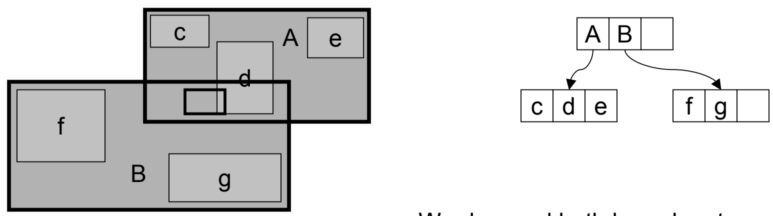
High-Dimensional Data

Topics

- Motivation
- Similarity Measures
- Index Structures

R trees, redux

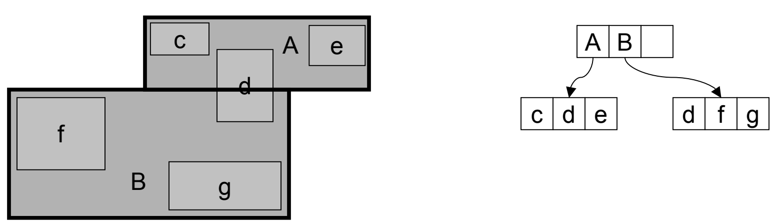
- We want to minimize *coverage* and *overlap*



We descend both branches to search for

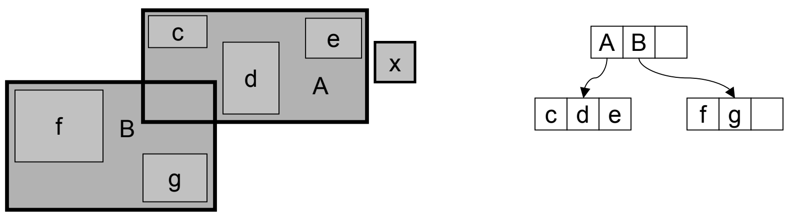
R+ Trees

- store d in both A and B
- like splitting d into two pieces



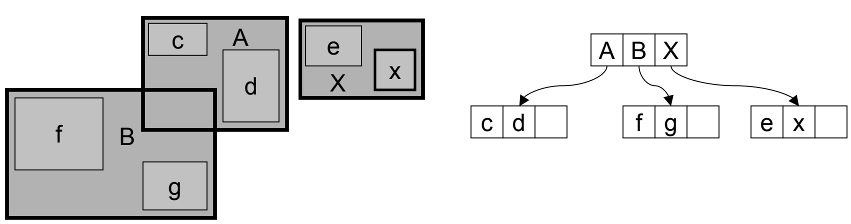
R* trees

- When a node overflows,
  - don't split it right away;
  - reinsert some of its nodes



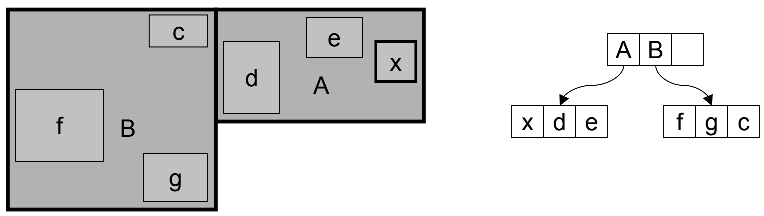
R* trees

- Normal Insertion:

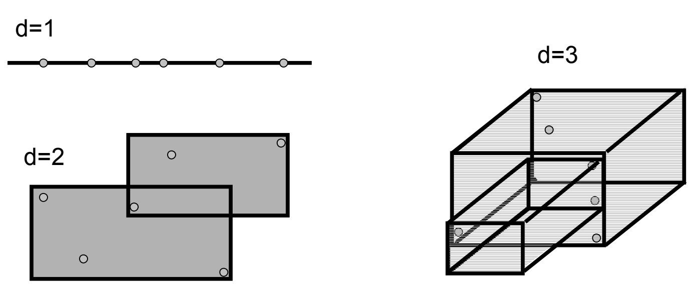


R* trees

- Reinsert c instead of splitting node



Curse of Dimensionality



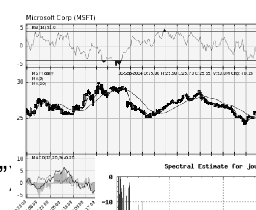
Coverage and *overlap* as a function of dimension?

