

REFRAME Project Presentation



<http://www.reframe-d2k.org/>

Rethinking the **E**ssence, **F**lexibility and **R**eusability of **A**dvanced **M**odel **E**xploitation

Cèsar Ferri
(Universitat Politècnica de València, Spain)

Istanbul, 5 March 2014

Outline

- ▶ Project information
- ▶ Scientific background
- ▶ Challenges and potential impact
- ▶ Workplan and working teams
- ▶ Activities Developed and results
- ▶ Management
- ▶ Dissemination
- ▶ Conclusions

Project title and dates

- ▶ REFRAME project
 - Rethinking the Essence, Flexibility and Reusability of Advanced Model Exploitation
- ▶ Dates:
 - START:
 - Officially: 1st October 2012 (kick-off meeting, Bristol)
 - Practically: 1st April 2013
 - END:
 - 31 March 2016 (30 months + 6 months extension granted)

Consortium

▶ BRIS: University of Bristol / Intelligent Systems Laboratory

- Peter Flach, Meelis Kull, Reem Al-Otaibi
- Provides the **smart electricity meters domain**.



▶ STRAS: University of Strasbourg / LSIIT

- Nicolas Lachiche, Agnes Braud, Chowdhury Farhan Ahmed, Clement Charnay
- Provides the **geographical domain**.



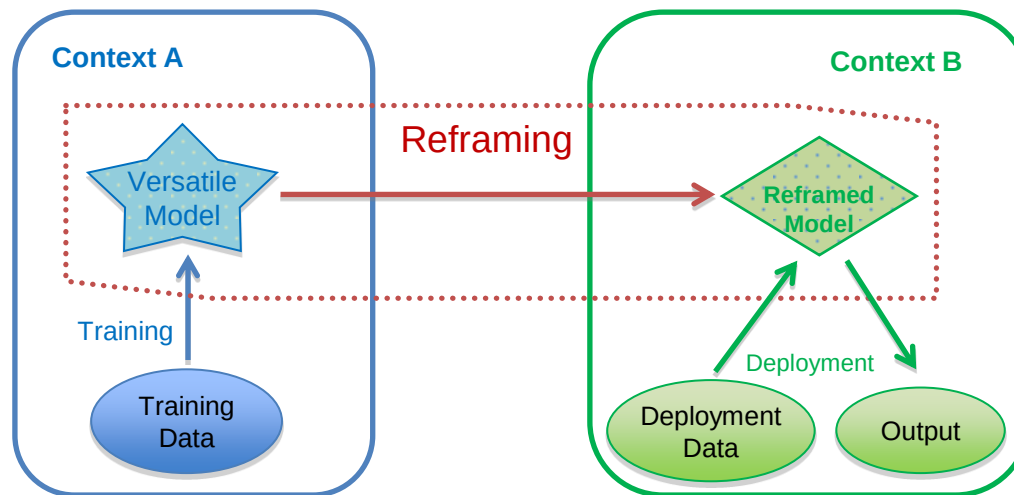
▶ VAL: Universitat Politècnica de València / DSIC

- Jose Hernandez-Orallo, Cesar Ferri, Maria-Jose Ramirez-Quintana
- Provides the **human genomics domain**.



Scientific background (1)

- ▶ What is “reframing”?
 - “process of applying an existing [machine learning/data mining] model to the new operating context by the proper transformation of inputs, outputs and patterns”.



Scientific background (2)

► Examples:

- “Model predicting sales in Strasbourg for the following week may fail in Bristol for next Wednesday. The operating context has changed in terms of location as well as resolution of the inputs and outputs.”
 - Use a general model that can be applied at different locations and resolutions.
- “20% of the population will have an uncommon disease in the following ten years. The concept and frequency of *uncommon disease* changes.”
 - A general model is defined in terms of the background knowledge (e.g., “uncommon disease”).
 - If the background knowledge changes, the model predictions change.

Scientific background (3)

► What is a “context”?

