Università di Genova
DIPARTIMENTO INFORMATICA, BIOINGEGNERIA, ROBOTICA E INGEGNERIA DEI SISTEMI

Computer Science Workshop
PhD program in Computer Science and Systems Engineering

Non-Discriminating Data Transformations

Chiara Accinelli, Barbara Catania, Giovanna Guerrini

A Coverage-based Approach to Data Transformations

Motivation and Context

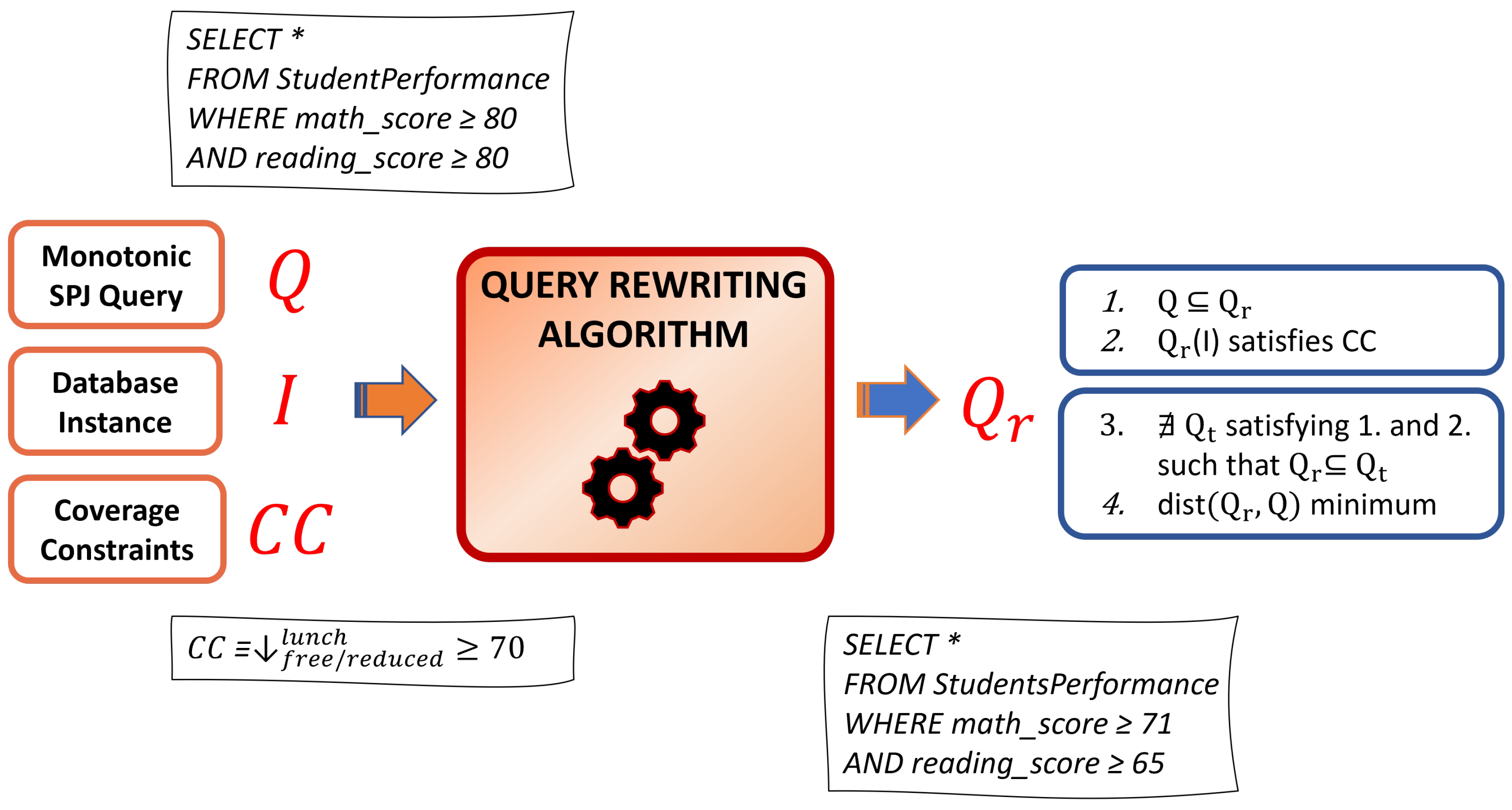
The development of technological solutions satisfying non-discrimination requirements is currently one of the main challenges in data processing [5]