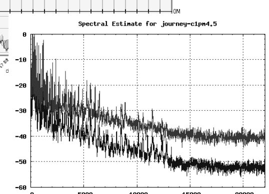
Curse of Dimensionality

- *Generally: exponential growth of the hypervolume as a function of dimension*

- Other manifestations:
  - number of samples required to maintain the same accuracy
  - number of nodes in a neural network required to “monitor” the input space
  - lots more

High-dimensional data





- Finance
- Multimedia
  - Sound
  - Music (“Query by humming”)
  - Images
  - Video
- Document Retrieval
- Biology/Medicine
  - DNA sequence matching
  - Medical imagery
- Moving Objects [(t0,x0,y0), (t1,x1,y1), …]
- High-Energy Physics

High-dimensional Access Methods

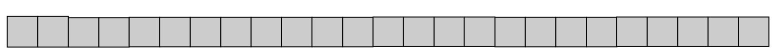
- Three components:
  - Similarity Measure
  - Index Structures
  - Search Strategy

we won't cover search strategy

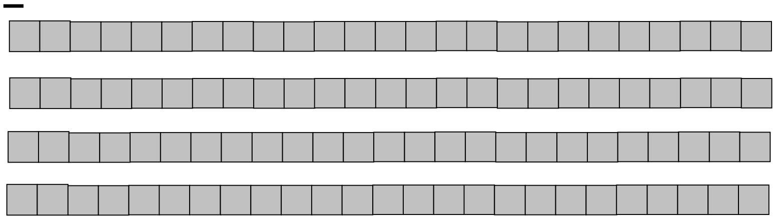
Similarity Measure

- When are two vectors similar?

Q =



DB =



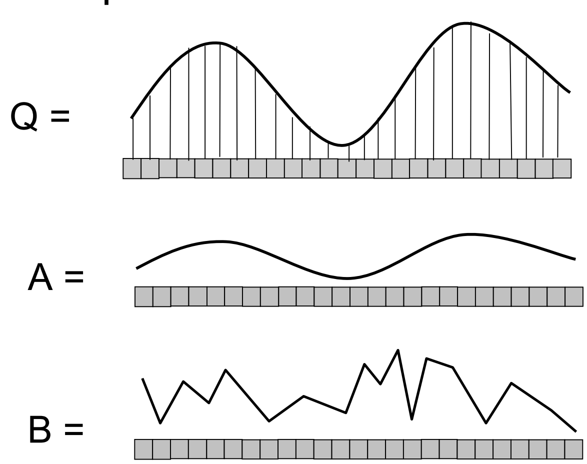
Similarity Measure

Define a function *s* : V → V → Real

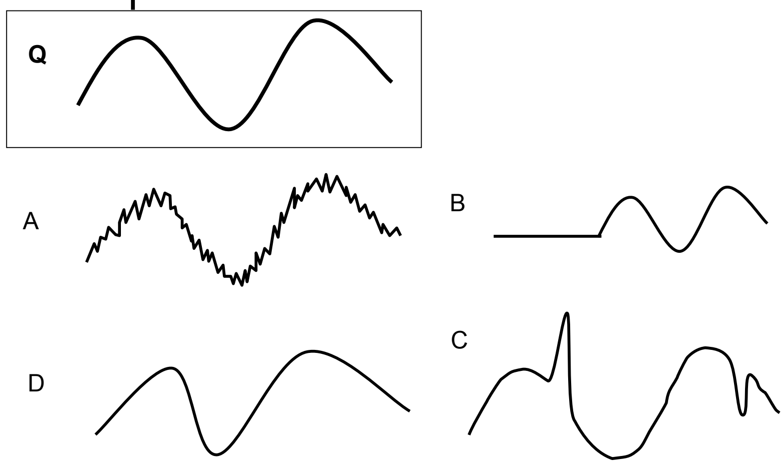
What properties should *s* have?

Reflexive:
  s(x,x) = 0 *// or infinity*
Symmetric:
  s(x,y) = s(y,x)
Triangle Inequality:
  s(x,y) + s(y,z) >= s(x,z)

