Compact Data Representations and their Applications

Moses Charikar
Princeton University

Lots and lots of data

- AT&T
- Information about who calls whom
- What information can be got from this data ?

- Network router
- Sees high speed stream of packets
- Detect DOS attacks ?
  fair resource allocation ?

Lots and lots of data

- Google search engine
- About 3 billion web pages
- Many many queries every day
- How to efficiently process data ?
  - Eliminate near duplicate web pages
  - Query log analysis

Sketching Paradigm

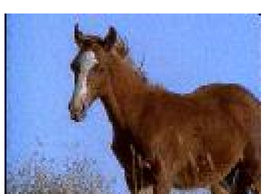
- Construct compact representation (sketch) of data such that

- Interesting functions of data can be ~~computed~~ estimated from compact representation

complex object

sketch

0 1 0 1 1 0 0 1 1 0



0 0 1 0 1 1 0 0 1 0

Why care about compact representations ?

- Practical motivations
  - Algorithmic techniques for massive data sets
  - Compact representations lead to reduced space, time requirements
  - Make impractical tasks feasible

- Theoretical Motivations
  - Interesting mathematical problems
  - Connections to many areas of research

Questions

- What is the data ?
- What functions do we want to compute on the data ?
- How do we estimate functions on the sketches ?

- Different considerations arise from different combinations of answers

- Compact representation schemes are functions of the requirements

What is the data ?

- Sets, vectors, points in Euclidean space, points in a metric space, vertices of a graph.

- Mathematical representation of objects (e.g. documents, images, customer profiles, queries).

What functions do we want to compute on the data ?

- Local functions : pairs of objects
  e.g. distance between objects
- Sketch of each object, such that function can be estimated from pairs of sketches

- Global functions : entire data set
  e.g. statistical properties of data
- Sketch of entire data set, ability to update, combine sketches

Local functions: distance/similarity

- Distance is a general metric, i.e satisfies triangle inequality

- Normed space
  $x = (x_1, x_2, \ldots, x_d)$ $y = (y_1, y_2, \ldots, y_d)$

$$d(x, y) = \left(\sum_{i=1}^{d} | x_i - y_i |^p \right)^{1/p}$$

$$L_p \text{ norm} \qquad L_1, L_2, L_\infty$$

- Other special metrics
  (e.g. Earth Mover Distance)

Estimating distance from sketches

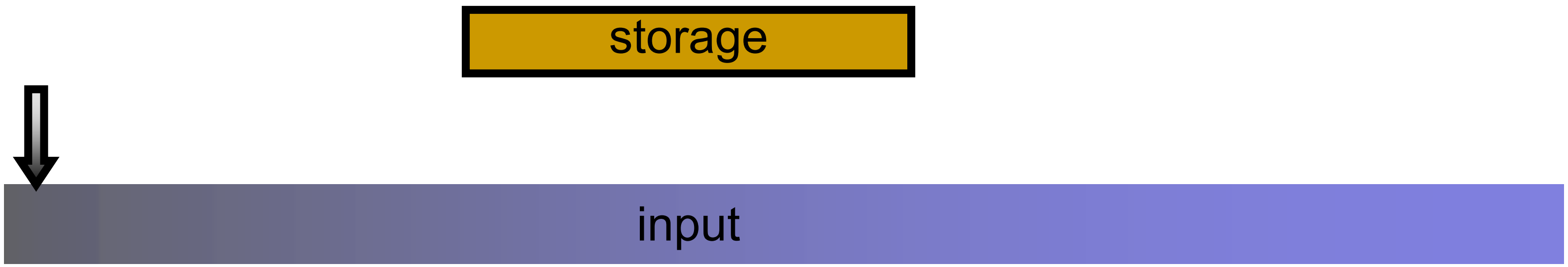
- Arbitrary function of sketches
  - Information theory, communication complexity question.
- Sketches are points in normed space
  - Embedding original distance function in normed space. [Bourgain '85] [Linial,London,Rabinovich '94]
- Original metric is (same) normed space
  - Original data points are high dimensional
  - Sketches are points low dimensions
  - Dimension reduction in normed spaces [Johnson Lindenstrauss '84]

Global functions

- Statistical properties of entire data set
- Frequency moments
- Sortedness of data
- Set membership
- Size of join of relations
- Histogram representation
- Most frequent items in data set
- Clustering of data

Streaming algorithms

- Perform computation in one (or constant) pass(es) over data using a small amount of storage space



- Availability of sketch function facilitates streaming algorithm
- Additional requirements - sketch should allow:
  - Update to incorporate new data items
  - Combination of sketches for different data sets

Goals

- Glimpse of compact representation techniques in the sketching and streaming domains.

- Basic ideas, no messy details

Talk Outline

- Classical techniques: spectral methods
- Dimension reduction
- Similarity preserving hash functions
  - sketching vector norms
  - sketching Earth Mover Distance (EMD)