- Diversity, fairness, protection of minorities, and transparency are becoming increasingly crucial
- Data pre-processing can introduce bias at different levels
- *Data transformations*: protected categories can be under-represented in the result of a Select-Project-Join (SPJ) query and this might introduce bias in the following analytical steps

Key Concepts

- Sensitive attribute: discrete valued attribute used for the identification of protected groups (e.g. gender or race)
- Data pre-processing operation: monotonic SPJ query that returns, among the others, at least one sensitive attribute in the project list
- Coverage constraint [1]: condition specifying how many instances of a protected group should be returned by a data transformation

The Problem



*SELECT **
FROM StudentPerformance
WHERE math_score ≥ 80
AND reading_score ≥ 80

1. $Q \subseteq Q_r$
2. $Q_r(I)$ satisfies CC

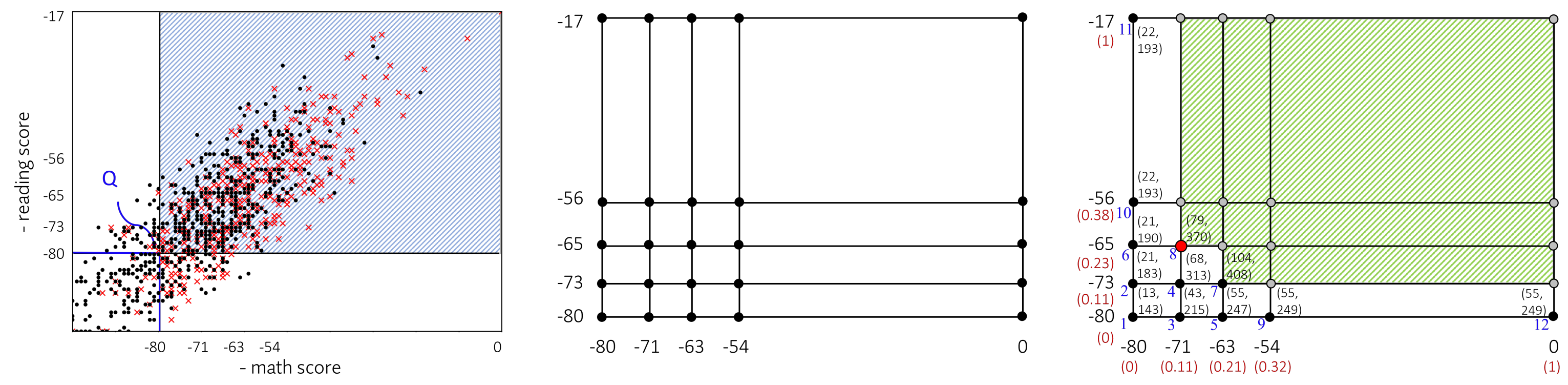
3. $\nexists\, Q_t$ satisfying 1. and 2. such that $Q_r \subseteq Q_t$
4. $\mathrm{dist}(Q_r, Q)$ minimum