Timeseries Indexing



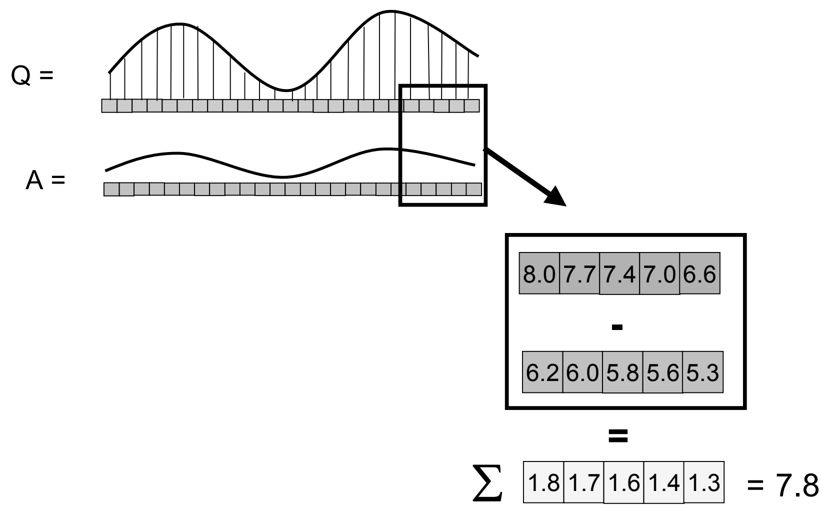
Timeseries Indexing



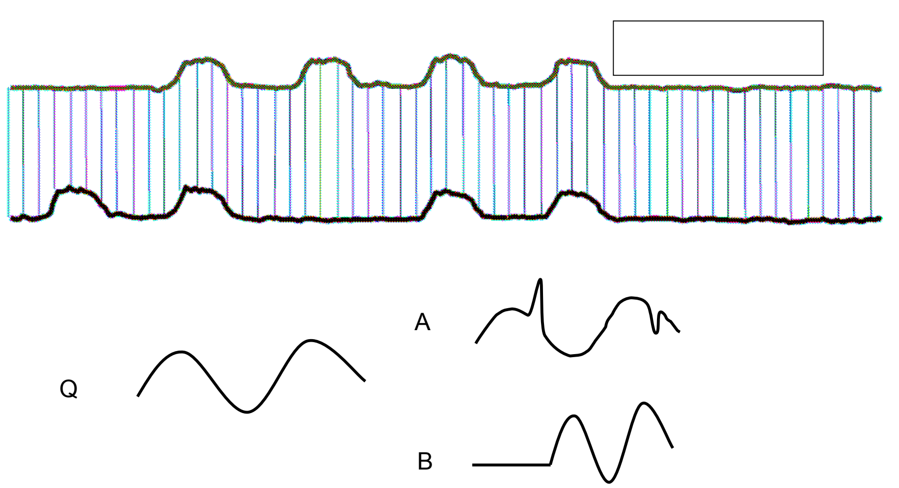
Timeseries Indexing

- Euclidean distance
- Dynamic Time Warping
  - Jagadish, Faloutsos 1998, Keogh 2002
- Wavelets
  - Miller 2003
- LCSS
  - Vlachos, Kollios, Gunopolos 2002
- EDR
  - Chen, Ozsu, Oria 2005

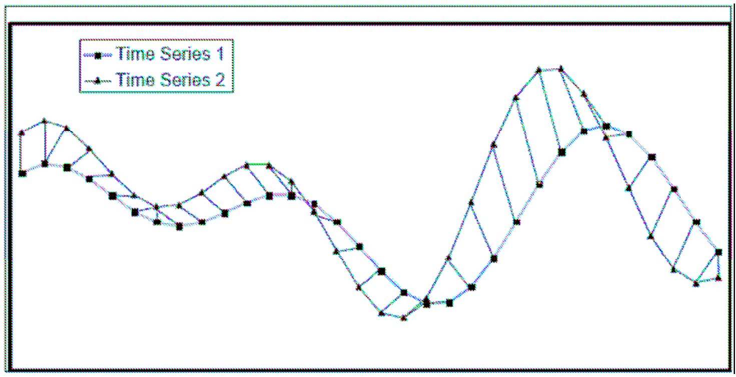
Euclidean Distance



Eclidean Distance (2)



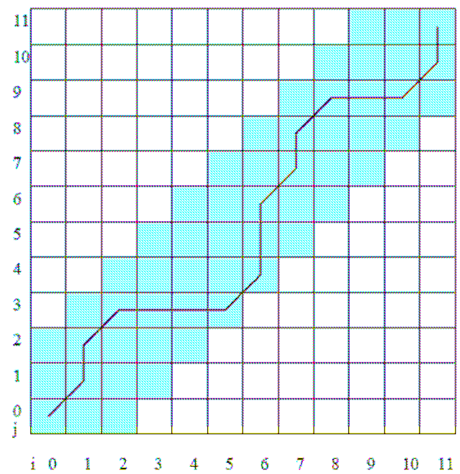
Dynamic Time Warping

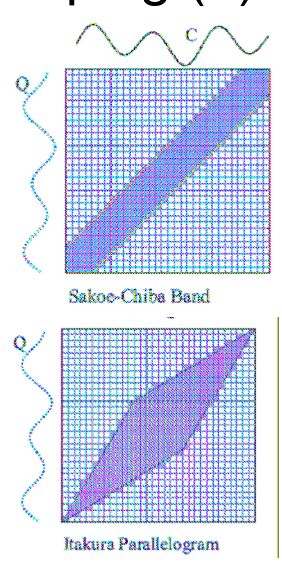


Dynamic Time Warping (2)