◦ Examples:

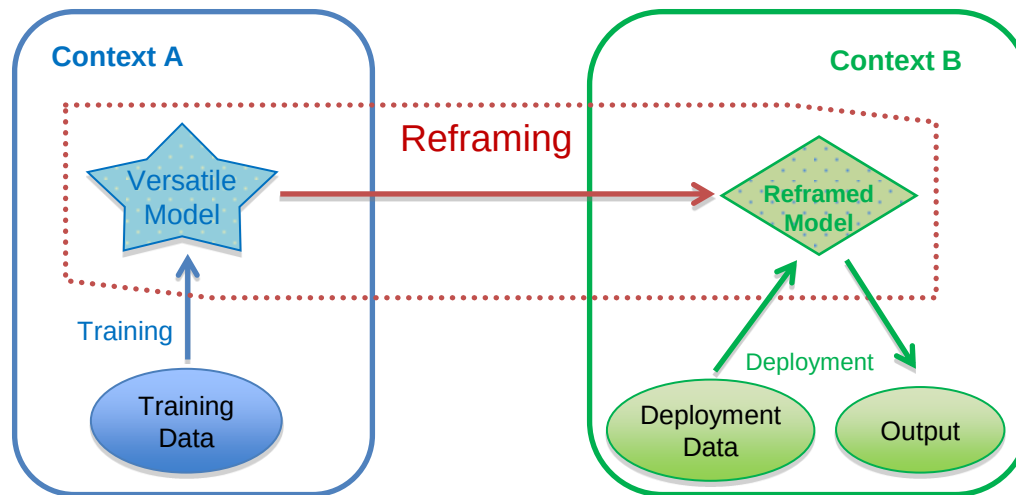
1. Costs (output, inputs, labelling, etc.).
2. Changes in data distribution.
 - Both 1 and 2 are integrated in tools such as ROC analysis.
3. Changes and incompleteness in background knowledge.
 - A model can depend on the background knowledge.
4. Granularity of feature (input) space or output space.
 - Multi-dimensional approach. Input features and output can be hierarchical.
5. More or fewer (or different) attributes, missing data.
6. Constraints

◦ One of the goals of the project is to generalise the notion of “operating context”.

- So this list could become much larger in the future.

Scientific background (4)

► “Versatile models”:



- The key issue is the construction of a “**versatile model**”, either directly or by “enrichment” of an existing model.
 - The versatile model may be more general (in structure, feature handling, output handling, knowledge dependencies, etc.) than really needed for context A.

Challenges and potential impact (1)

- ▶ Usual D2K process:
 - Models are trained in a context.
 - This commonly results in overfitting to the training context.
 - Restricts model applicability whenever the training conditions change.
 - Models must be discarded and retrained repeatedly.

Inefficient and unreliable process.

Challenges and potential impact (2)

- ▶ We challenge that process:
 - We do not want to build a set of models for a range of operating contexts, or a very general, but inflexible model.
 - We aim at building **versatile models** that can be properly deployed in a range of operating contexts.

Any advance in this regard constitutes an important innovation and can have a strong impact on the way models are trained and deployed.

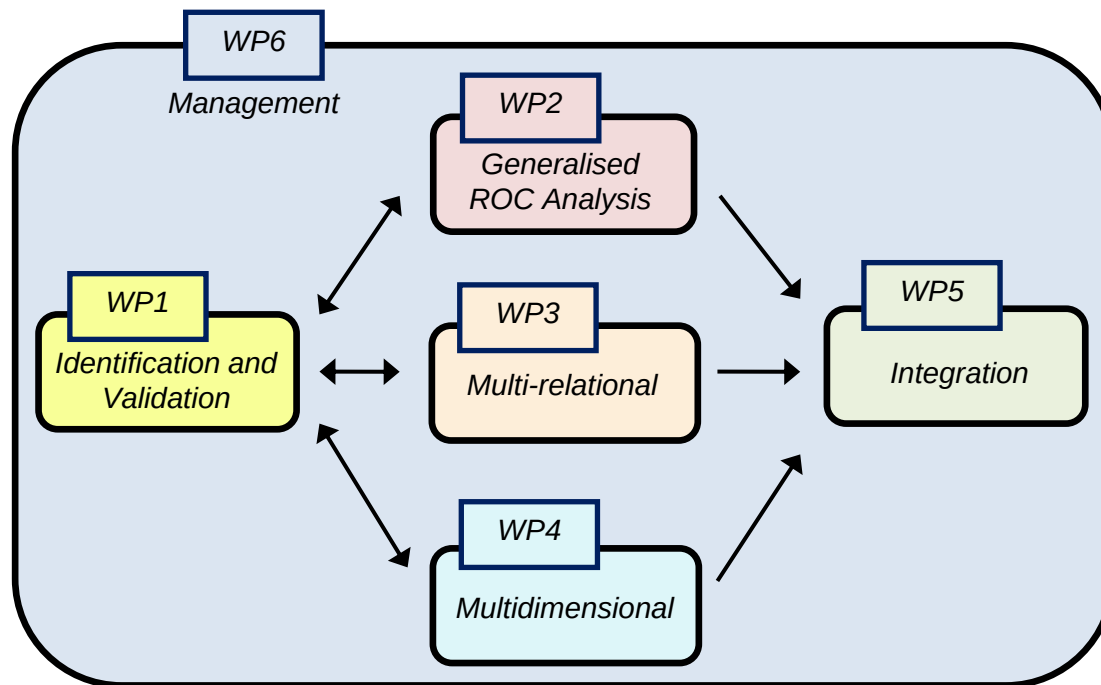
Challenges and potential impact (3)

