The use of Data Mining for the Monitoring and Control of Anaerobic Waste Water Treatment Plants

Julian Gallop & Simon Lambert
Business & Information Technology Department
CCLRC Rutherford Appleton Laboratory, UK

Maurice Dixon & Jerome Healy
Dept of Computing, Communications Technology & Mathematics
London Metropolitan University, UK

Laurent Lardon & Jean-Phillipe Steyer
Laboratoire de Biotechnologie de l'Environnement (LBE) - INRA, France

TELEMAC project

- TELEMAC is a project under the EC 5th Framework IST programme http://www.ercim.org/telemac
  - Overall goal:
    - To improve the efficiency and reliability of operation of anaerobic wastewater treatment plants (WWTP) used for treating waste products of alcoholic beverage production (wine, whisky, tequila, …)
- Multiple approaches to the problem, unified by:
  - Promise of telecontrol
  - Leveraging of knowledge
    - Of experts at telecontrol centre or remote locations
    - From past data

Roles in TELEMAC (1)

- Several sites run WWTPs of different types and scales – at industrial scale:

- e.g. in Spain (32000 litres)

and in Mexico

Roles in TELEMAC (2)

- ***INRA LBE***
  - fully instrumented experimental facility (500 litres)
  - Fault detection, diagnosis and identificatrion
- Other partners run simulations
- Other partners are responsible for software integration
- ***CCLRC RAL and London Metropolitan:***
  - Data mining

Outline

- Introduction – anaerobic digesters
- Sensor depletion
- Linear regression and neural nets
- Cluster analysis
- Decision support
- Summary

The basics of anaerobic digestion (1)

- A high-yield biological process for organic pollutants
- Advantages
  - Rapid
  - Can degrade concentrated and difficult substrates
  - Energy can be recovered (cogeneration)
  - Produces very little sludge
- Drawbacks
  - Unstable
  - The biological/chemical variables are difficult to measure (lack of sensors)
    - An expert is required …
    - … and/or the plant is run at much less than optimal performance.

The basics of anaerobic digestion (2)

Two main steps

- Acidogenesis:

