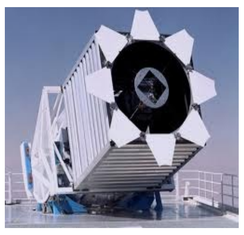
Scalable Interactive Exploration over Large Databases

Liping Peng*, Enhui Huang [+], Yuqing Xing*, Anna Liu*, Yanlei Diao* [+]
[+]: Ecole Polytechnique, France ; *: University of Massachusetts Amherst, USA

Interactive Data Exploration

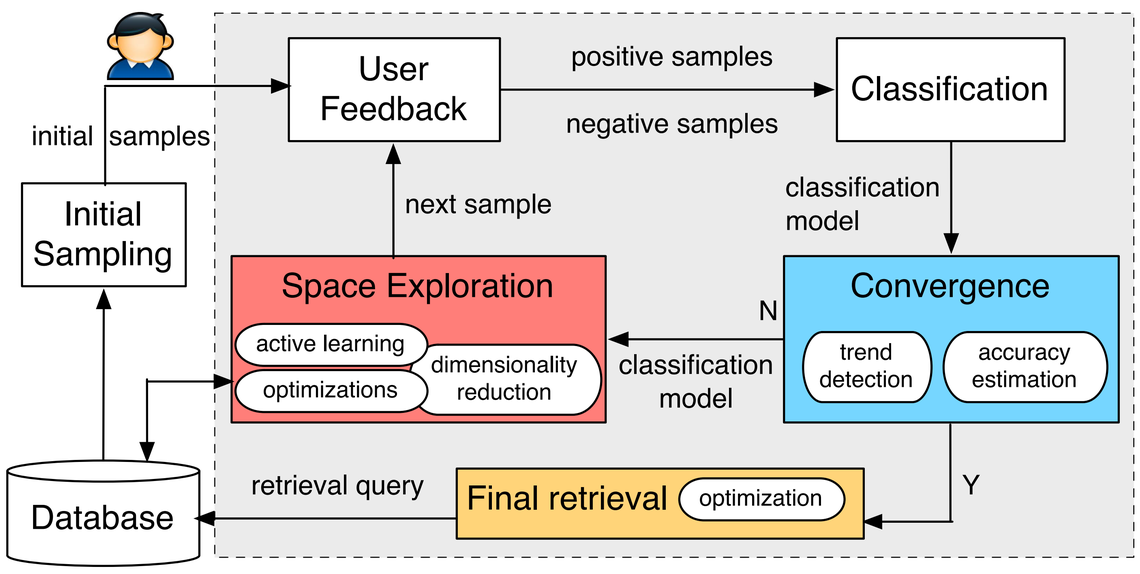
- Human-in-the-loop applications that search big datasets to discover interesting information
- Need system-assisted exploration tools to accelerate information discovery

Medical Applications



Scientific Applications

An "Explore-by-Example" Approach



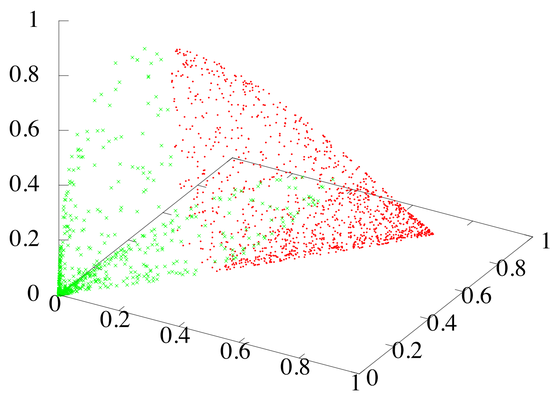
System architecture for explore by example

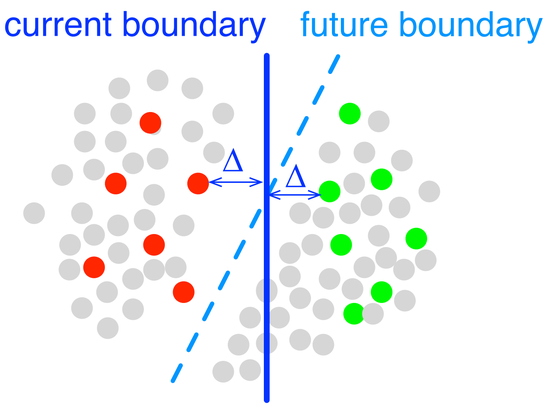
User Interface

- **Scenario:**
  - ➢ Interactive Linear/Non-Linear Exploration
  - ➢ Linear/Non-linear Exploration with pre-defined queries
  - ➢ Comparison to Manual Exploration
- **Database: SDSS (Sloan Digital Sky Survey), Housing**

SVM-based Active Learning

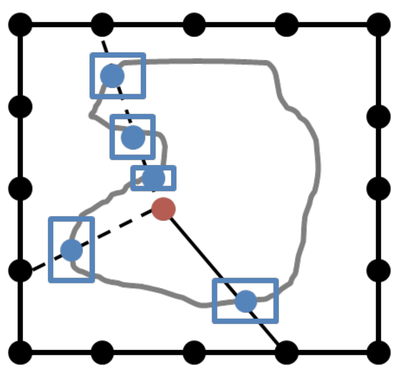
- **Kernel-based SVM for classification**:
  Kernel function implicitly maps the labeled examples to a higher-dimensional feature space where examples of different classes are linearly separable
- **Exploration based on active learning theory**:
  To quickly improve the accuracy of the current model, choose the most informative example which is closest to the current decision boundary as the next to-be-labeled example

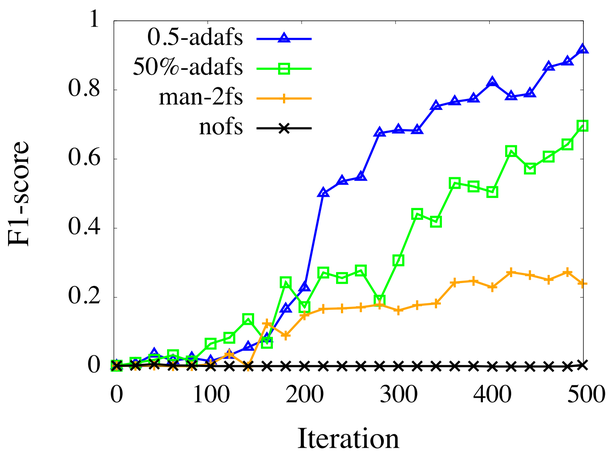




Optimizations

- **Solver method for retrieving samples**:
  Given two points with opposite labels, find a point on the decision boundary through a solver of the boundary condition, y(x) = 0, and then a database example locally near this point
- **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)





Comparison with Alternative Systems