

A Non-Linear Dimensionality-Reduction Technique for Similarity Search in Large Databases

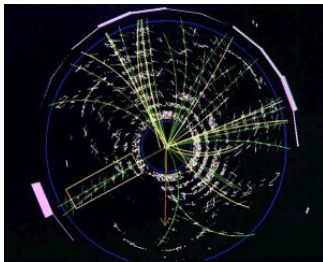
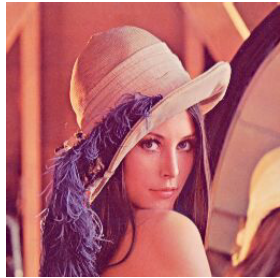
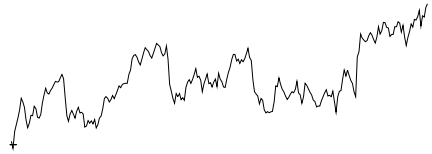
Khanh Vu, Kien A. Hua, Hao Cheng, and Sheau-Dong Lang

School of Electrical Engineering and Computer Science
University of Central Florida

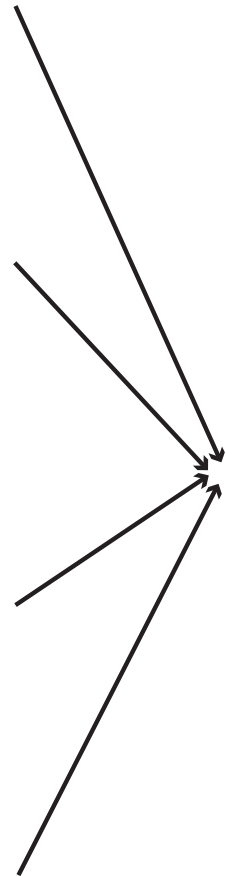
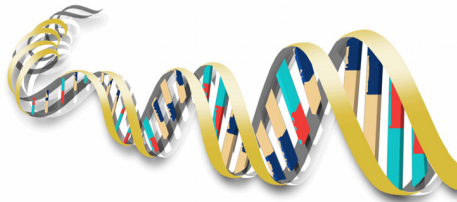
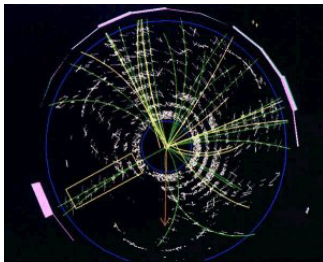
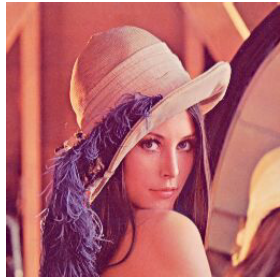
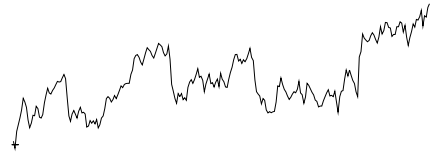
Outline

- Similarity Search
- Motivation and Solution
- The Proposed Technique
- Some Results
- Concluding Remarks

Similarity Search

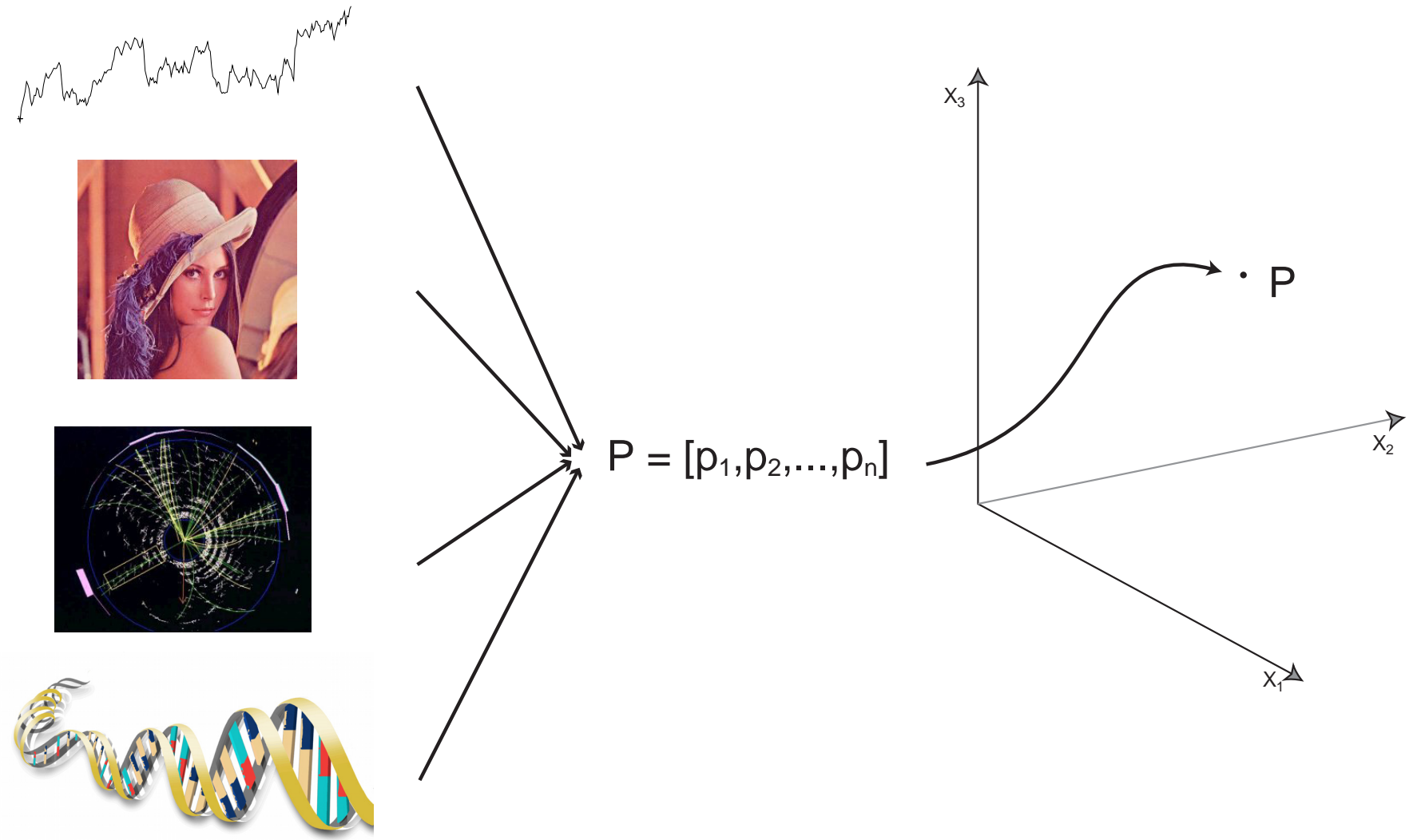


Similarity Search

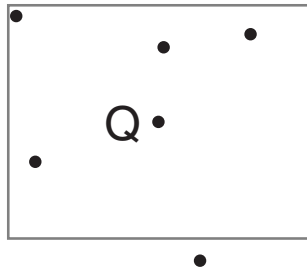


$$P = [p_1, p_2, \dots, p_n]$$

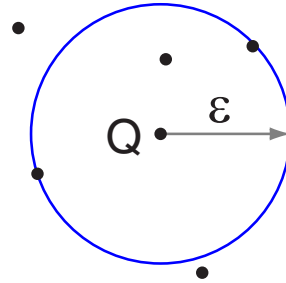
Similarity Search



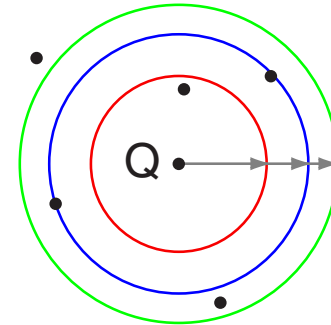
Similarity Search: Query Types



Window query

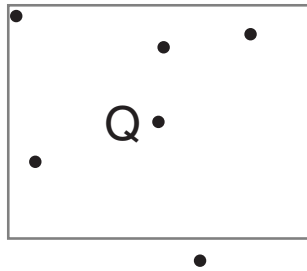


Spherical range query

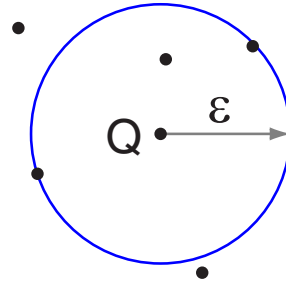


k-NN query

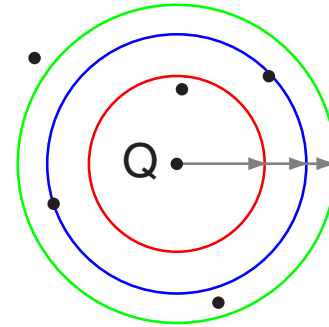
Similarity Search: Query Types



Window query



Spherical range query



k-NN query

We focus on the spherical range query:

- Window queries can be executed very efficiently.
- k-NN search can be implemented by means of spherical range queries with *expected* k-NN search radii.