Chemical Oxygen Demand—a measure of water quality

COD $\xrightarrow{\text{Acidogenic bacteria}}$ VFA + CO_2

Fast

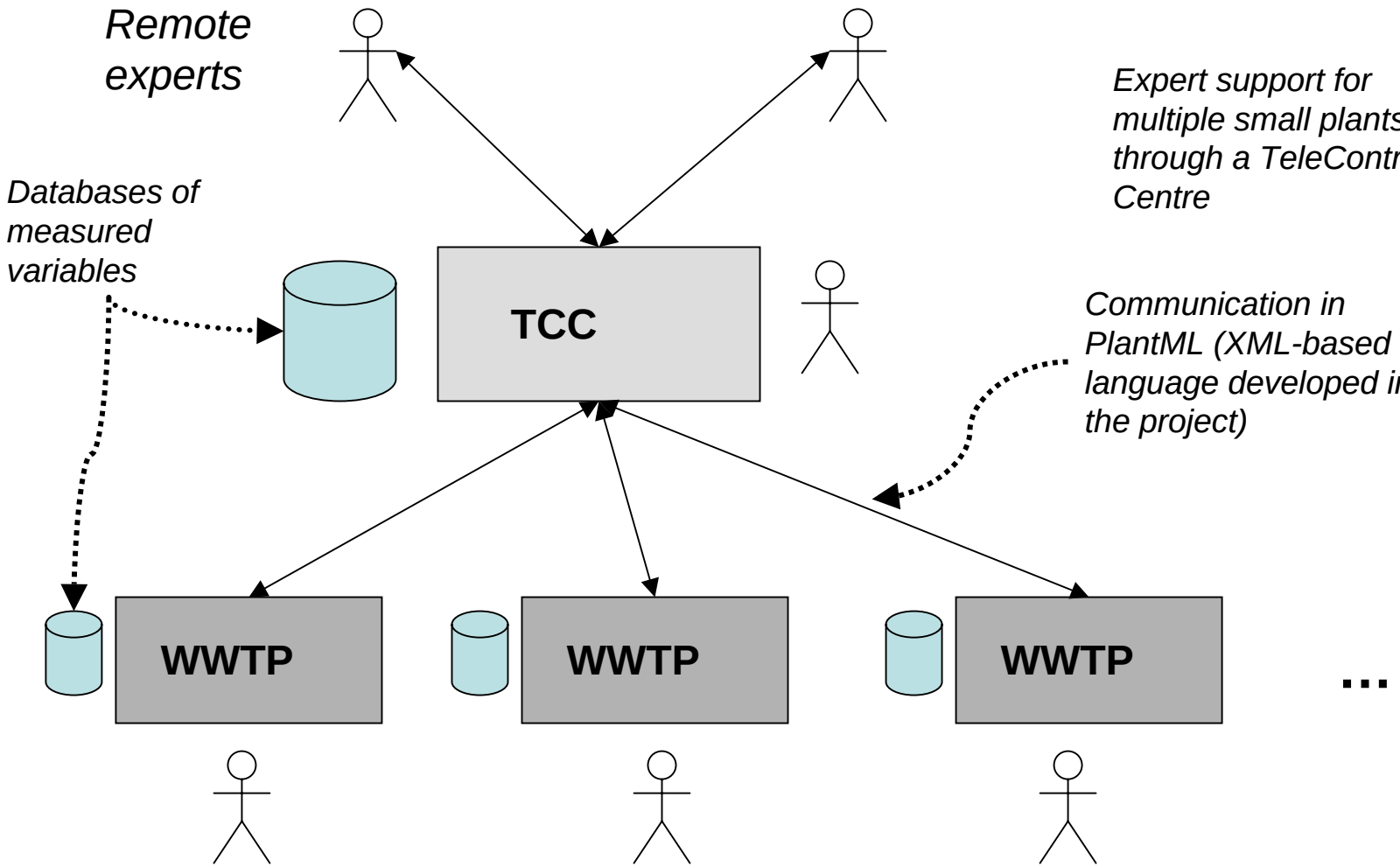
Inhibition

- Methanogenesis:

VFA $\xrightarrow{\text{Methanogenic bacteria}}$ CH_4 + CO_2

Slow

Proposed outline TELEMAC system



Role of data mining in TELEMAC

- Data mining seeks to discover ‘valid, novel, useful and understandable patterns in data’.

- Parallel tracks in TELEMAC
  - Use of linear regression and neural nets
    - To determine sensor relationships
  - Use of clustering and rule induction
    - To strengthen fault diagnosis and identification (FDI) and identify major process states

- Can any of the insights be generalised across plants?

Sensor ranking

- Sensor depletion
  - A production WWTP may have few sensors (expense). The absent sensors are often those most indicative of imminent trouble
  - Sensors may fail

temp dig / ºC ; **q in / (l/hour) ;** **pH dig**	rank 1	Most readily available
CO2 Gas / %age ; **q gas / (l/hour)**	2	
VFA dig / (g/l)	3	
TOC dig / (g/l) ; **COD dig / (g/l)**	4	Least readily available

- Therefore can we use some sensors to emulate others?

Modelling sensors of high rank with sensors of low rank

- Methods used:
  - Linear regression
    - Given a target variable,
    - Identify the variable that fits it best (maximum R^2)
    - Continue to add the next variable that fits best when combined with those already used to the model
  - Creation of a model using neural net architecture
    - One hidden layer

Sensor ranking initial results

- Whether linear regression or neural nets are used, a rank 4 variable can be modelled with variables of ranks 3,2 and 1
- A rank 4 variable:
  - Cannot be well modelled using variables of ranks 2 and 1 using linear regression (R^2=0.29)
  - Can be modelled using variables of ranks 2 and 1 using neural nets (R^2=0.92)
- Similar results are obtained for a rank 3 variable in terms of variables of ranks 2 and 1.

