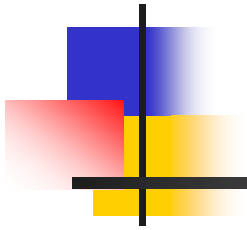
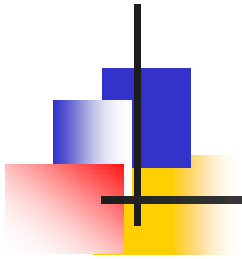


ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΑΤΡΩΝ - ΤΜΗΥΠ
ΒΑΣΕΙΣ ΔΕΔΟΜΕΝΩΝ ΙΙ

Β. Μεγαλοοικονόμου

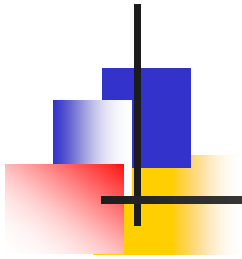
Δεικτοδότηση Πολυμέσων

(παρουσίαση βασισμένη εν μέρη σε σημειώσεις των Silberchatz, Korth και Sudarshan και του C. Faloutsos)



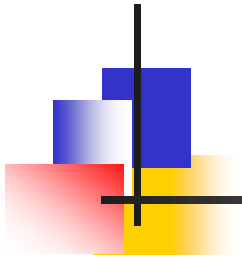
Γενική Θεώρηση

- Σχεσιακό μοντέλο – SQL, σχεδιασμός ΒΔ
- Δεικτοδότηση, Q-opt, Επεξεργασία δοσοληψιών
- Προχωρημένα θέματα
  - Κατανεμημένες Βάσεις
  - RAID
  - Authorization / Stat. DB
  - Spatial Access Methods
  - Δεικτοδότηση Πολυμέσων



Πολυμέσα- λεπτομερώς

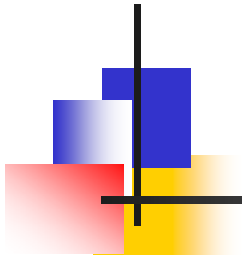
- Πολυμέσα
  - Motivation / ορισμός προβλήματος
  - Κύρια ιδέα / time sequences
  - εικόνες
  - sub-pattern matching
  - Αυτόματη εξαγωγή χαρακτηριστικών / FastMap



Πρόβλημα

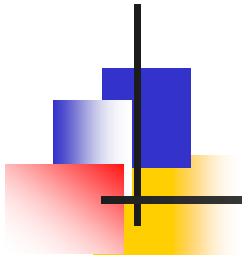
Δοθείσας μίας μεγάλης συλλογής (πολυμεσικών) εγγραφών (πχ. μετοχές)

Επιτρέπει γρήγορα, ερωτήματα ομοιότητας



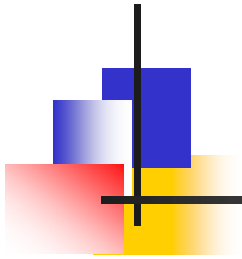
Εφαρμογές

- time series: χρηματοοικονομικά, marketing (click-streams!), ECGs, ήχος;
- εικόνες: ιατρική, ψηφιακές βιβλιοθήκες, εκπαίδευση, τέχνη
- higher-d σήματα: επιστημονικές ΒΔ (πχ. αστροφυσική, ιατρική (MRI ακτινογραφίες), ψυχαγωγία (video)



Παραδείγματα Ερωτημάτων

- Βρες ιατρικές υποθέσεις παρόμοιες και του κ. Παπαδόπουλου
- Βρες ζεύγη μετοχών που κινούνται με συγχρονισμό
- Βρες ζεύγη εγγράφων που είναι παρόμοια (λογοκλοπία;)
- Βρες πρόσωπα παρόμοια με του 'Tiger Woods'

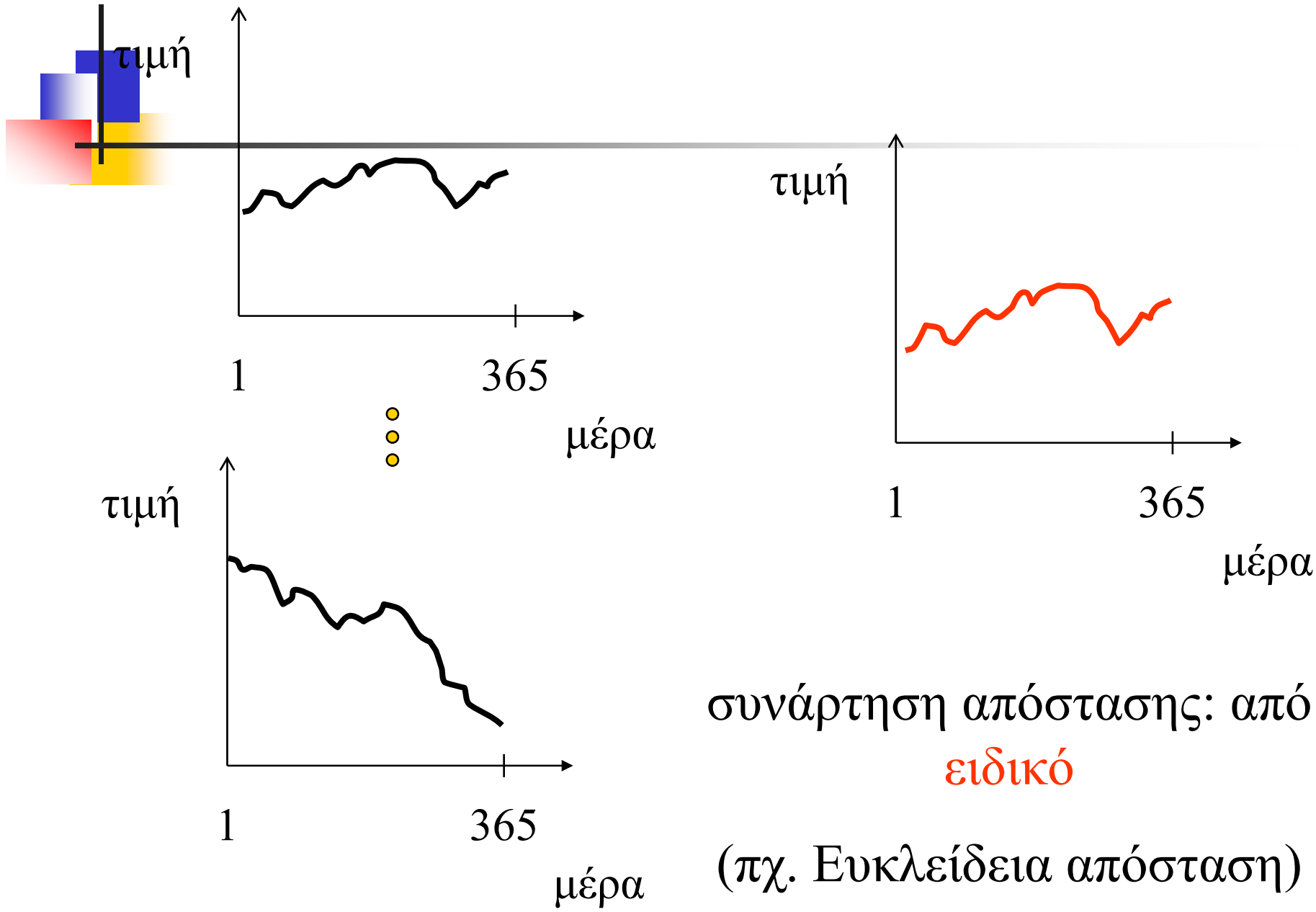


Λεπτομ. ορισμός προβλήματος:

Πρόβλημα:

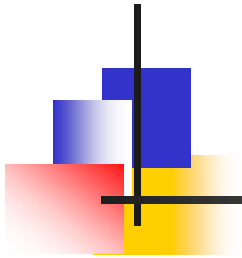
- δοθείσας μιας συλλογής πολυμεσικών αντικειμένων,
- βρες αυτά που είναι παρόμοια με ένα επιθυμητό αντικείμενο-ερώτημα

- για παράδειγμα:



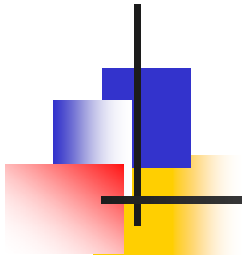
συνάρτηση απόστασης: από ειδικό

(πχ. Ευκλείδεια απόσταση)



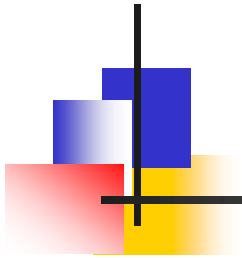
Τύποι ερωτημάτων

- Ολική ταύτιση εν. sub-pattern match
- ερωτήματα εύρους εν. πλησιέστερων γειτόνων
- όλα τα ζεύγη ερώτημα (all pairs queries or spatial joins)



Στόχοι σχεδίου

- Γρήγορα (γρηγορότερα από σειρ. αναζήτηση)
- ‘ορθό’ (πχ., όχι ψεύτικοι συναγερμοί, όχι λάθος απορρίψεις)

