# Optimization for Active Learning-based Interactive Database Exploration

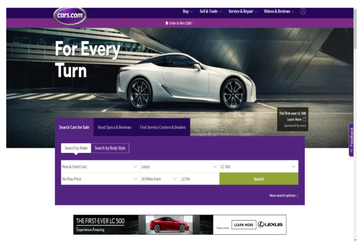
Enhui Huang[+], Liping Peng[*], Luciano Di Palma[+], Ahmed Abdelkafi[+], Anna Liu[*], Yanlei Diao[*+]
[+]: Ecole Polytechnique, France ; [*]: University of Massachusetts Amherst, USA

## Interactive Data Exploration

- Human-in-the-loop applications that search big datasets to discover interesting information.
- Need system-assisted exploration tools to accelerate information discovery.

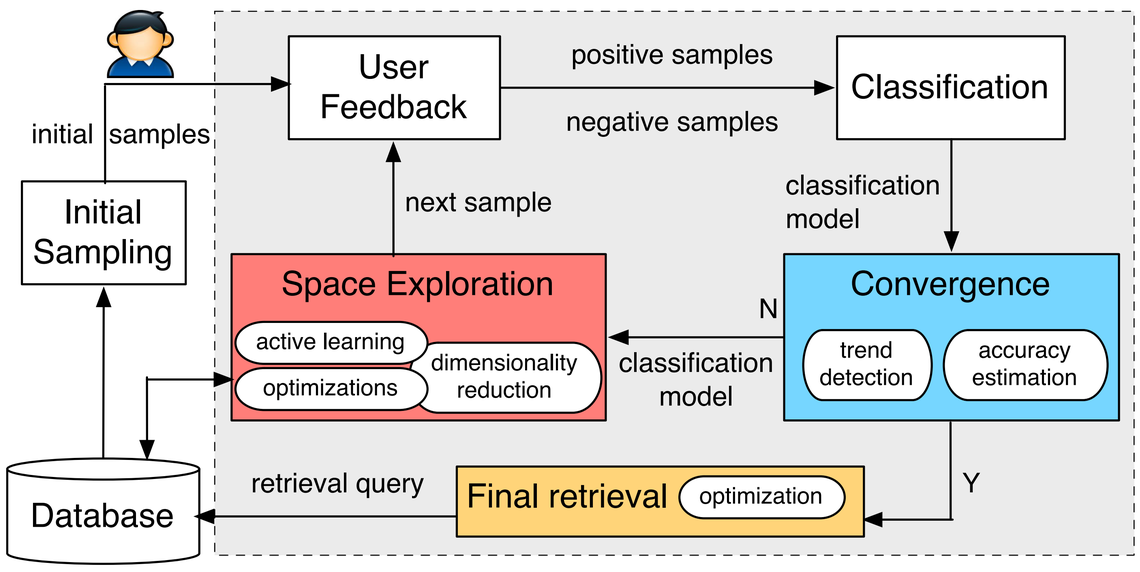

**Medical Applications**

**Scientific Applications**




**Web Applications**

## An "Explore-by-Example" Approach

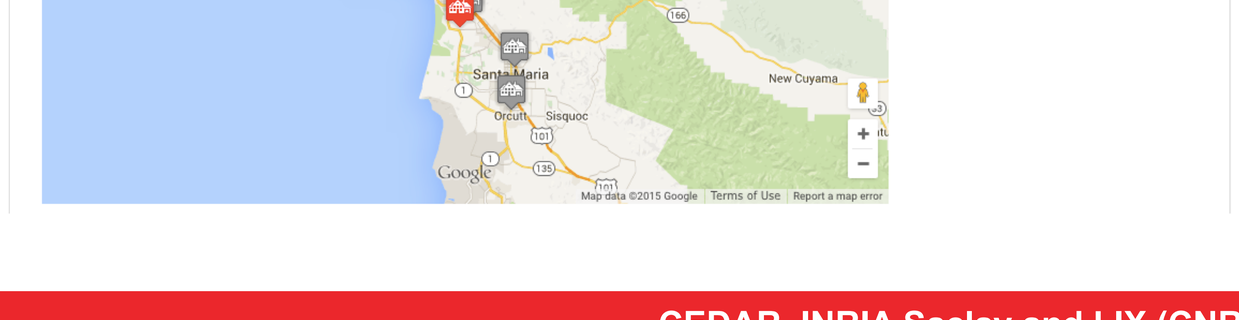



**System architecture for explore by example**

## User Interface

- **Scenario:**
  - ➢ Interactive Exploration with user-generated queries
  - ➢ Interactive Exploration with pre-defined queries
  - ➢ Comparison to Manual Exploration

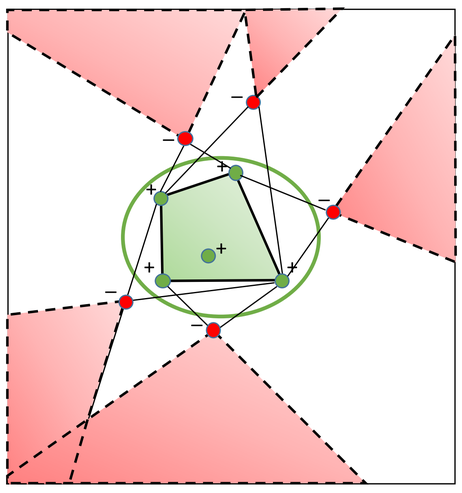
- **Database:** SDSS (Sloan Digital Sky Survey), Housing, Cars

