AIDE: An Automatic User Navigation System for Interactive Data Exploration

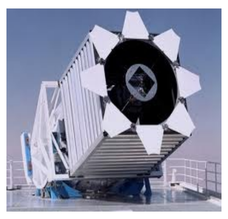
Yanlei Diao*, Kyriaki Dimitriadou+, Zhan Li+, Wenzhao Liu*, Olga Papaemmanouil+, Kemi Peng+, Liping Peng*

*: University of Massachusetts Amherst +: Brandeis University

Interactive Data Exploration

- ❖ Human-in-the-loop applications that search big datasets to discover interesting information
- ❖ A long-running, multi-step process with end-goals not stated explicitly

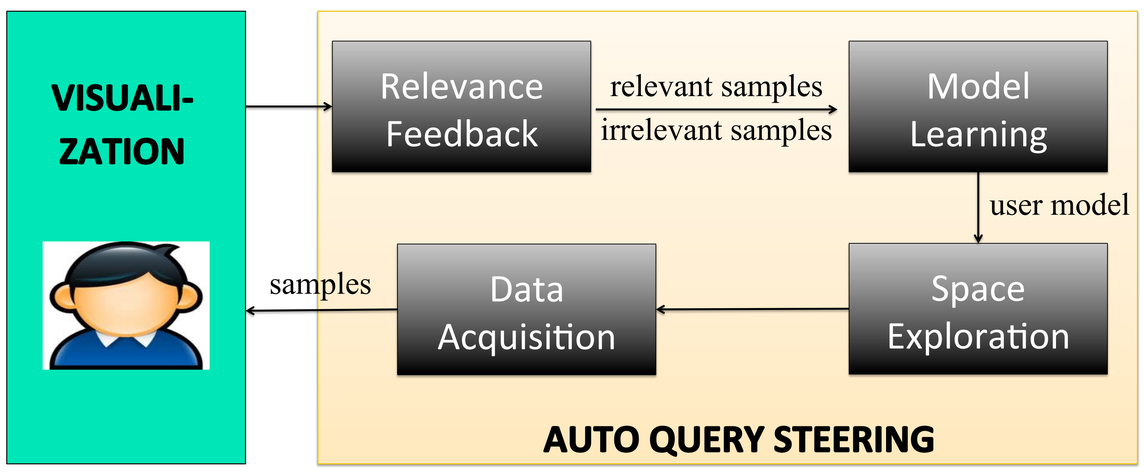
Medical Applications



Scientific Applications

FrameWork Overview and Challenges

AIDE: Automatic Interactive Data Exploration



➢Effective Data Exploration: Which data samples to show to the user?

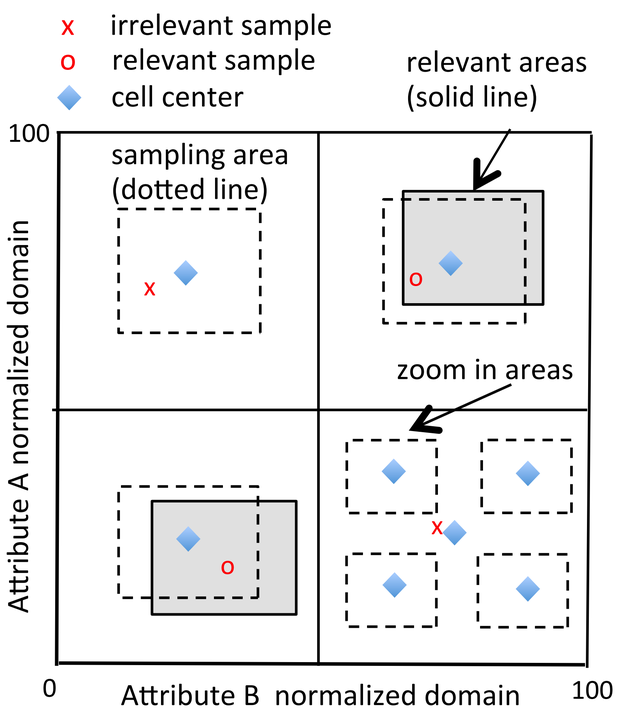
- Unknown user interests

➢Efficient Sample Acquisition: How to minimize acquisition cost?