- ▶ The purpose is to ensure “model reuse”.
 - Not the reuse of parts of old, existing models
 - The use of a general, versatile model for each possible context
 - Appropriate reframing procedures to apply it to any possible context.
 - Because of its long-life, the models can be robust and implemented in a more resource-efficient way.

Workplan (1)

► Overview

- Six work packages:



Working Teams

- ▶ WT-A) Operating context
- ▶ WT-B) Attribute costs
- ▶ WT-C) Cardinality, complex aggregates and continuous input attribute
- ▶ WT-D) Multidimensional schemas for the domain
- ▶ WT-E) Reframing for regression
- ▶ WT-F) Multi-class reframing
- ▶ WT-G) Multi-label classification with decision trees
- ▶ WT-H) Reframing review/survey paper
- ▶ WT-I) Multidimensional versatile models, aggregation and disaggregation

Activities Developed(1)

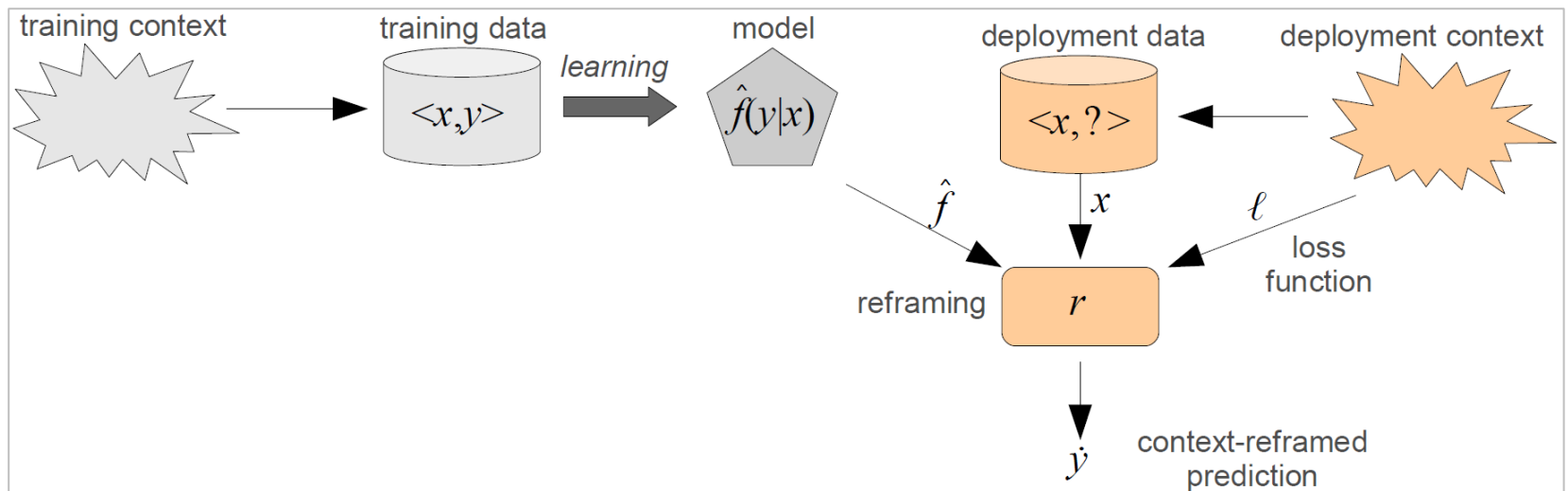
- ▶ WP1) Identification of context changes and solution validation. (WT-A,H)
 - Each partner collected information from its domain
 - Refine the terminology and different kinds of operating context changes.
 - It motivated a broader survey of the relationships with related topics such as transfer learning, domain adaptation
- ▶ WP2) Generalised ROC Analysis for Model Reframing. (WT-B,E,F,G)
 - We have explored operating conditions on non-binary target variables
 - Operating conditions on input variables have been analysed
- ▶ WP3) Reframing in the Multi-Relational Setting with Background Knowledge (WT-C):
 - The work on cardinality is in progress.
 - We addressed three aspects of complex aggregates:
 - Incremental generation
 - Complex aggregates seen as continuous input attributes
 - Reframing of continuous input attributes