$$D^2_{DTW}(x,y) = D^2(First(x), First(y)) + min \begin{cases} D^2_{DTW}(x, Rest(y)) \\ D^2_{DTW}(y, Rest(x)) \\ D^2_{DTW}(Rest(x), Rest(y)) \end{cases}$$

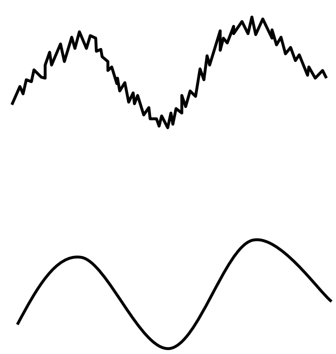
Dynamic Time Warping (3)



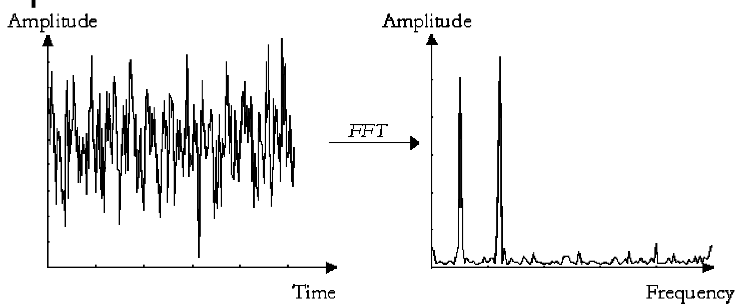


Dynamic Time Warping (4)

- Drawbacks:
  - Sensitive to noise
  - expensive to compute



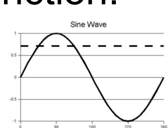
Wavelets



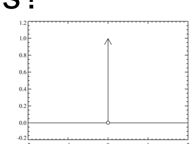
- Fourier Transform
  - Represents a timeseries as a sum of sine waves
  - The coefficients of the constituent waves indicate the dominant structure

Wavelets (2)

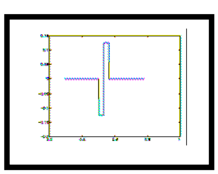
- Same trick, different basis function:
  - Sum of sine waves?

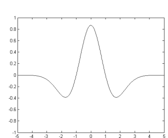


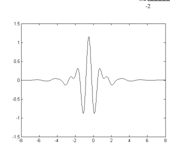
  - Sum of Dirac delta functions?



  - Sum of …







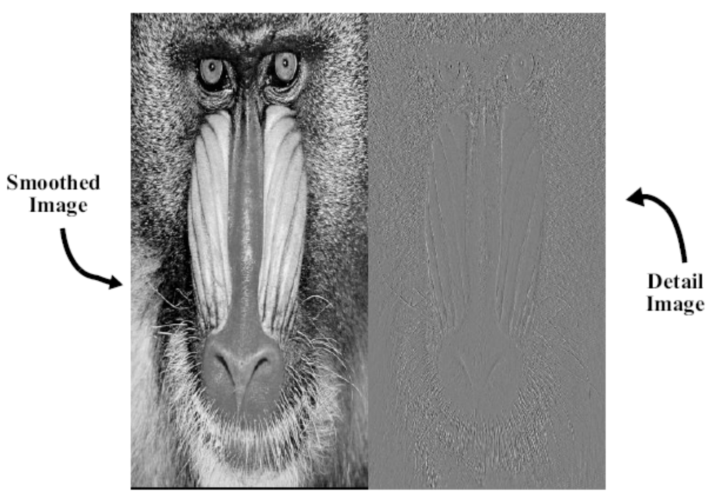
Wavelets (3)

Haar wavelet transform

	Original Signal	2	7	11	8	11	5	0	3
$s_i + s_{i+1}$									
$s_i - s_{i+1}$	Detail Signal (1)	-5		3		6		-3	
	Detail Signal (2)	-10				13			
	Smoothed Signal (3)	47							
	Detail Signal (3)	9							