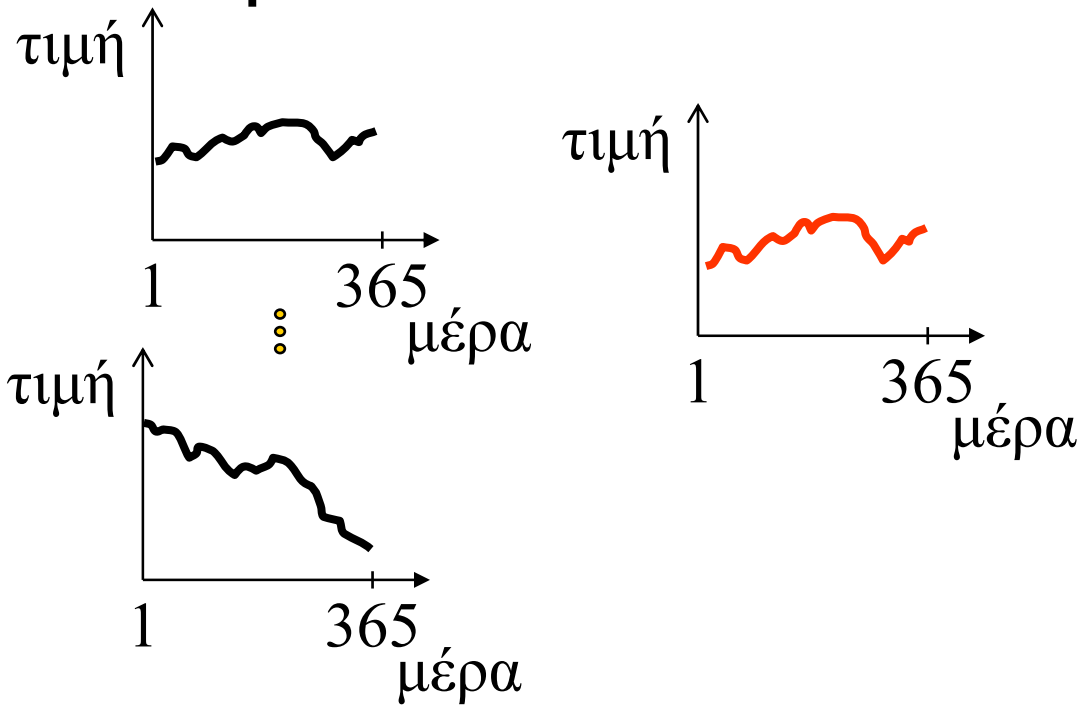


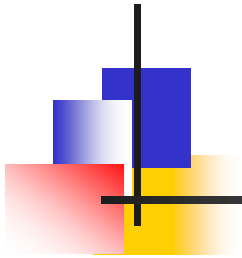
Πολυμέσα- λεπτομερώς

- Πολυμέσα
  - Motivation / ορισμός προβλήματος
  - Κύρια ιδέα / time sequences
  - εικόνες
  - sub-pattern matching
  - Αυτόματη εξαγωγή χαρακτηριστικών / FastMap

Κεντρική ιδέα

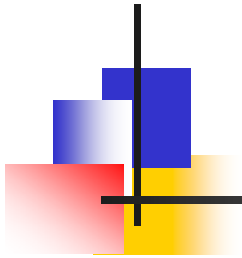
- Πχ., χρονικές ακολουθίες, 'ολική ταύτιση', ερωτήματα εύρους, Ευκλείδεια απόσταση





Κεντρική ιδέα

- Η ακολουθιακή αναζήτηση δουλεύει – πώς μπορεί να γίνει πιο γρήγορα;

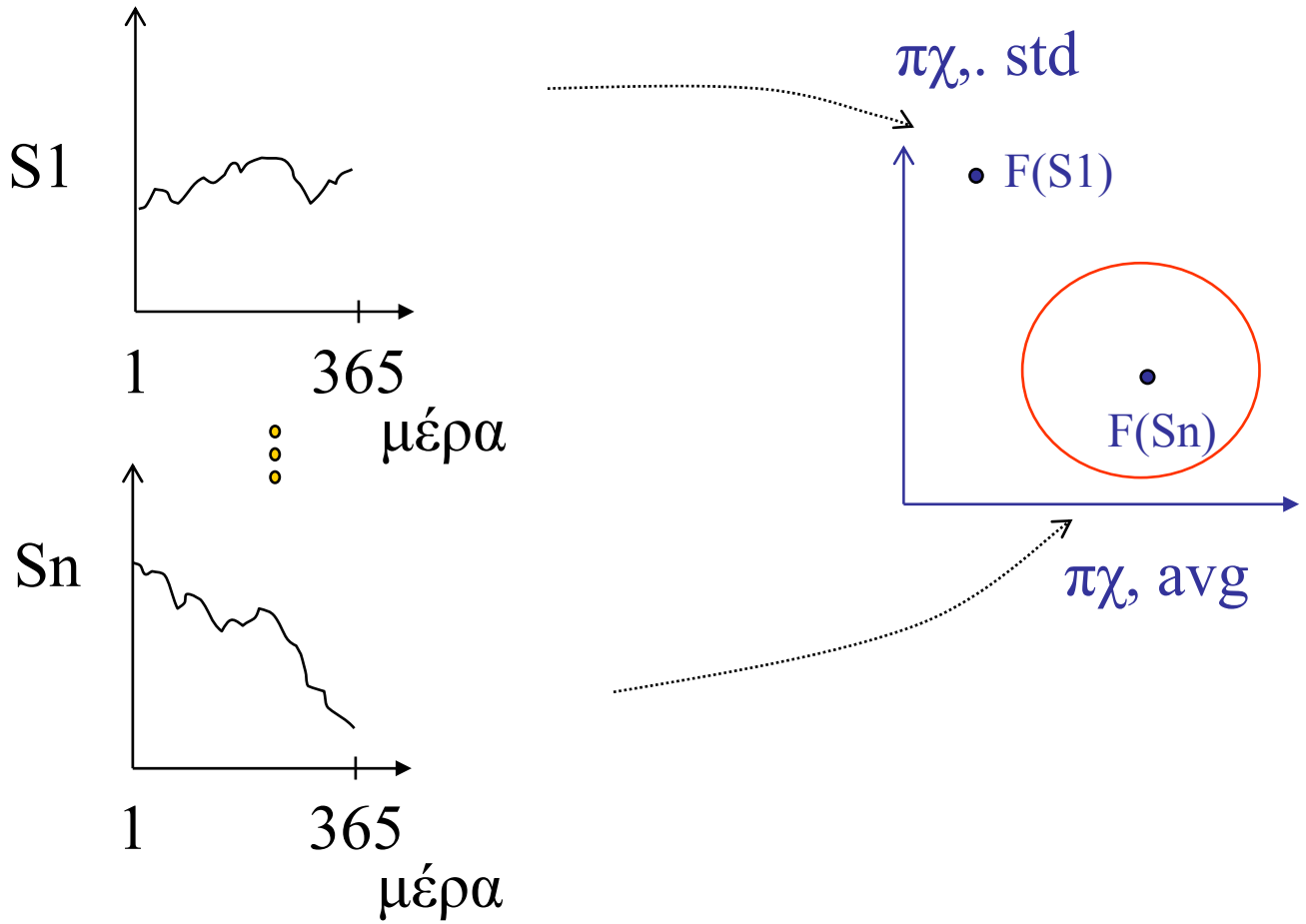


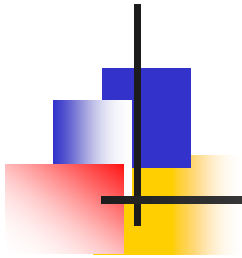
Ιδέα: ‘GEMINI’

(GEneric Multimedia INdexIng)

Εξήγαγε μερικά αριθμητικά χαρακτηριστικά, για ‘γρήγορο και πρόχειρο’ έλεγχο

‘GEMINI’ - Παραστατικά

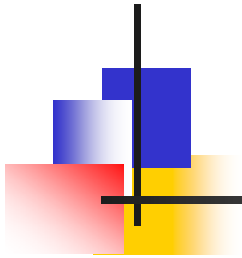




GEMINI

Λύση: 'Γρήγορο-και-πρόχειρο' φίλτρο:

- εξήγαγε *n* χαρακτηριστικά (αριθμούς, πχ., avg, κτλ.)
- πρόβαλε σε ένα σημείο στο *n*-διάστατο χώρο χαρακτηριστικών
- οργάνωσε τα σημεία με έτοιμη spatial access μέθοδο ('SAM')
- Απόρριψε false alarms

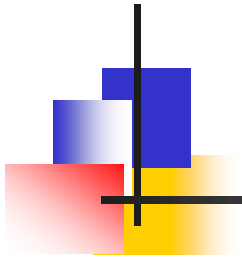


GEMINI

Σημαντικό: Ε: πώς να εγγυηθεί απουσία ψευδών απορρίψεων;

Α1: διατήρηση αποστάσεων (αλλά: δύσκολο/ακατόρθωτο)

Α2: Lower-bounding λήμμα: αν η αποτύπωση ‘κάνει τα πράγματα να είναι εγγύτερα’, τότε δεν υπάρχουν ψευδείς απορρίψεις

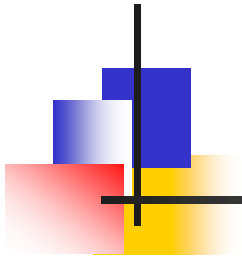


GEMINI

Σημαντικό :

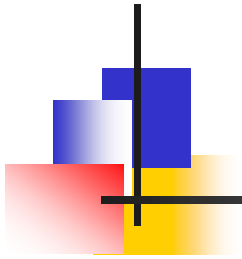
Q: πώς να εξάγουμε τα χαρακτηριστικά?

A: “Εάν έχω μόνο έναν αριθμό για να περιγράψω το αντικείμενο μου ποιο θα έπρεπε να είναι αυτό?”



Χρονικές Ακολουθίες

Q: ποια χαρακτηριστικά?



Χρονικές Ακολουθίες

Q: ποια χαρακτηριστικά?