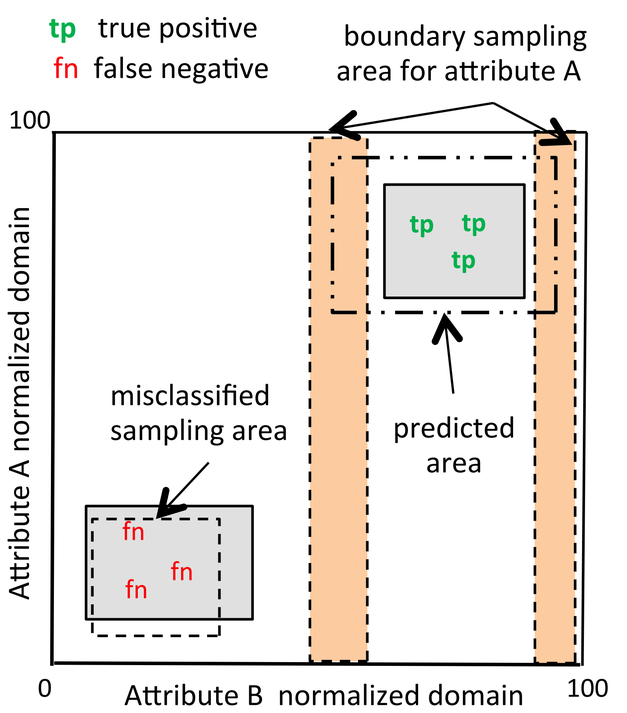
- High sample acquisition cost on big data sets
- Accuracy vs efficiency trade-off
- Model learning & sample acquisition need to be coupled

Decision-tree-based Linear Pattern Discovery

- **Relevant object discovery**:
  Clustering of exploration space to address skewed data sets
- **Misclassified exploitation**:
  Clustering false negatives to reduce sampling queries
- **Boundary exploitation**:
  Leverage decision tree conditions to identify and eliminate overlapping sampling areas



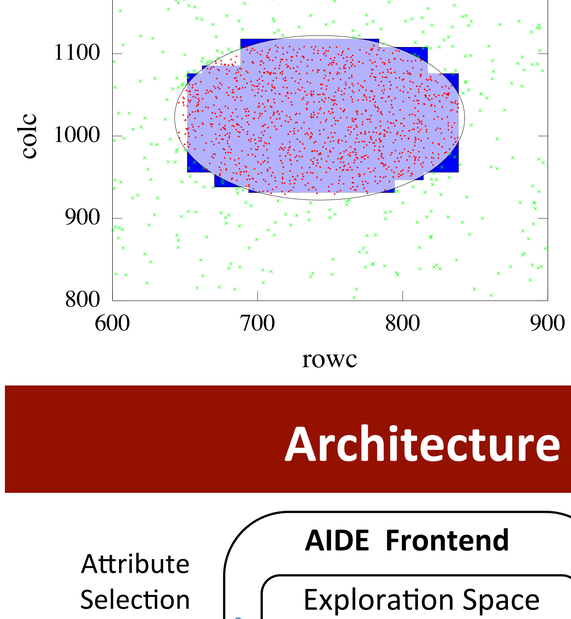
(a) Relevant Object Discovery

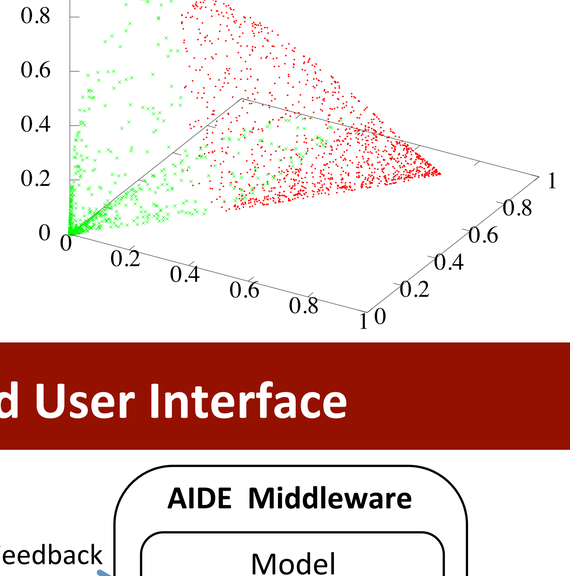


(b) Misclassified and Boundary Exploitation

SVM-based Non-linear Pattern Discovery

- **Classification based on SVM**:
  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 example closest to the current decision boundary as the next to-be-labeled example





Architecture and User Interface