Hierarchical decomposition allows fine-tuning

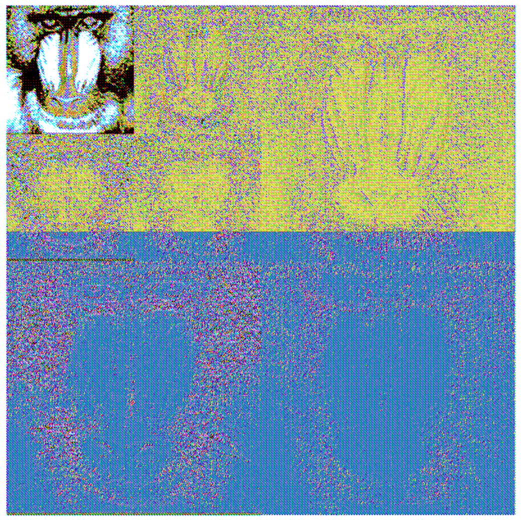
Wavelets (4)



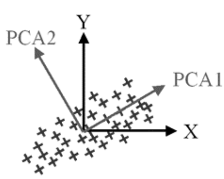
After one Horizontal filtering

Wavelets (5)

- After two vertical and horizontal filterings



Wavelets (6)



- Wavelets can *reduce dimensionality,* like
  - Principal Component Analysis (PCA),
  - Singular Value Decomposition (SVD),
  - others
- Indexing in the reduced feature space
  - False positives ok, False negatives aren't
  - Use a more refined similarity measure to eliminate false positives

Other measures

- Longest Common Subsequence

$$LCSS(R,S) = \begin{cases} 0 & \text{if } m = 0 \text{ or } n = 0 \\ LCSS(Rest(R), Rest(S)) + 1 & \text{if } |r_{1,x} - s_{1,x}| \leq \epsilon \\ & \& |r_{1,y} - s_{1,y}| \leq \epsilon \\ max\{LCSS(Rest(R), S), LCSS(R, Rest(S))\} & \text{otherwise} \end{cases}$$

- Edit Distance on Real sequence

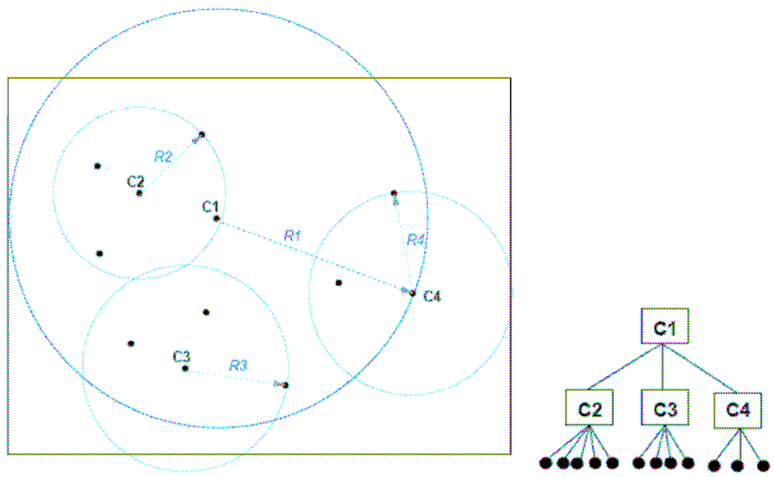
$$EDR(R,S) = \begin{cases} n & \text{if } m = 0 \\ m & \text{if } n = 0 \\ min\{EDR(Rest(R), Rest(S)) + subcost, \\ \quad EDR(Rest(R), S) + 1, EDR(R, Rest(S)) + 1\} & \text{otherwise} \end{cases}$$

Index Structures

- SS-Tree [White, Jain 96]
  - *R*-Tree using Minimum Bounding Spheres*
- SR-Tree [Katayama, Satoh 97]
  - *Uses MBR during construction,*
  - *but MBS during lookup*
- X-Tree [Berchtold, Kreim, Kriegel 96]
  - *R*-Tree using extended nodes to avoid splits and control maximum overlap*
- M-Tree [Ciaccia, Patella 00]
  - *Build tree based on representative points*
- TV-tree [Lin, Jagadish, Faloutsos 94]

SR-Tree and M-Tree appear to outperform others

M-Tree



M-Tree partitioning of Ω_d

Associated M-Tree

Telscoping Vector Tree (TV)

- node = (center, radius)
- dim(center) >= # of “active dimensions”