Similarity Search: Distance Measure

We use the Euclidean distance (\mathcal{L}_2 norm):

- It is used extensively in research.
- Any finite metric space can be embedded into normed space \mathcal{L}_2 .
- It is possible to convert \mathcal{L}_p -based queries to \mathcal{L}_2 -based queries.

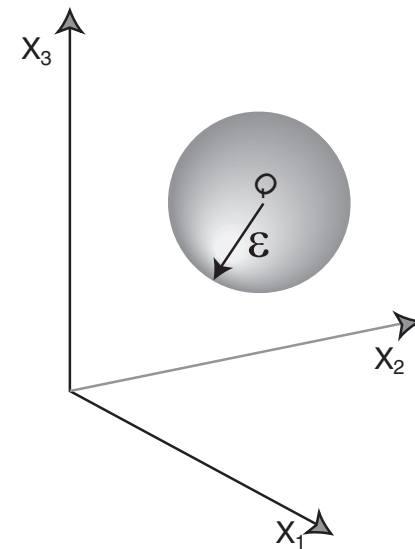
Similarity Search: Distance Measure

We use the Euclidean distance (\mathcal{L}_2 norm):

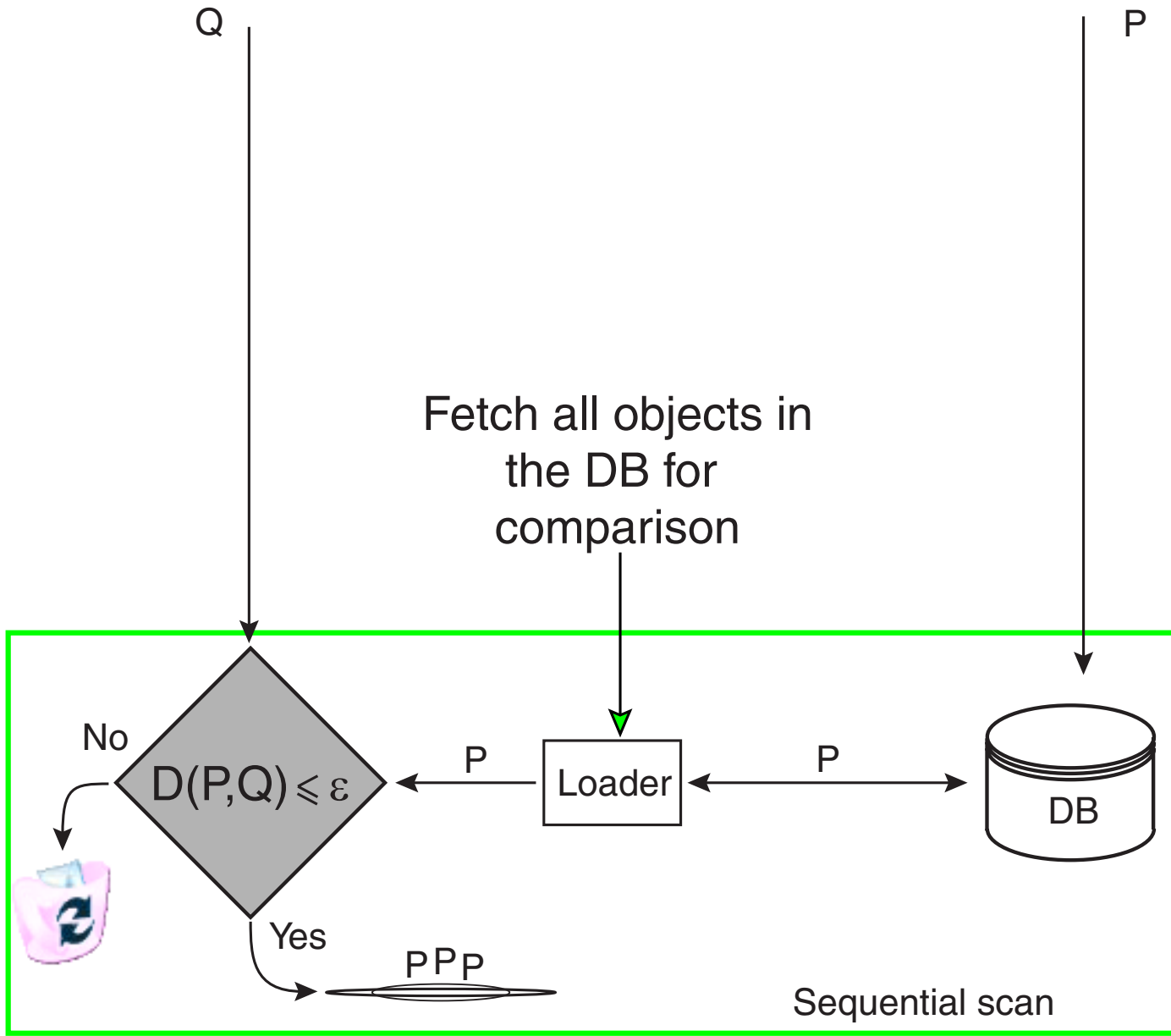
- It is used extensively in research.
- Any finite metric space can be embedded into normed space \mathcal{L}_2 .
- It is possible to convert \mathcal{L}_p -based queries to \mathcal{L}_2 -based queries.

Spherical range query: *Given point Q , find points P such that*

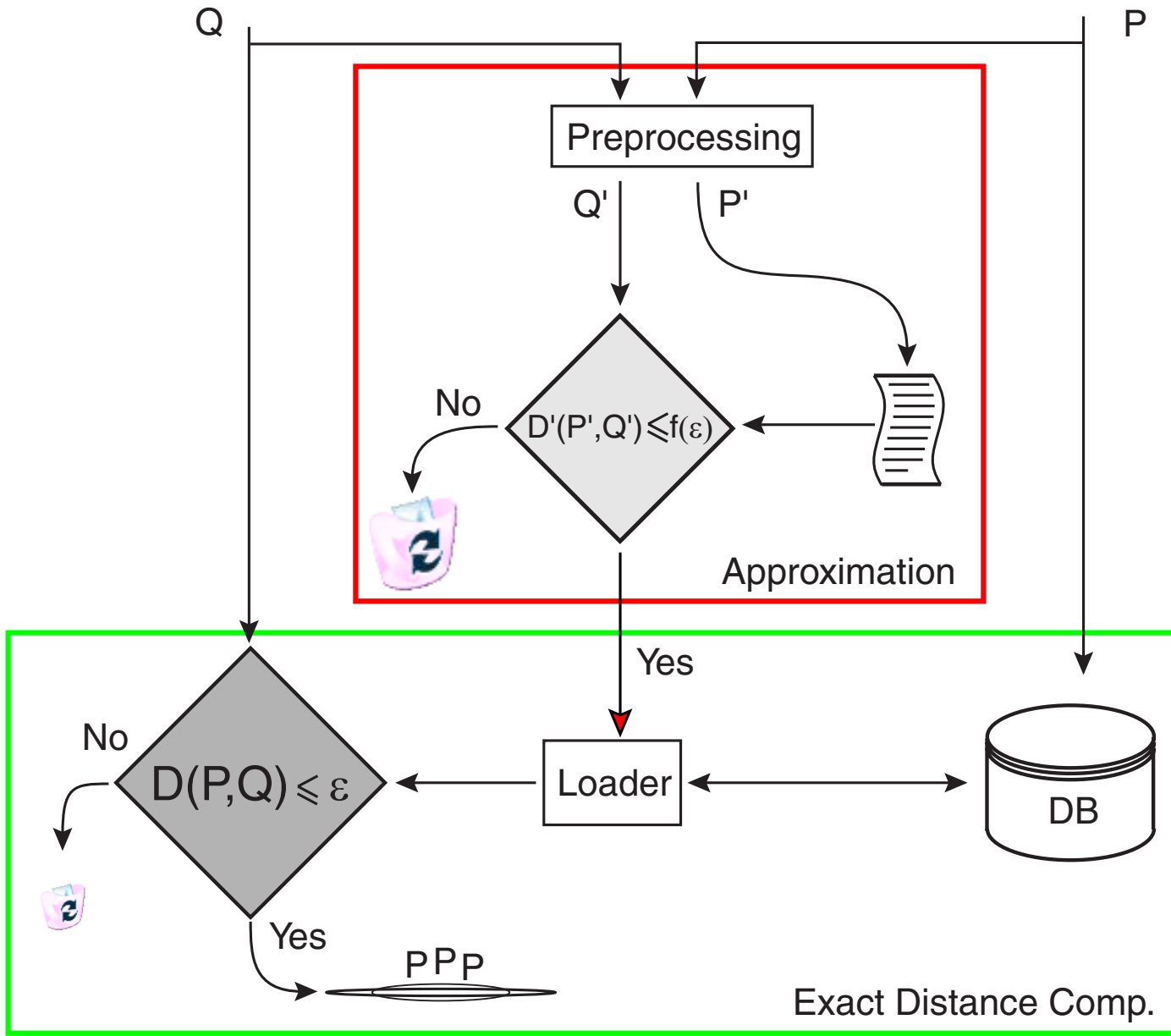
$$\mathcal{D}(P, Q) = \sqrt{\sum_{i=1}^n (p_i - q_i)^2} \leq \varepsilon .$$



