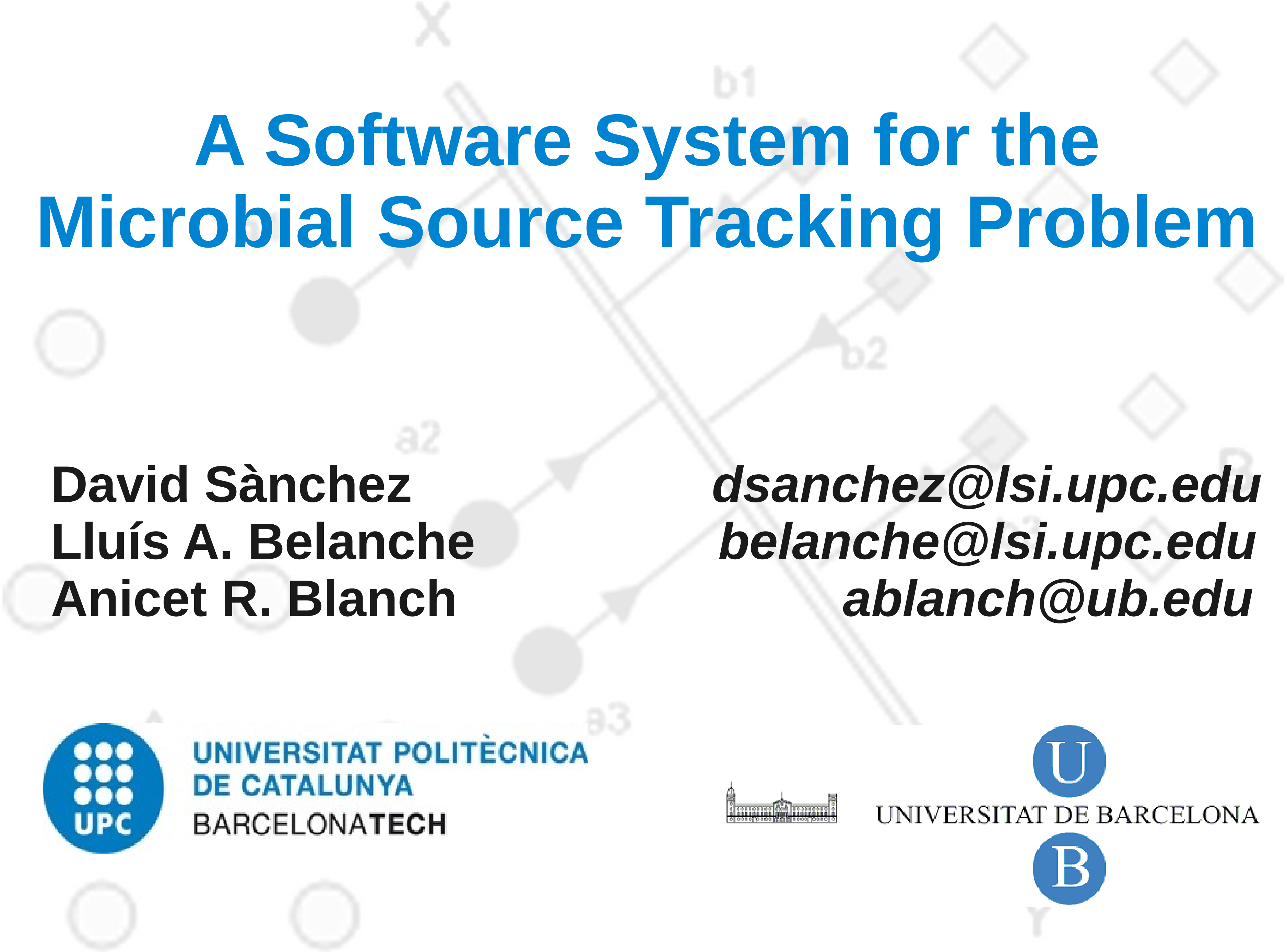
A Software System for the Microbial Source Tracking Problem

David Sànchez dsanchez@lsi.upc.edu
Lluís A. Belanche belanche@lsi.upc.edu
Anicet R. Blanch ablanch@ub.edu

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH

UNIVERSITAT DE BARCELONA



Contents

- What is the Microbial Source Tracking Problem? Why is it important?
- Our Contribution.
- Problems, challenges and solutions.
- System overview.
- System validation.
- Conclusions and further research.

