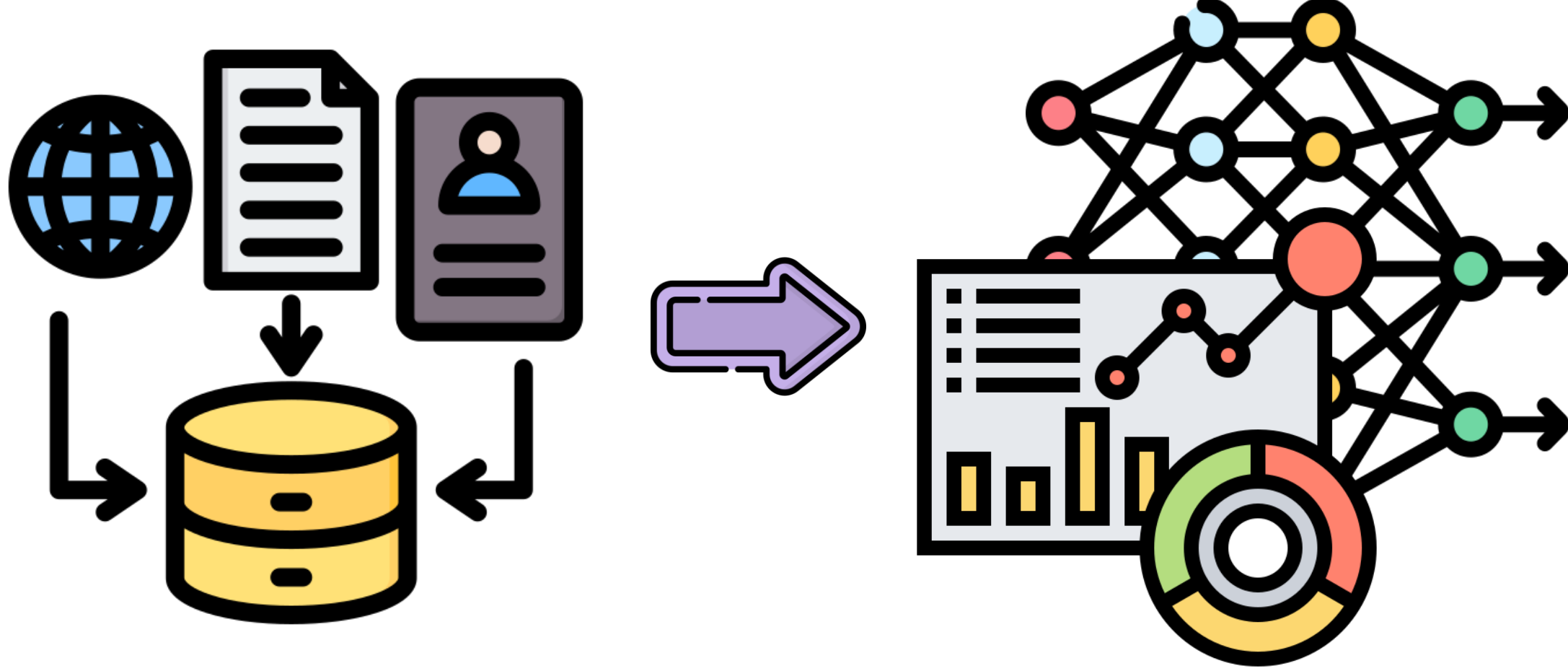# Big Data Compression Tool using Attribute-based Signatures

Constantinos Costa[1], Panos K. Chrysanthis[1], Herodotos Herodotou[2], Marios Costa[1], Efstathios Stavrakis[3], Nicolas Nicolaou[3]

[1]Rinnoco Ltd, [2]Cyprus University of Technology, [3]Algolysis Ltd

Motivation

Data* is *continuously collected* to meet the demands of *advanced analytical* and *AI applications.