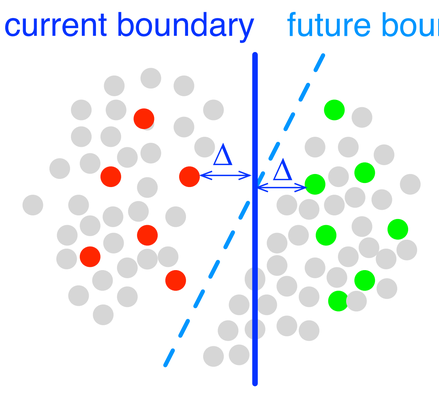



## Dual-Space Model (DSM)

- **Data-Space Model (Three-Set Partition)**
  At each iteration, all available labeled examples are leveraged to build a partitioning function of the data space, dividing the data space into three disjoint regions.
  I. Positive region ($R^+$): a convex polytope
  II. Negative region ($R^-$): the union of negative convex cones
  III. Unknown region ($R^u$): $R^u = \mathbb{R}^d - R^+ - R^-$

- **SVM-based active learning**
  To quickly improve the accuracy of the current model, choose the most informative example which is closest to the current decision boundary as the next to-be-labeled example.



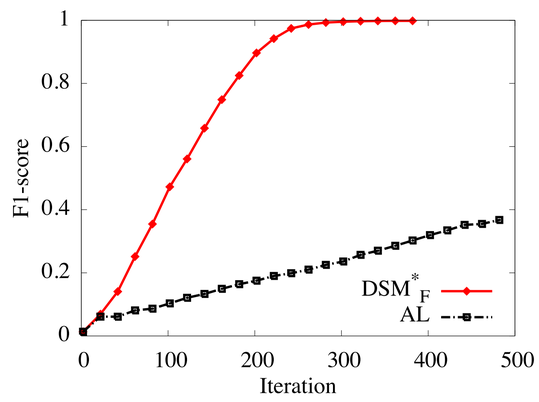


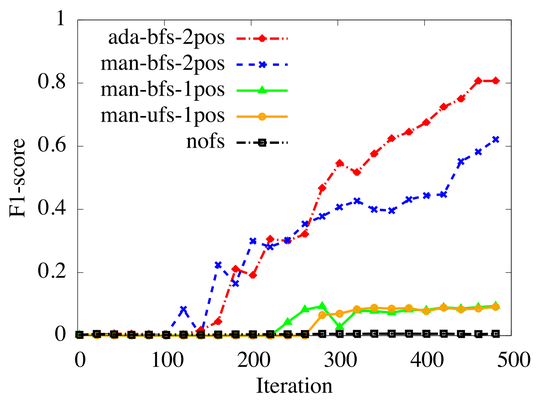

## Optimizations

- **Factorization on feature space**
  With increased dimensionality, the volume of the uncertain region may grow fast. This problem, referred to as slow convergence, can be addressed by factorizing a high-dimensional data space into a set of low-dimensional spaces and combining DSMs built in each subspace together by some rules.

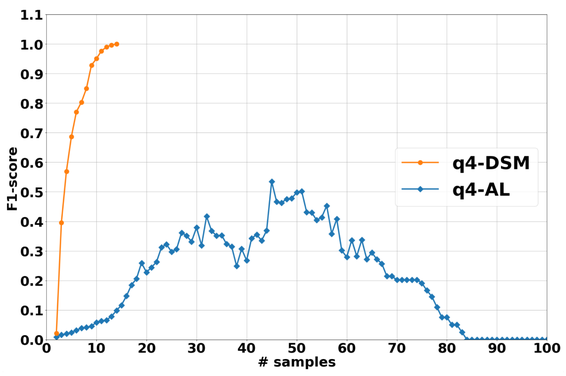
- **GBRT-based dimensionality reduction**
  Adaptive strategy of using Gradient Boosting Regression Trees (GBRT) to choose top-k features from the original features based on feature importance scores.

- **Final result retrieval**
  To expedite the retrieval of the final results, build R-tree as the index over the database, and perform a top-down search in a depth-first fashion (Branch and Bound).

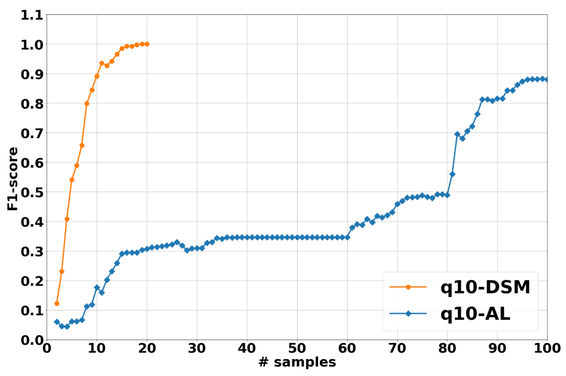



## User Study using a Car Database




Accuracy for Q4 (0.249% selectivity)




Accuracy for Q10 (0.356% selectivity)