# REFRAME Project Presentation


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

Rethinking the Essence, Flexibility and Reusability of Advanced Model Exploitation

Peter Flach (University of Bristol, UK),
with input from José Hernández-Orallo
(Universitat Politècnica de València, Spain)

Brussels, 27 March 2013

# Outline

- Project title and dates
- Scientific background
- Challenges and potential impact
- Consortium
- Workplan
- Management
- Results and dissemination
- Sustainability / valorisation

# Project title and dates

- REFRAME project
  - **R**ethinking the **E**ssence, **F**lexibility and **R**eusability of **A**dvanced **M**odel **E**xploitation
- Dates:
  - START:
    - **Officially: 1st October 2012** (kick-off meeting, Bristol)
    - **Practically: 1st April 2013**
  - END:
    - **31 March 2016** (30 months + 6 months extension granted)

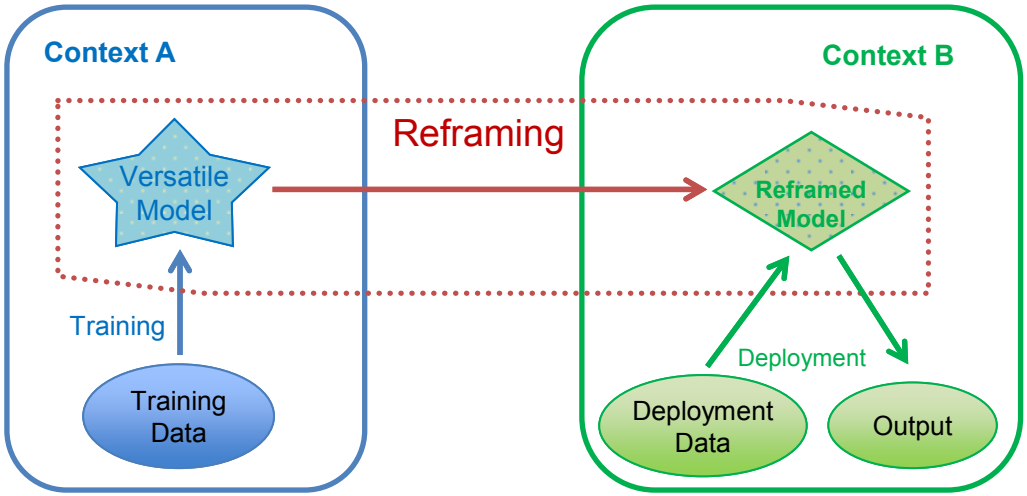


chist-era

REFRAME

# 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”.






chist-era

REFRAME

# 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.



chist-era

REFRAME

# 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.

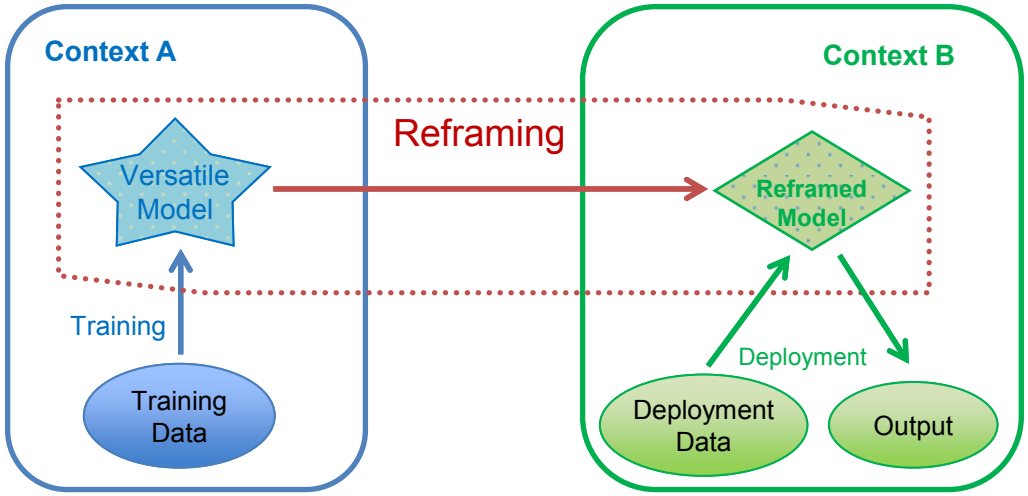


chist-era

REFRAME

# 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.



chist-era

REFRAME

# Scientific background (5)

- Relevance to the topics addressed in the call
  - topic 3: generic models and systems for processing highly heterogeneous data, especially involving different levels or scales
  - topic 4: systems able to know when they don't know and dynamically cope with unpredicted input data
  - topic 6: generic methodologies, tools and formats to ease the exchange of data and models