Microbial Source Tracking

- Determination of the origin of faecal pollution in water by the use of microbial or chemical indicators.
- Faecal pollution in water is one of the main causes of health problems in the world.
- Scientific term: models should use a minimum number of variables.
- Legal term: who should clean polluted waters?

Our Contribution

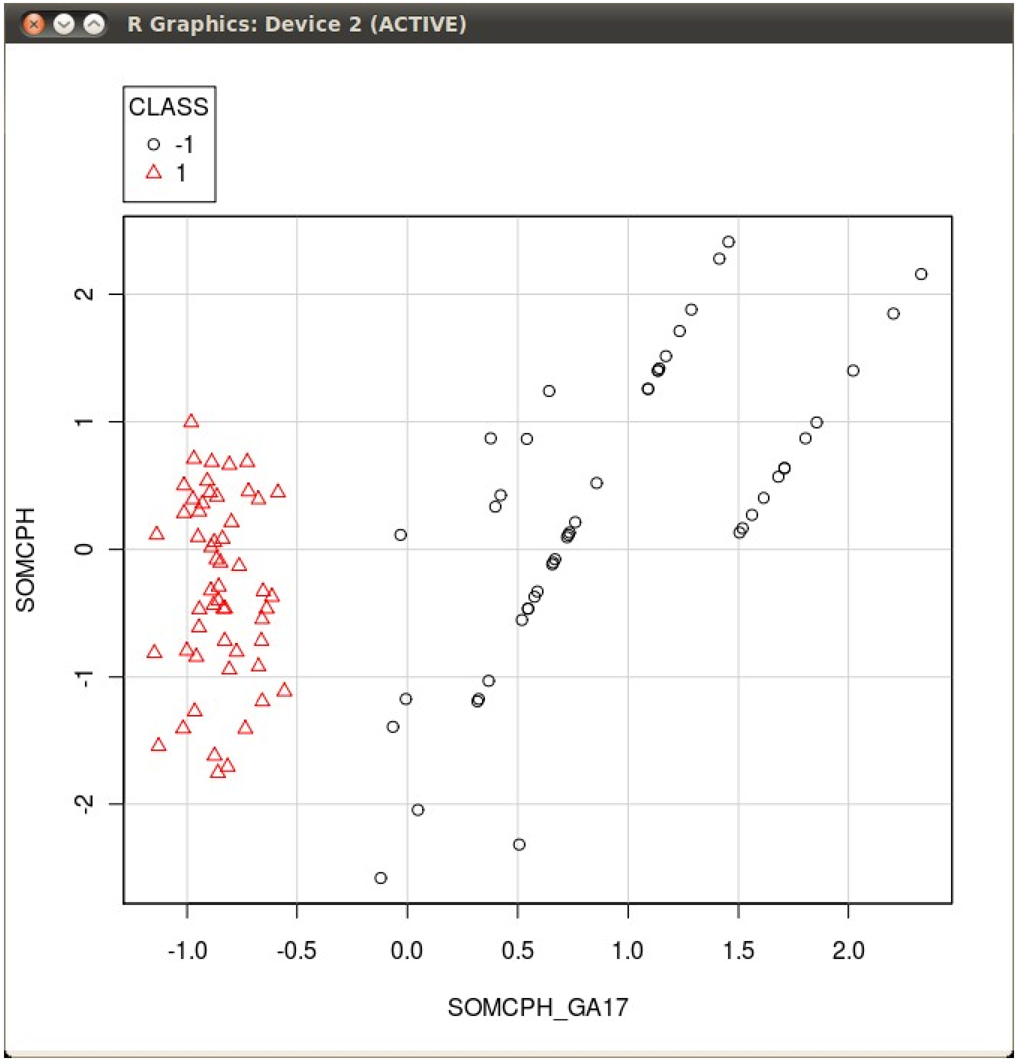
- Already **solved** MST instance assumes data is expressed at **point source**.
- Our system makes no assumption about it, thus, system accepts:
  - Examples showing different concentrations levels (**dilution**)
  - Examples with different environmental persistence (**ageing**)
- Dilution and ageing are independent processes.

