Learning Set Cardinality in Distance Nearest Neighbours

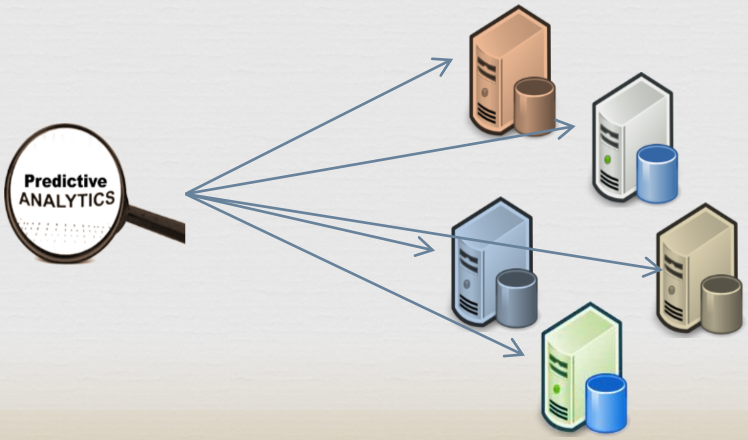
Christos Anagnostopoulos & Peter Triantafillou

School of Computing Science
University of Glasgow, UK

IEEE ICDM 2015; Atlantic City, NJ, USA

Observation

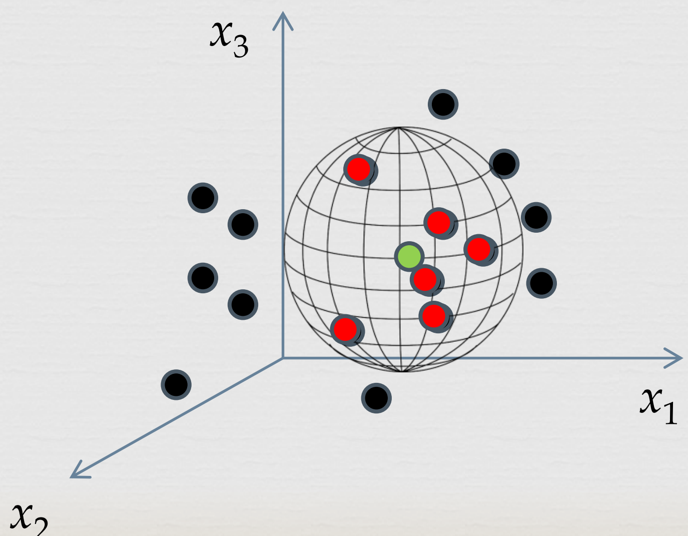
- Consider a **federation** of data nodes storing large datasets.
- Data nodes *locally* execute **cardinality-based nearest neighbours queries**, *e.g.*, *k*-NN, *distance-based*-NN.
- **Exploit** knowledge derived **only** from the queries for **predictive analytics** and **data exploration**.



Distance Nearest Neighbours Query

- Consider a point **x** in R^d, a radius θ, and a dataset *B*.
- A ***d*NN** query **q** = [**x**, θ] finds all points **x′** in dataset:

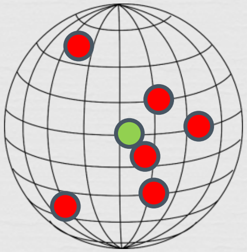
$$\mathbf{x}' \in B : \left\| \mathbf{x} - \mathbf{x}' \right\|_p \leq \theta$$



dataset *B*

*d*NN Query Cardinality

- Cardinality y is the *number* of points $\mathbf{x}'$ in the answer set of the query $\mathbf{q}$, with $0 \leq y \leq |B|$



$y = 6$

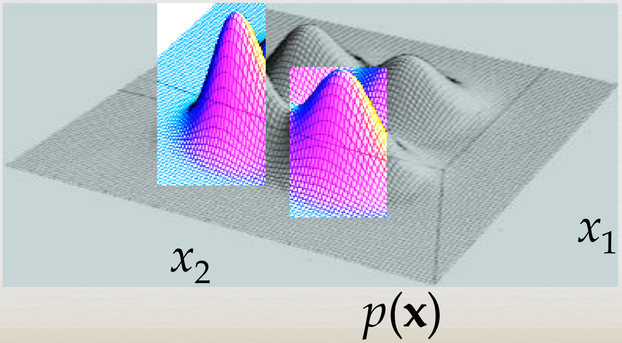
Set Cardinality Prediction (SCP)

- **Predict** the cardinality y of a dNN query $\mathbf{q}$ issued over a dataset B **without** executing the query.