A: Συντελεστές Fourier (θα τους δούμε στην συνέχεια)

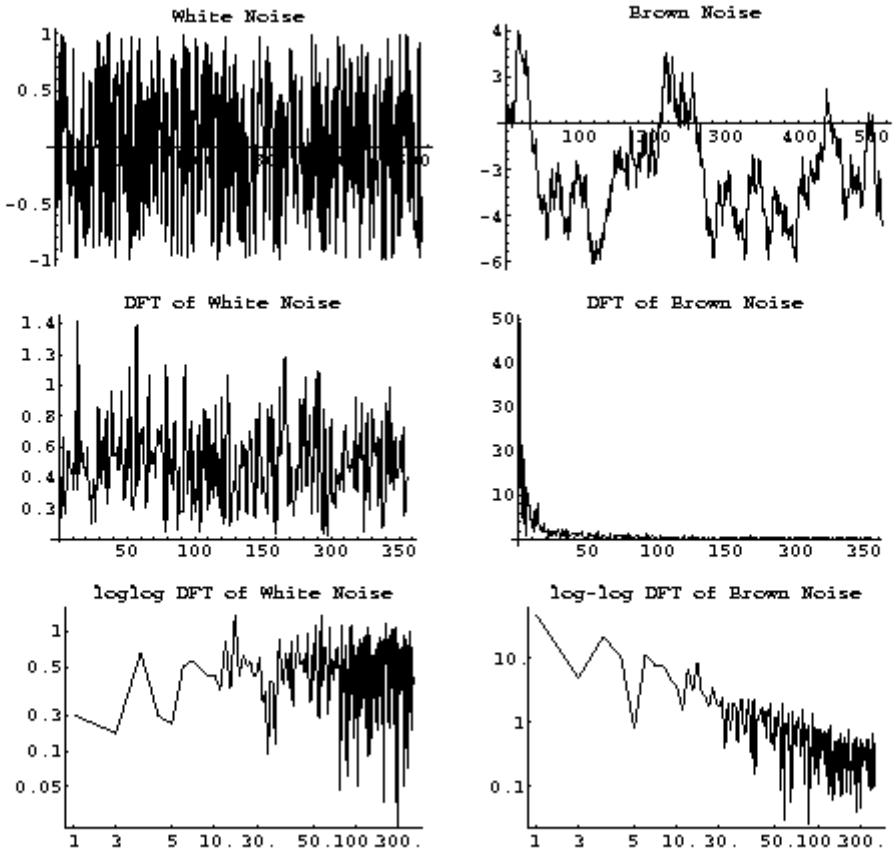
Χρονικές Ακολουθίες

white noise

brown noise

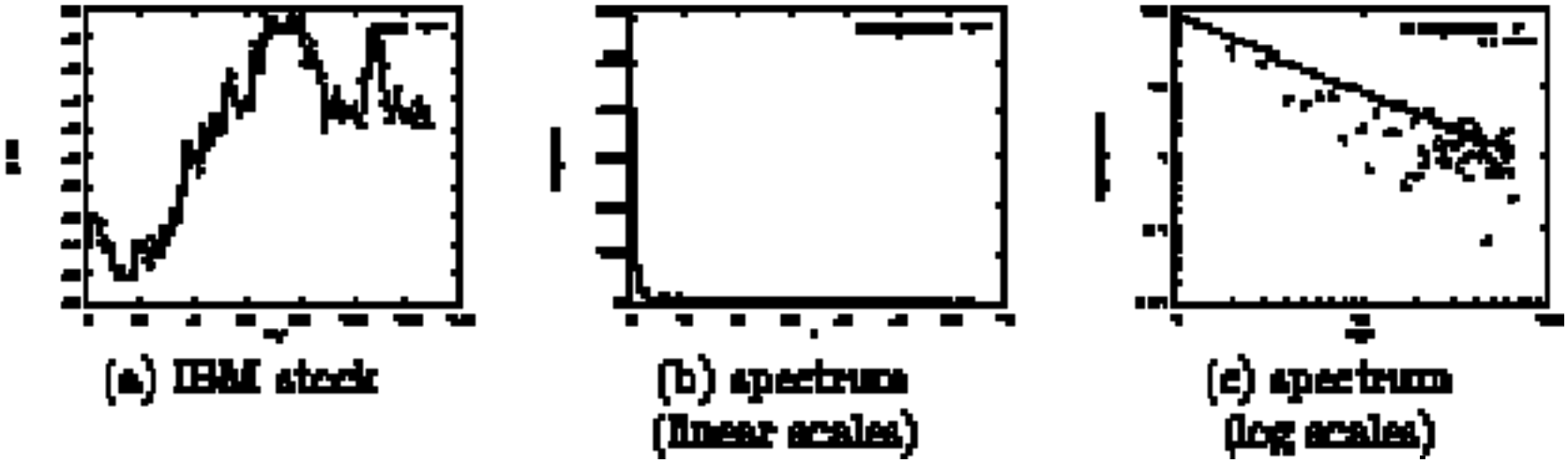
Fourier spectrum

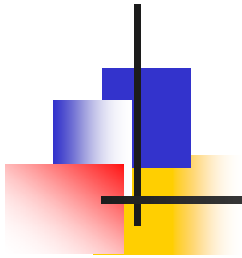
... in log-log



Χρονικές Ακολουθίες

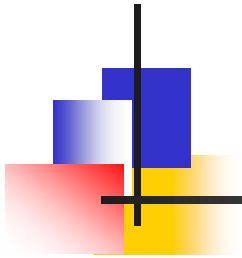
- Eg.:





Χρονικές Ακολουθίες

- Συμπέρασμα: οι χρωματικοί θόρυβοι προσεγγίζονται από τους πρώτους Fourier συντελεστές
- Οι χρωματικοί θόρυβοι εμφανίζονται στην φύση

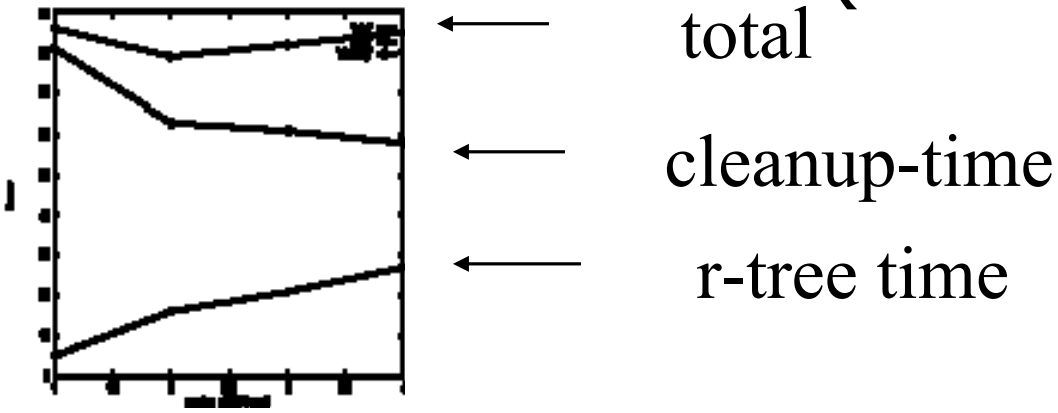


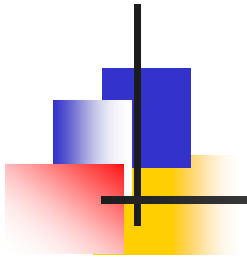
Χρονικές Ακολουθίες

- brown noise: τιμές μετοχών($1/f^2$ energy spectrum)
- pink noise: works of art (1/f spectrum)
- black noises: δεξαμενές νερού ($1/f^b$, b>2)
- (slope: related to 'Hurst exponent', for self-similar traffic, like, eg. Ethernet/web [Schroeder], [Leland+]

Χρονικές Ακολουθίες-Αποτελέσματα

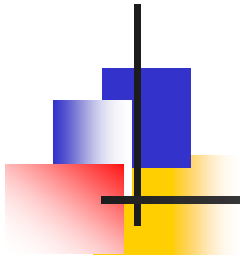
- κράτησε τους πρώτους 2-3 Fourier συντελεστές
- Πιο γρήγορας από ακολουθιακή αναζήτηση
- NO false dismissals (see book)





Χρονικές Ακολουθίες - βελτιστοποιήσεις

- Βελτιστοποιήσεις/παραλλαγές: [Kanellakis+Goldin], [Mendelzon+Rafiei]
- Μπορούν να χρησιμοποιηθούν Wavelets, or DCT
- Μπορούν να χρησιμοποιηθούν segment averages [Yi+2000]