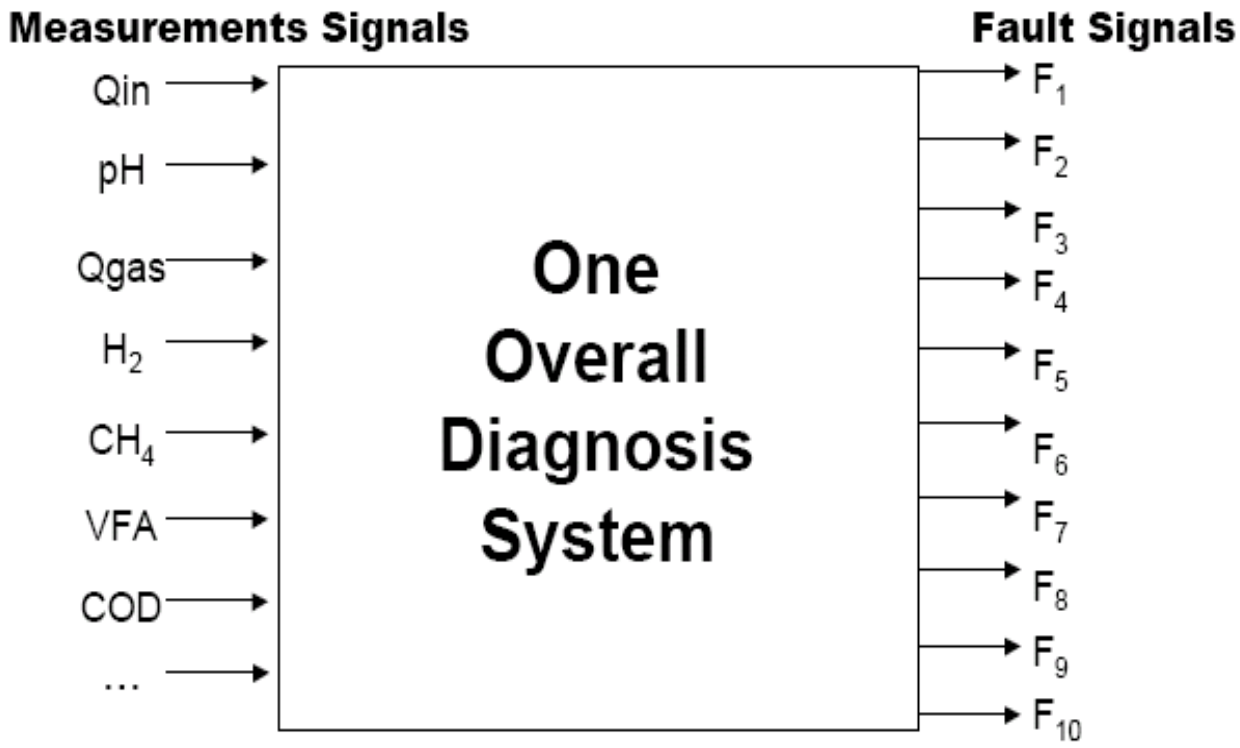
Process states and fault identification

- To guide an operator of a WWTP, it needs to be straightforward but sufficiently informative, e.g. a condition recognisable to the operator such as *organic overload*

- To reach this a modular scheme is used for better traceability.

- So ….