Set Cardinality Prediction in Predictive Analytics

- Data analysts *define* data subspaces of **interest** through *d*NN queries,
  - *e.g.*, *local* statistics, data exploration tasks.

- **Not all** dataspaces are of the *same* interest to analysts
  - *Specific* regions of datasets are worth exploring.



Data-driven SCP

- A *data-driven* SCP, *e.g.*, histogram, sketching, sampling, **relies on the data**,
  - *i.e.*, requires *full access* to the data $\mathbf{x}$.

- For instance, an histogram estimates the *underlying* data probability distribution $p(\mathbf{x})$ for SCP.

Motivation

- **However**, access to nodes' data may be **restricted**:
  - **Confidentiality/security** reasons, *e.g.*, medical databases,
  - **Costly** data accesses, *e.g.*, in Cloud deployments, for maintaining accurate statistical structures,
  - In modern Big Data Systems the query processing engines **do not own** the data:
    - They are **oblivious** to updates/insertions/deletions;
    - It is **impossible** to **maintain** the statistical structures up-to-date (required by data-driven SCP, *e.g.*, histograms)

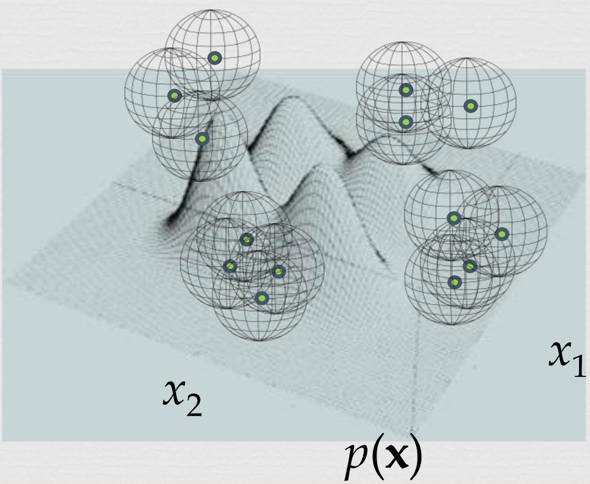
Idea: Query-driven SCP

- A *query-driven* SCP **extracts** knowledge about the data **without** accessing the data, but **only** from the *queries* and their *answers*.

- The only **available** knowledge is:
  - pairs of (query, answer)
  - *In our case*: (dNN query $\mathbf{q}$, answer set cardinality y)

SCP as a Machine Learning Problem

- Given a series of *past* pairs ($\mathbf{q}_i$, y_i) learn:
  - the *query patterns space* **instead** of the *data space* to identify **the areas of interest** to the users and
  - the *association* between query & cardinality