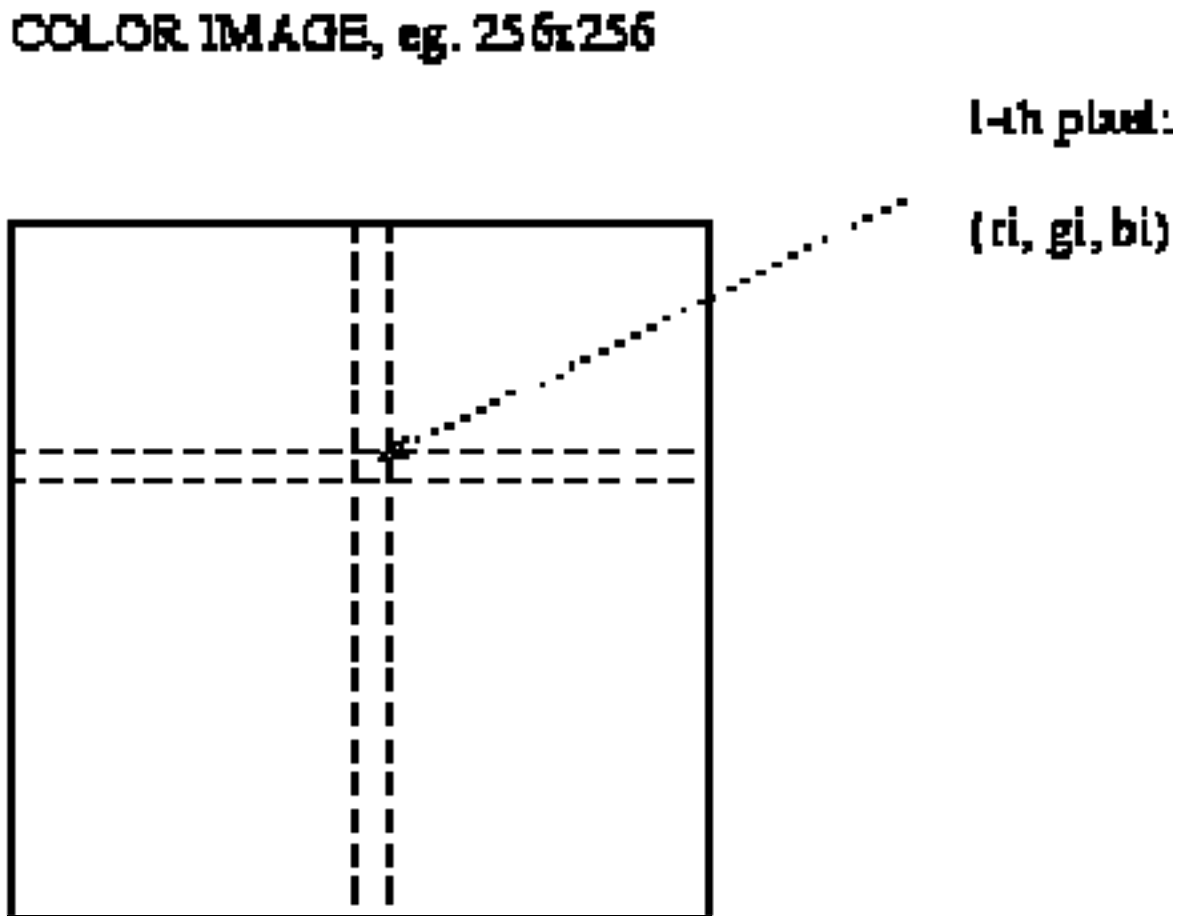
Πολυμεσικά Δεδομένα

- Πολυμεσικά δεδομένα
  - Κίνητρο/ ορισμός προβλήματος
  - Κύρια ιδέα/ χρονικές ακολουθίες
  - εικόνες (χρώμα, σχήμα)
  - sub-pattern matching
  - Αυτόματη εξαγωγή χαρακτηριστικών/ FastMap

Images - color

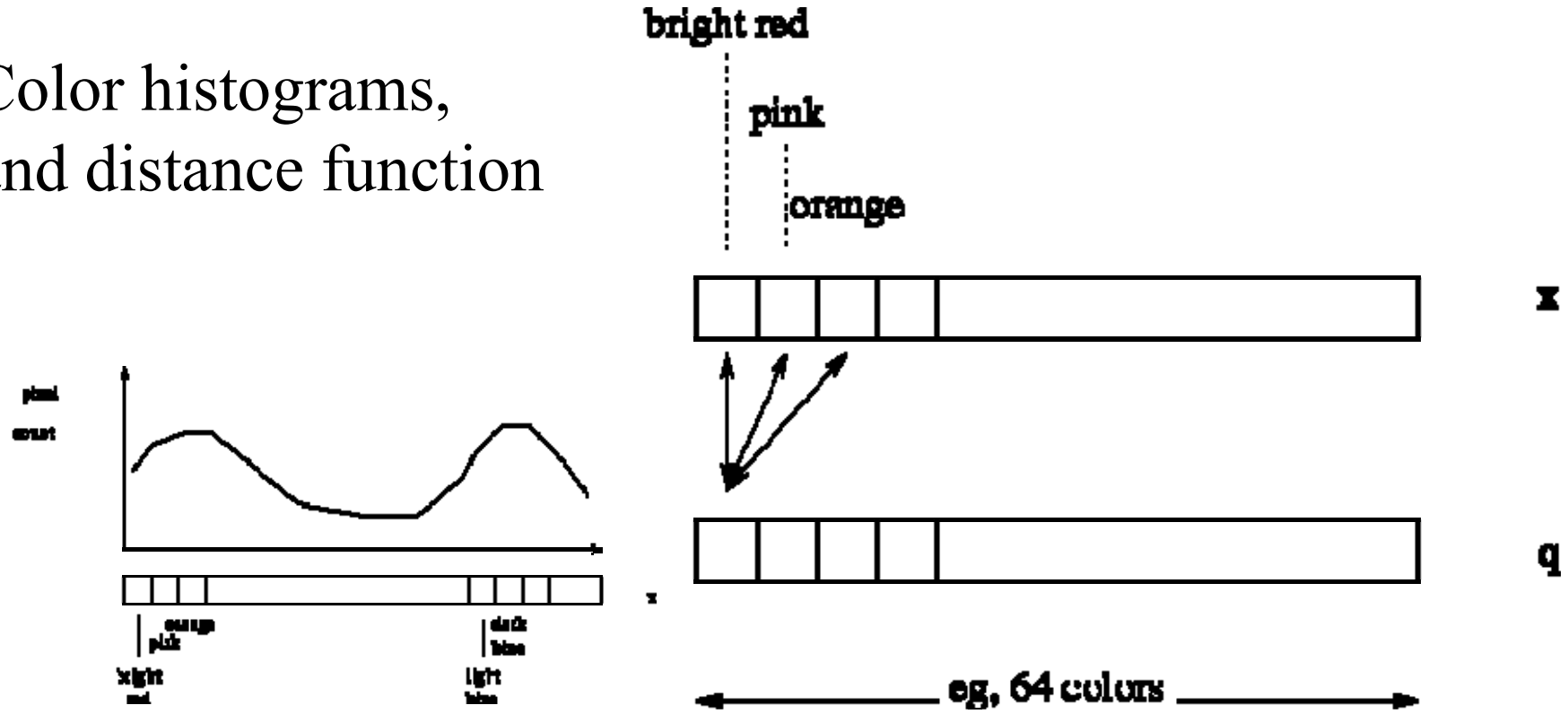
what is an image?

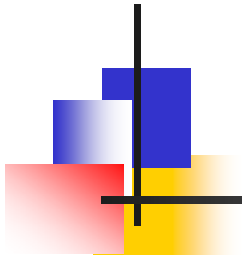
A: 2-d array



Images - color

Color histograms,
and distance function

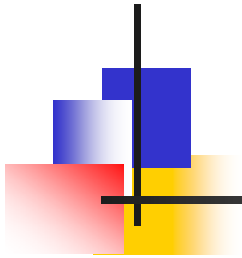




Images - color

Mathematically, the distance function is:

$$\begin{aligned} distance_{histogram}(\vec{I}, \vec{I}') &= (\vec{I} - \vec{I}') \begin{bmatrix} a_{RR} & a_{RP} & \dots \\ a_{PR} & a_{PP} & \dots \\ \dots & \dots & \dots \end{bmatrix} (\vec{I} - \vec{I}')^t \\ \dots &= (\vec{I} - \vec{I}') A (\vec{I} - \vec{I}')^t \end{aligned}$$

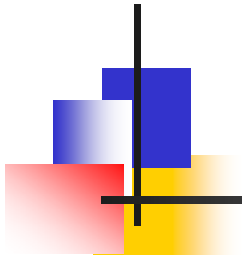


Images - color

Problem: ‘cross-talk’:

- Features are not orthogonal ->
- SAMs will not work properly

- Q: what to do?
- A: feature-extraction question



Images - color

possible answers:

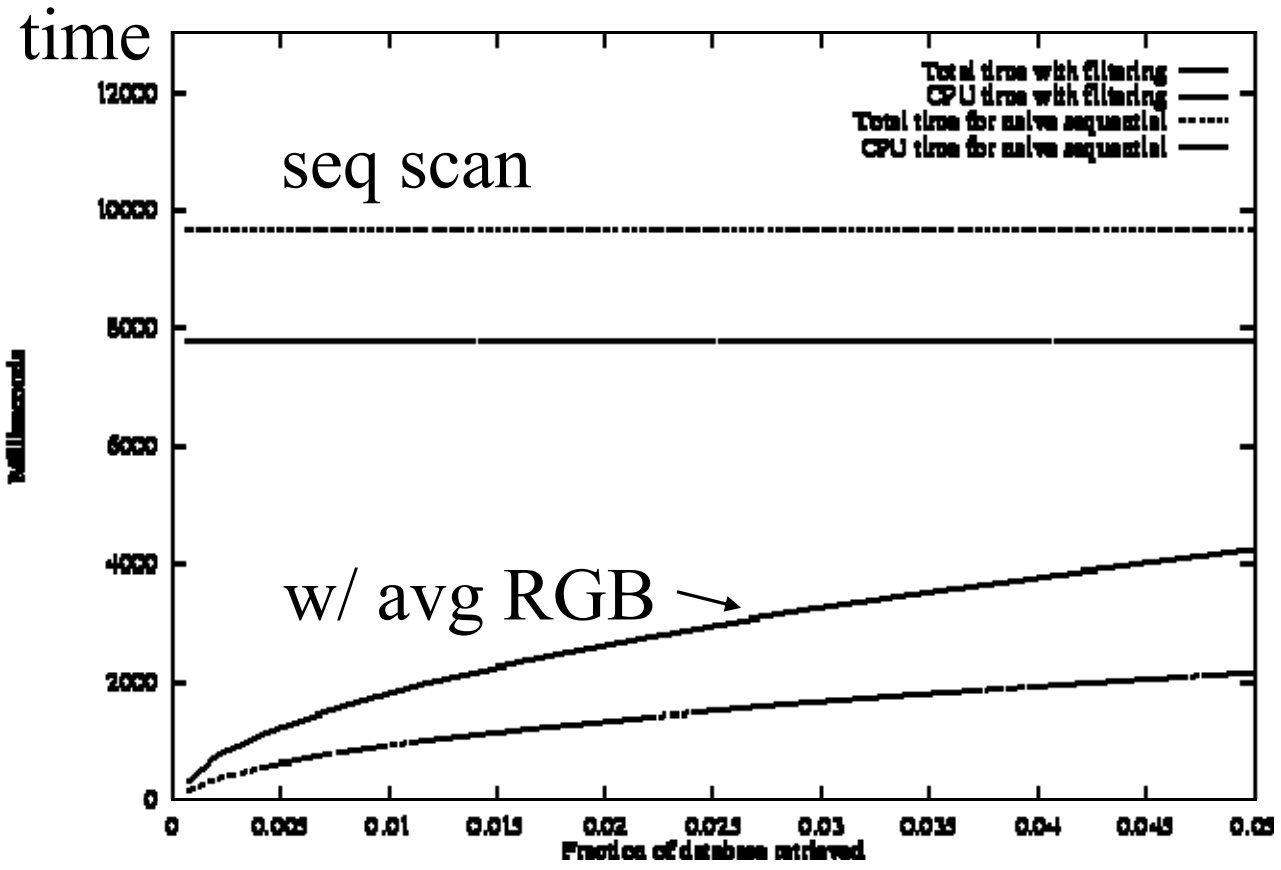
- avg red, avg green, avg blue

it turns out that this lower-bounds the histogram distance ->

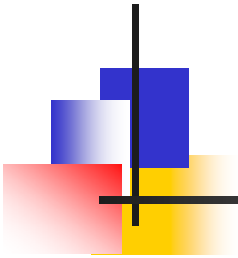
- no cross-talk
- SAMs are applicable

Images - color

performance:

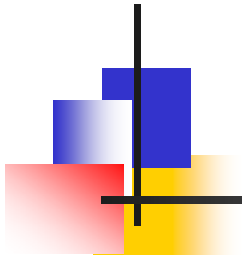


selectivity



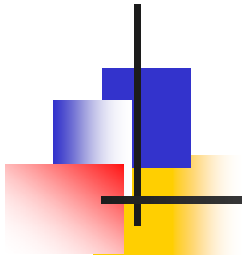
Multimedia - Detailed outline

- multimedia
  - Motivation / problem definition
  - Main idea / time sequences
  - images (color; shape)
  - sub-pattern matching
  - automatic feature extraction / FastMap



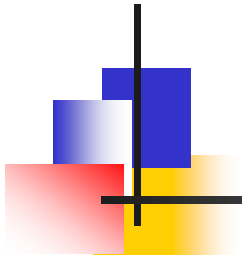
Images - shapes

- distance function: Euclidean, on the area, perimeter, and 20 ‘moments’
- (Q: how to normalize them?



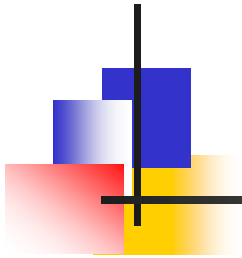
Images - shapes

- distance function: Euclidean, on the area, perimeter, and 20 ‘moments’
- (Q: how to normalize them?
- A: divide by standard deviation)



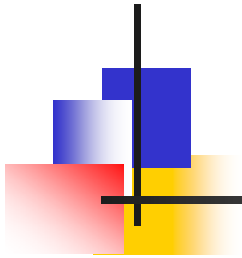
Images - shapes

- distance function: Euclidean, on the area, perimeter, and 20 ‘moments’
- (Q: other ‘features’ / distance functions?



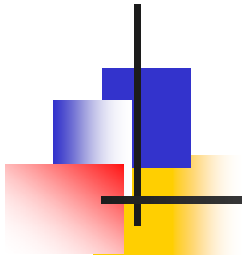
Images - shapes

- distance function: Euclidean, on the area, perimeter, and 20 ‘moments’
- (Q: other ‘features’ / distance functions?
- A1: turning angle
- A2: dilations/erosions
- A3: ...)



Images - shapes

- distance function: Euclidean, on the area, perimeter, and 20 ‘moments’
- Q: how to do dim. reduction?



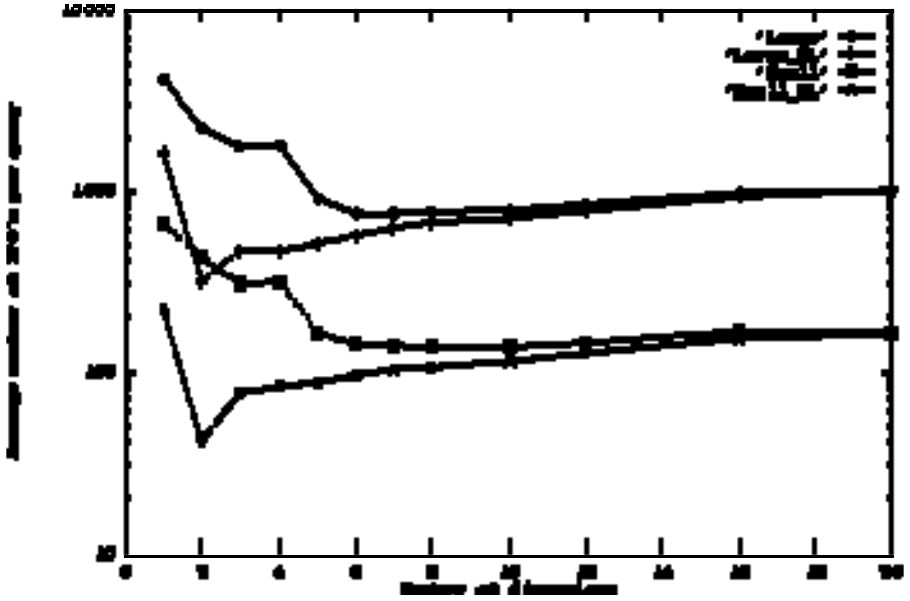
Images - shapes

- distance function: Euclidean, on the area, perimeter, and 20 ‘moments’
- Q: how to do dim. reduction?
- A: Karhunen-Loeve (= centered PCA/SVD)

Images - shapes

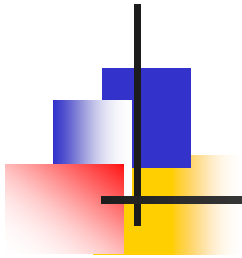
- Performance: ~10x faster

log(# of I/Os)



← all kept

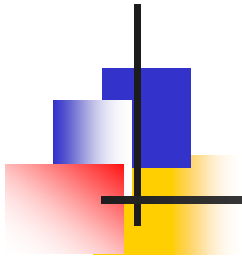
# of features kept



Case study: Informedia

- Video database system, developed at CMU
- 2+ TB of video data (broadcast news)
- retrieval by text, image and face similarity