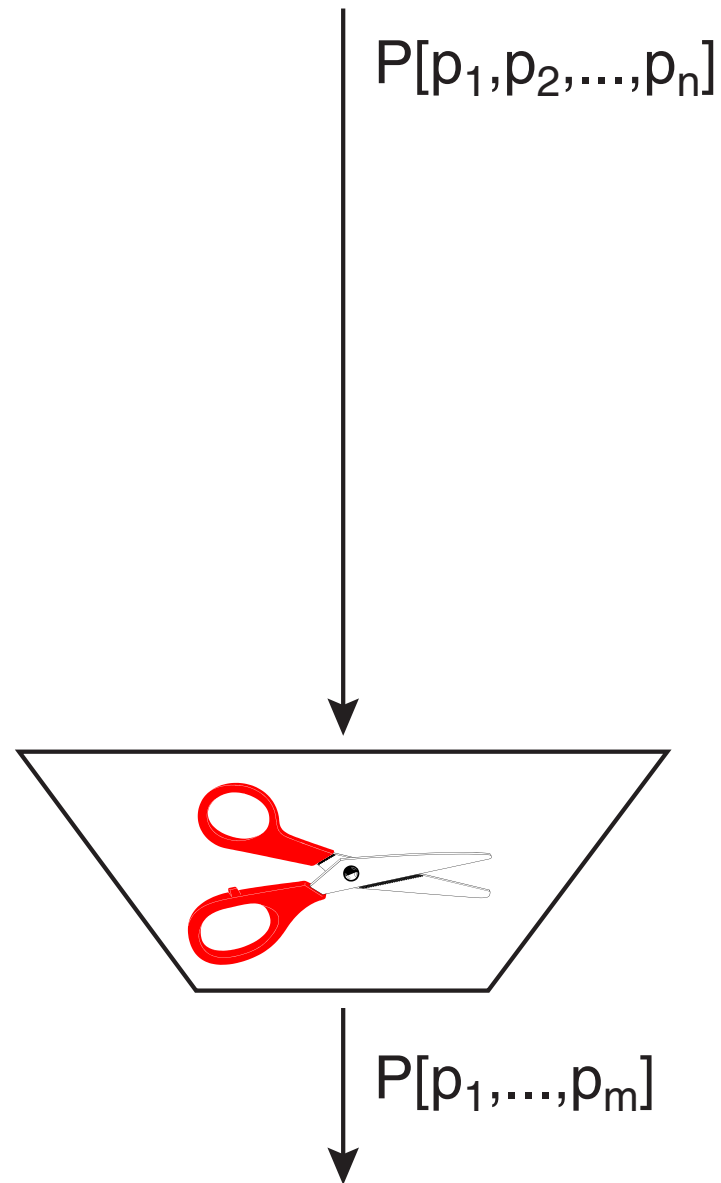
Similarity Search: 2-Step Search



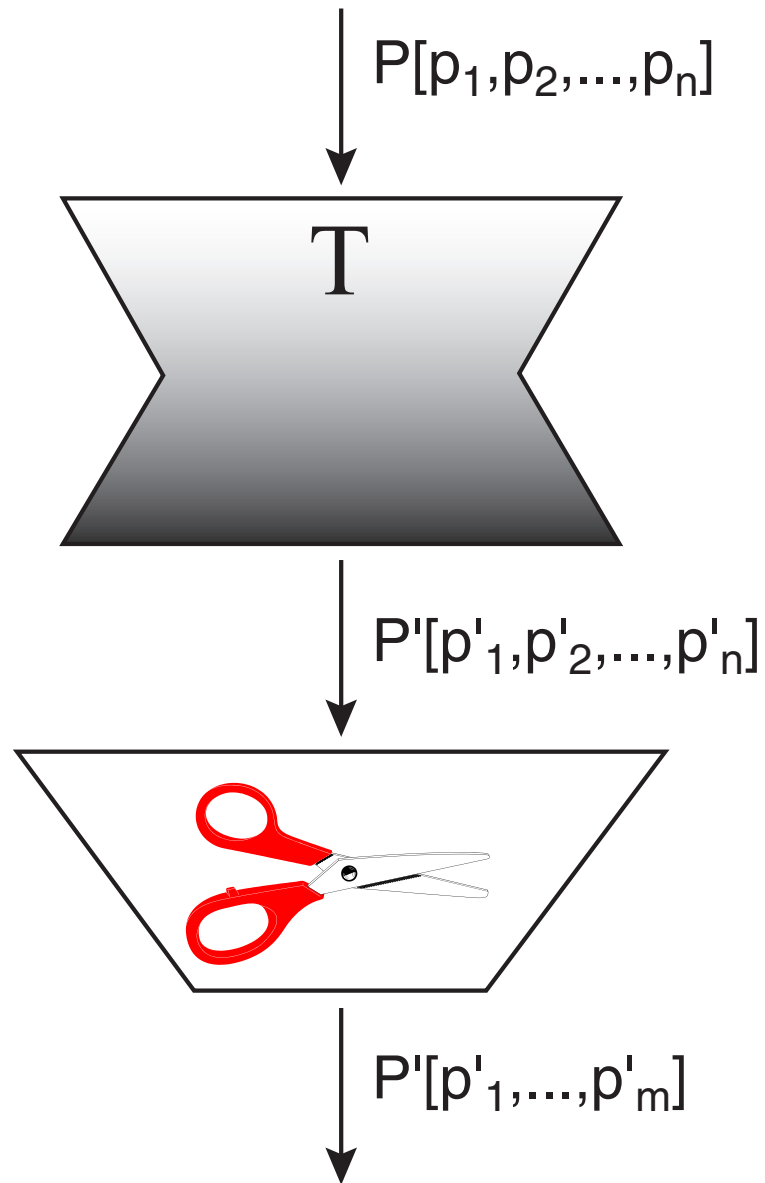
Similarity Search: 2-Step Search



Similarity Search: Dimensionality Reduction



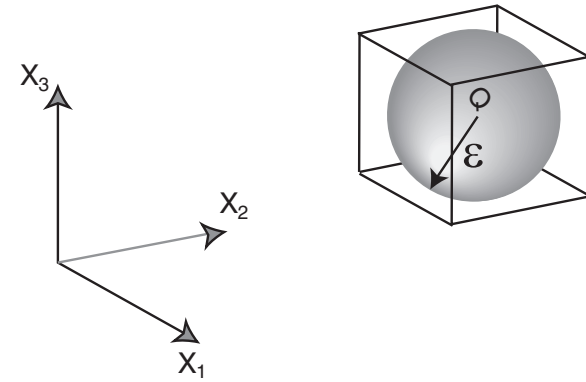
Similarity Search: Dimensionality Reduction



Similarity Search: Unbounded Approximation

Example: window approximation with no transformation

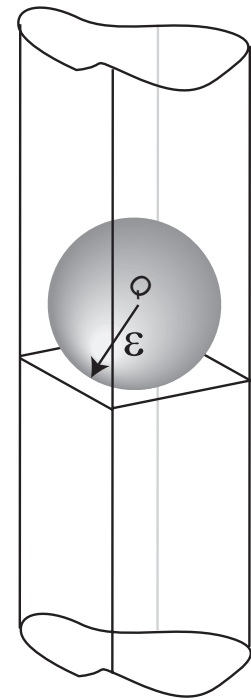
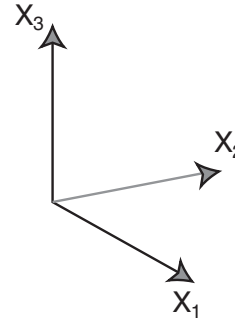
- An n -D hypercube:



Similarity Search: Unbounded Approximation

Example: window approximation with no transformation

- An n -D hypercube:
- An m -D hyperrectangle, $m < n$:



State-of-the-Art Techniques

Dimensionality reduction

- Transformation & reduction (DFT, DWT, SVD, etc.)
- Transformation (PAA, Segmented Means, etc.)