How do dilution and ageing behave?

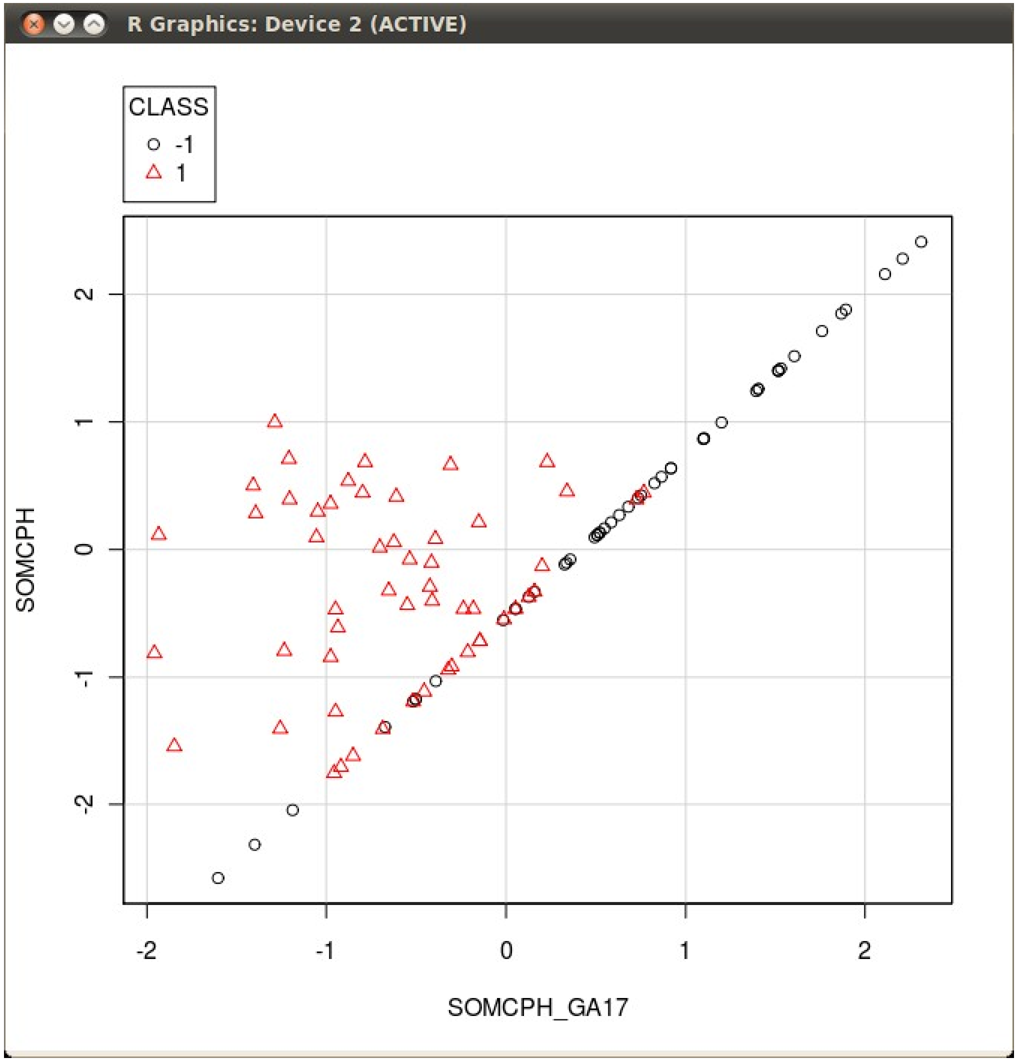
- **Dilution**: A dilution factor of **d** represents that the theoretical value is divided by **d**.
  - If diluted value falls below attribute threshold the value will be constant and equal to the detection threshold.