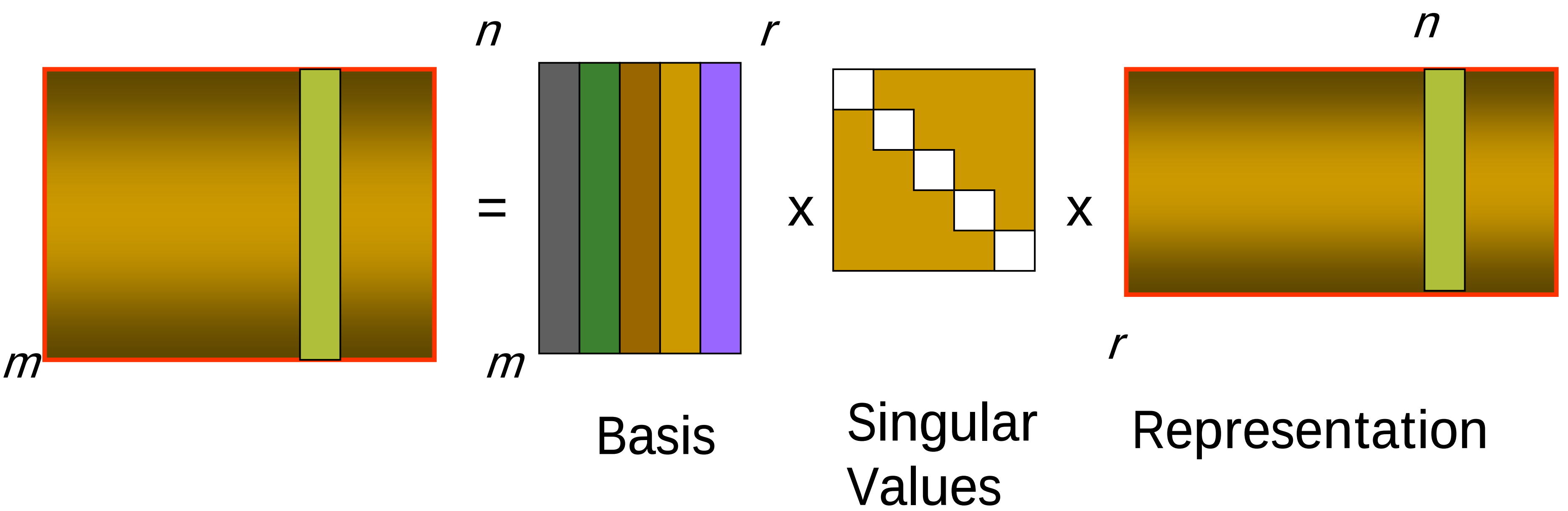
Spectral methods: approximating matrices

- SVD: Singular Value Decomposition
  LSI: Latent Semantic Indexing

- Related to
  PCA: Principal Component Analysis
  MDS: MultiDimensional Scaling

SVD Matrix Factorization

$X \quad = \quad U \quad \times \quad \Sigma \quad \times \quad V^T$



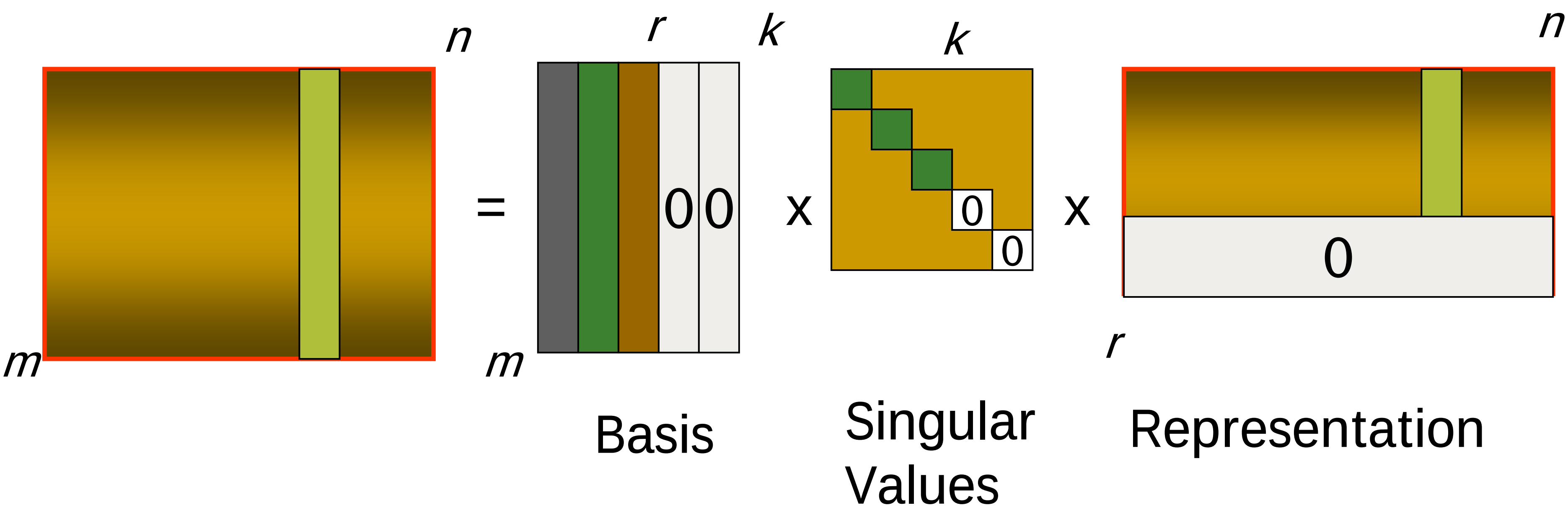
Restrictions on representation: U, V orthonormal; Σ diagonal

Matrix approximation

- $X = \sum_i \mathbf{u}_i s_i \mathbf{v}_i^T$
- $X^{(k)} = \sum_{i=1}^{k} \mathbf{u}_i s_i \mathbf{v}_i^T$
- $X^{(k)}$ is best rank k approximation to X minimizes $\sum_{ij} |x_{ij} - x^{(k)}_{ij}|^2$

Dimension Reduction

$$X_r = U \times \Sigma_r \times V^T$$



The columns of X_r represent the docs, but in $r \ll m$ dimensions
Best rank r approximation according to 2-norm

Closely related notions

- Singular Value Decomposition
- Karhunen-Loeve (KL) Transform
- Principal Component Analysis (PCA)
- Latent Semantic Indexing (LSI)
  - Information retrieval

SVD complexity

- $O(\min(nm^2, mn^2))$
- Less work
  - if we want just eigenvalues
  - if we want first k eignevectors
  - if matrix is sparse
- Implemented in any linear algebra package (LINPACK, matlab, Splus, mathematica,…)

Applications

- Image processing and compression
  - low rank approximation leads to compressed representation, noise reduction
- Molecular dynamics
  - characterizing protein molecular dynamics
  - higher prinicipal components correspond to large scale motions

Applications

- Information retrieval
  - LSI: Latent semantic indexing
    SVD applied to term document matrix
  - compute best rank k approximation
  - eigenvectors correspond to linguistic concepts

- Gene expression data analysis
  - SVD useful preprocessing step
  - grouping genes by transcriptional response,
    grouping assays by expression profile