Problems

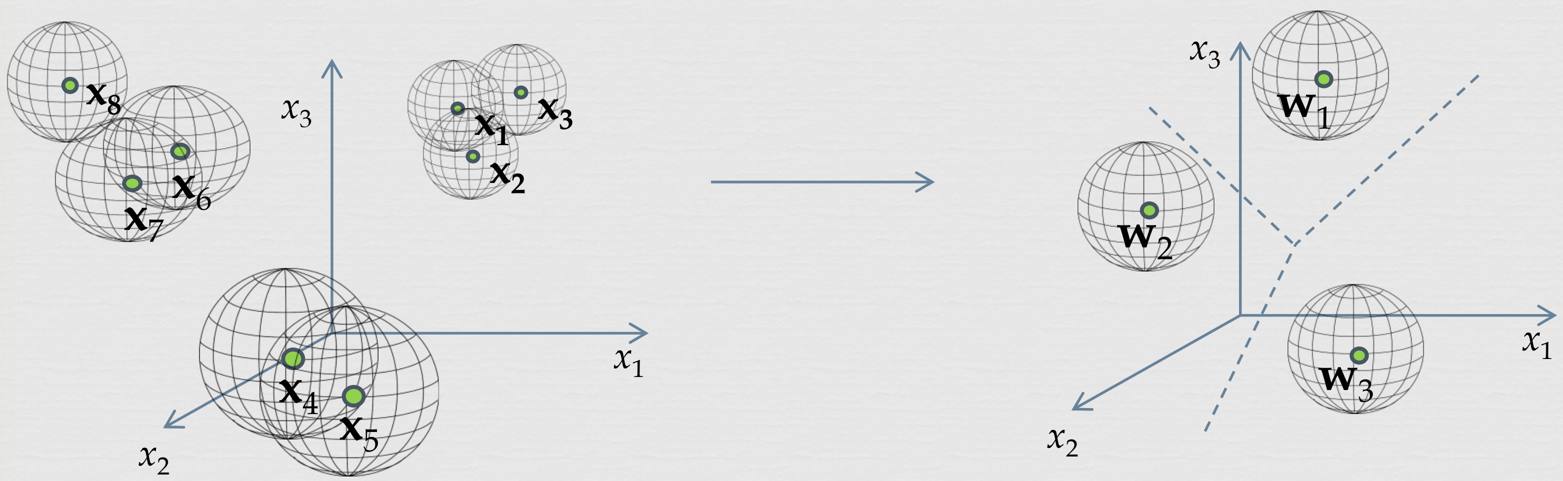
- **Problem 1.** Given an unseen *d*NN query $\mathbf{q}$, predict its cardinality y based **only** on the pairs $(\mathbf{q}_i, y_i)$ and **not** on the data $\mathbf{x}$.

- **Problem 2.** Enhance the query-driven SCP to *adapt* and *learn* on-the-fly the **new** query patterns.

- **Problem 3.** Enhance the query-driven SCP to incrementally *adapt* to **updates** of the data.

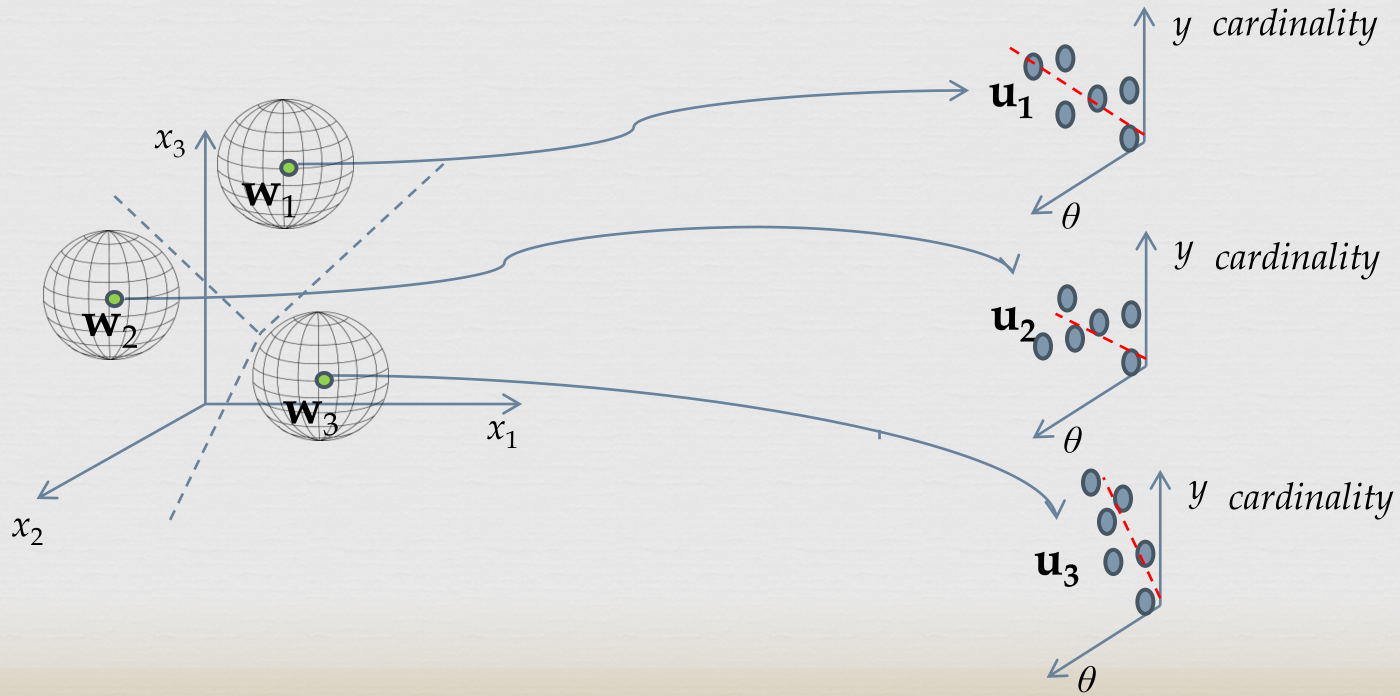
Unsupervised Regression

- **Learning Task 1 (Unsupervised):** partition the query space to identify query representatives *w.r.t. query similarity*.