$CC \equiv \downarrow^{lunch}_{free/reduced} \geq 70$

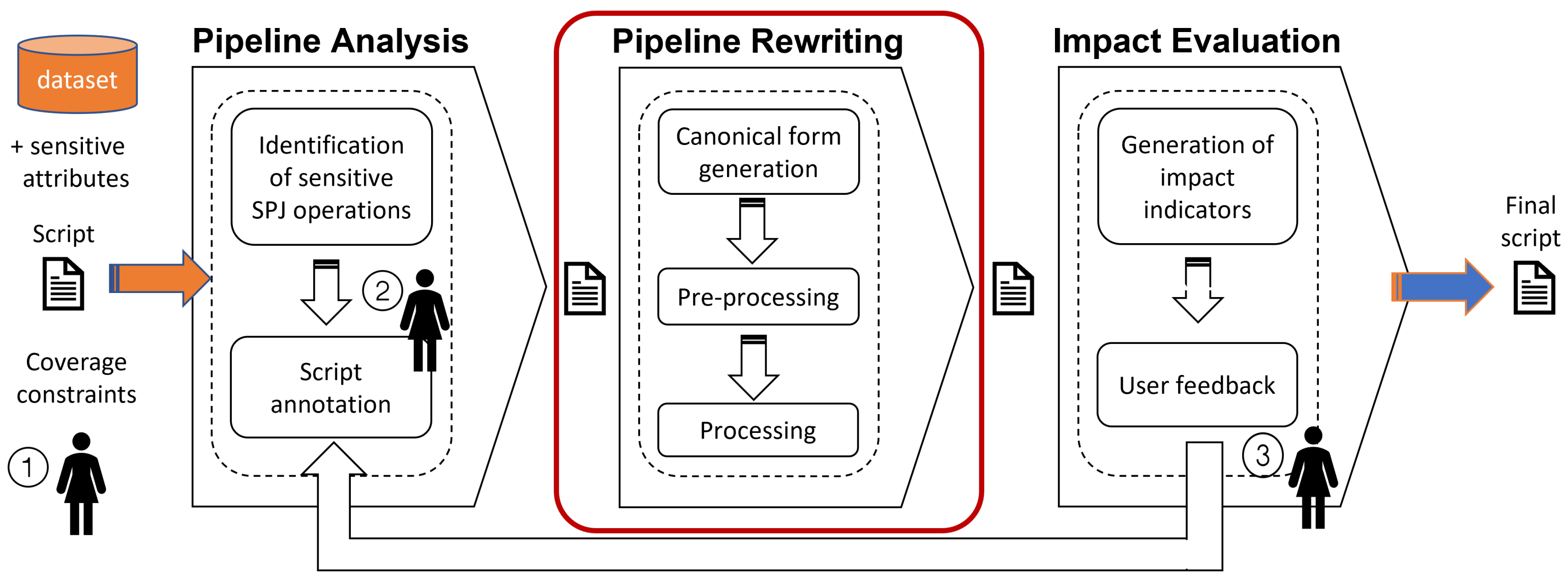
*SELECT **
FROM StudentsPerformance
WHERE math_score ≥ 71
AND reading_score ≥ 65

Design Choices

- Rewriting-based approach to guarantee transparency
- Canonical query representation as a point in the multidimensional space defined by query selection attributes
- Pre-processing: search space *approximation* as a *multidimensional grid* through traditional bucketing approaches (e.g., equi-depth and equi-width)
- Processing: visit of the grid from the input query, optimized through *pruning* and *iteration*
- Sample-based cardinality estimation for each visited point, to guarantee fast and accurate minimality and coverage constraint checking
- Grid-based, sample-based, and solution-based accuracy measures