State-of-the-Art Techniques

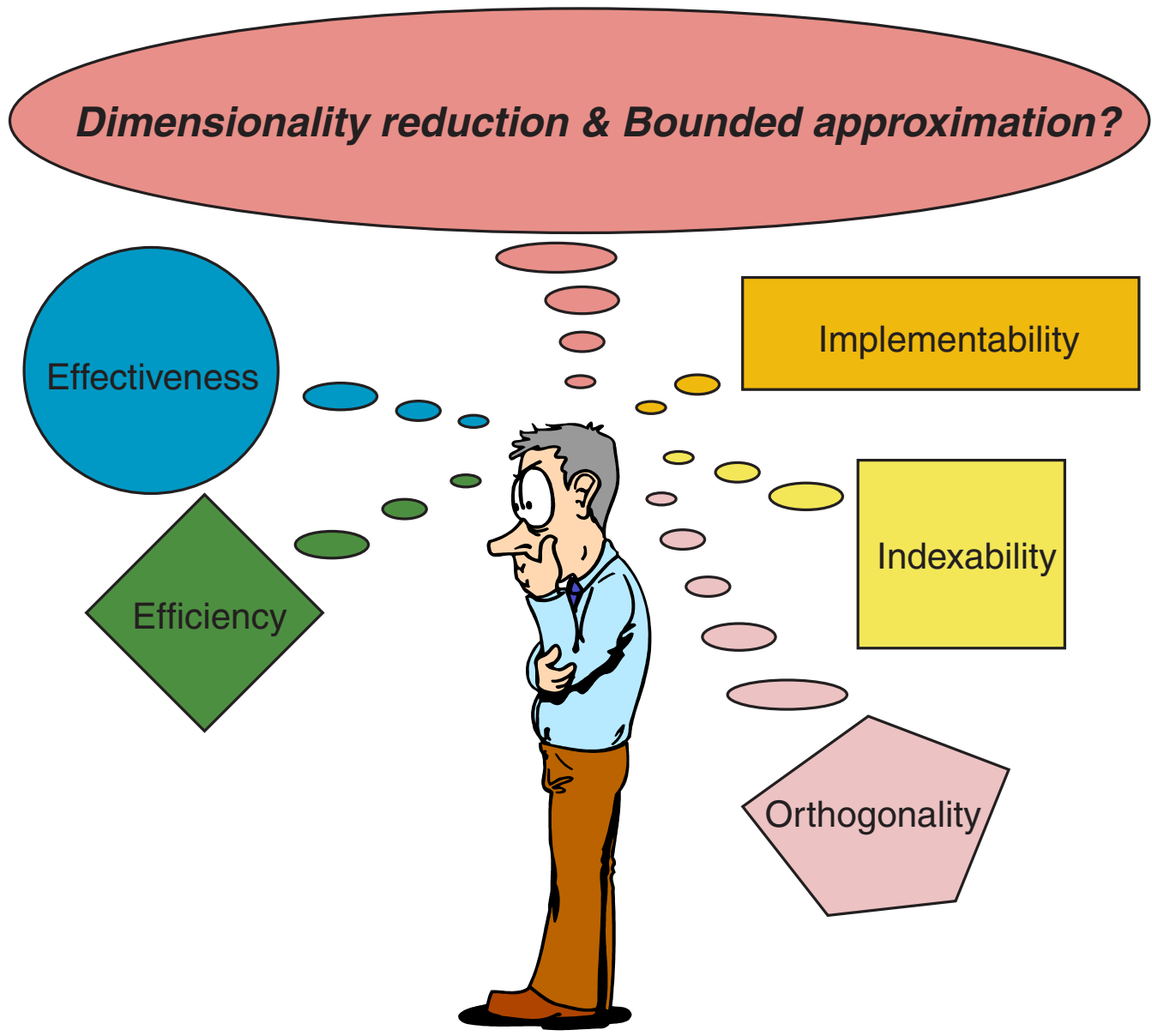
Dimensionality reduction

- Transformation & reduction (DFT, DWT, SVD, etc.)
- Transformation (PAA, Segmented Means, etc.)

Effects

- Unbounded
- Infinite volume
- Degraded performance

Similarity Search: Challenges



Non-Linear Transformation

Consider point P , define:

$$\mu_P = \frac{\sum_{i=1}^n p_i}{n}$$

$$\sigma_P = \left[\frac{\sum_{i=1}^n p_i^2}{n} - \mu_P^2 \right]^{(1/2)}$$

Non-Linear Transformation

Consider point P , define:

$$\mu_P = \frac{\sum_{i=1}^n p_i}{n}$$

$$\sigma_P = \left[\frac{\sum_{i=1}^n p_i^2}{n} - \mu_P^2 \right]^{(1/2)}$$

μ_P : *Mean*

σ_P : *Standard Deviation*

Bounds

If

$$\mathcal{D}(P, Q) = \sqrt{\sum_{i=1}^n (p_i - q_i)^2} \leq \varepsilon,$$

then

$$|\mu_P - \mu_Q| \leq \frac{\varepsilon}{\sqrt{n}}, \quad (1)$$

$$|\sigma_P - \sigma_Q| \leq \frac{\varepsilon}{\sqrt{n}}, \quad (2)$$

$$\left[(\mu_P - \mu_Q)^2 + (\sigma_P - \sigma_Q)^2 \right]^{1/2} \leq \frac{\varepsilon}{\sqrt{n}}. \quad (3)$$

Approximation of Search Sphere

- Using

$$|\mu_P - \mu_Q| \leq \frac{\varepsilon}{\sqrt{n}},$$

$$|\sigma_P - \sigma_Q| \leq \frac{\varepsilon}{\sqrt{n}}.$$

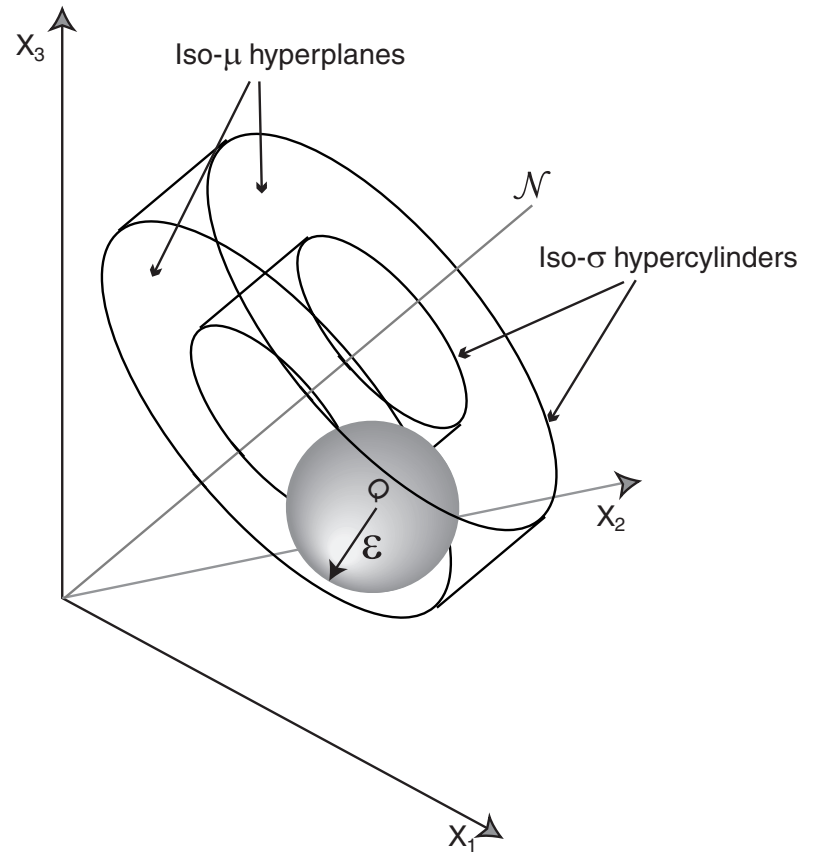
Approximation of Search Sphere

- Using

$$|\mu_P - \mu_Q| \leq \frac{\varepsilon}{\sqrt{n}},$$

$$|\sigma_P - \sigma_Q| \leq \frac{\varepsilon}{\sqrt{n}}.$$

- A bounded volume



Approximation of Search Sphere (2)

