# Auto Query Steering for Interactive Data Exploration Applications

Kyriaki Dimitriadou
Brandeis University
kiki@cs.brandeis.edu

Olga Papaemmanouil
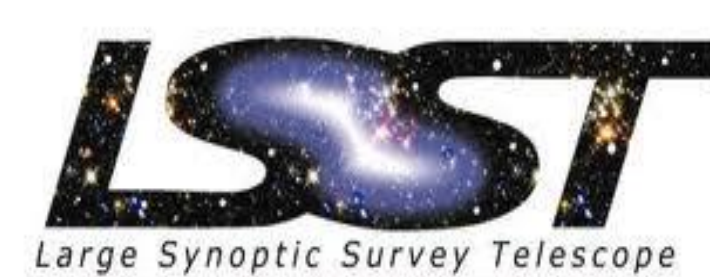Brandeis University
olga@cs.brandeis.edu

Yanlei Diao
UMass Amherst
yanlei@cs.umass.edu

## Interactive Data Exploration

• Human-in-the-loop applications that search enormous datasets to discover interesting information.
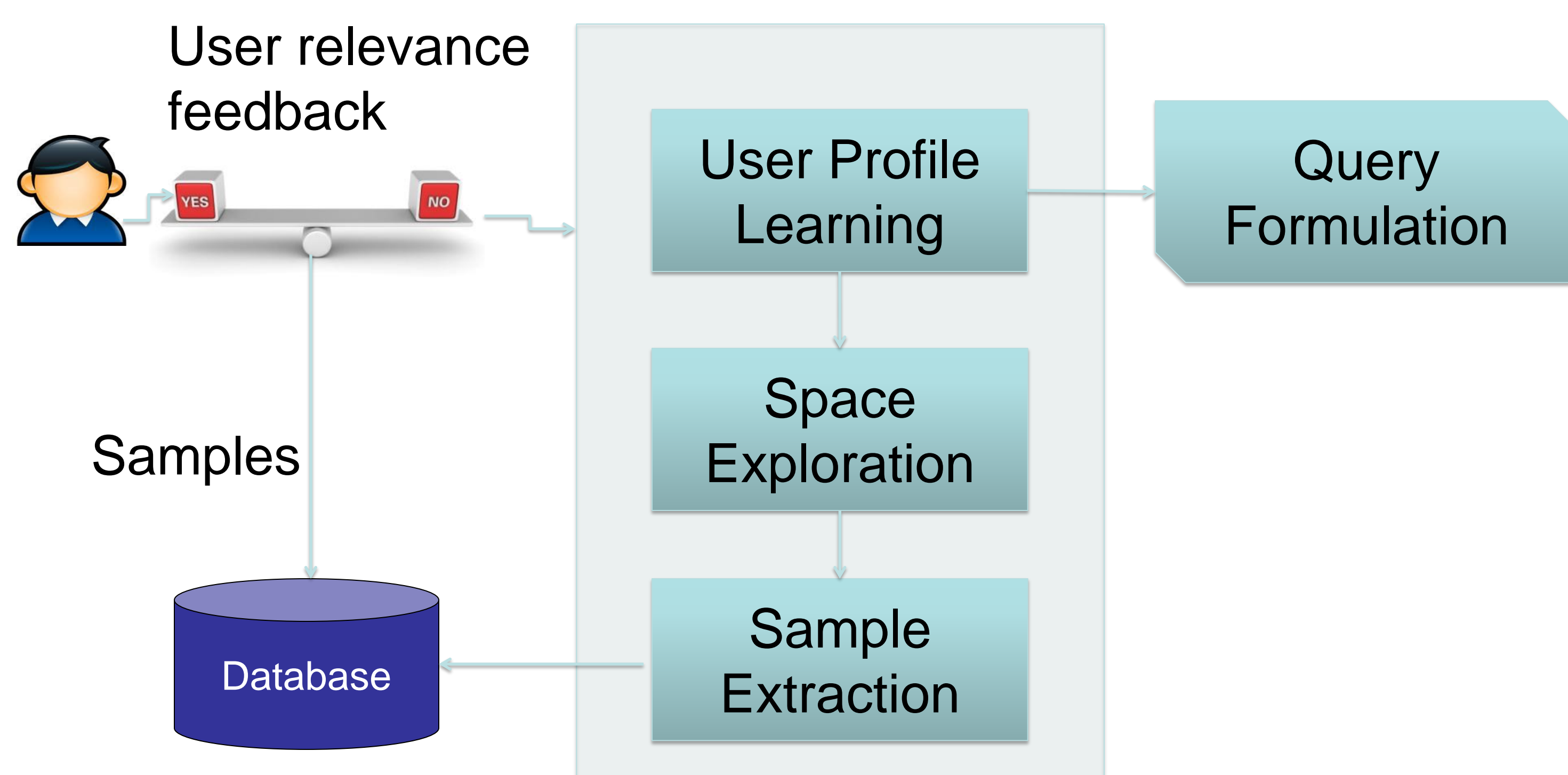
Medical Applications