Data storage* remains a challenge despite the advances in *storage technologies* and the *reduction of capacity costs.
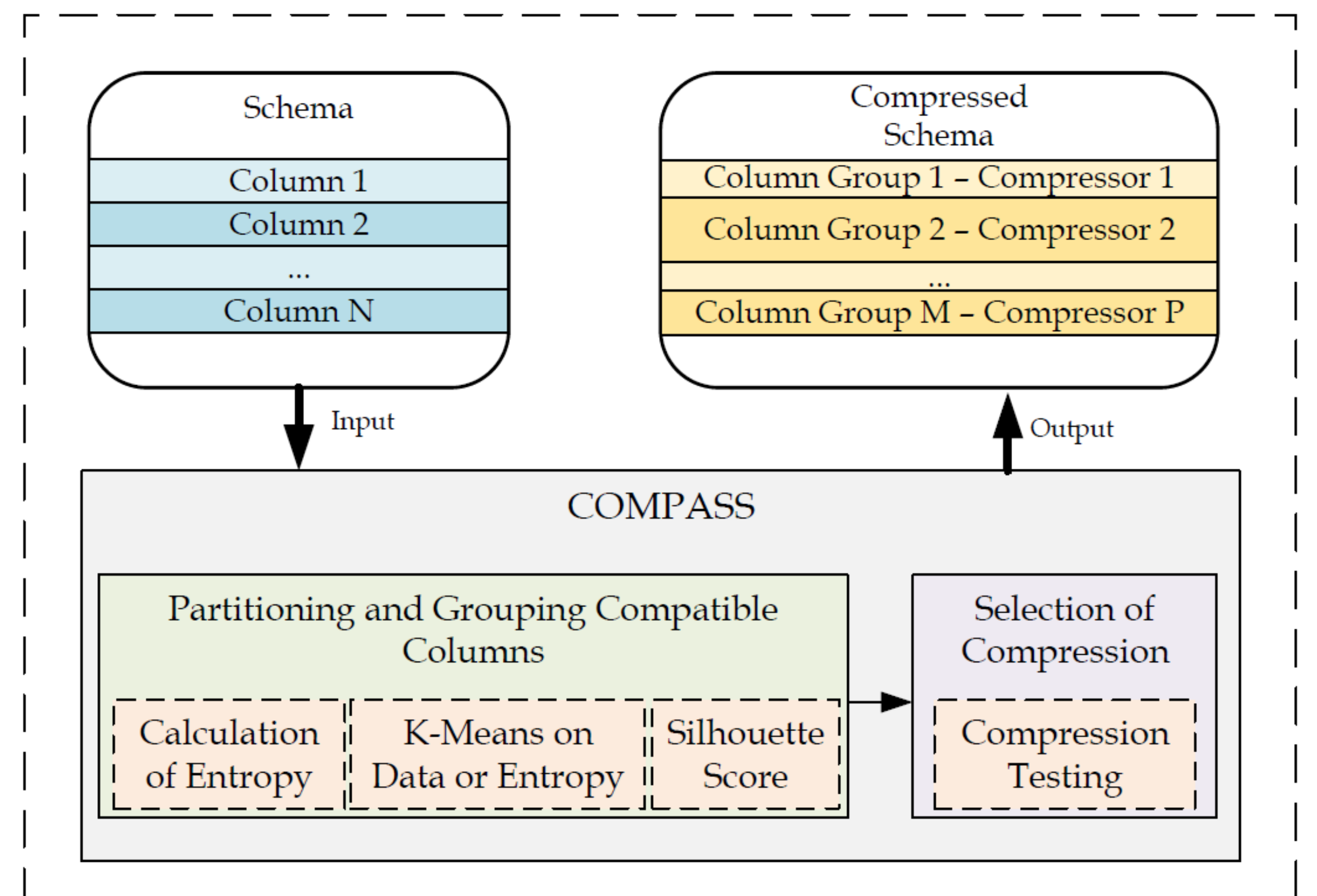
Data Storage

- Although the volume of electronically stored data **doubles yearly**, storage capacity costs **decline** only at a rate of less than **1/5 per year**.
  - For this reason, both **lossless** and **lossy** compression techniques are used to reduce the data storage.
- The current approach to database compression is typically **"monolithic"**, using a **single**, lossless method to reduce the database's disk space usage.