- Using

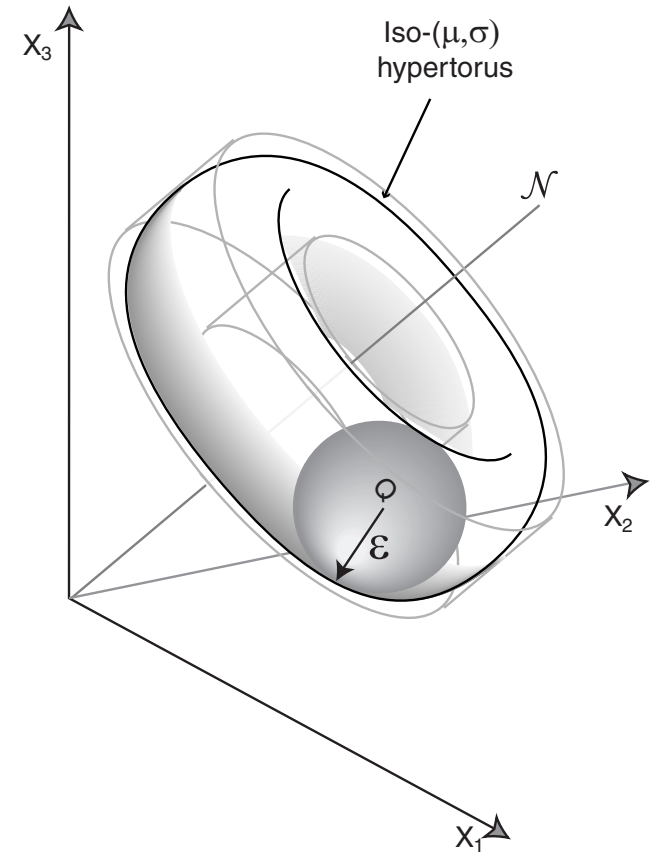
$$\left[(\mu_P - \mu_Q)^2 + (\sigma_P - \sigma_Q)^2 \right]^{1/2} \leq \frac{\varepsilon}{\sqrt{n}}.$$

Approximation of Search Sphere (2)

- Using

$$\left[(\mu_P - \mu_Q)^2 + (\sigma_P - \sigma_Q)^2 \right]^{1/2} \leq \frac{\varepsilon}{\sqrt{n}}.$$

- A bounded volume



Analysis

Equation (6):

$$(\mu_P - \mu_Q)^2 + (\sigma_P - \sigma_Q)^2 \leq \frac{\varepsilon^2}{n}.$$

When $\sigma_Q = 0$:

$$(\mu_P - \mu_Q)^2 + (\sigma_P - \sigma_Q)^2 = \frac{\sum_{i=1}^n (p_i - q_i)^2}{n}.$$

Analysis

Equation (6):

$$(\mu_P - \mu_Q)^2 + (\sigma_P - \sigma_Q)^2 \leq \frac{\varepsilon^2}{n}.$$

When $\sigma_Q = 0$:

$$(\mu_P - \mu_Q)^2 + (\sigma_P - \sigma_Q)^2 = \frac{\sum_{i=1}^n (p_i - q_i)^2}{n}.$$

Eq. (6) retrieves the exact set of qualifying points P !

Analysis

Equation (6):

$$(\mu_P - \mu_Q)^2 + (\sigma_P - \sigma_Q)^2 \leq \frac{\varepsilon^2}{n}.$$

When $\sigma_Q = 0$:

$$(\mu_P - \mu_Q)^2 + (\sigma_P - \sigma_Q)^2 = \frac{\sum_{i=1}^n (p_i - q_i)^2}{n}.$$

Eq. (6) retrieves the exact set of qualifying points P !

- Left side: only two dimensions

Analysis

Equation (6):

$$(\mu_P - \mu_Q)^2 + (\sigma_P - \sigma_Q)^2 \leq \frac{\varepsilon^2}{n}.$$

When $\sigma_Q = 0$:

$$(\mu_P - \mu_Q)^2 + (\sigma_P - \sigma_Q)^2 = \frac{\sum_{i=1}^n (p_i - q_i)^2}{n}.$$

Eq. (6) retrieves the exact set of qualifying points P !

- Left side: only two dimensions
- Right side: n is arbitrary.

Analysis

Equation (6):

$$(\mu_P - \mu_Q)^2 + (\sigma_P - \sigma_Q)^2 \leq \frac{\varepsilon^2}{n}.$$

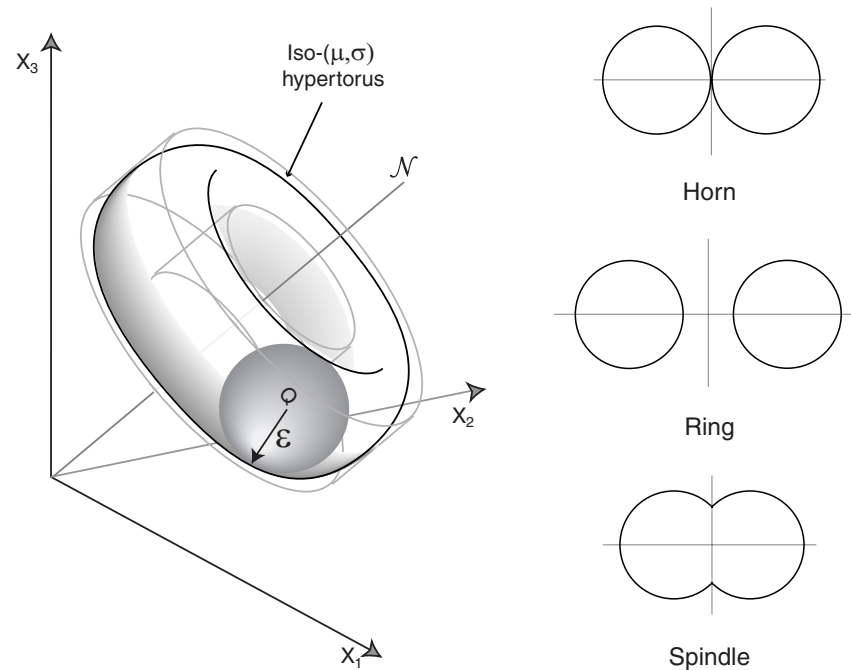
When $\sigma_Q = 0$:

$$(\mu_P - \mu_Q)^2 + (\sigma_P - \sigma_Q)^2 = \frac{\sum_{i=1}^n (p_i - q_i)^2}{n}.$$

Eq. (6) retrieves the exact set of qualifying points P !

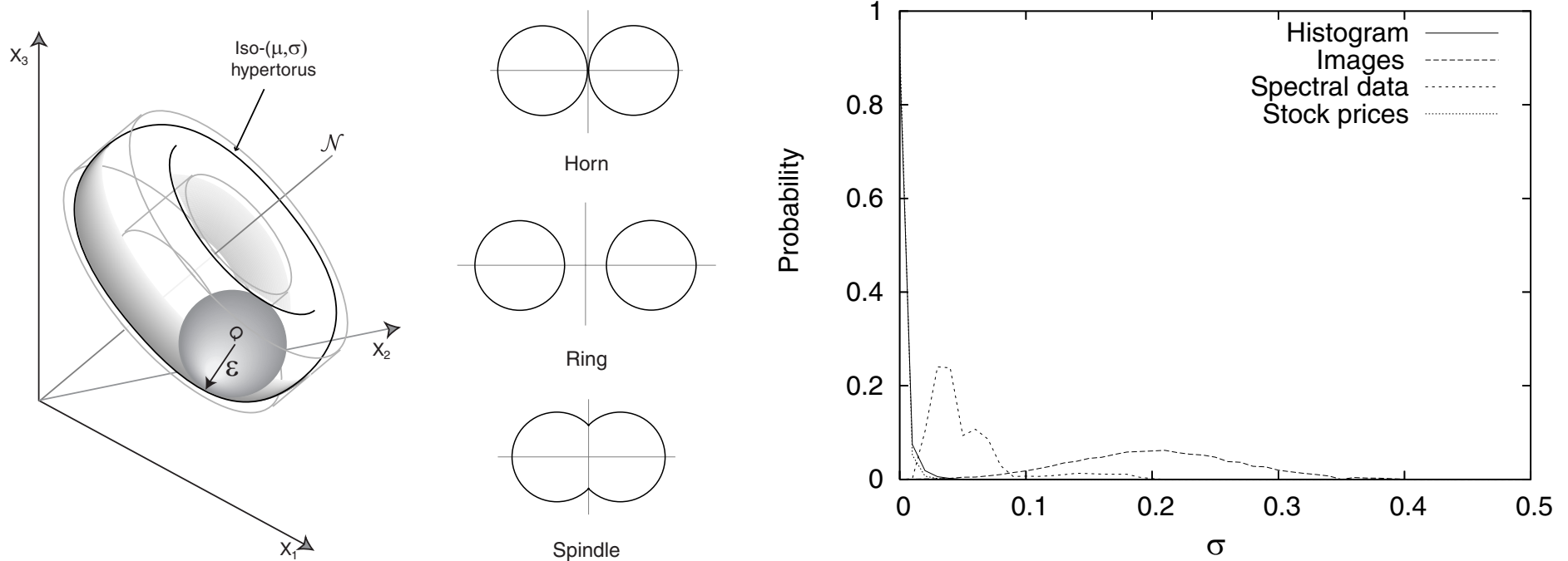
- Left side: only two dimensions
- Right side: n is arbitrary.
- Any data distribution