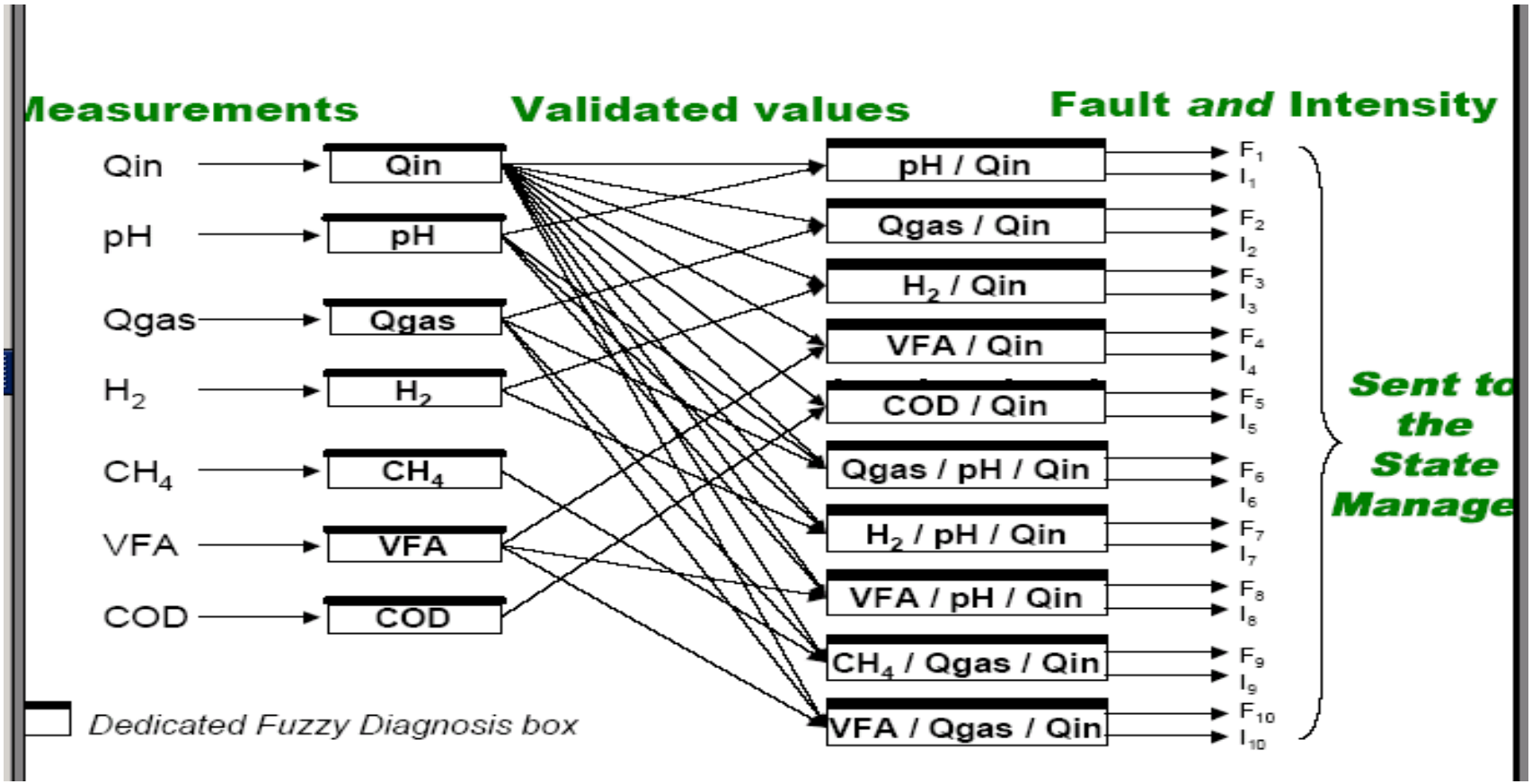
Instead of using a monolithic scheme, such as ……