Microarray gene expression data

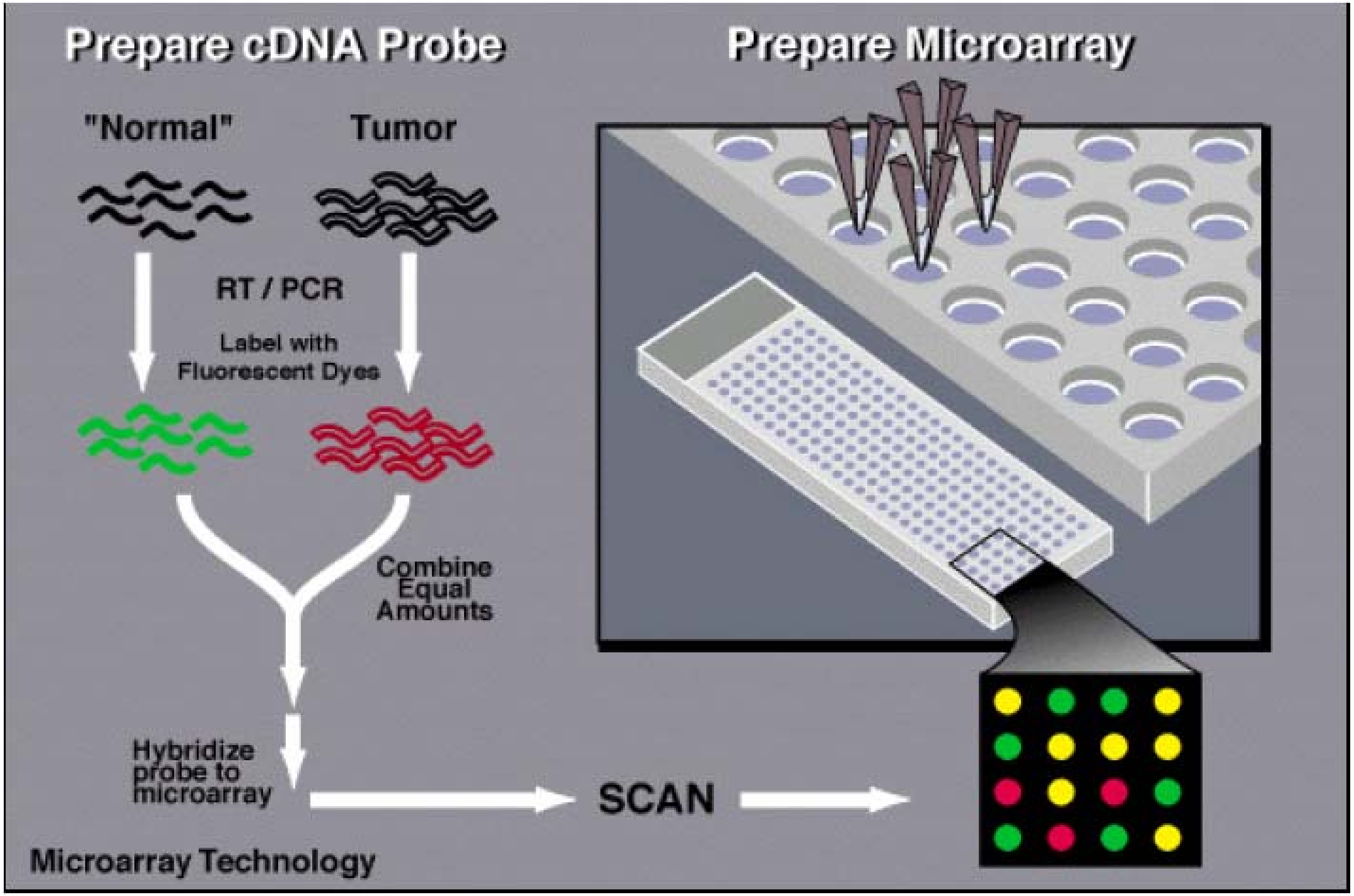
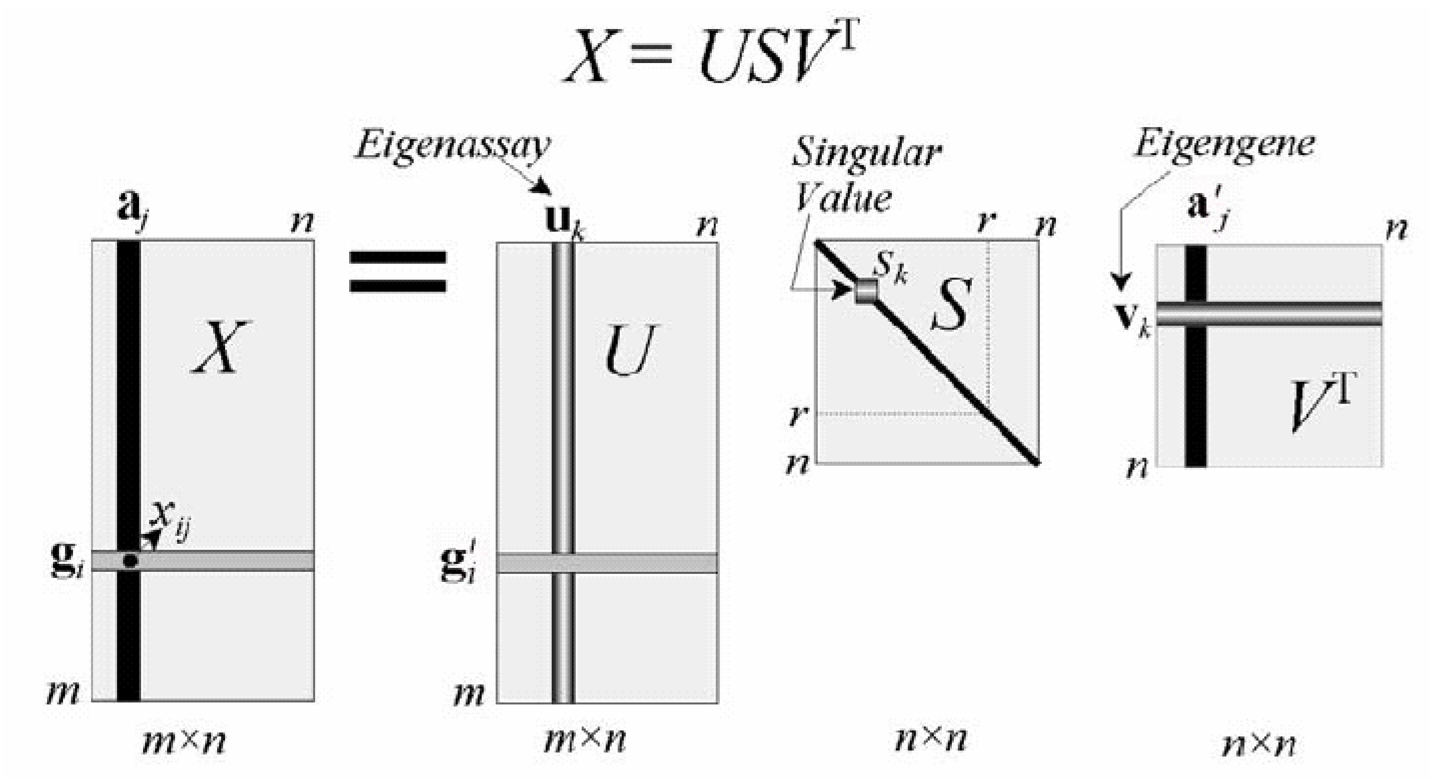