Analysis (2)



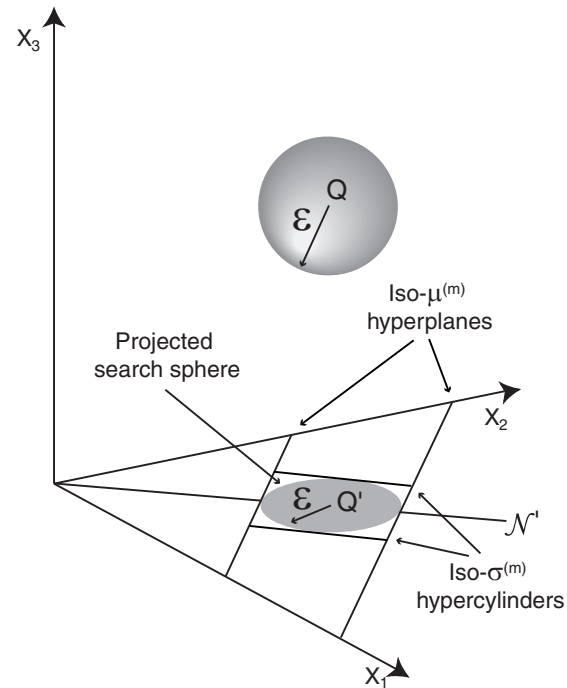
- $\sigma_Q = 0$, the approximation volume is identical to the search sphere.
- $\sigma_Q \rightarrow 0$, the approximation approaches the exact set.

Analysis (2)



- $\sigma_Q = 0$, the approximation volume is identical to the search sphere.
- $\sigma_Q \rightarrow 0$, the approximation approaches the exact set.
- Large portion of data have small σ

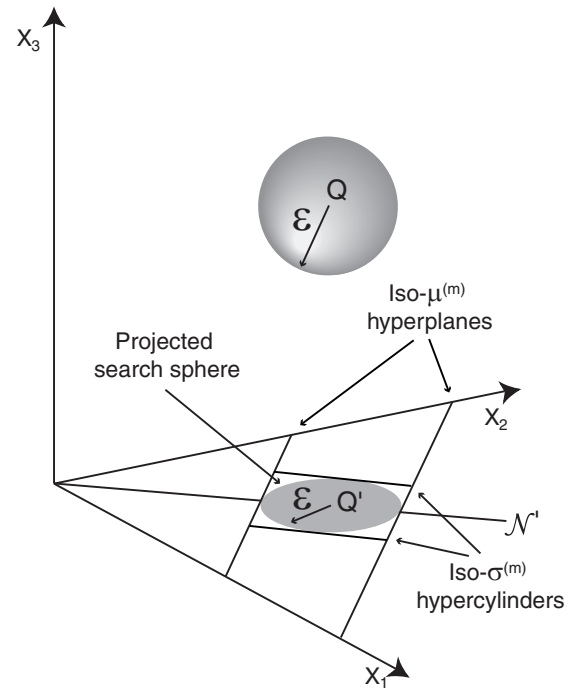
Approximation in Subspace



For $1 \leq m \leq n$:

$$\left[\sum_{i=1}^m (p_i - q_i)^2 \right]^{(1/2)} \leq \left[\sum_{i=1}^n (p_i - q_i)^2 \right]^{(1/2)} \leq \epsilon .$$

Approximation in Subspace



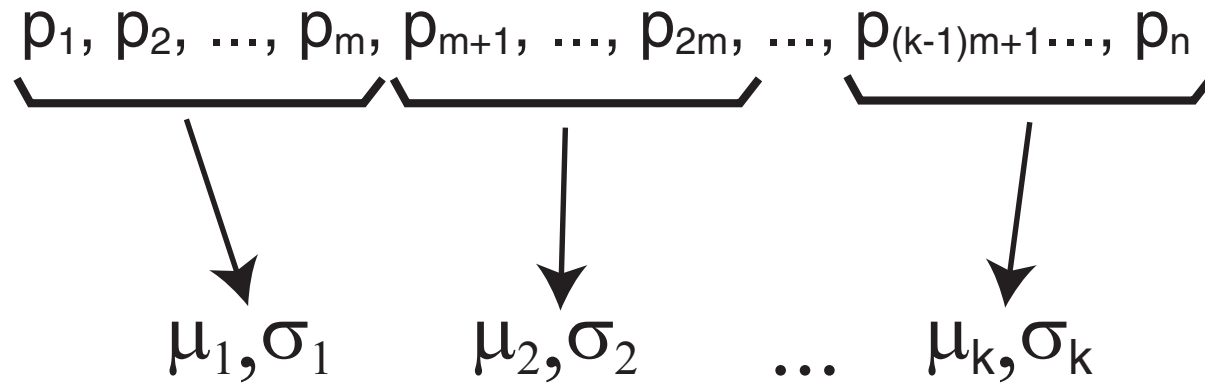
For $1 \leq m \leq n$:

$$\left[\sum_{i=1}^m (p_i - q_i)^2 \right]^{(1/2)} \leq \epsilon .$$

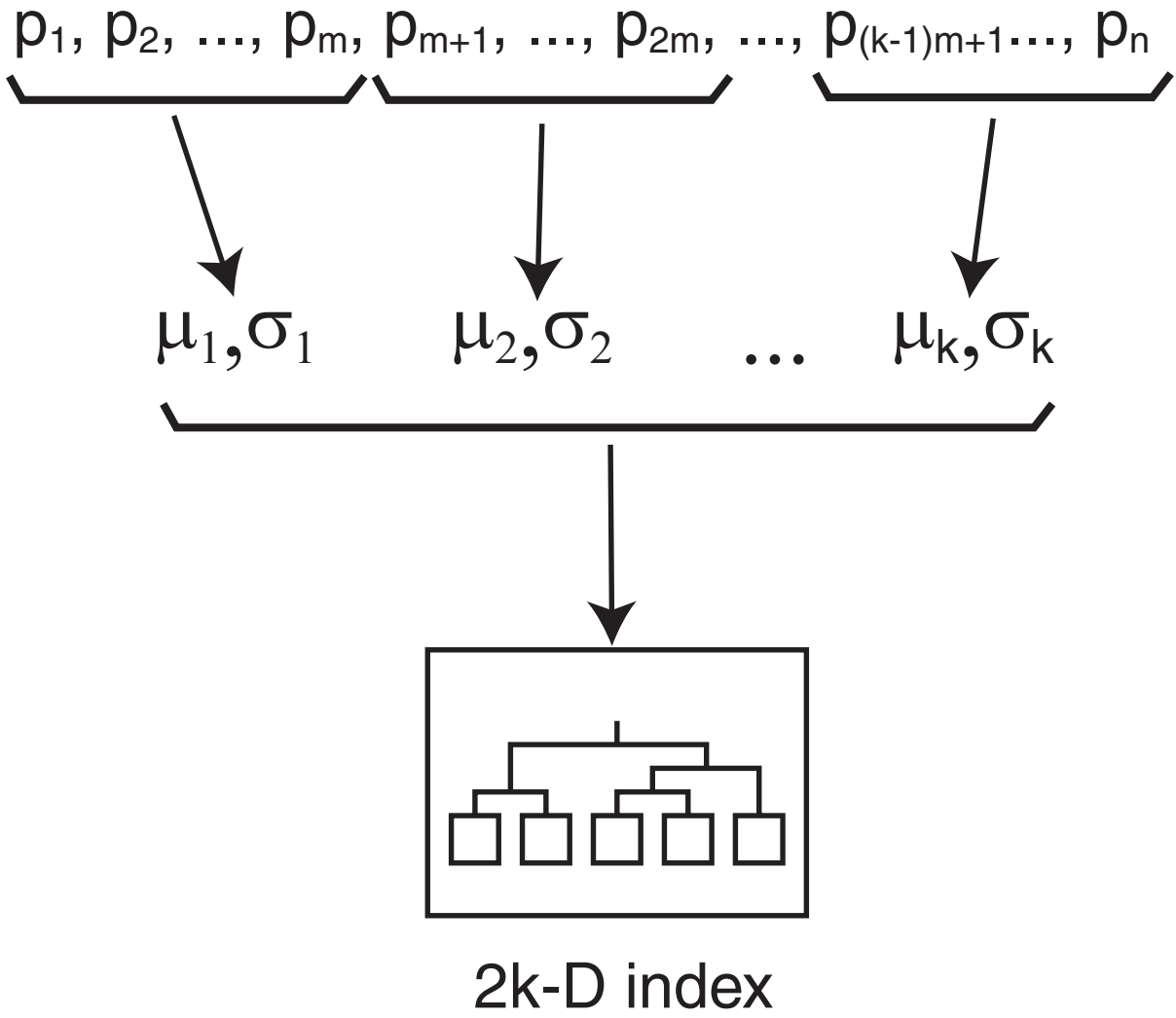
Indexing

$p_1, p_2, \dots, p_m, p_{m+1}, \dots, p_{2m}, \dots, p_{(k-1)m+1}, \dots, p_n$

Indexing



Indexing



2-Phased Approximation

Idea:

1. Retrieve points P if

$$|\mu_P - \mu_Q| \leq \frac{\varepsilon}{\sqrt{n}} \quad \text{and}$$

$$|\sigma_P - \sigma_Q| \leq \frac{\varepsilon}{\sqrt{n}}.$$

2-Phased Approximation

Idea:

1. Retrieve points P if

$$|\mu_P - \mu_Q| \leq \frac{\varepsilon}{\sqrt{n}} \quad \text{and}$$

$$|\sigma_P - \sigma_Q| \leq \frac{\varepsilon}{\sqrt{n}}.$$

2. For returned points P, keep P if

$$\left[(\mu_P - \mu_Q)^2 + (\sigma_P - \sigma_Q)^2 \right]^{1/2} \leq \frac{\varepsilon}{\sqrt{n}}.$$

Searching (2)

To retrieve points within radius ε of query point Q ,

- Approximation:

Searching (2)

To retrieve points within radius ε of query point Q ,

- Approximation:
 - Group the dimensions of Q in the same manner.

Searching (2)

To retrieve points within radius ε of query point Q ,

- Approximation:
 - Group the dimensions of Q in the same manner.
 - Compute k pairs of μ, σ for Q .

Searching (2)

To retrieve points within radius ε of query point Q ,

- Approximation:
 - Group the dimensions of Q in the same manner.
 - Compute k pairs of μ, σ for Q .
 - Perform window search of range $\varepsilon \sqrt{\frac{k}{n}}$ centered at Q .

Searching (2)

To retrieve points within radius ε of query point Q ,

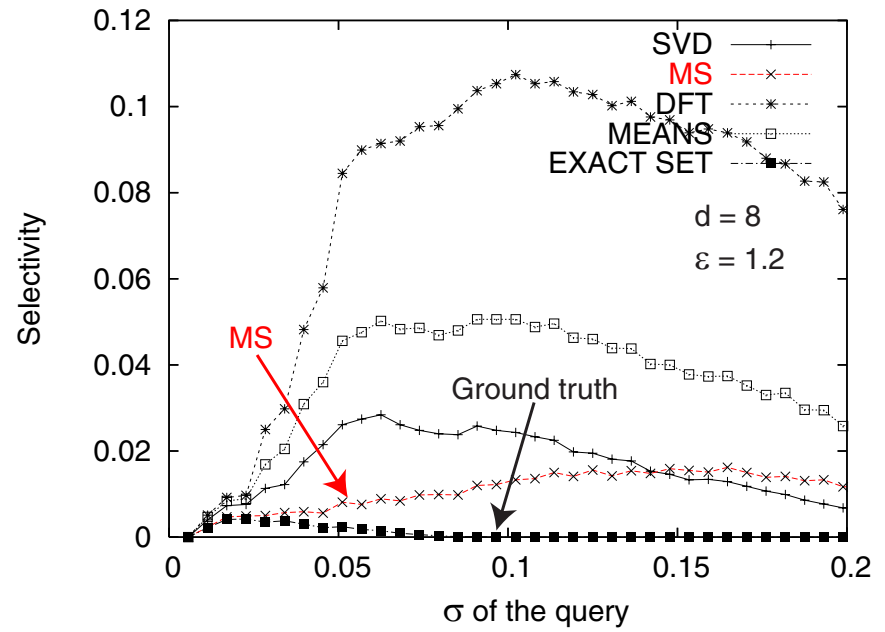
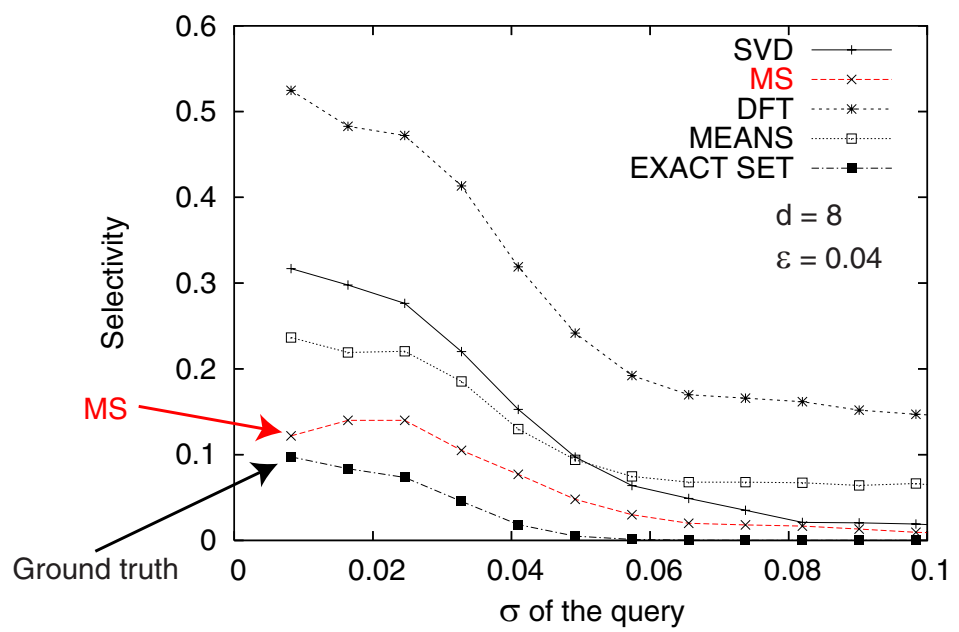
- Approximation:
 - Group the dimensions of Q in the same manner.
 - Compute k pairs of μ, σ for Q .
 - Perform window search of range $\varepsilon \sqrt{\frac{k}{n}}$ centered at Q .
 - Filter returned points using lower-bounding formula.