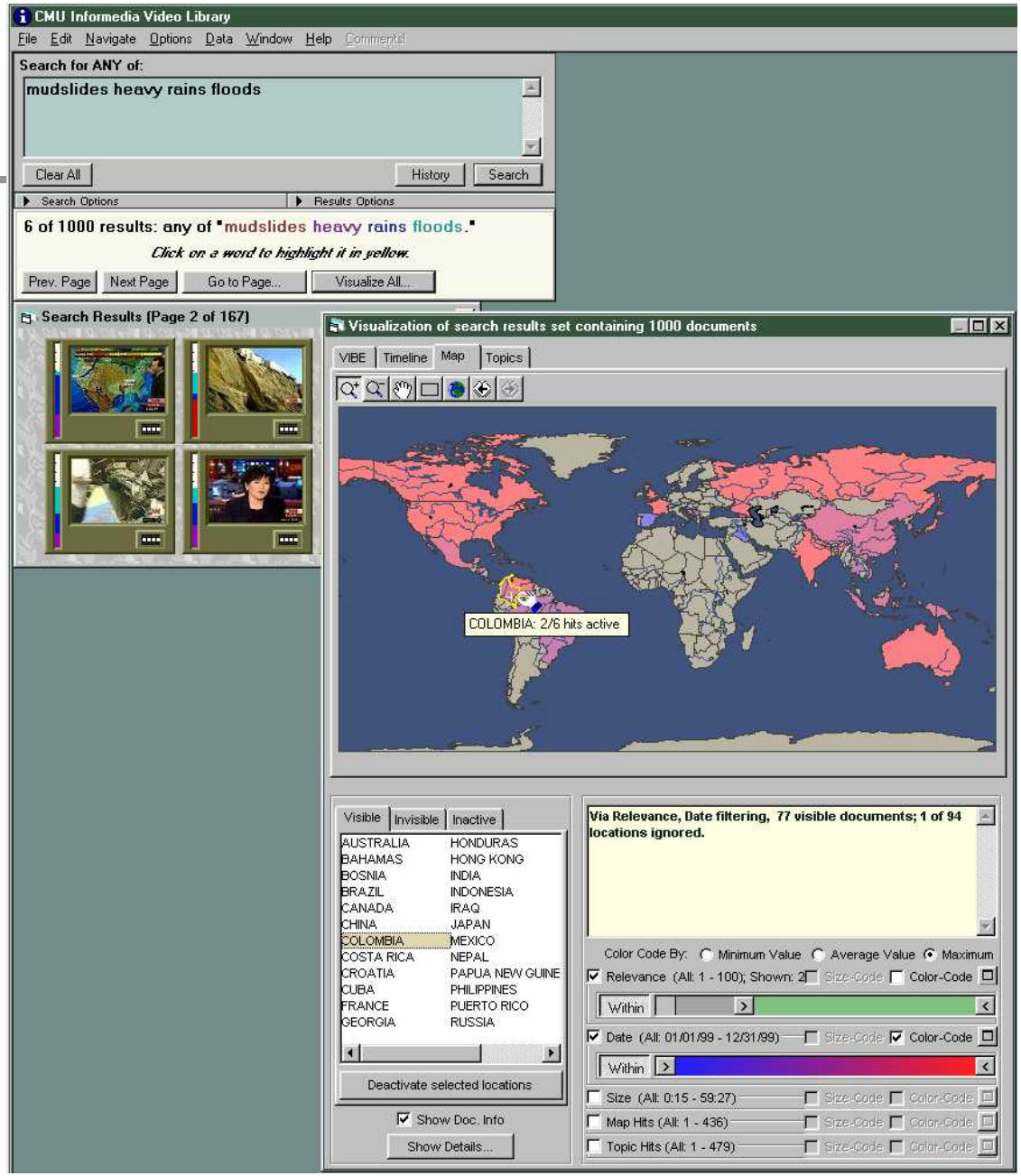
www.informedia.cs.cmu.edu/



Case study: Informedia

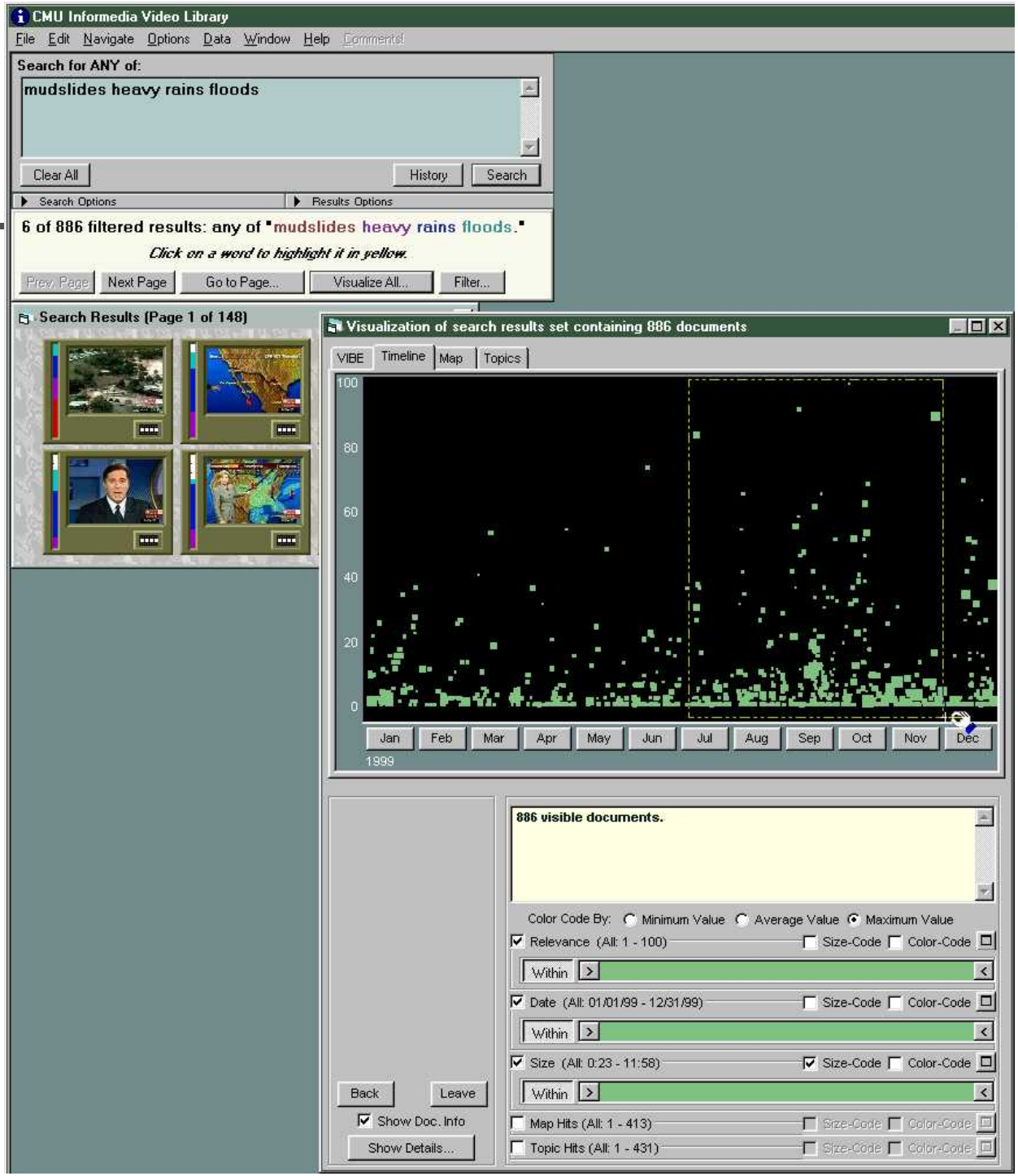
- next foils: visualization features
  - by space
  - by time
  - by concept

- geo mapping
- automatic place recognition
- ambiguity resol. +
- lookup



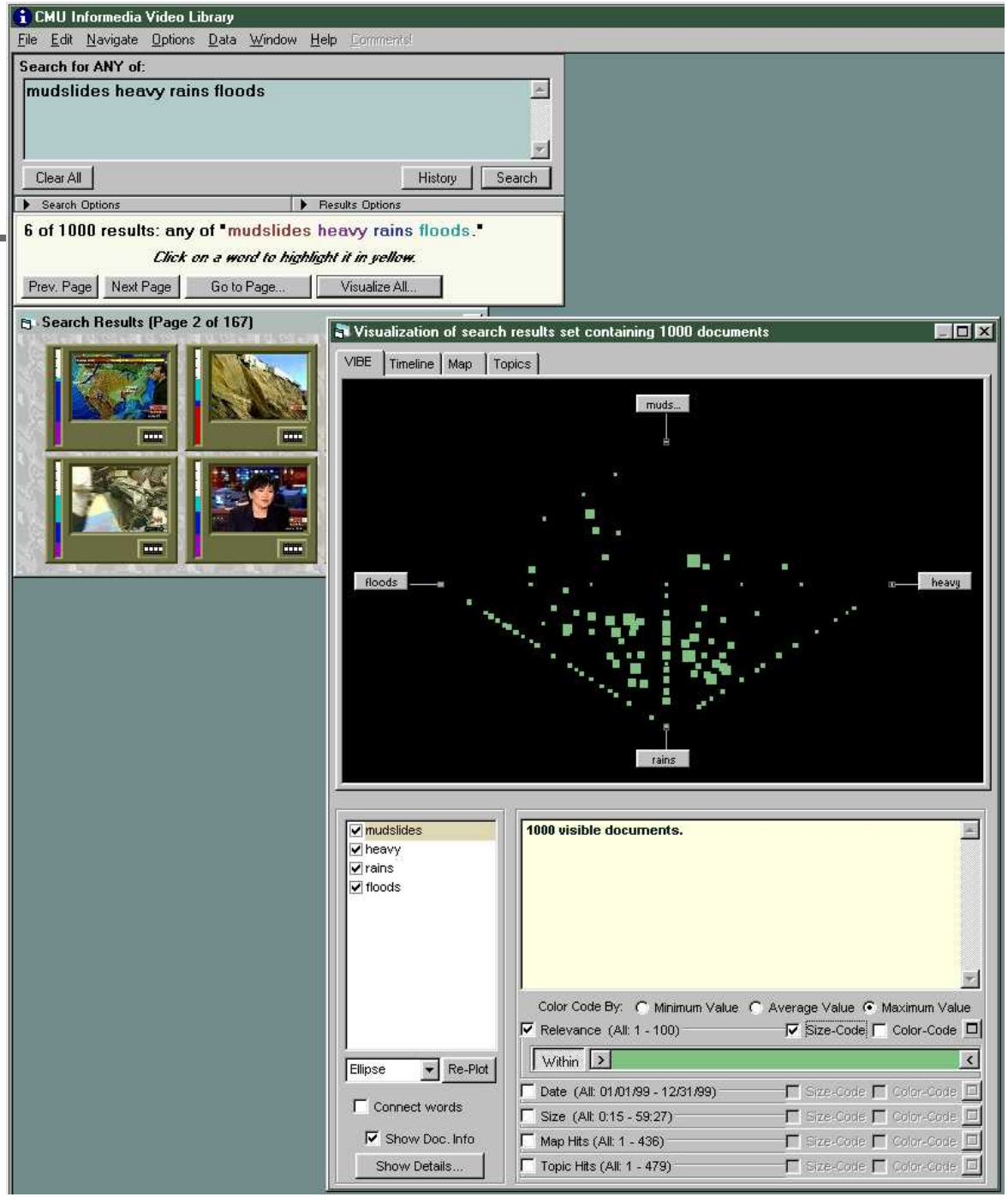

CMU Informedia Video Library
File Edit Navigate Options Data Window Help Comments!
Search for ANY of:
mudslides heavy rains
Clear All
History
Search
Search Options
Results Options
6 of 287 results: any of "mudslides heavy rains."
Click on a word to highlight it in yellow.
Prev. Page
Next Page
Go to Page...
Visualize All...
Search Results (Page 1 of 48)
6 of total
Map with data from: Moscow's Endgame Isn't Clear Yet; Xerox S...
84.838135W , 33.806364N
Texas
Mexico
Mexico City
Scale 1":9
Label
Cities
NAME
Moscow's Endgame Isn't Clear Yet; Xerox Shares we...
CNN
RESET
PREV.
ALL
NEXT
in a massive mudslide. It happened in the southeastern part of the country, triggered by days of torrential rains and flooding. At least 230 people have died. Cnn's Mexico city bureau chief Harris Whitbeck reports. Wearing borrowed clothes, he has nothing left. He stares at the remnants of his house buried in a sea of mud. He's convinced his 15-year-old daughter was swept under the mud as well. [speaking spanish] a neighbor told him how seconds before the mud av latch he saw her in front of the house folding her umbrella as she arrived from school. Nearly a week of heavy rains caused more than 80 mud slides in five Mexican states.