Figure 2.1: *Microarray process.*
The illustration is "Courtesy of the National Human Genome Research Institute/National Institutes of Health.

SVD applied to gene expression data



Information retrieval

- X is term document matrix
  - m terms, n documents
  - entry (t,d) for term t and document d is function of how many times t occurs d
- SVD of X gives low dimensional representation X_r
  - Latent Semantic Indexing
- $X_r^T X_r$ is matrix of document similarities
- Columns of X_r represent the documents, but in $r << m$ dimensions

Semi-precise intuition

- We accomplish more than dimension reduction here:
  - Docs with lots of overlapping terms stay together
  - Terms from these docs also get pulled together.
- Thus ***car*** and ***automobile*** get pulled together because both co-occur in docs with ***tires, radiator, cylinder***, etc.

Query processing

- View a query as a (short) doc:
  - call it column 0 of X_r.
- Now the entries in column 0 of $X_r^T X_r$ give the similarities of the query with each doc.
- Entry $(j,0)$ is the score of doc j on the query.

Talk Outline

- Dimension reduction

- Similarity preserving hash functions
  - sketching vector norms
  - sketching Earth Mover Distance (EMD)

Low Distortion Embeddings

- Given metric spaces (X_1,d_1) & (X_2,d_2), embedding f: $X_1 \rightarrow X_2$ has distortion D if ratio of distances changes by at most D

http://www.physast.uga.edu/~jss/1010/ch10/earth.jpg

http://humanities.ucsd.edu/courses/kuchtahum4/pix/earth.jpg

- "Dimension Reduction" –
  - Original space high dimensional
  - Make target space be of "low" dimension, while maintaining small distortion

Dimension Reduction in L_2

- n points in Euclidean space (L_2 norm) can be mapped down to $O((\log n)/\varepsilon^2)$ dimensions with distortion at most $1+\varepsilon$.
[Johnson Lindenstrauss '84]

- Two interesting properties:
  - Linear mapping
  - Oblivious – choice of linear mapping does not depend on point set
  - Quite simple [JL84, FM88, IM98, DG99, Ach01]: Even a random +1/-1 matrix works…
- Many applications…

Dimension reduction for L_1

- [C,Sahai '02]
  Linear embeddings are not good for dimension reduction in L_1

- There exist $O(n)$ points in L_1 in n dimensions, such that any *linear mapping* with distortion δ needs n/δ^2 dimensions

Dimension reduction for L_1

- [C, Brinkman '03]
  Strong lower bounds for dimension reduction in L_1
- There exist n points in L_1 , such that *any embedding* with constant distortion δ needs n^{1/δ^2} dimensions
- Simpler proof by [Lee,Naor '04]
- Does not rule out other sketching techniques

Talk Outline

- Dimension reduction

- Similarity preserving hash functions
  - sketching vector norms
  - sketching Earth Mover Distance (EMD)

Similarity Preserving Hash Functions

complex object

sketch

→ 0 1 0 1 1 0 0 1 1 0

→ 0 0 1 0 1 1 0 0 1 0

- Similarity function *sim(x,y),* distance *d(x,y)*
- Family of hash functions *F* with probability distribution such that

$$\Pr_{h\in F}[h(x)=h(y)]=sim(x,y)$$

$$\Pr_{h\in F}[h(x)\neq h(y)]=d(x,y)$$

Applications

- Compact representation scheme for estimating similarity

$$x \longrightarrow (h_1(x), h_2(x), \ldots, h_k(x))$$

$$y \longrightarrow (h_1(y), h_2(y), \ldots, h_k(y))$$

- Approximate nearest neighbor search [Indyk,Motwani '98] [Kushilevitz,Ostrovsky,Rabani '98]

Relaxations of SPH

- Estimate distance measure, not similarity measure in [0,1].
- Measure E[f(h(x),h(y))].

$$\Pr_{h \in F}[h(x) \neq h(y)] = d(x, y)$$

$$\mathrm{E}[f(h(x), h(y))] = d(x, y)$$

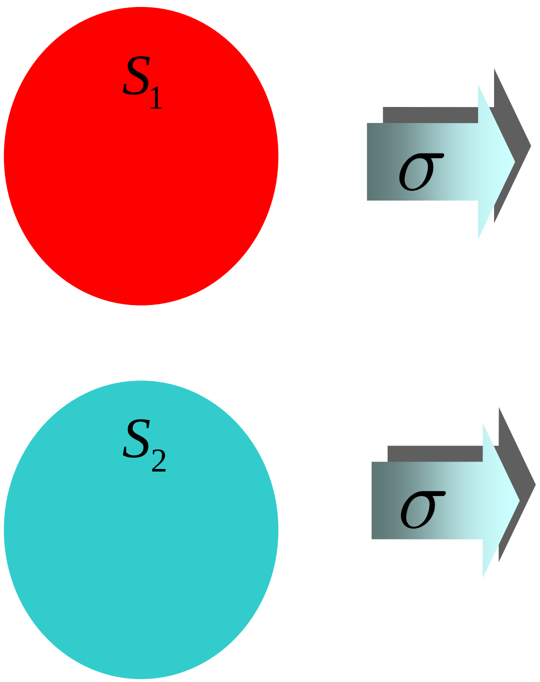
- Estimator will approximate distance function.