Unsupervised Regression

- **Learning Task 2 (Supervised):** associate with each query prototype, a *localized regression coefficient* over *cardinality* and *radius* domain.



Competitive Learning Model

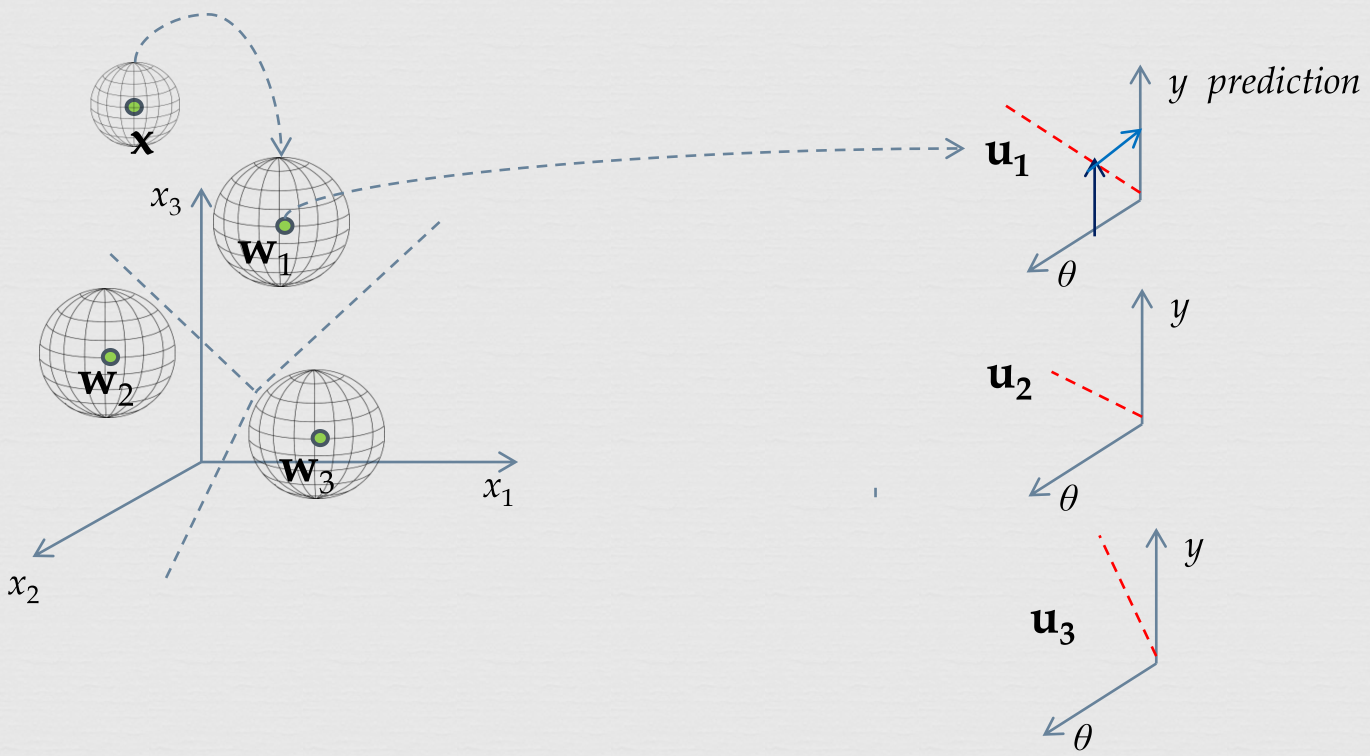
- The optimization function is:

$$J(\{\mathbf{w}_i\},\{\mathbf{u}_i\}) = E\left[\min_i \|\mathbf{q}-\mathbf{w}_i\|_{(p,1)}\right] + E\left[\left(y-\mathbf{u}_j^T\theta\right)^2 \mid j\right]$$

$$j = \arg\min_i \|\mathbf{q}-\mathbf{w}_i\|_{(p,1)}$$ refers to the *closest* query prototype

$$\|\mathbf{q}-\mathbf{w}_i\|_{(p,1)} = \frac{1}{2}\left(d^{-\frac{1}{p}}\|\mathbf{x}-\mathbf{x}_i\|_p + |\theta-\theta_i|\right)$$ refers to *distance* between queries