chist-era

REFRAME

# 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.



chist-era

REFRAME

# Challenges and potential impact (3)

- The purpose is to ensure “model reuse”.
  - Not the reuse of parts of old, existing models, but the use of a general, versatile model for each possible context, along with appropriate reframing procedures to apply it to any possible context.
    - The versatile model is validated and enhanced/refined/enriched by its application over several contexts.
      - Because of its long-life, the models can be robust, validated and implemented in a more resource-efficient way.
      - Contexts are where the model is deployed, but also where the model learns how it can be further generalised.



chist-era

REFRAME

# Consortium (1)

- BRIS: University of Bristol / Intelligent Systems Laboratory
  - Peter Flach, Meelis Kull
  - An internationally renowned centre of excellence, performing leading research in machine learning and data mining and developing applications to web intelligence, bioinformatics, semantic image analysis, and other areas.
    - Overall project coordinator.
    - Focus on generalised ROC analysis, relational data mining for reframing, and management.
    - Provides the **smart electricity meters domain**.

chist-era

REFRAME

# Consortium (2)

- STRAS: University of Strasbourg / LSIIT
  - Nicolas Lachiche, Agnes Braud
  - Integrates experts in relational data mining, clustering and geography, and its applications to real application domains, CRM, chemistry, ...
    - In charge of WP1 (application domains), and multi-relational and background knowledge aspects.
    - Involved in many other tasks and work packages of the project, including dissemination tasks.
    - Provides the **geographical domain**.



chist-era

REFRAME

# Consortium (3)

- VAL: Universitat Politècnica de València / DSIC
  - Jose Hernandez-Orallo, Cesar Ferri, Maria-Jose Ramirez-Quintana
  - Expertise in the general area of data mining: classification, model combination, hierarchical clustering, regression, inductive (logic) programming, decision tree learning, calibration, data warehousing, ROC analysis, quantification, etc.
    - Lead package on multidimensional data.
    - Active role in the other packages.
    - Provides the **human genomics domain**.