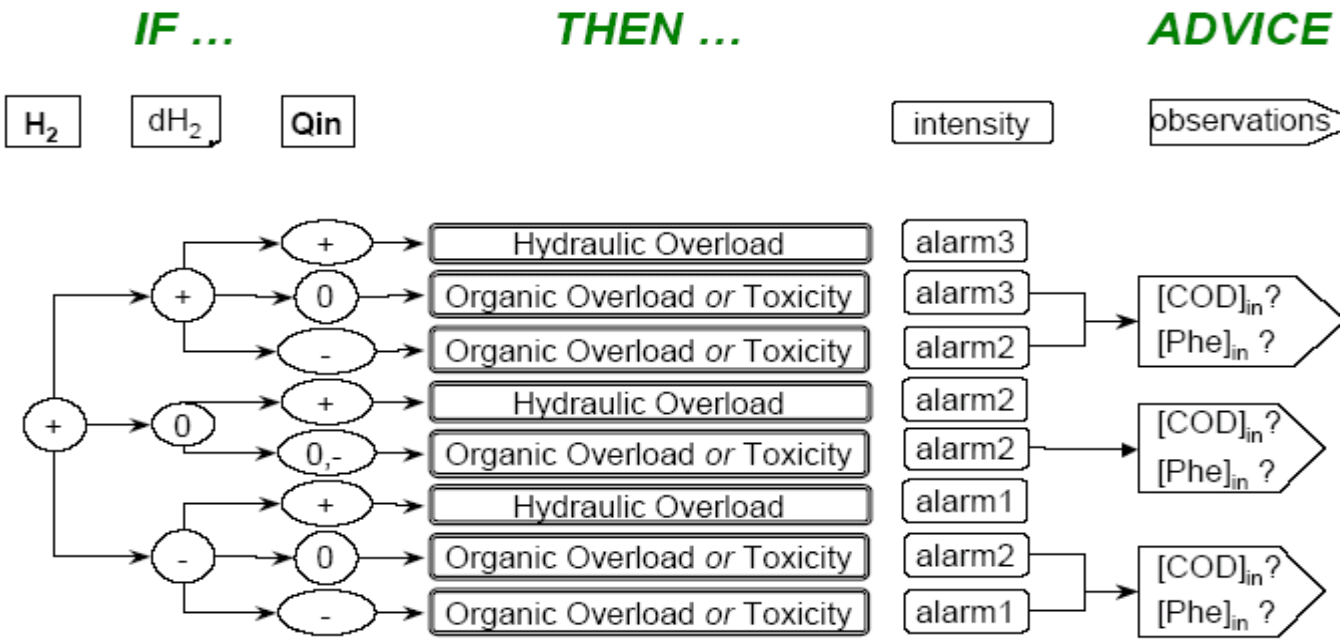
Instead, use a modular scheme

Modular here means that:

• a small number (no more than three) measurements are combined in a single fault calculation

