The Problem

- Similarity joins are a key tool in analyzing and processing data.
- Some standalone Similarity Join algorithms have been proposed.
- Little work on implementing Similarity Joins as physical database operators has been done.

Our Contribution

- DBSimJoin, a general Similarity Join database operator for metric spaces implemented inside PostgreSQL.
  - Non-blocking behavior
  - Prioritizes early generation of results
  - Fully supports the iterator interface

- We show how this operator can be used in real-world data analysis scenarios:
  - Identify similar images (vectors)
  - Identify similar publications (strings)

DBSimJoin Algorithm

- Partitions data in successive rounds until the partitions are small enough to be joined with a nested loop.
- Partitioning is done in a series of rounds.
- The algorithm is structured as a finite-state machine in order to support the database iterator interface.

DBSimJoin Rounds

- The first round partitions the input data. All partitions too large to be processed immediately in-memory are stored on-disk.
- Additional rounds re-partition partitions that have been stored on disk.

Performance

- Increasing Scale Factor SynthData[6D], Eps:2.5%
- Increasing Epsilon SynthData[6D], SF1

Exploiting Database Similarity Joins for Metric Spaces
Yasin N. Silva, Spencer S. Pearson
Arizona State University