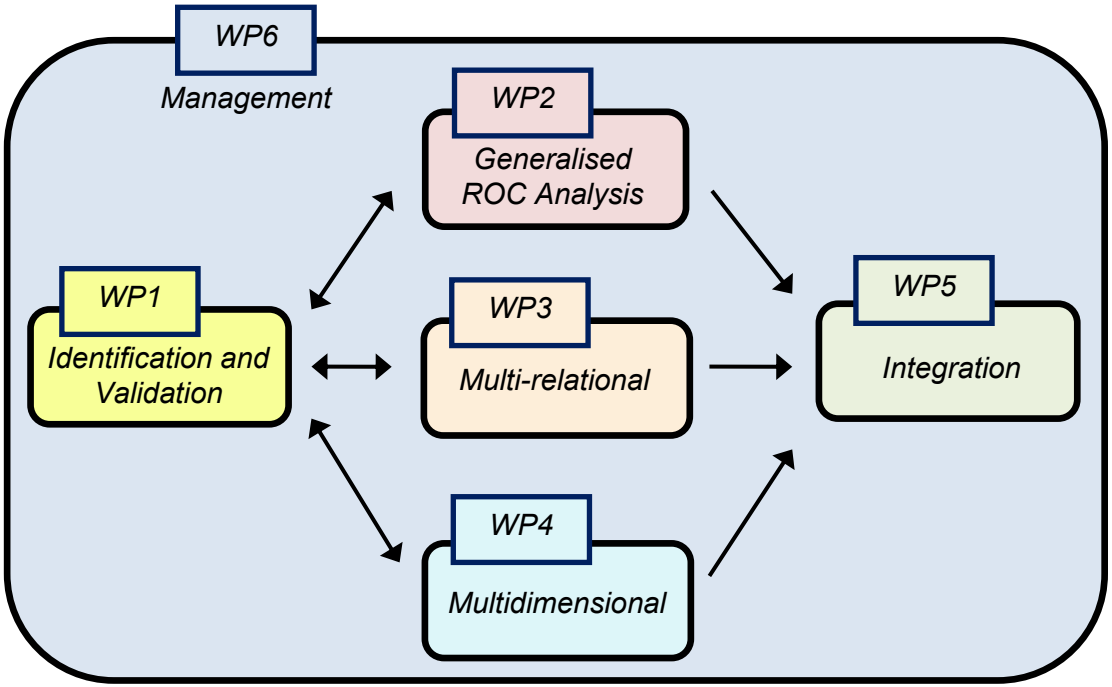


chist-era

REFRAME

# Workplan (1)

- Overview
  - Six work packages:



# Workplan (2)

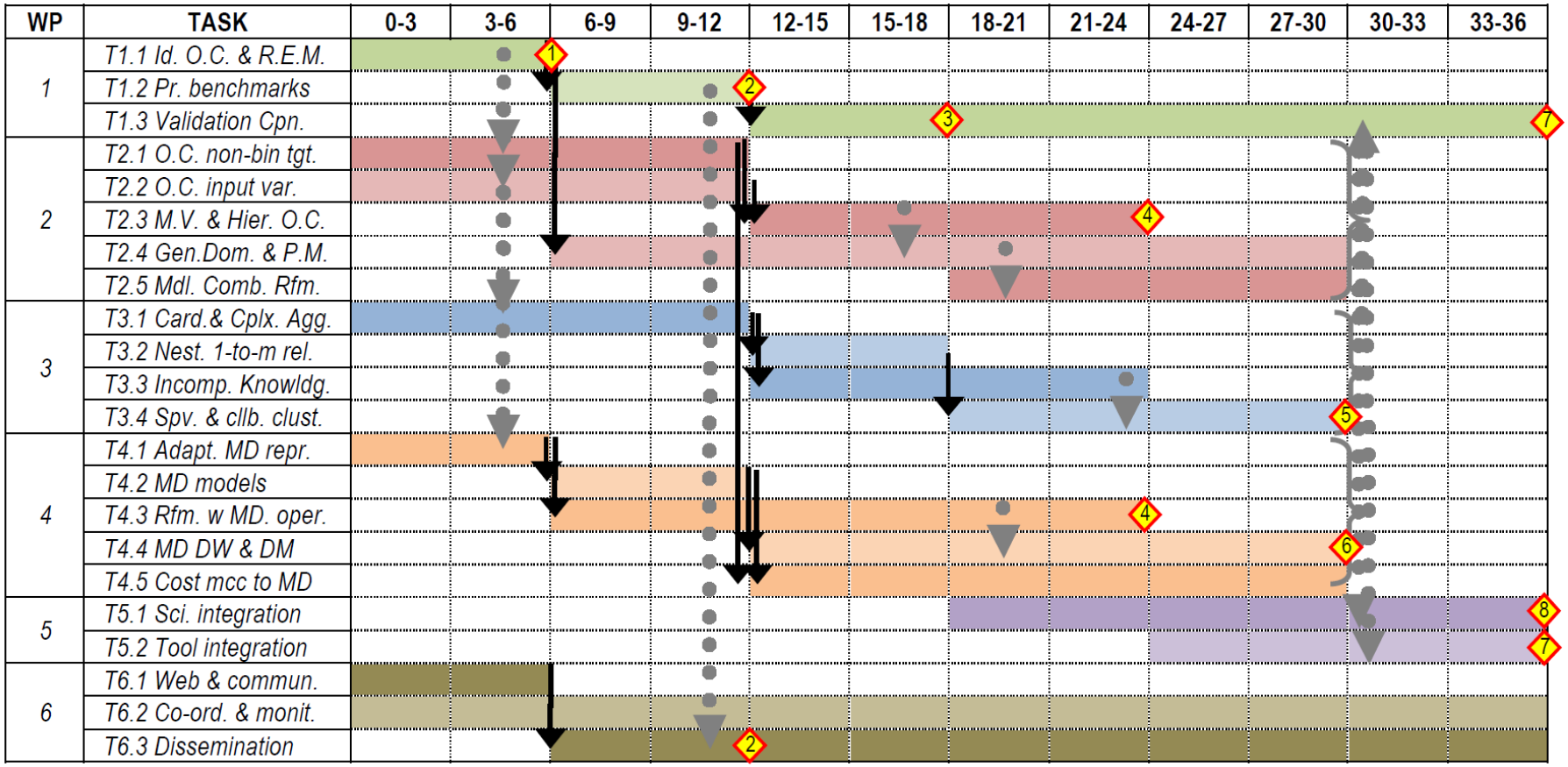
- WP1) Identification of context changes and solution validation.
  - Application domains are used to analyse different kinds of operating context changes, as a testbed for techniques and methods for WP2, WP3 and WP4.
  - Validation criteria linked to performance metrics.
- WP2) Generalised ROC Analysis for Model Reframing:
  - Generalisation of ROC analysis over different (and generalised) types of input and output data, by extending notions of dominance, hybrid models and evaluation metrics.
- WP3) Reframing in the Multi-Relational Setting with Background Knowledge:
  - Context changes originated by a single one-to-many relationship and nested one-to-many relationships, or the change in previous knowledge.
  - Task change, such as supervised clustering, and the appearance of new classes.
- WP4) Hierarchical and Multidimensional Reframing:
  - Multidimensional data representations can be used to properly specify operating contexts
  - A change of granularity is seen as a change in context.
  - Aggregation and disaggregation operators in the hierarchy as reframing operators.
- WP5) Integration:
  - Puts together the results in terms of theory, methods, technologies and methodologies.
  - Comprehensive new view of the knowledge generation and deployment pipelines.
  - The package also integrates tools and prototypes.
- WP6) Management:
  - Monitors the work plan and objectives, co-ordinates the interaction between partners, organises workshops and meetings, evaluates risks, develops websites and publicity, fosters contacts with industries, performs publication follow-up and dissemination.