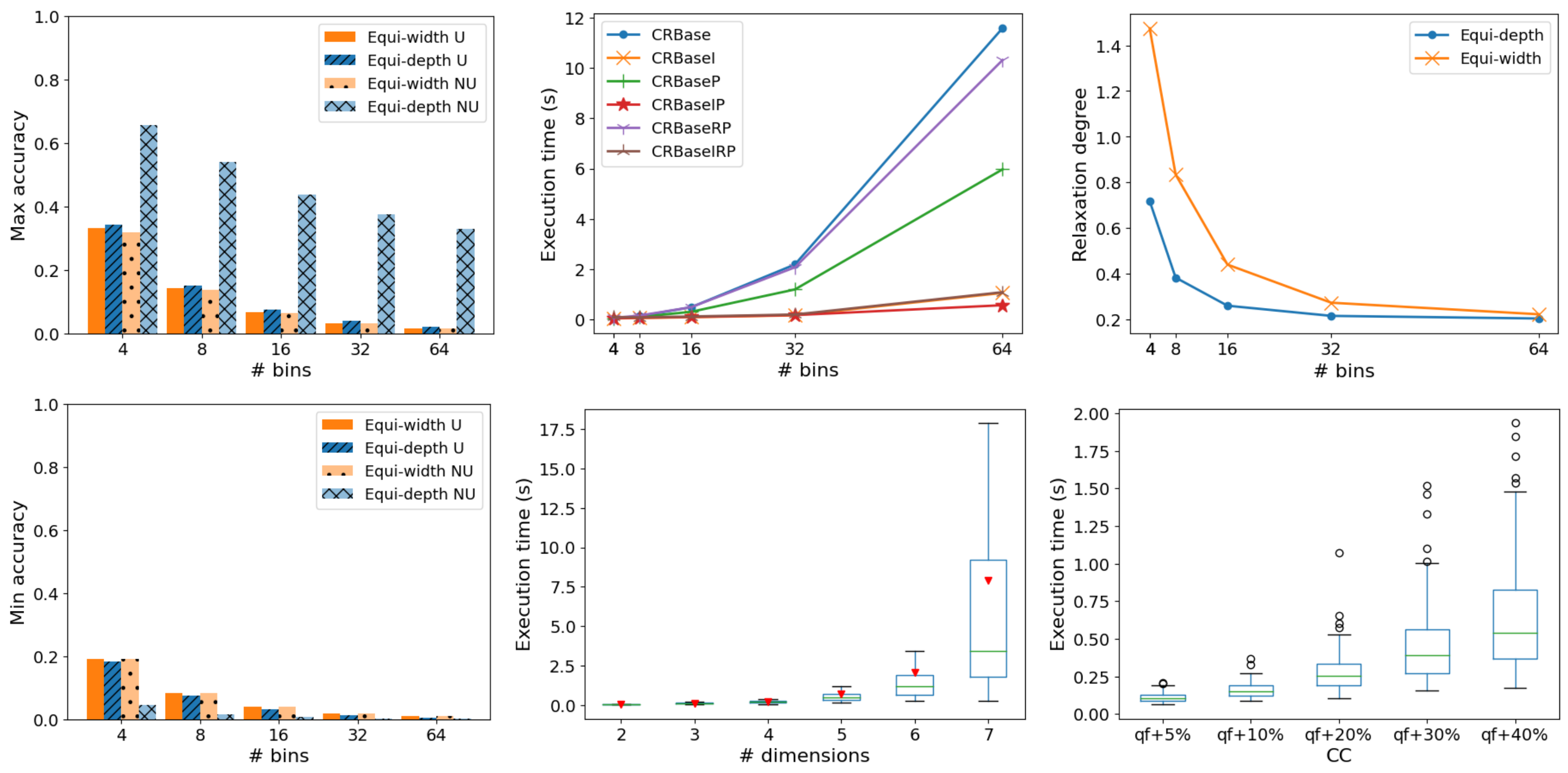


The covRew Python Toolkit



Results

- Synthetic and real datasets (US Adult Census, Diabetes US)
- Different distributions of sensitive and selection attribute values



- Pre-processing impacts accuracy
- Pruning and iteration greatly improve processing performance
- Execution time is affected by the curse of dimensionality problem
- Trade-off between efficiency and accuracy
- The coverage constraint threshold impacts both execution time and relaxation degree
- Execution time linearly depends on the number of coverage constraints

Forthcoming Research

- Optimizations: materialization of estimation results to be shared by similar queries; space dimensionality reduction through subspace selection
- Coverage-based rewriting as a new relational operation to be taken into account during all the query processing steps
- Coverage+fairness-based rewriting as a constrained optimization problem

References

[1] A. Asudeh et al. Assessing and remedying coverage for a given dataset. *ICDE*, 2019.
[2] C. Accinelli et al. Coverage-based rewriting for data preparation. *EDBT/ICDT Workshops*, 2020.
[3] C. Accinelli et al. covRew: a python toolkit for pre-processing pipeline rewriting ensuring coverage constraint satisfaction. *EDBT*, 2021.
[4] C. Accinelli et al. The impact of rewriting on coverage constraint satisfaction. *EDBT/ICDT Workshops*, 2021.
[5] J. Stoyanovich et al. Responsible data management. *PVLDB*, 2020.

Data Management and Analysis Research Group (DaMa)

CONTACTS

Chiara Accinelli
chiara.accinelli@dibris.unige.it

Barbara Catania
barbara.catania@unige.it

Giovanna Guerrini
giovanna.guerrini@unige.it