Sketching Set Similarity: Minwise Independent Permutations

[Broder,Manasse,Glassman,Zweig ‘97]

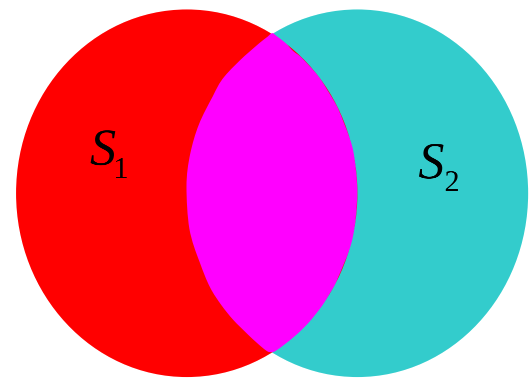
[Broder,C,Frieze,Mitzenmacher ‘98]

$$\text{similarity} = \frac{|S_1 \cap S_2|}{|S_1 \cup S_2|}$$



$\min(\sigma(S_1))$

$\min(\sigma(S_2))$



$$\text{prob}(\min(\sigma(S_1)) = \min(\sigma(S_2))) = \frac{|S_1 \cap S_2|}{|S_1 \cup S_2|}$$

Other similarity functions ?

[C'02]

- Necessary conditions for existence of similarity preserving hash functions.
  - SPH does not exist for Dice coefficient and Overlap coefficient.
- SPH schemes from rounding algorithms
  - Hash function for vectors based on random hyperplane rounding.

Existence of SPH schemes

- *sim(x,y)* admits an SPH scheme if $\exists$ family of hash functions *F* such that

$$\Pr_{h \in F}[h(x) = h(y)] = sim(x, y)$$

Theorem: If *sim(x,y)* admits an SPH scheme then *1-sim(x,y)* satisfies triangle inequality.

Proof:

$$1 - sim(x,y) = \Pr_{h \in F}(h(x) \neq h(y))$$

$\Delta_h(x,y):$ indicator variable for $h(x) \neq h(y)$

$$\Delta_h(x,y) + \Delta_h(y,z) \geq \Delta_h(x,z)$$

$$1 - sim(x,y) = \mathrm{E}_{h \in F}[\Delta_h(x,y)]$$

Non-existence of SPH

$$sim_{Dice}(A,B) = \frac{|A \cap B|}{\frac{1}{2}(|A| + |B|)}$$: Dice's coefficient

$$sim_{Ovl}(A,B) = \frac{|A \cap B|}{\min(|A|,|B|)}$$: Overlap coefficient

Triangle inequality violated for:

$$A = \{a\},\ B = \{b\},\ C = \{a,b\}$$

$$1 - sim(A,C) + 1 - sim(C,B) \ < \ 1 - sim(A,B)$$

Stronger Condition

Theorem: If *sim(x,y)* admits an SPH scheme then *(1+sim(x,y))/2* has an SPH scheme with hash functions mapping objects to {0,1}.

Theorem: If *sim(x,y)* admits an SPH scheme then *1-sim(x,y)* is isometrically embeddable in the Hamming cube.

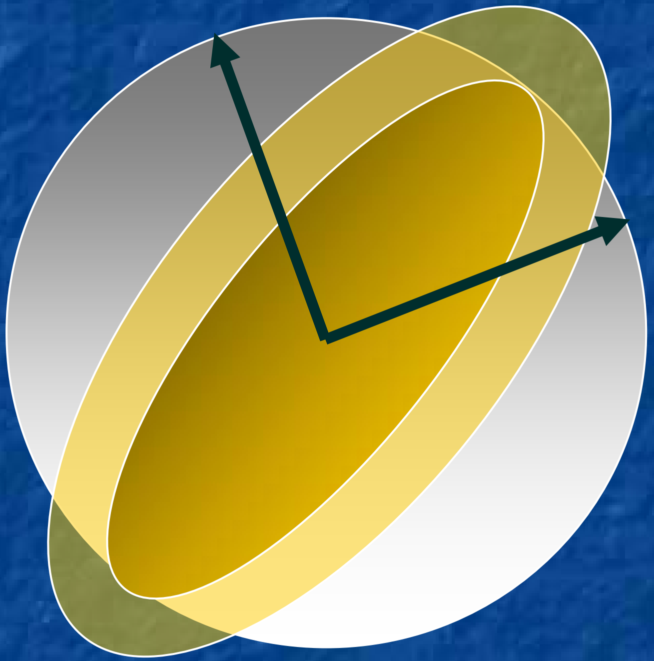
Random Hyperplane Rounding based SPH

- Collection of vectors

$$sim(\vec{u},\vec{v}) = 1 - \frac{\measuredangle(\vec{u},\vec{v})}{\pi}$$

- Pick random hyperplane through origin (normal $\vec{r}$)

$$h_{\vec{r}}(\vec{u}) = \begin{cases} 1 & \text{if } \vec{r}\cdot\vec{u} \geq 0 \\ 0 & \text{if } \vec{r}\cdot\vec{u} < 0 \end{cases}$$



- [Goemans,Williamson]