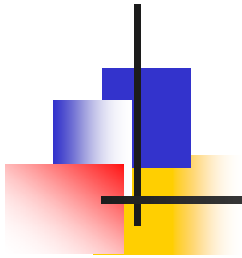
time
line


CMU Informedia Video Library
File Edit Navigate Options Data Window Help Comments!
Search for ANY of:
mudslides heavy rains floods
Clear All
History
Search
Search Options
Results Options
6 of 886 filtered results: any of "mudslides heavy rains floods."
Click on a word to highlight it in yellow.
Prev. Page
Next Page
Go to Page...
Visualize All...
Filter...
Search Results (Page 1 of 148)
Visualization of search results set containing 886 documents
VIBE
Timeline
Map
Topics
100
80
60
40
20
0
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec
1999
886 visible documents.
Color Code By: Minimum Value Average Value Maximum Value
Relevance (All: 1 - 100)
Size-Code
Color-Code
Within
Date (All: 01/01/99 - 12/31/99)
Size (All: 0:23 - 11:58)
Map Hits (All: 1 - 413)
Topic Hits (All: 1 - 431)
Back
Leave
Show Doc. Info
Show Details...

concept

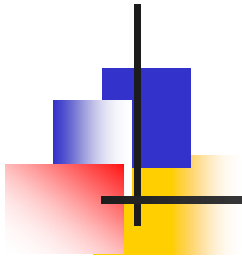
space





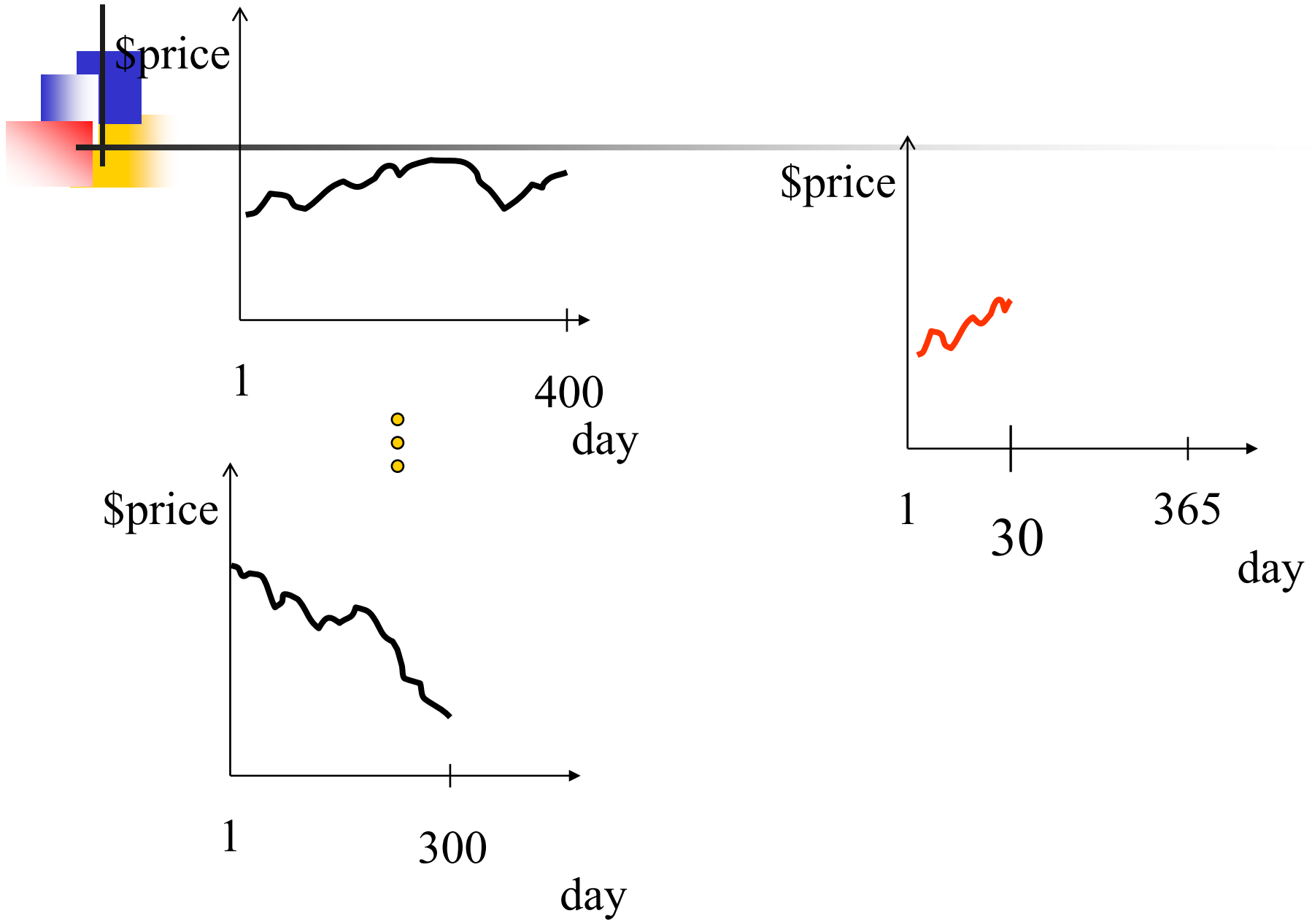
Multimedia - Detailed outline

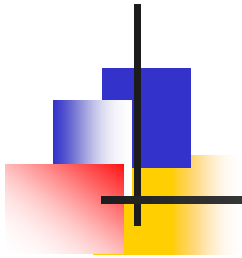
- multimedia
  - Motivation / problem definition
  - Main idea / time sequences
  - images (color; shape)
  - sub-pattern matching
  - automatic feature extraction / FastMap



Sub-pattern matching

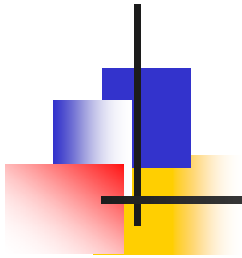
- Problem: find **sub**-sequences that match the given query pattern


$price
1
400
day
$price
1
30
365
day
$price
1
300
day



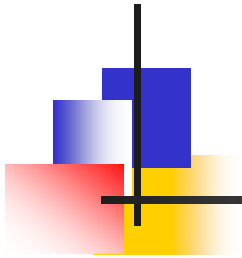
Sub-pattern matching

- Q: how to proceed?
- Hint: try to turn it into a ‘whole-matching’ problem (how?)



Sub-pattern matching

- Assume that queries have minimum duration *w;* (eg., *w*=7 days)
- divide data sequences into windows of width *w* (overlapping, or not?)

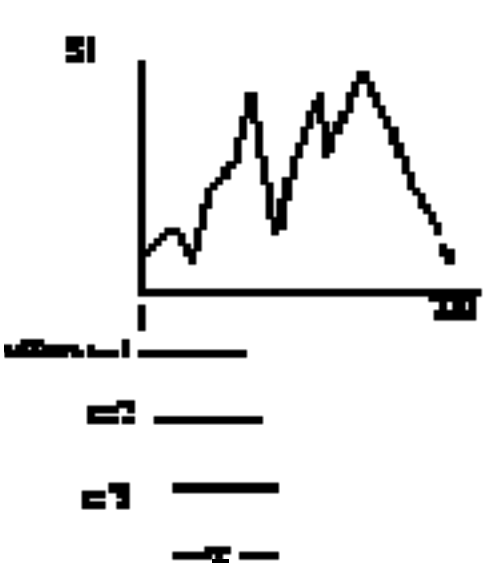


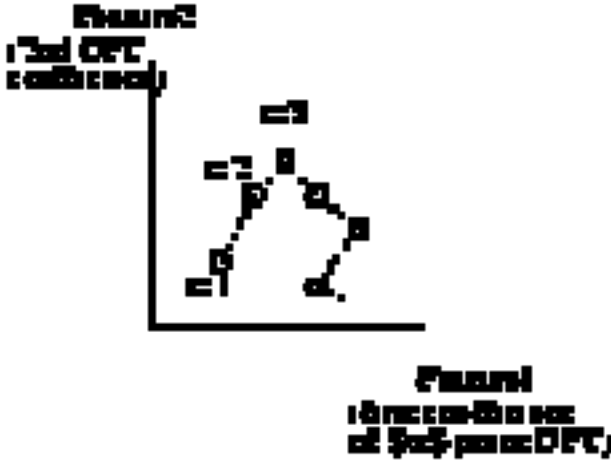
Sub-pattern matching

- Assume that queries have minimum duration *w;* (eg., *w*=7 days)
- divide data sequences into windows of width *w* (overlapping, or not?)
- A: sliding, overlapping windows. Thus: trails