Activities Developed (2)

- ▶ WP4) Hierarchical and Multidimensional Reframing (WT-D,I):
 - Each partner provided a multidimensional schema for its domain
 - WP affected by the delay of Spanish funding that prevented hiring a third post-doc
- ▶ WP5) Integration:
 - Second part of the project
- ▶ WP6) Management:
 - A wiki has been set up: <http://reframe-d2k.org>
 - 4 plenary meetings
 - Submitted a workshop proposal on reframing and related topics

Results(1)

► Reframing framework



Results(2)

- ▶ Exploring classification decision thresholds
 - A model can behave differently if the threshold is fixed, uniform or depends on the operating context.
 - The expected loss of a model depends on the decision rule that is used and the contextual information
 - Interesting results about relations between different evaluation metrics
 - New graphical representations:
 - Brier Curves, Rate-driven Curves and Kendall Curves.
- ▶ Partial information about the rate context
 - Expected loss in partial regions.
 - Partial area for rate-driven cost curves and derived a rate-weighted AUC

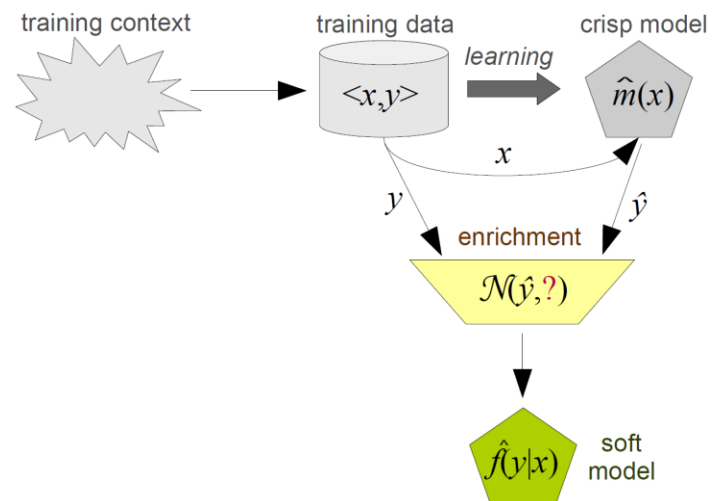
Results(3)

- ▶ Reframing in regression:
 - New techniques for analysing the quality of a regression model when this operating condition can change.
 - The notion of *shift* (a constant that can be added or subtracted to each example):
 - New ROC analysis for regression, whose area has been shown to be the error variance.
 - Reunderstading each model prediction as the output of a probability distribution:
 - Regression models are much more versatile and can be reframed in more different ways.
 - We can apply them for different types of contexts: asymmetric costs, reject rules and auction bids

Results(4)

► Model enrichment

- We have explored the creation of models that capture more information than initially required in the original context.
- We have also analysed the possibility of taking a not versatile model and try to enrich the model and make it versatile
 - Calibration, reunderstanding each model prediction as the output of a probability distribution



Results(5)