- For *n* vectors, random hyperplane can be chosen using *$O(\log^2 n)$* random bits. [Indyk], [Engebretson,Indyk,O'Donnell]
- Alternate similarity measure for sets

$$sim(A,B) = 1 - \frac{\theta}{\pi}$$

$$\theta = \cos^{-1} \frac{|A \cap B|}{|A \cup B|}$$

Sketching L_1

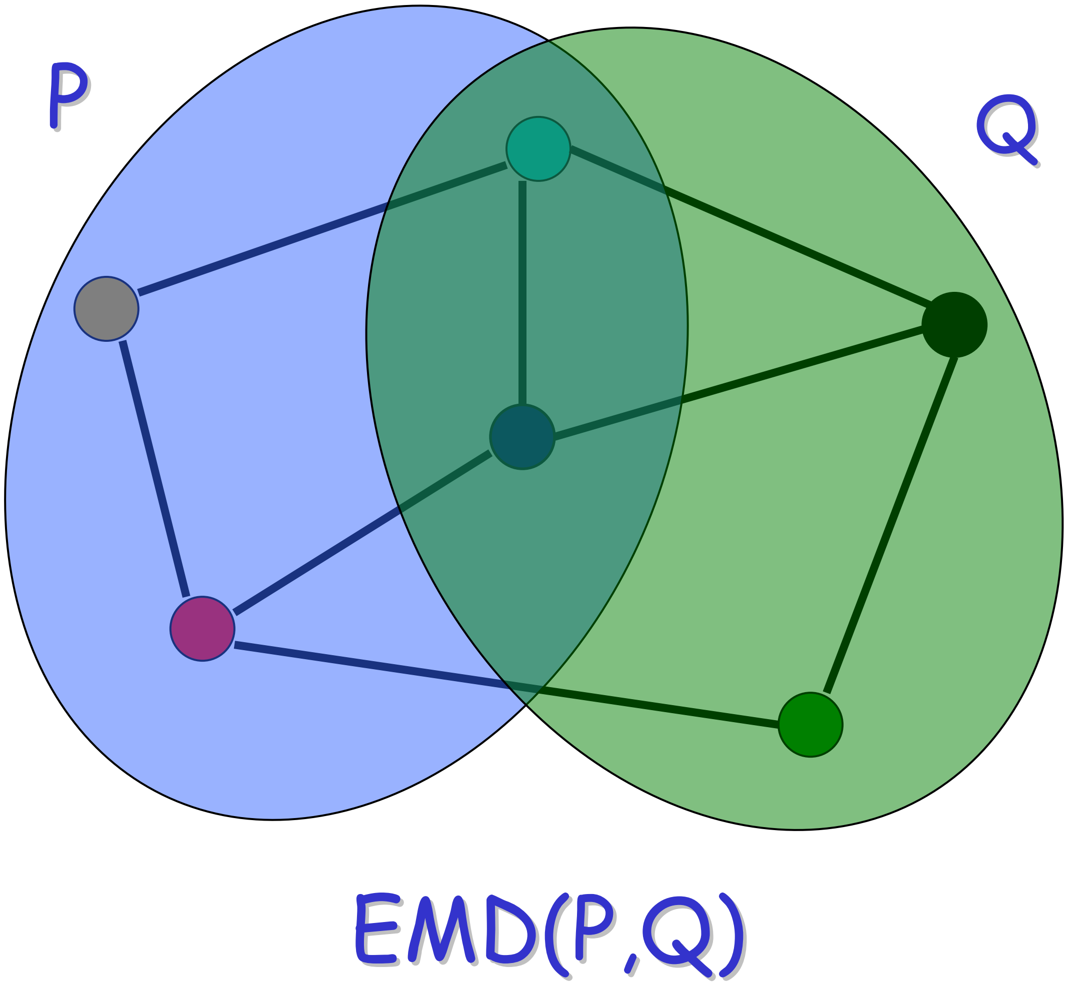
- Design sketch for vectors to estimate L_1 norm
- Hash function to distinguish between small and large distances [KOR '98]
  - Map L_1 to Hamming space
  - Bit vectors $a=(a_1,a_2,\ldots,a_n)$ and $b=(b_1,b_2,\ldots,b_n)$
  - Distinguish between distances $\leq (1-\varepsilon)n/k$ versus $\geq (1+\varepsilon)n/k$
  - XOR random set of k bits
  - $\Pr[h(a)=h(b)]$ differs by constant in two cases

Sketching L_1 via stable distributions

- $a=(a_1,a_2,\ldots,a_n)$ and $b=(b_1,b_2,\ldots,b_n)$
- Sketching L_2
  - $f(a) = \Sigma_i\, a_i\, X_i \quad f(b) = \Sigma_i\, b_i\, X_i$
    X_i independent Gaussian
  - $f(a)-f(b)$ has Gaussian distribution scaled by $|a-b|_2$
  - Form many coordinates, estimate $|a-b|_2$ by taking L_2 norm

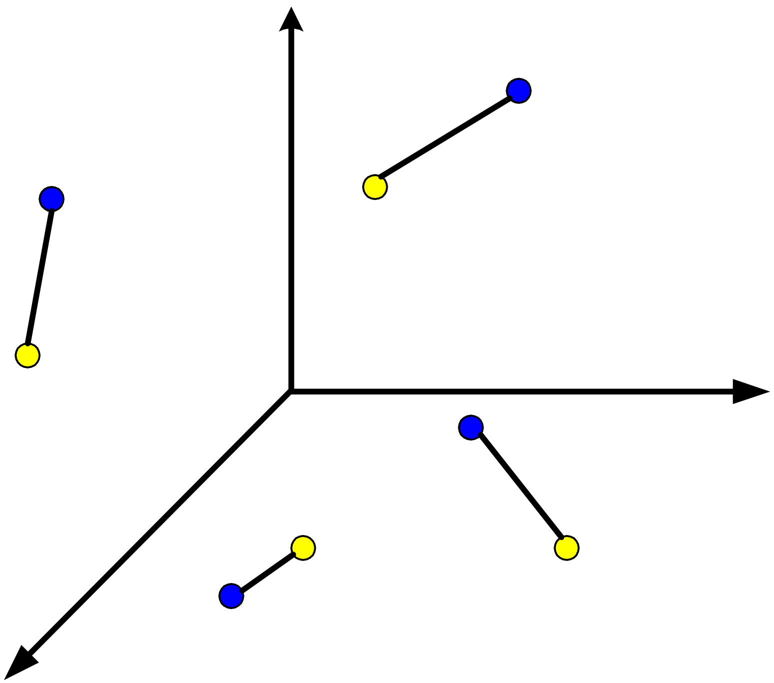
- Sketching L_1
  - $f(a) = \Sigma_i\, a_i\, X_i \quad f(b) = \Sigma_i\, b_i\, X_i$
    X_i independent Cauchy distributed
  - $f(a)-f(b)$ has Cauchy distribution scaled by $|a-b|_1$
  - Form many coordinates, estimate $|a-b|_1$ by taking median
    [Indyk '00] -- streaming applications