Searching (2)

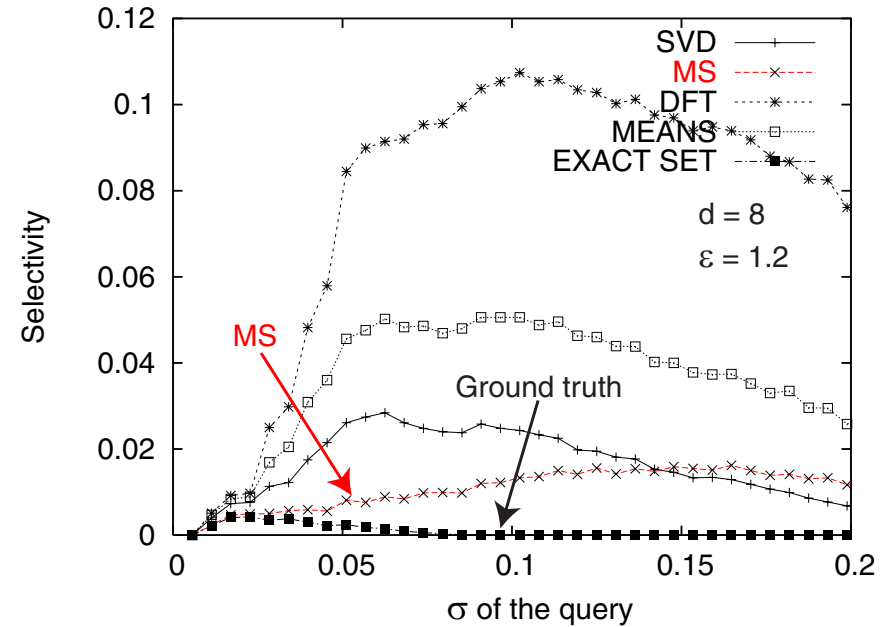
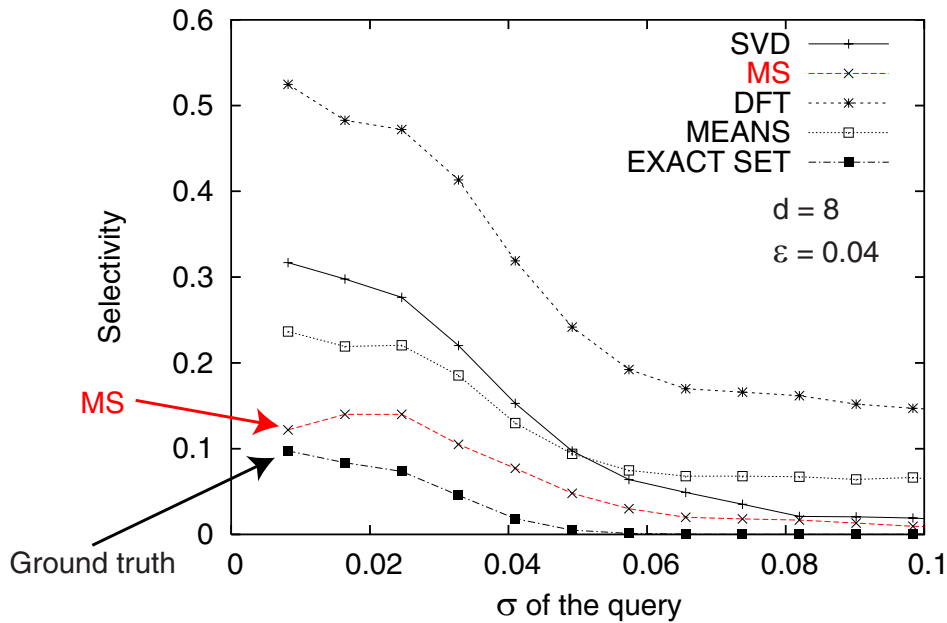
To retrieve points within radius ε of query point Q ,

- Approximation:
 - Group the dimensions of Q in the same manner.
 - Compute k pairs of μ, σ for Q .
 - Perform window search of range $\varepsilon \sqrt{\frac{k}{n}}$ centered at Q .
 - Filter returned points using lower-bounding formula.
- Discard irrelevant points using the exact distance.

Performance



Performance



- Approximation set approaches the exact set as $\sigma \rightarrow 0$.
- MS outperforms SOTA techniques by a wide margin.

Performance: Phase 1

Average *Selectivity* of SOTA techniques relative to *MS* after Phase 1

Dataset	Size	n	d	Search range	SVD	OMNI	DFT	PAA/MEANS
Histograms	15,766	256	8	0.005–0.2	0.35	16.63	0.75	0.21
16x16 images	15,766	256	8	0.05–2.0	-0.26	33.58	2.94	1.13
8x8 images	12,000	64	2	0.05–2.0	0.18	1.40	1.63	0.51
Stock Prices	6,500	256	8	0.005–0.2	-0.27	1.12	0.29	0.31
Spectral data	4,435	32	4	0.005–0.2	-0.08	0.59	1.86	0.78

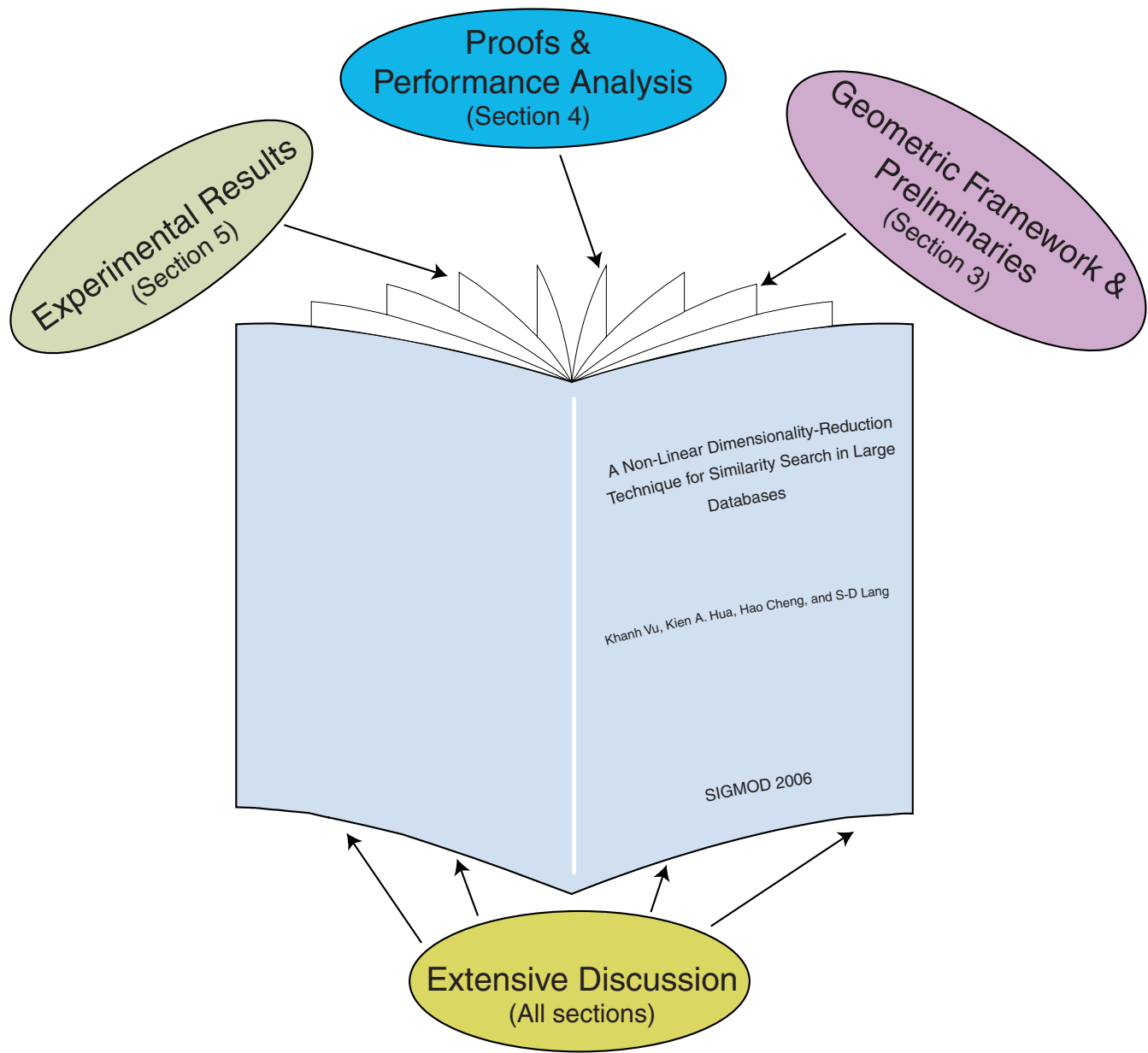
Example: On average, DFT returns 0.29 = 29% more points than MS for stock data.

Performance: Phase 2

Average *Selectivity* of SOTA techniques relative to *MS* after Phase 2

Dataset	Size	n	d	Search range	SVD	OMNI	DFT	PAA/MEANS
Histograms	15,766	256	8	0.005–0.2	0.49	56.10	1.05	0.25
16x16 images	15,766	256	8	0.05–2.0	-0.21	160.93	6.47	1.40
8x8 images	12,000	64	2	0.05–2.0	0.21	1.93	2.23	0.50
Stock Prices	6,500	256	8	0.005–0.2	-0.02	3.33	0.14	-0.01
Spectral data	4,435	32	4	0.005–0.2	-0.25	3.16	2.21	0.80

Additional Material



Concluding Remarks