Scientific Applications

Financial Applications

• Labor intensive process:
- Users iteratively strive to improve query predicates.
• Resource demanding exploration:
- Ineffective, non optimized query sequences that miss relevant or return irrelevant information.

## Automatic Query Steering Framework

Objective: Automatically capture the user's interest and steer him towards interesting data areas.

User relevance feedback

YES    NO

User Profile Learning → Query Formulation

Space Exploration

Samples
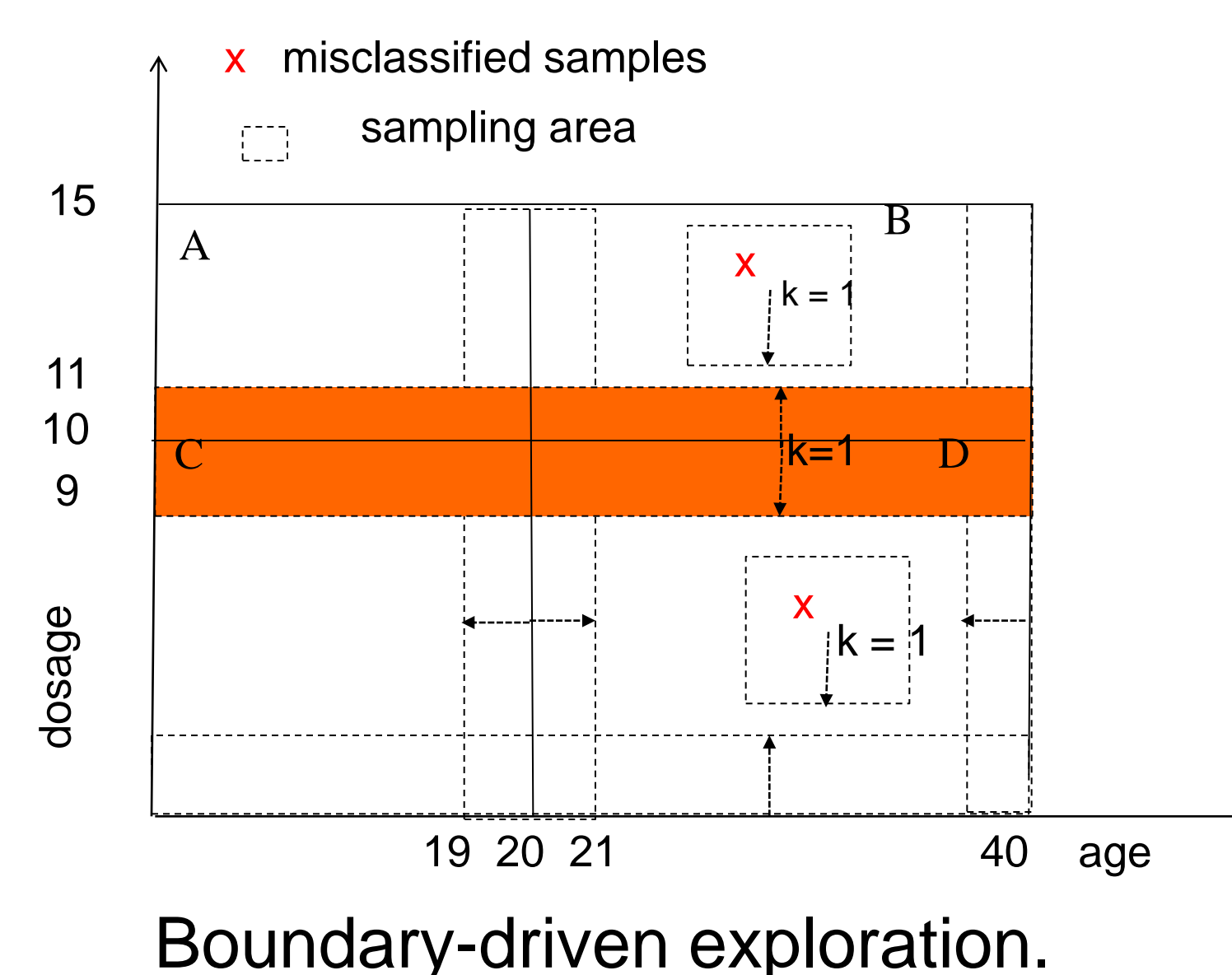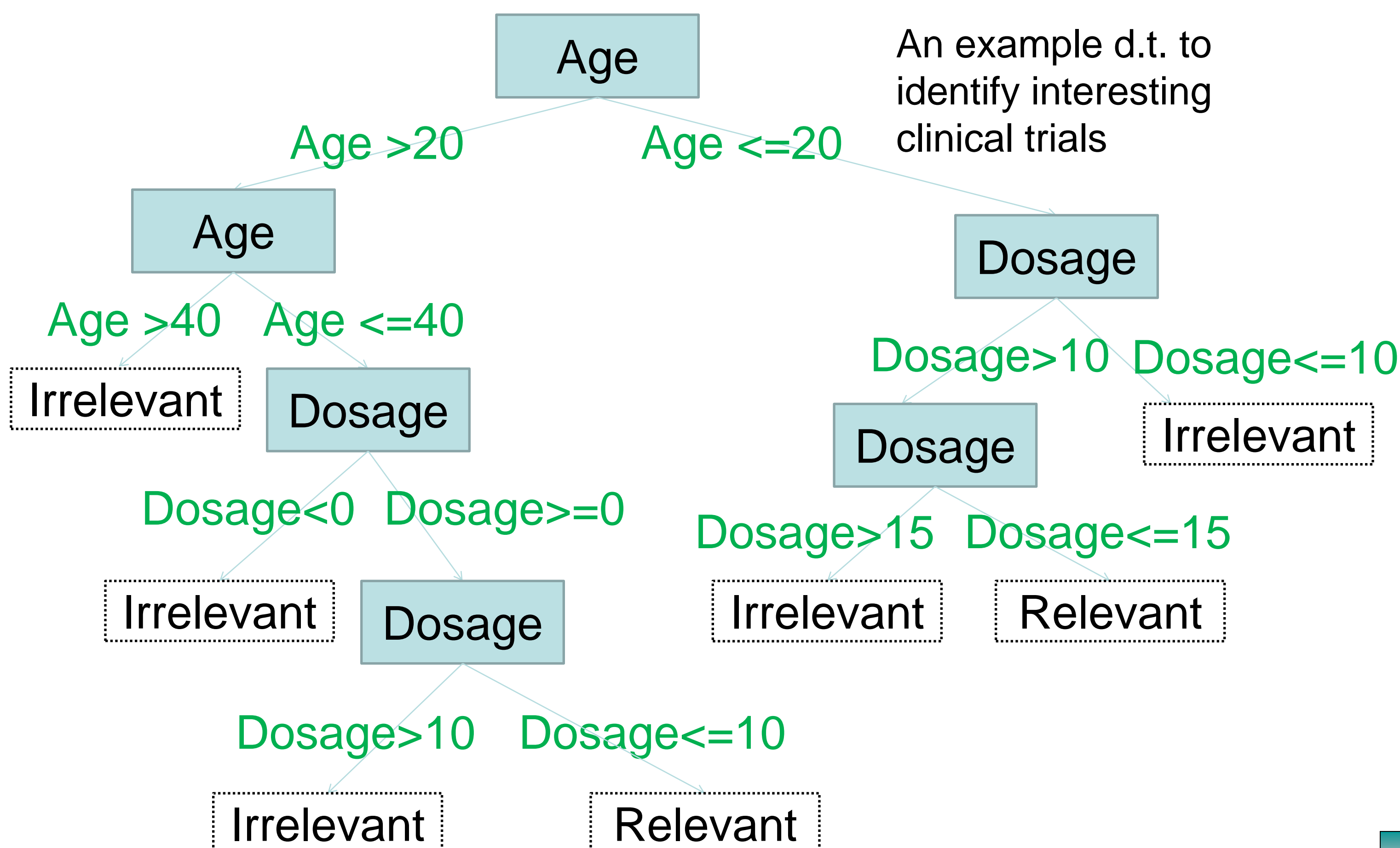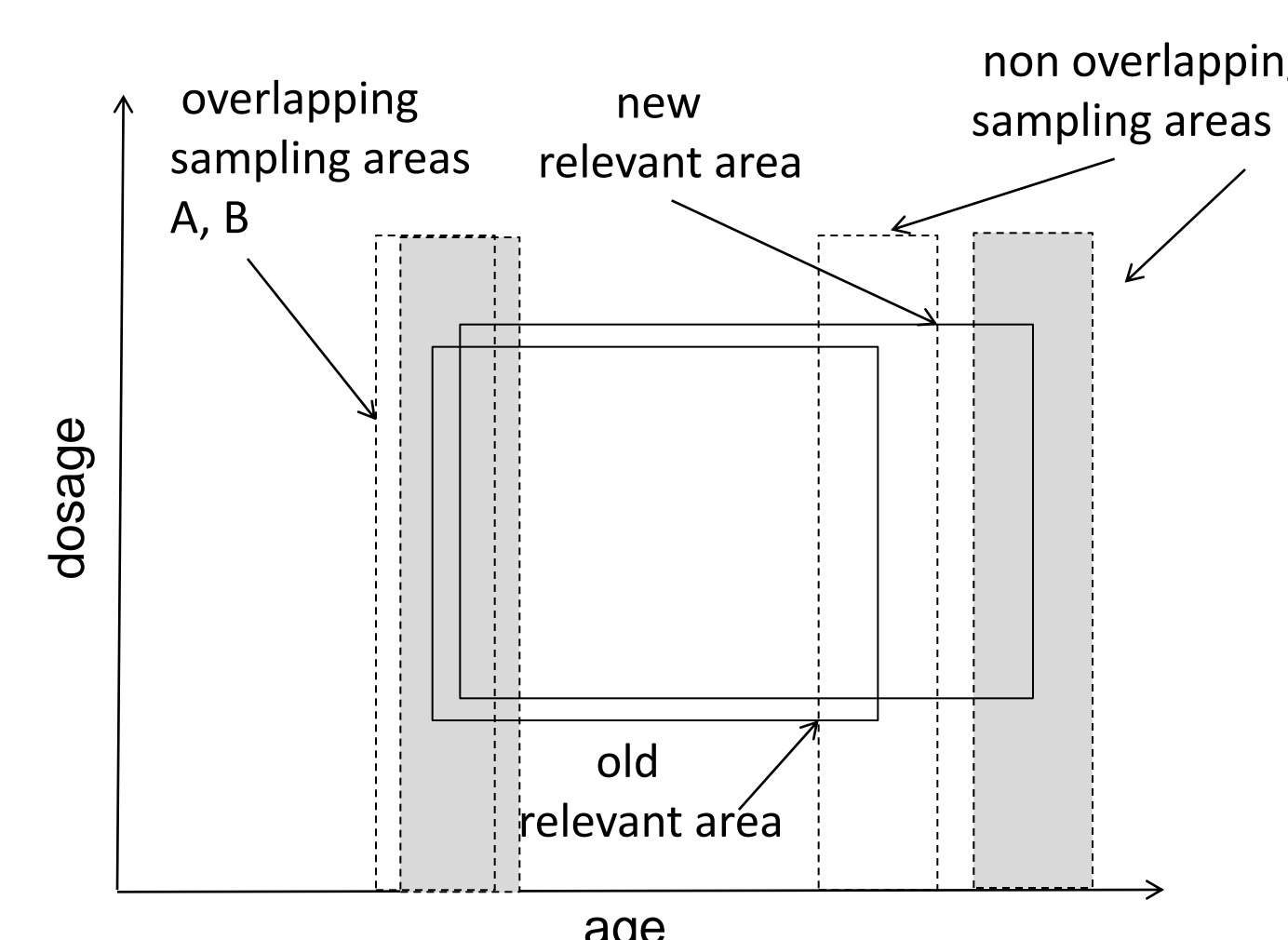
Sample Extraction