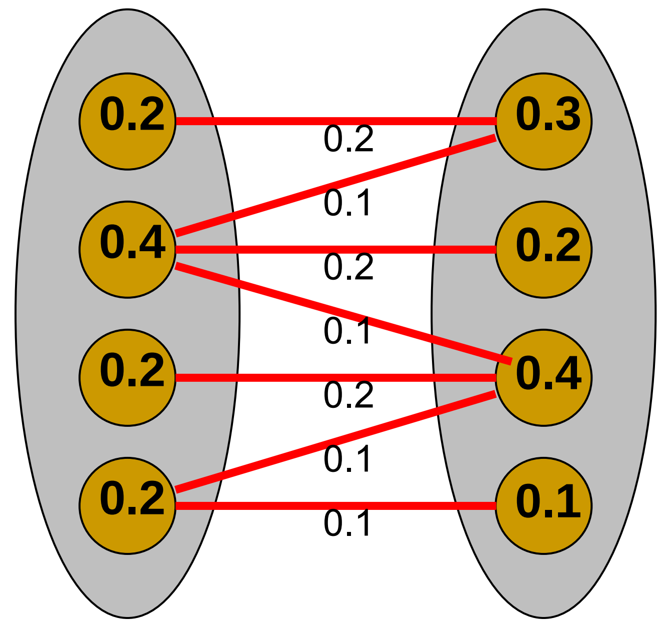
Earth Mover Distance (EMD)



Bipartite/Bichromatic matching

- Minimum cost matching between two sets of points.
- Point weights ≡ multiple copies of points

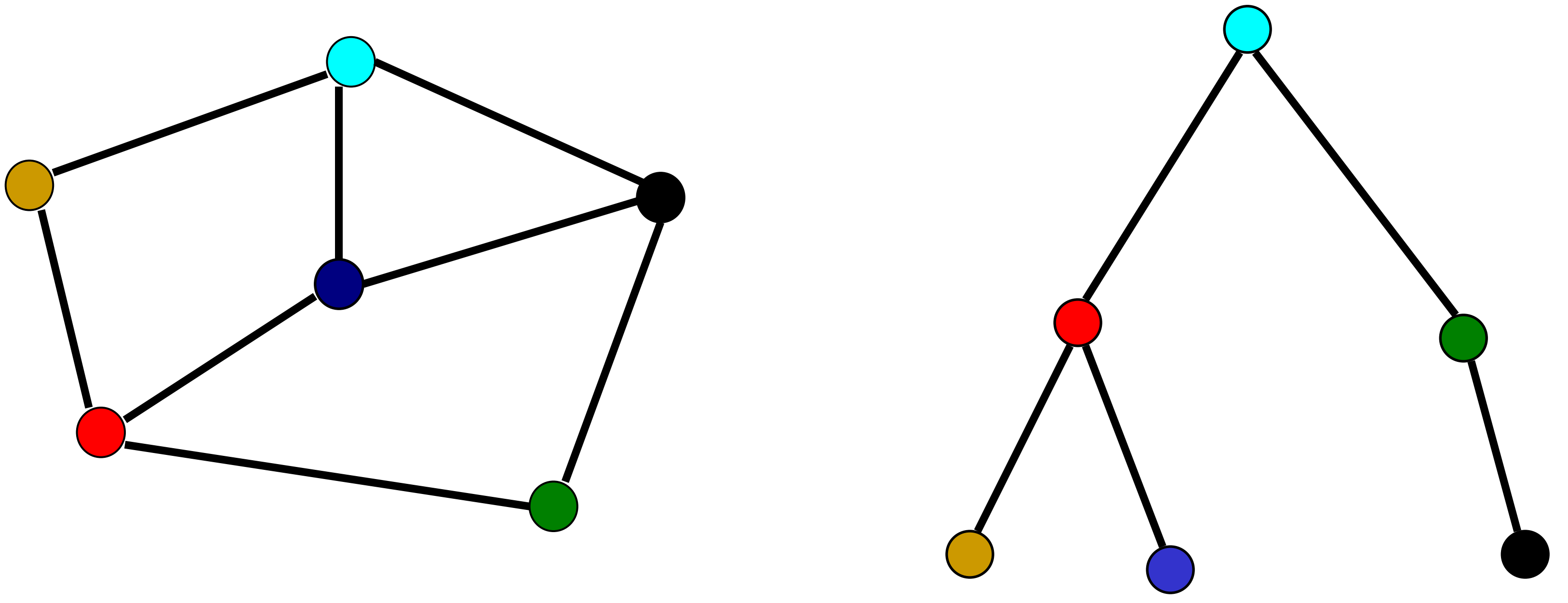




Fast estimation of bipartite matching [Agarwal,Varadarajan ’04]