chist-era

REFRAME

# Workplan (3)






chist-era

REFRAME

# Workplan (4)

- First deliverables
  - Month 6 (now September 2013):
    - D1.1. **List of operating context changes specific and common to the three application domains and of the relevant evaluation measures (T1.1) → M1**
    - D3.1. Techniques for reframing cardinalities (T3.1)
    - D4.1. Multidimensional schemas for input and output data (T4.1)
    - D6.1: Website and methods for communication, and data exchange (T6.1).
    - D6.2: Meeting reports and state of deliverables and milestones
  - Month 12 (now March 2014):
    - D1.2: **Benchmarks (T1.2) → M2**
    - D2.1: OCs on multi-label and real-valued target variables (T2.1)
    - D2.2: OCs and model dominance for categorical input vars. (T2.2, T2.4)
    - D3.2. Techniques for complex aggregates reframing (T3.1)
    - D4.2. Concept of versatile model in the multidimensional setting (T4.2)
    - D4.3. Aggregation operators (T4.3)
    - D6.3: One-year report (delayed to month 16).



chist-era

REFRAME

# Management (1)

- Goals:
  - Ensure smooth interactions between each partner.
  - Organise meetings, workshops, stays and general agenda.
  - Ensure that milestones and deliverables are met.
  - Evaluate risks at 6 monthly intervals.
  - Develop website for sharing data, reports, and internal and external dissemination.
  - Boost the connections with other academia and industry through open workshops.
  - Ensure that conflicts of interest and other conflicts are solved swiftly.