Database

## Research Challenges

• Effective Data Exploration:
- Given a classification model identify new training samples to improve the accuracy of the system.
• Efficient Sample Acquisition:
- Reduce the total cost of the query steering process:
→ sampling overhead.
→ number of steering iterations.

## Our Data Exploration Approach

• Use decision trees to build classification model.
• Exploit node split rules to identify exploration areas.

Age

An example d.t. to identify interesting clinical trials

Age >20          Age <=20

Age          Dosage

Age >40    Age <=40          Dosage>10    Dosage<=10

Irrelevant    Dosage          Dosage    Irrelevant

Dosage<0    Dosage>=0          Dosage>15    Dosage<=15

Irrelevant    Dosage          Irrelevant    Relevant

Dosage>10    Dosage<=10

Irrelevant    Relevant
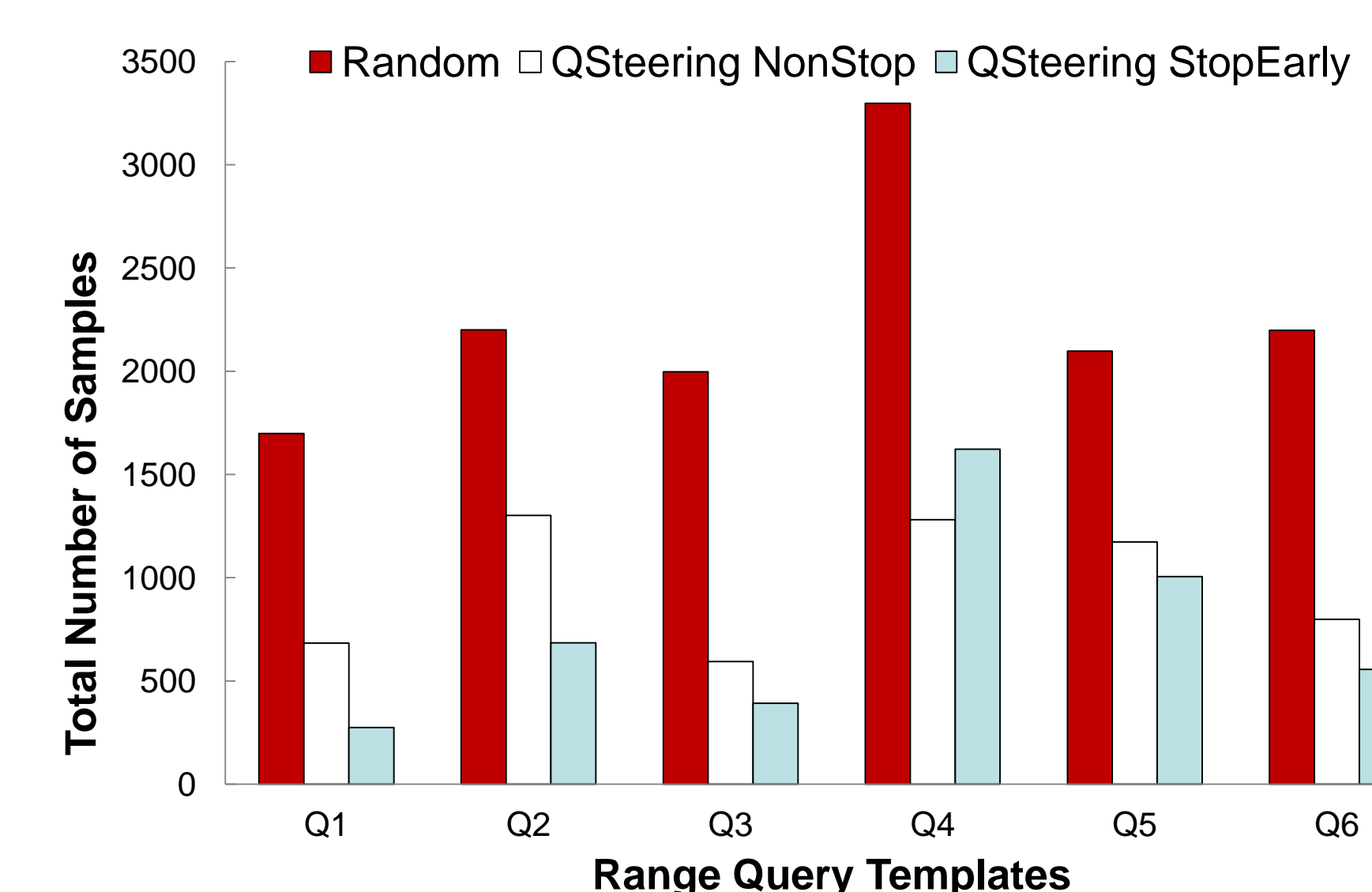
x misclassified samples
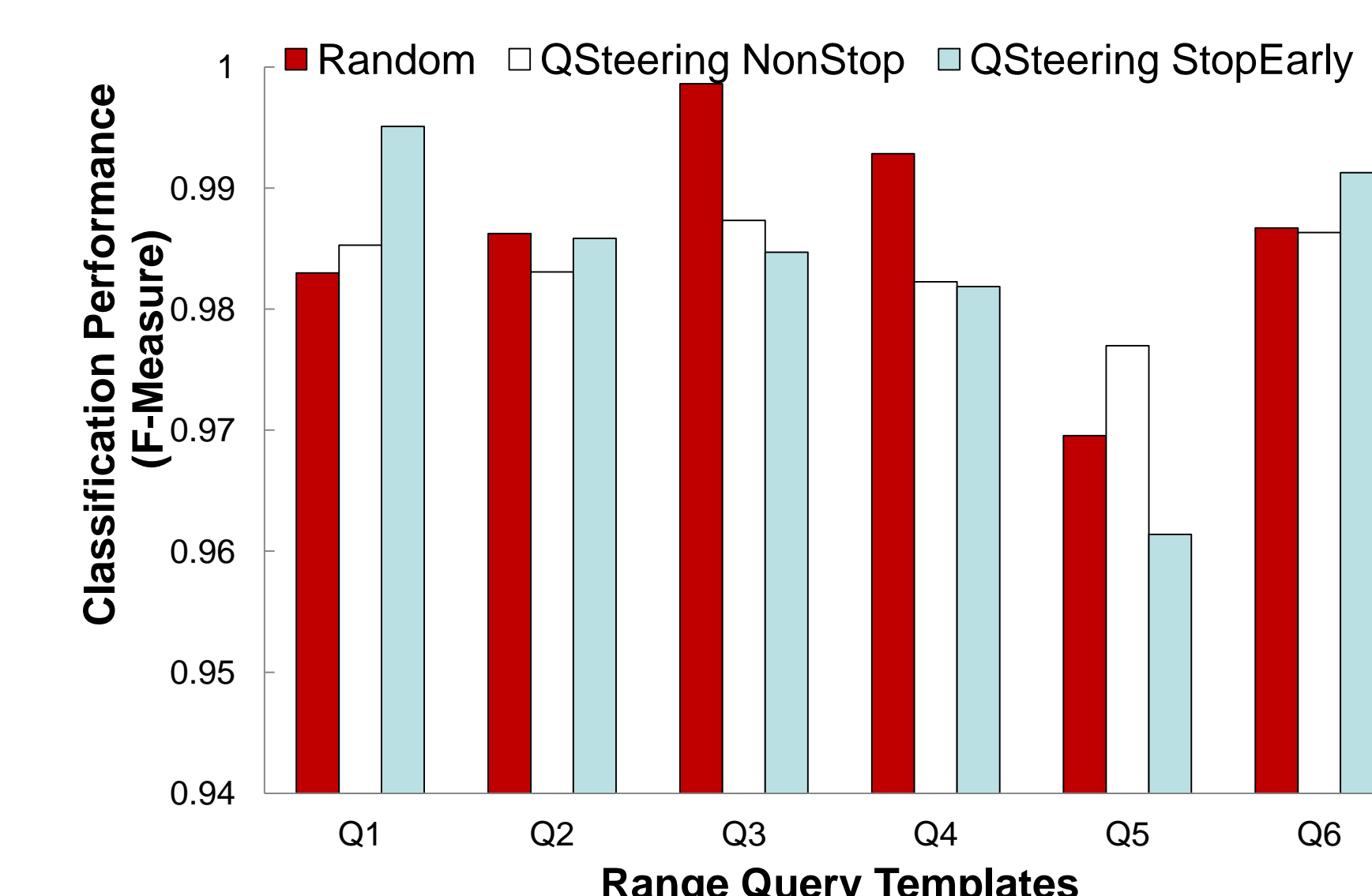⬚ sampling area

Boundary-driven exploration.