"These systems can yield significant storage cost savings by selecting the optimal compression scheme for a given dataset."
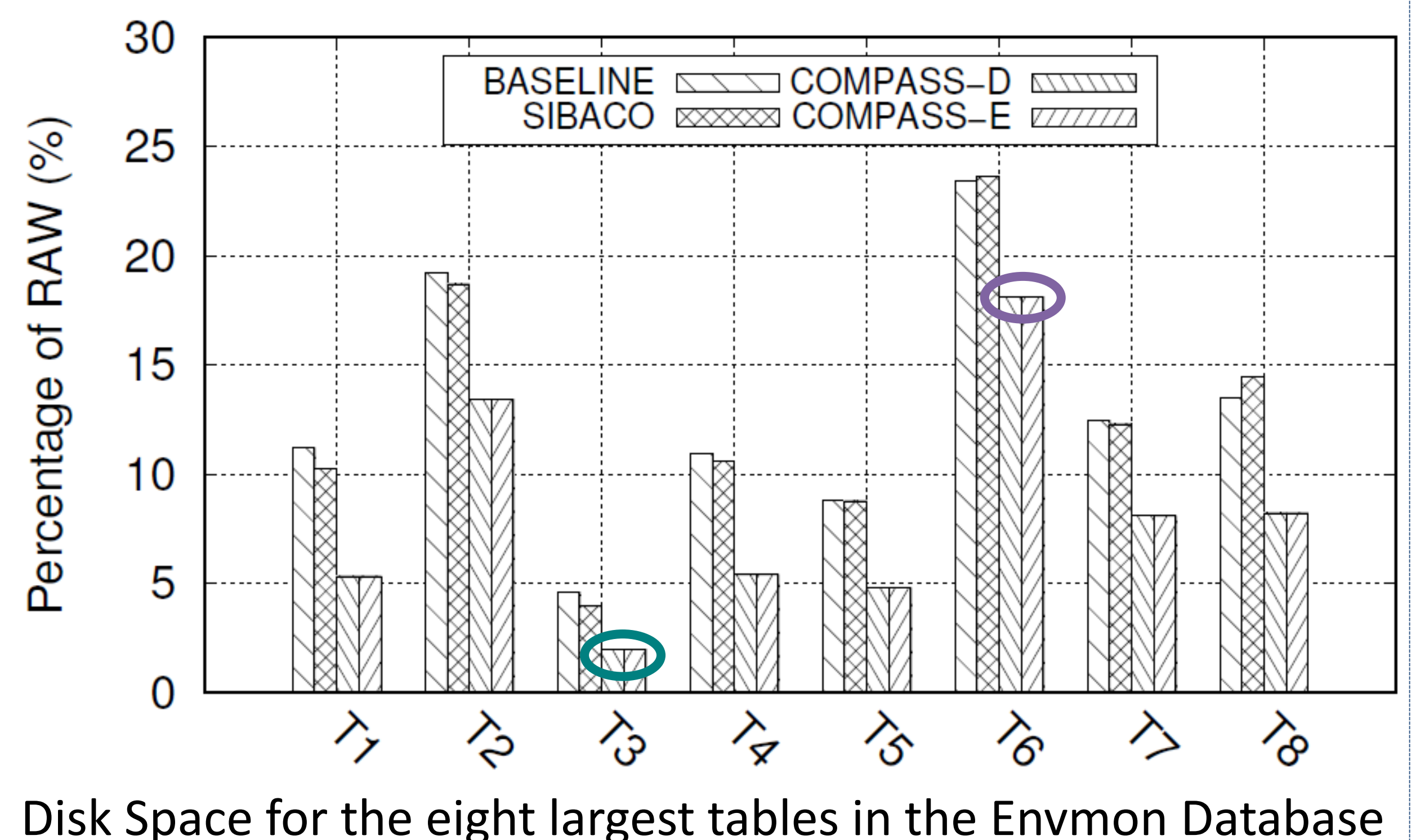
COMPASS Overview

- **COMPASS** consists of two stages:
  (i) partitioning and grouping; and
  (ii) selection of the best performing compression.

  - Stage I: **COMPASS** uses **K-means** clustering to **groups** similar columns together, based on their **data values** or **entropy**,
  - Stage II: applies the **best compression technique** for each cluster.

COMPASS Evaluation

- Experimental data indicates that **COMPASS** provides **significant reductions** in storage requirements compared to traditional "monolithic" methods
- COMPASS-D and COMPASS-E outperform the BASELINE and SIBACO (our prior work) techniques in terms of disk storage savings by more than **22%** in the worst case and **56% (i.e., ~2✕)** in the best case.

Disk Space for the eight largest tables in the Envmon Database