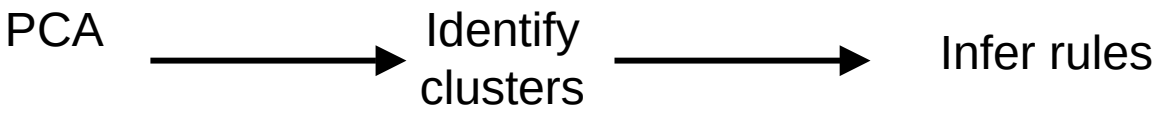


- Measurements are quantised
- Quantised values are used to determine a fault



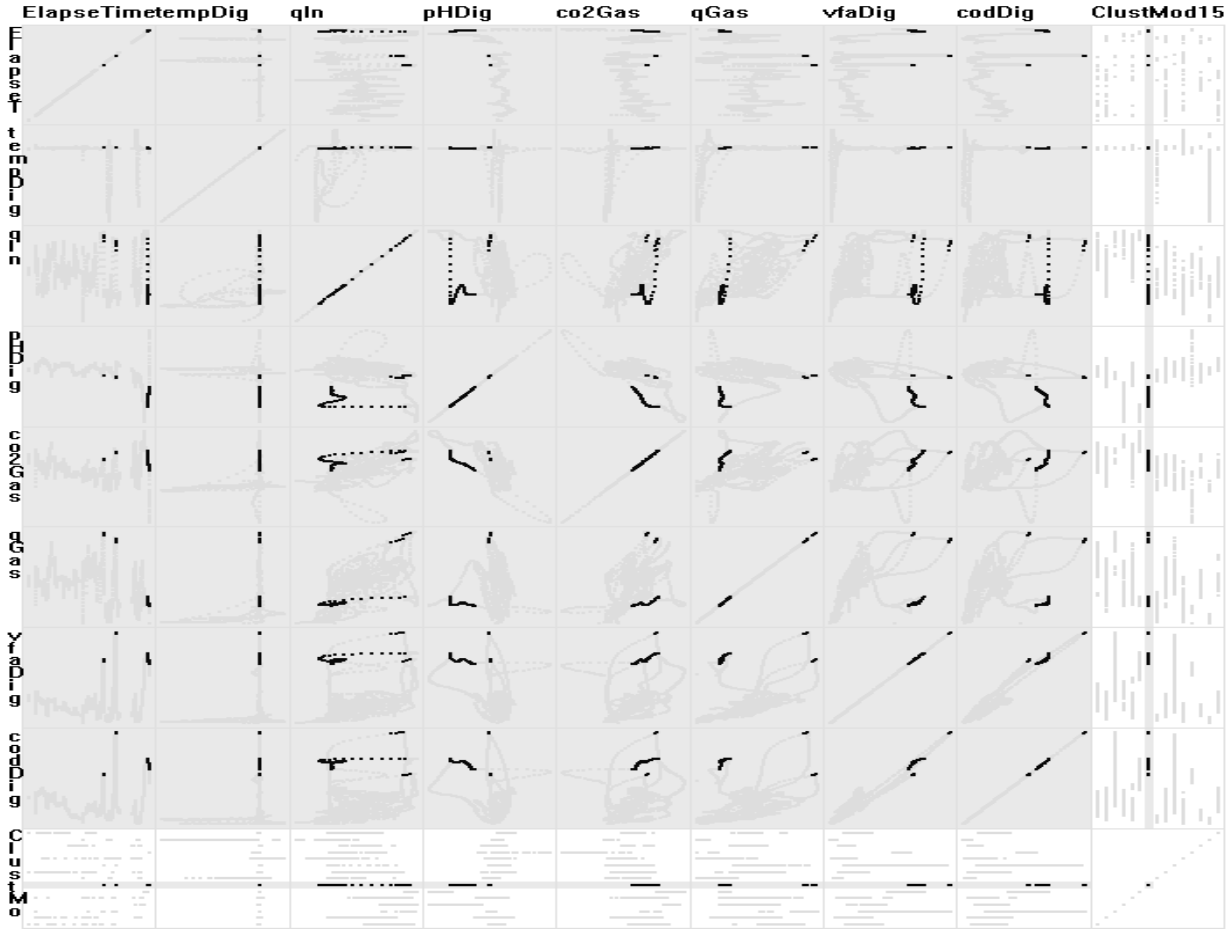
Deriving the rules

- Currently the thresholds and rules are derived by expert knowledge
- Can we improve or strengthen them by data mining
- Since we are concerned with rules, it is natural to consider rule inference.
- But we need coherent groups of data on which to infer rules … so consider cluster analysis
- But we do not want to bias the clusters … so consider Principal Components Analysis.



Multiway plot of time, sensor variables and cluster i.d.

- Some rules can be sketched out by hand, as a confidence test



Cluster properties

- To provide a check on whether the cluster produces reasonable groups of data, we define some properties:
  - Inclusion
    - A cluster analysis algorithm can be directed to produce a given number of clusters. If each of the numerous clusters is largely contained within just one of the less numerous ones, then we can gain some confidence in the stability of the cluster algorithm. In general, we find that with few exceptions, the percentage of inclusion is higher than 85%
  - Compactness
    - We would expect that a cluster should be compact in at least one of the variables and we check that this is the case.
  - Precedence
    - By analysing which cluster precedes another, we can identify the characteristics which appear to lead to a significant event

Use of data mining in decision support (1)

- The TELEMAC system is a complex assembly of interacting components.
- It needs to be customised to a new WWTP.
- Many of the TELEMAC components run in real time, whether at the WWTP itself or the telecontrol centre.
- However, the role of data mining might be more as part of the customisation process …
- … and then possibly occasional recalibrations when needed.
  - For example, if there is a major change in the operating regime of the plant.