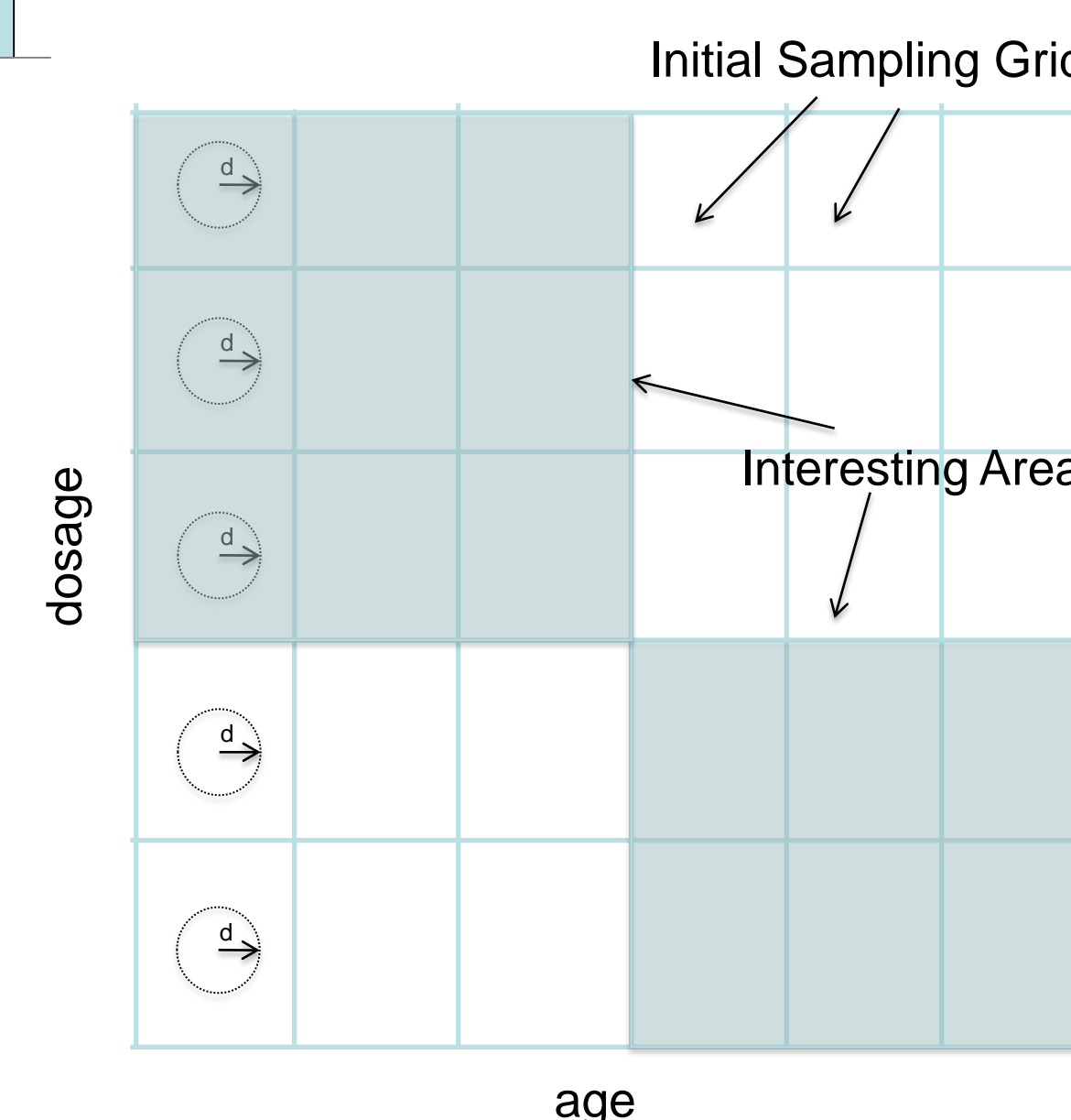Overlap of sampling areas for similar decision trees.

## Preliminary Results

User effort decreased by **58-68%** depending on the initial sampling technique.

Classification performance is higher than **96%**.

Initial Sampling:
→ QSteering StopEarly
→ QSteering NonStop

Initial Sampling Grids

Interesting Areas

## Ongoing Work – Sample Acquisition

• Optimizations on exploratory workloads:
- Avoiding sparse regions using histograms.
- Across-iteration optimization.
• Profile driven optimizations:
- Leverage histories of user interactions (*exploration profiles*) :
→ predict exploration trajectories.
→ pre-fetch samples.
→ identify interesting attributes.