Cardinality Prediction



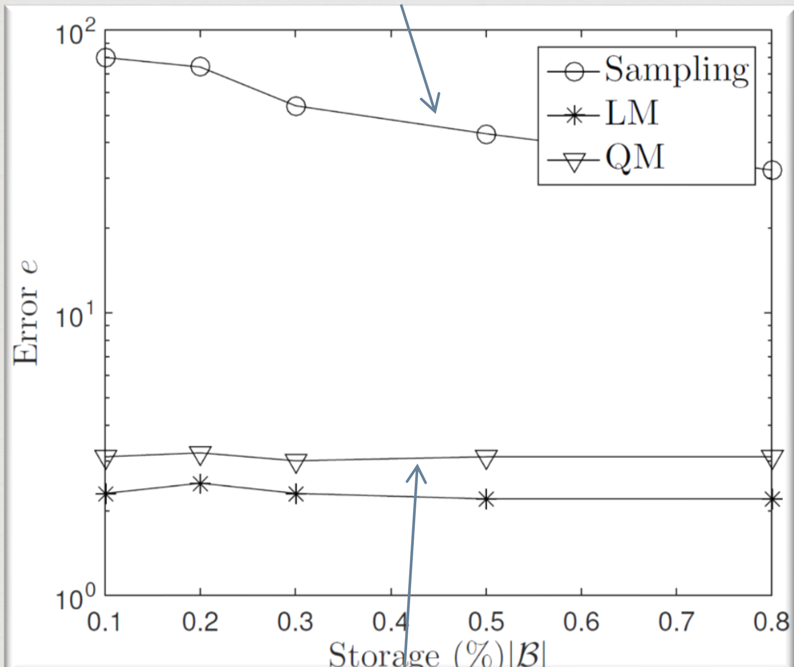
Performance Evaluation

- **Metrics:**
  - SCP accuracy (absolute relative error) *vs.* storage requirements (prototypes)

- **Comparative assessment**:
  - Data-centric SCP: *sampling, histograms, self-tuning histograms, Power-method*

Performance Evaluation

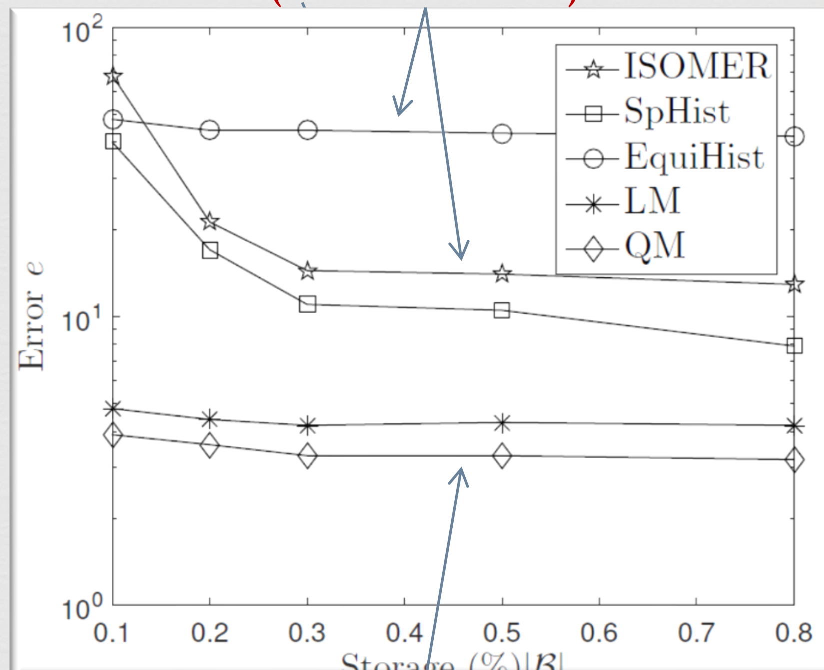
Sampling (Data-driven SCP)



