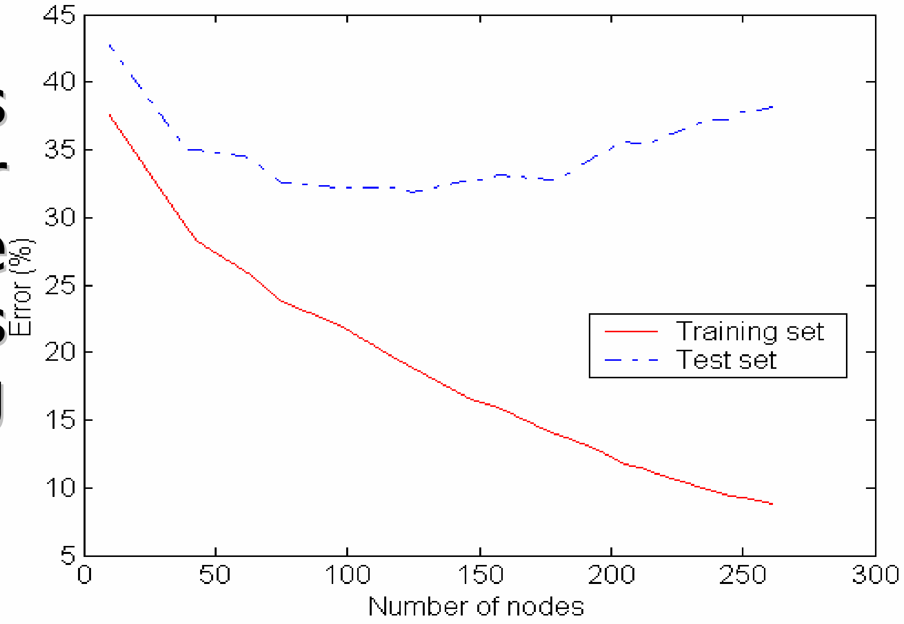
Modellization needs to manage noisy environments:

Relevant improvements arise by combining robust estimators and neural networks.

Data Mining. A way for mix strategies

Modellization needs to manage non uniform environments:

Mixed approaches are required for helping to produce good quality samples to be used during model building.



Ordieres, J.B., Vergara, E.P., Capuz, R.S., Salazar, R.E. **"Neural network prediction model for fine particulate matter (PM2.5) on the US-Mexico border in El Paso (Texas) and Ciudad Juárez (Chihuahua)".** *Environmental Modelling and Software. doi:10.1016/j.envsoft.2004.03.010.*

UNIVERSIDAD
DE LA RIOJA

App. Data Mining. Table of Contents

- Introduction
- Interest for using DM in Ind. Env.
- Opportunity regarding EU market
- Main concerns
- Dealing with massive data
- Dealing with outliers
- A way for mixed strategies
- Conclusion

App. Data Mining. Conclusion

- Some industrial environments are suitable for improving using DM.
- The cost of data gathering is normally lower than in other fields.
- There are financial support (at least in EU) for such common efforts.
- Real problems make possible, also, to produce scientific research including new algorithms.
- Mixing technologies produces very good results and improves the results of isolated techniques. Currently we are working by combining **Betti's invariants** (from geometry) and clustering strategies in order to add some semantic knowledge by taking shape into account.