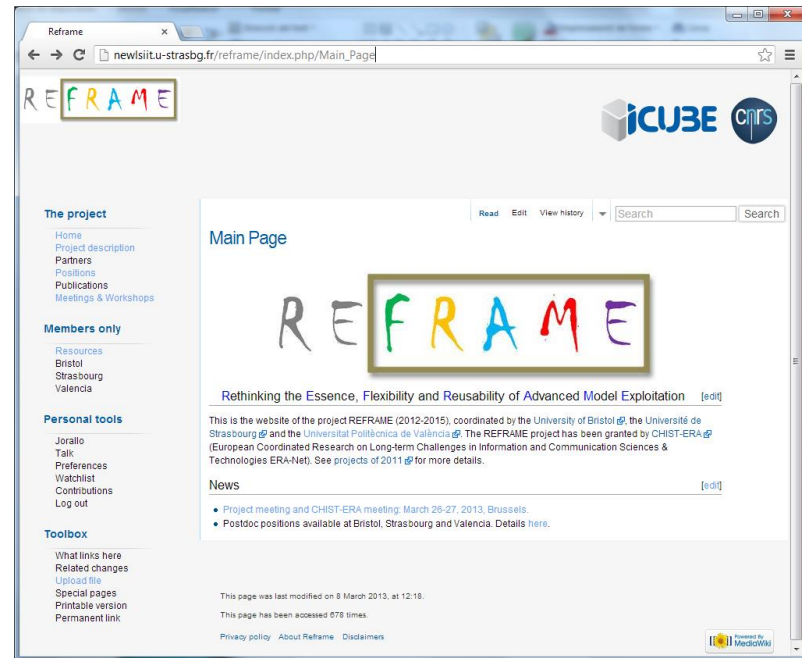
- ▶ **Multi-label classification:**
 - Developed a hybrid multi-label decision tree using a label covariance splitting criterion.
 - It can be refine predictions if the covariance matrices are kept and new contextual information is given
- ▶ **Multiclass classification:**
 - New approach based on pairwise classification where threshold optimisation is performed for each of the pairwise classifiers.
 - This optimisation can be performed when the context description changes.
 - A different approach based on the idea of estimating the reliability of the probability estimations for each instance.
 - This information can be used to construct a distribution for each example and make the decisions that minimise a given loss function.

Results(6)

- ▶ **Complex aggregates**
 - Novel idea where relational features go beyond existential quantifier and simple aggregates and can combine conditions and aggregation functions.
 - Models must then be modified in order to account for this versatility.
- ▶ **Aggregative quantification**
 - First methodology for regression quantification
 - Many classification problems are ultimately regression quantification problems when the number (or proportion) of positive cases are aggregated.
 - We focus on quantifiers that output a single value or a distribution (more versatile).

Management (1)

- ▶ Website/Wiki:
<http://www.reframe-d2k.org/>



- ▶ Communication:
 - Project mailing lists (internal and external)
 - Videoconferences. One Google hang-out per week
 - Sharing data: Dropbox and Google Drive

Management (2)

- ▶ Internal project meetings:
 - Bristol, October 2012 (Kick-off)
 - Brussels, March 2013
 - Valencia, June 2013
 - Strasbourg, October 2013
 - Valencia, March 2014
- ▶ Research stays
- ▶ Consortium agreement

Dissemination (1)

► Publications

◦ Publications to date:

- "A Unified View of Performance Metrics: Translating Threshold Choice into Expected Classification Loss" Journal of Machine Learning Research, 2012.
- "ROC Curves in Cost Space" Machine Learning Journal, 2013
- "Aggregative Quantification for Regression", Data Mining and Knowledge Discovery, 2014.

Dissemination (2)

► Publications

◦ Publications to date:

- "Probabilistic reframing for cost-sensitive regression", ACM Transactions on Knowledge Discovery from Data, 2014.
- "ROC Curves for Regression", Pattern Recognition, 2013.
- "Incremental Construction of Complex Aggregates: Counting over a Secondary Table", International Conference on Inductive Logic Programming (ILP), 2013.
- "Pairwise Optimization of Bayesian Classifiers for Multi-Class Cost-Sensitive Learning", IEEE International Conference on Tools with Artificial Intelligence (ICTAI), 2013.

Dissemination (3)

► Publications

◦ Publications submitted:

- “Learning Hybrid Multi-Label Decision Trees using Label Covariance as Splitting Criterion”, AAAI Conference on Artificial Intelligence (AAAI).
- “Reframing Continuous Input Attributes”, AAAI Conference on Artificial Intelligence (AAAI).
- “Learning and exploiting the reliability of instance-wise probability estimates in multi-class classification”, International Conference on Machine Learning (ICML).
- “Rate-constrained ranking and the rate-weighted AUC”, International Conference on Machine Learning (ICML).

Dissemination (4)

- ▶ Workshop on "Model Generalisation and Reuse over Multiple Contexts"
 - To be held during one of the forthcoming international or European machine learning conferences.

Conclusions

- ▶ Project as a short-term goal.
- ▶ More ambitious long-term goals:
 - Establish a solid and stable research effort to re-interpret the whole process of knowledge discovery from data.
 - Consolidate a research cluster in Europe as a catalyst for the area of machine learning evaluation and reuse worldwide.
- ▶ Dissemination and obtaining additional funding is key to build this cluster and make it sustainable.