- **Ageing**: Distinct variables follow different ageing processes.
  - Only empirical measurements at different stipulated time are available for some indicators.

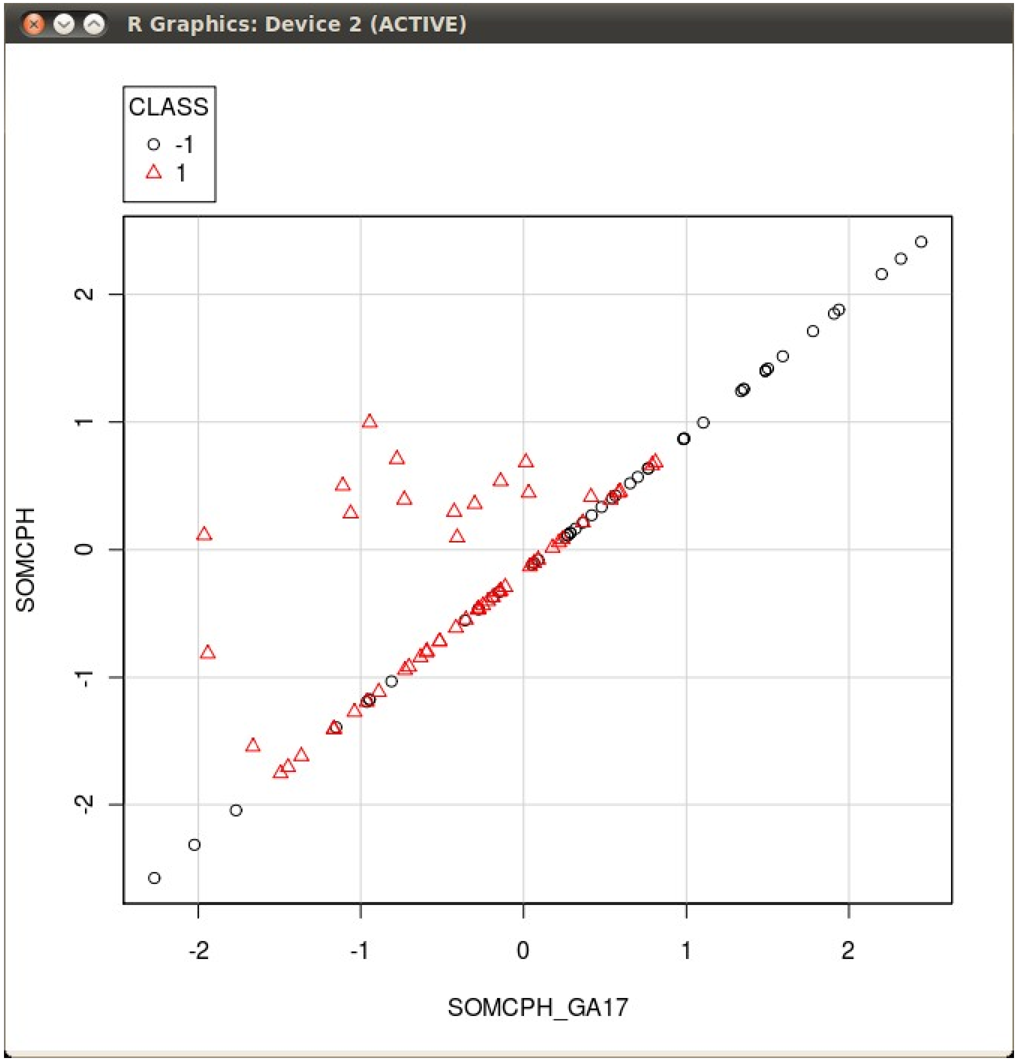
Dilution and Ageing Effects (1)