Query-driven SCP

2-*dim*. data

Self-tuning histogram (Data-driven SCP)



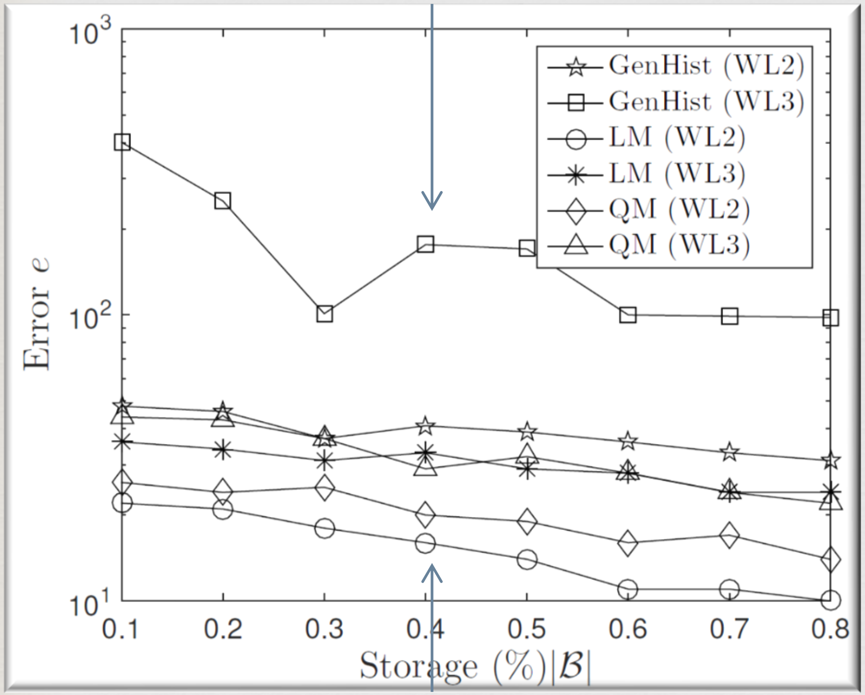
Query-driven SCP

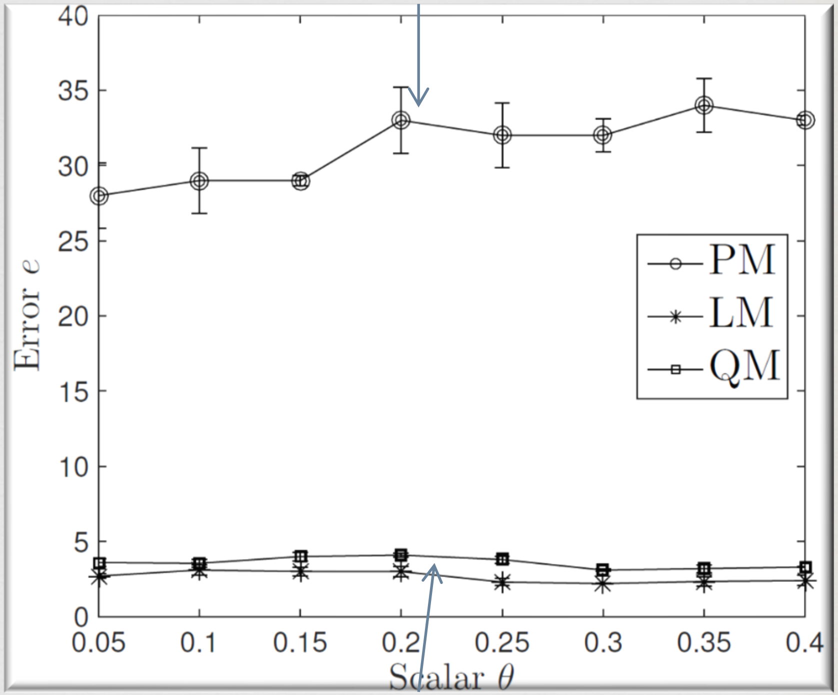
3-*dim*. data

Performance Evaluation

Histograms (Data-driven SCP)

Power-method (Data-driven SCP)





Query-driven SCP

10-*dim*. data

Query-driven SCP

2-*dim*. data

Thank you