Use of data mining in decision support (2)

- The TELEMAC integrated system for a typical plant

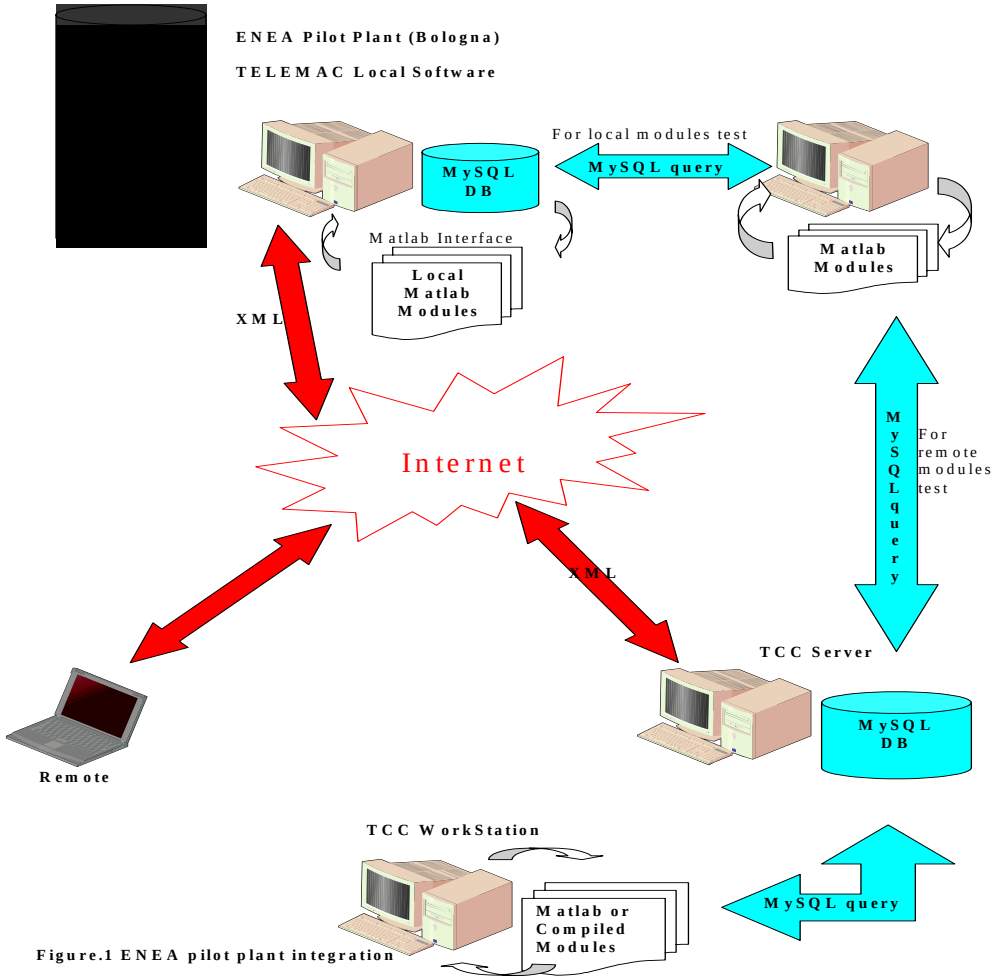


Figure.1 ENEA pilot plant integration

Summary

- The work on linear regression and neural nets is producing results, which show which sensors are needed in order to model other chemical variables
- Cluster analysis and rule induction are likely to lead to a useful characterisation of process state and training and validation will be carried out.
- Based on some work on fault diagnosis and identification, it appears that some results could be generalised to other digesters, with different thresholds, but this needs further work