Dilution and Ageing Effects (2)



Dilution and Ageing Effects (3)



Problems, Challenges and Solutions (1)

- Starting point: 103 examples by 26 indicators.
- Why not an straight solution?
  - Examples in the data matrix are expressed at **point source** (no dilution).
  - Examples in the data matrix are expressed at **zero-time** (no ageing).
  - Data matrix should be regarded as maximal, only a fraction on indicators will be supplied on prediction.

Problems, Challenges and Solutions (2)

- Consider a set of empirical measurement on one indicator:

$$S_\alpha = \{(x_1, \log_{10}(y_1/\alpha)), \ldots, (x_n, \log_{10}(y_n/\alpha))\}$$

- Consider its regression:

$$f_\alpha(x) = ax + b - \log_{10}(\alpha)$$

- Consider the theoretical representation of a supplied indicator:

$$\log_{10}(\frac{\widetilde{v_i}}{\alpha}) + a_i t = v_i$$

- If we subtract two of the equations we arrive at:

$$(a_i - a_j)t + \log_{10}(\widetilde{v_i}) - \log_{10}(\widetilde{v_j}) = v_i - v_j$$

Problems, Challenges and Solutions (3)

- Once an estimation for the elapsed time is known, an estimation for the dilution factor can be obtained by:

$$\log_{10}(\alpha^*) = t^* a_i + b_i - v_i$$

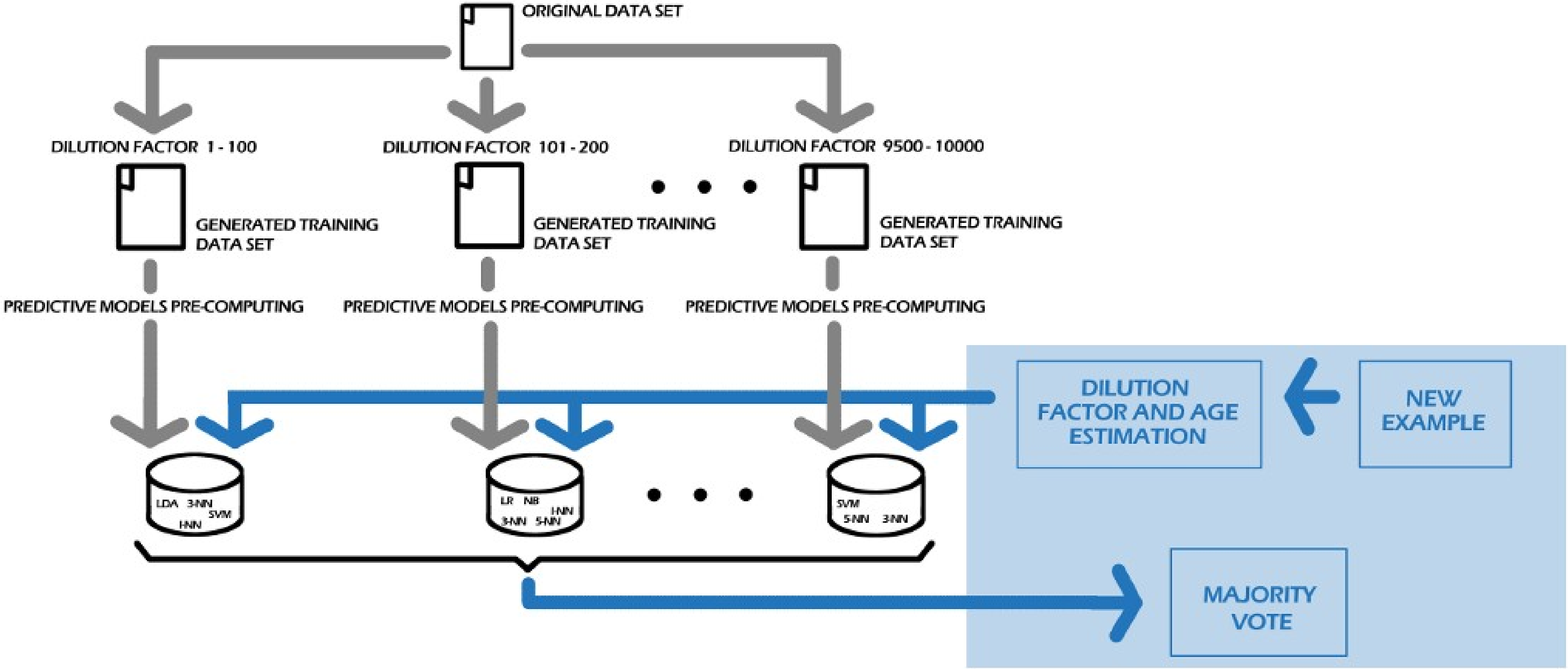
- Reversing time on whole example is also possible by using:

$$V_i\colon \widehat{v_i} = v_i - a_i t^*,\ 1 \leq i \leq N$$

Problems, Challenges and Solutions (4)

- Only a **varying fraction** of indicators will be supplied:
  - Best subsets of variables will depend on dilution factor and age.
  - Independents training processes for different values of equidistant dilution factors in range [1,500]
  - For each subinterval **d**: data matrix is diluted to **d**, different models are developed using this diluted matrix.
  - All possible 2 and 3-sized combinations of indicators.
- Result: sets of models trained to respond to different ranges of dilution.

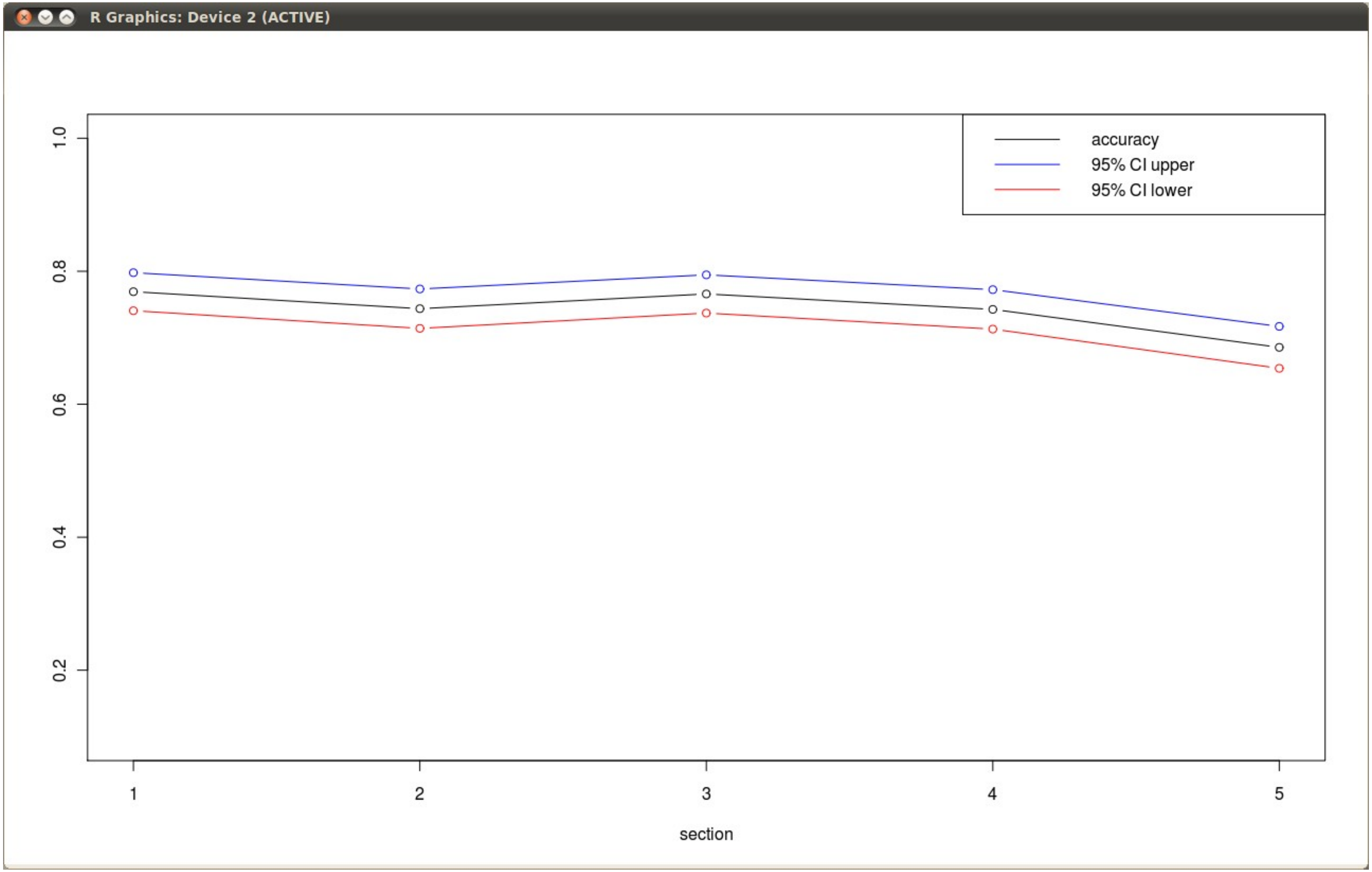
System Overview


ORIGINAL DATA SET
DILUTION FACTOR 1 - 100
DILUTION FACTOR 101 - 200
DILUTION FACTOR 9500 - 10000
GENERATED TRAINING DATA SET
GENERATED TRAINING DATA SET
GENERATED TRAINING DATA SET
PREDICTIVE MODELS PRE-COMPUTING
PREDICTIVE MODELS PRE-COMPUTING
PREDICTIVE MODELS PRE-COMPUTING
DILUTION FACTOR AND AGE ESTIMATION
NEW EXAMPLE
LDA 3-NN SVM 1-NN
LR NB 1-NN 3-NN 5-NN
SVM 5-NN 3-NN
MAJORITY VOTE

System Validation (1)

- Test set generated from original matrix:
  - aged from 0 to 150 hours.
  - diluted by a factor up to 500.
- Prediction accuracy depends of:
  - number and composition of indicators.
  - true dilution and age of the example.
- Estimated performance: 75 – 80% correct classification.
- Promising results due to:
  - majority class has probability 52.4%
  - great number of approximations and estimations system does.

System Validation (2)



Conclusions and Further Research

- ICHNAEA: a prototype computer-based system for predictions on MST.
  - System can be trained by user with their own data.
  - Accuracy and prediction precision is given, as well as the estimated degree of dilution and age for the analysed example.
  - Complementary MST indicators are suggested to improve MST prediction confidence.
- Analysing the presence of several distinct animal species (multi-class problem).
- Providing posterior probabilities for each class.

Thank you very much!

David Sànchez
Lluís A. Belanche
Anicet R. Blanch

dsanchez@lsi.upc.edu
belanche@lsi.upc.edu
ablanch@ub.edu

UPC
UNIVERSITAT POLITÈCNICA
DE CATALUNYA
BARCELONATECH

UB
UNIVERSITAT DE BARCELONA