Goal: Sketch point set to enable estimation of min cost matching

Approximating metrics by trees



Single tree may have high distortion

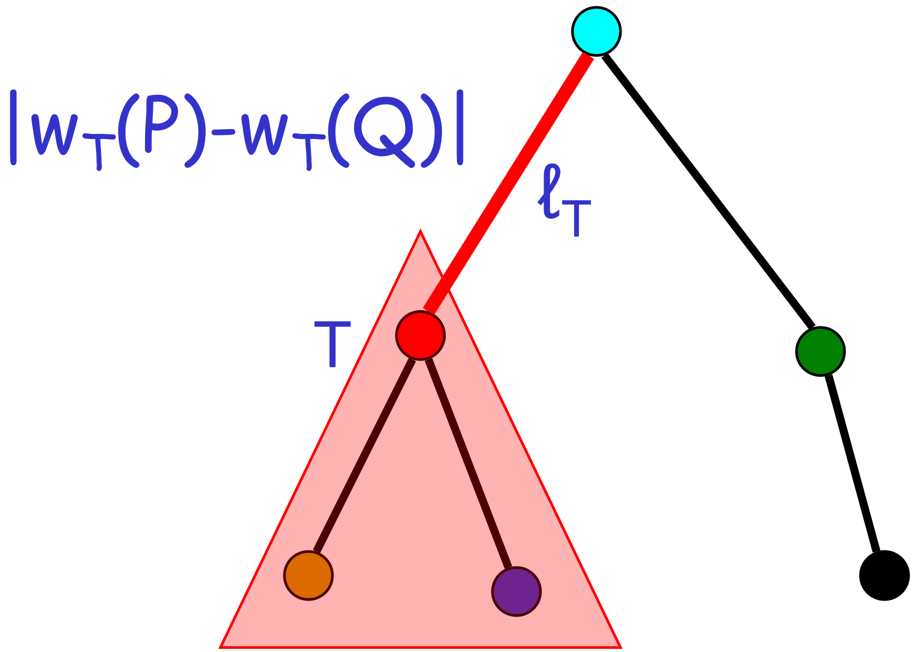
Use probability distribution over trees

$$d(u,v) \leq E[d_T(u,v)] \leq O(\log n)\, d(u,v)$$

[Bartal ’96,’98, FRT ’03]

EMD on trees: embedding into L_1

[suggested by Piotr Indyk]



$$EMD(P,Q) = \sum_T \ell_T |w_T(P) - w_T(Q)|$$

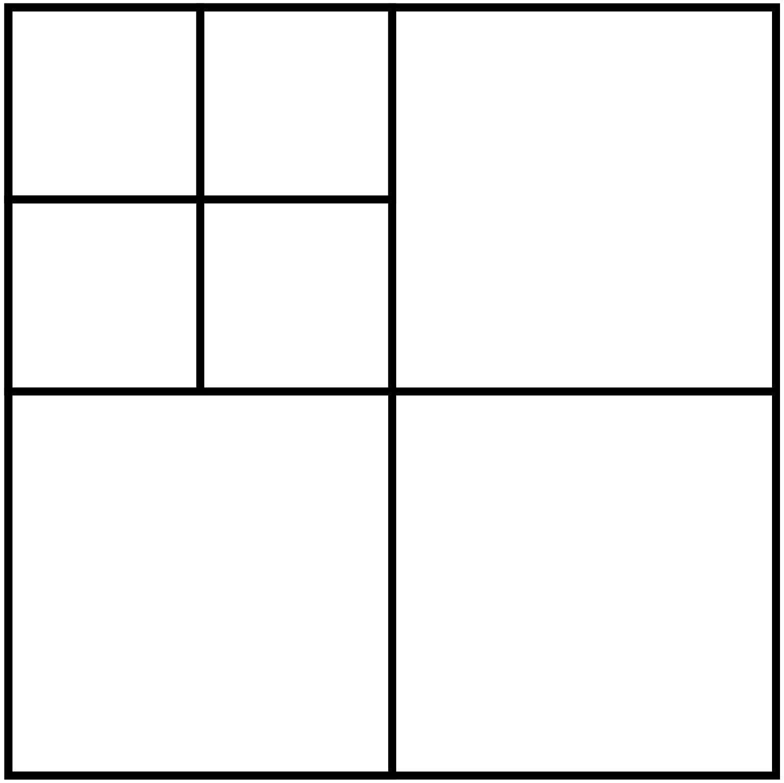
$$v(P) = \{\ell_T w_T(P)\}_T \qquad v(Q) = \{\ell_T w_T(Q)\}_T$$

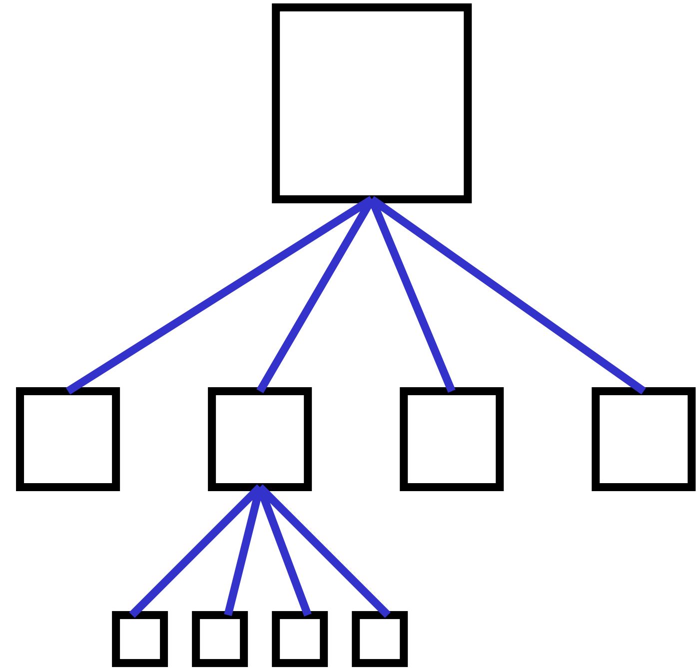
$$EMD(P,Q) = |v(P) - v(Q)|_1$$

EMD on general metrics

- Approximate metric by probability distribution on trees
- Sample tree from distribution and compute L_1 representation
- $EMD(P,Q) \leq E[d(v(P),v(Q))] \leq O(\log n)\ EMD(P,Q)$

Tree approximations for Euclidean points





distortion $O(d \log \Delta)$ [Bartal '96, CCGGP '98]

proposed by [Indyk,Thaper '03] for estimating EMD

Conclusions

- Compact representations at the heart of several algorithmic techniques for large data sets
  - Compact representations tailored to applications
  - Effective for region based image retrieval

ISOMAP and LLE

- Nonlinear dimension reduction methods
- “Learn” hidden structure in data

- See slides of Chan-Su Lee and Rong Xu from Michael Littman’s course at Rutgers
- http://www.cs.rutgers.edu/~mlittman/courses/lightai03/chansu.ppt
- http://www.cs.rutgers.edu/~mlittman/courses/lightai03/rongxu.ppt