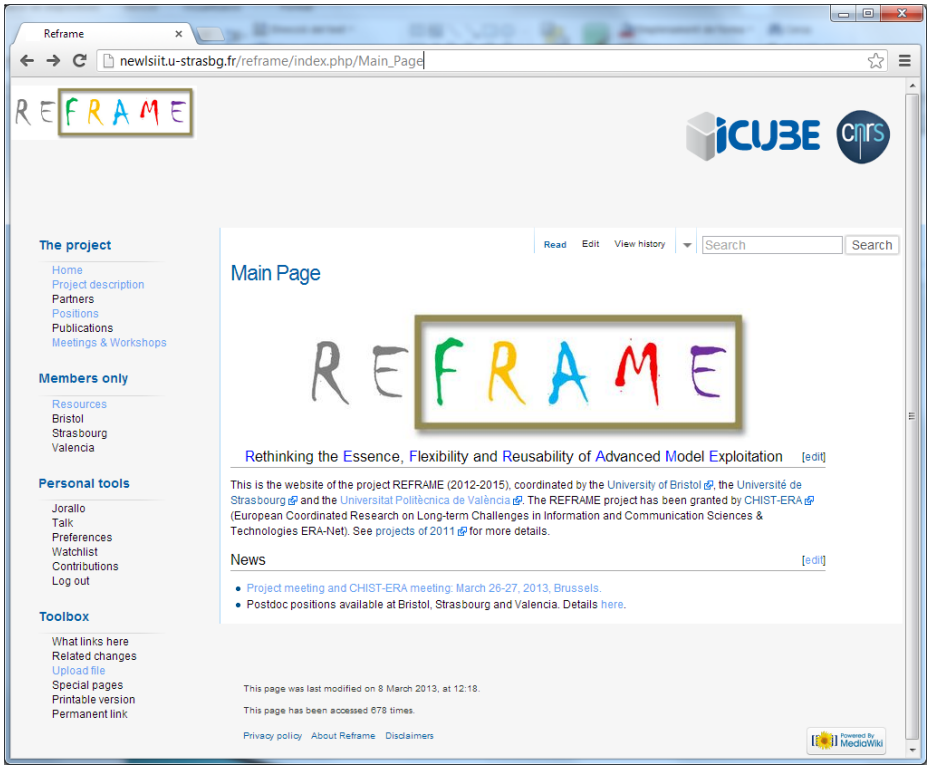


chist-era

REFRAME

# Management (2)

- Website/Wiki:
  http://www.reframe-d2k.org/




- Communication:
  - Project mailing lists (internal and external)
  - Videoconferences.



chist-era

REFRAME

# Management (3)

- Internal project meetings:
  - Kick-off (Bristol, October 2012)
  - This one (Brussels, March 2013)
  - Valencia (June-July 2013)
  - Prague (September 2013)
- Student/academic exchanges

- In progress: Consortium agreement



chist-era

REFRAME

# Results and dissemination (1)

- Main on-going work:
  - multi-label classification, cost-sensitive learning, regression, …
  - Identifying new contexts in the three working domains.
- Outputs (publications, conferences)
  - Common "CHIST-ERA Acknowledgment" to appear in any publication related to the project.
  - Publications to date:
    - J. Hernández-Orallo, P. Flach, C. Ferri "A Unified View of Performance Metrics: Translating Threshold Choice into Expected Classification Loss" Journal of Machine Learning Research (JMLR), 13(Oct):2813−2869, 2012.
    - J. Hernández-Orallo, P. Flach, C. Ferri "ROC Curves in Cost Space" Machine Learning Journal, to appear, 2013
    - A. Bella, C. Ferri, J. Hernández-Orallo, M.J. Ramírez-Quintana "Aggregative Quantification for Regression", Data Mining and Knowledge Discovery, to appear, 2013.

# Results and dissemination (2)

- Dissemination
  - Flyers at conferences
  - Workshops to be organised in Y2 and Y3

- Other funding:
  - Large project on sensor systems for health to start in Bristol
  - Potential follow-on project to continue collaborating with domain experts in smart electricity metering



chist-era

REFRAME

# Sustainability / valorisation

- Project as a short-term goal.
- More ambitious mid-term and long-term goals:
  - The (mid-term) aim is to establish a solid and stable research effort to re-interpret the whole process of knowledge discovery from data.
  - The ultimate (long-term) goal of the project is to 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.