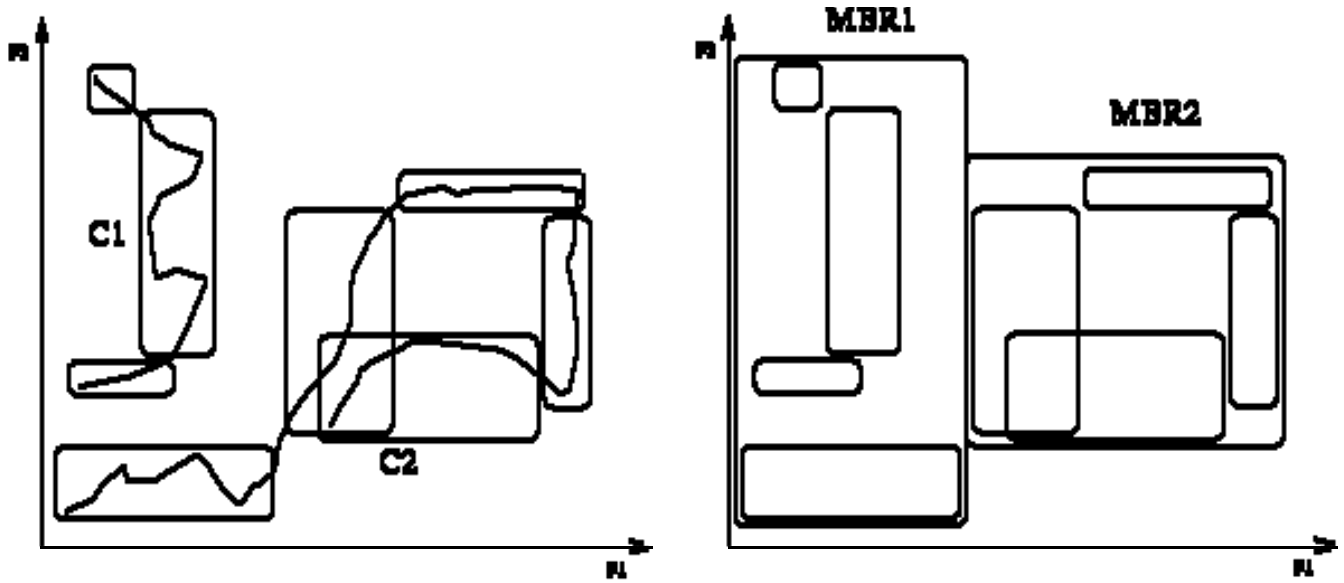
Pictorially:

Sub-pattern matching



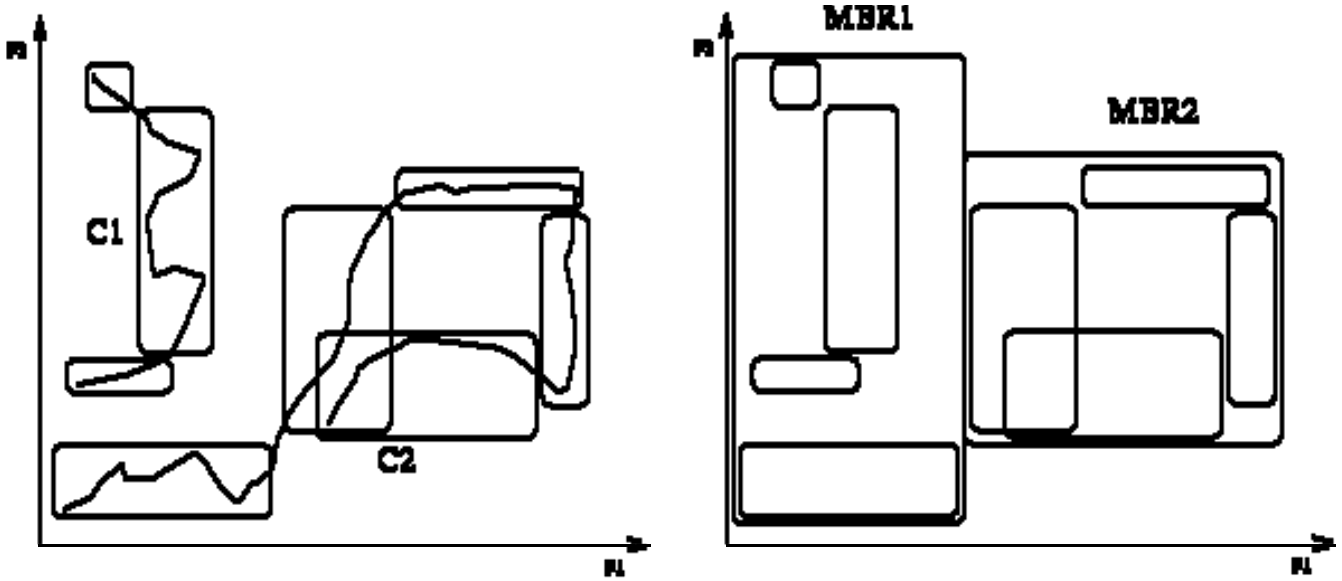


Sub-pattern matching



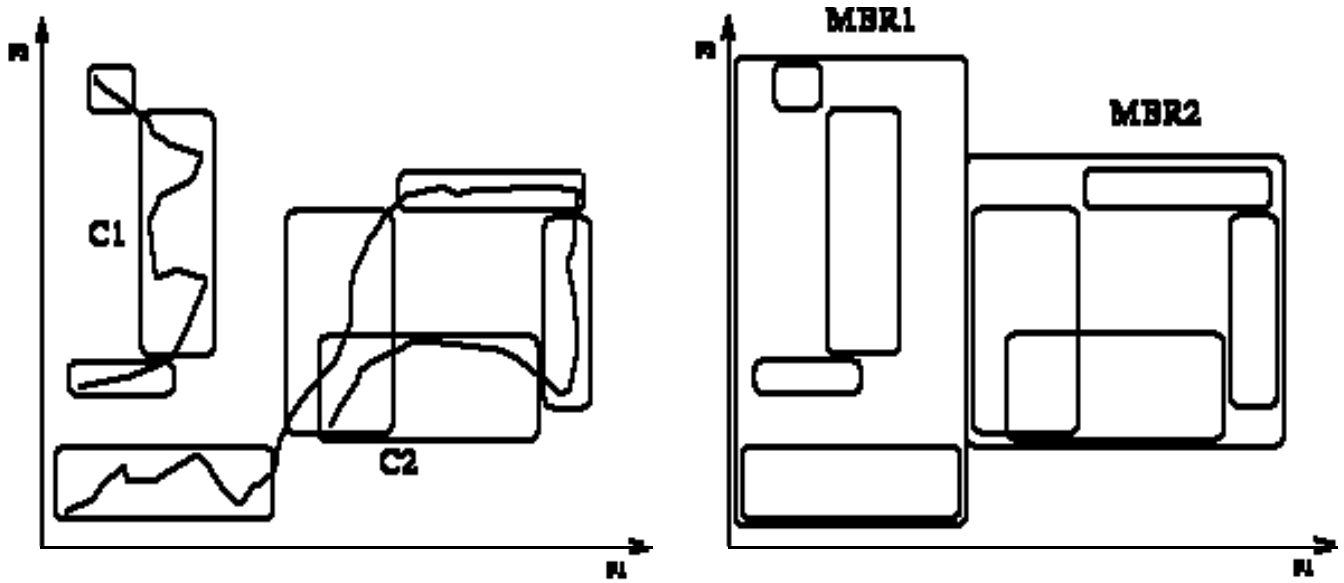
sequences -> trails -> MBRs in feature space

Sub-pattern matching

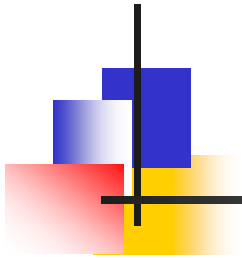


Q: do we store all points? why not?

Sub-pattern matching



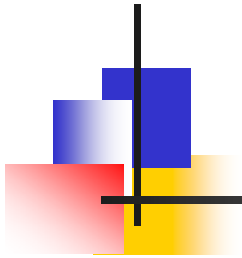
Q: how to do range queries of duration *w*?



Sub-pattern matching

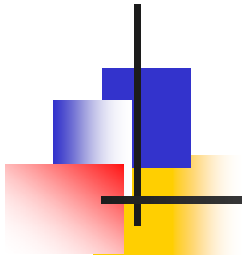
(very recent improvement [Moon+2001])

- use non-overlapping windows, for data



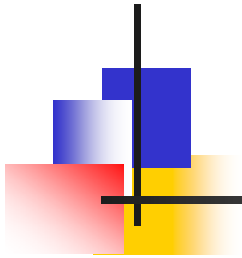
Conclusions

- GEMINI works for any setting (time sequences, images, etc)
- uses a ‘quick and dirty’ filter
- faster than seq. scan
- (but: how to extract features automatically?)



Multimedia - Detailed outline

- multimedia
  - Motivation / problem definition
  - Main idea / time sequences
  - images (color; shape)
  - sub-pattern matching
  - automatic feature extraction / FastMap



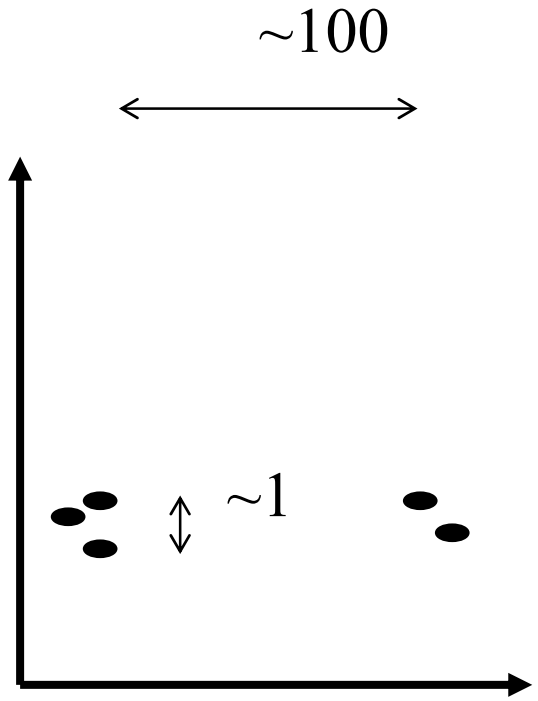
FastMap

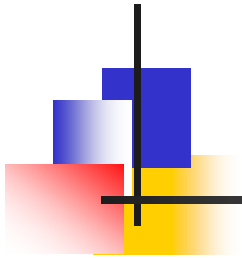
Automatic feature extraction:

- Given a dissimilarity function of objects
- Quickly map the objects to a (k-d) `feature' space.
- (goals: indexing and/or visualization)

FastMap

	O1	O2	O3	O4	O5
O1	0	1	1	100	100
O2	1	0	1	100	100
O3	1	1	0	100	100
O4	100	100	100	0	1
O5	100	100	100	1	0



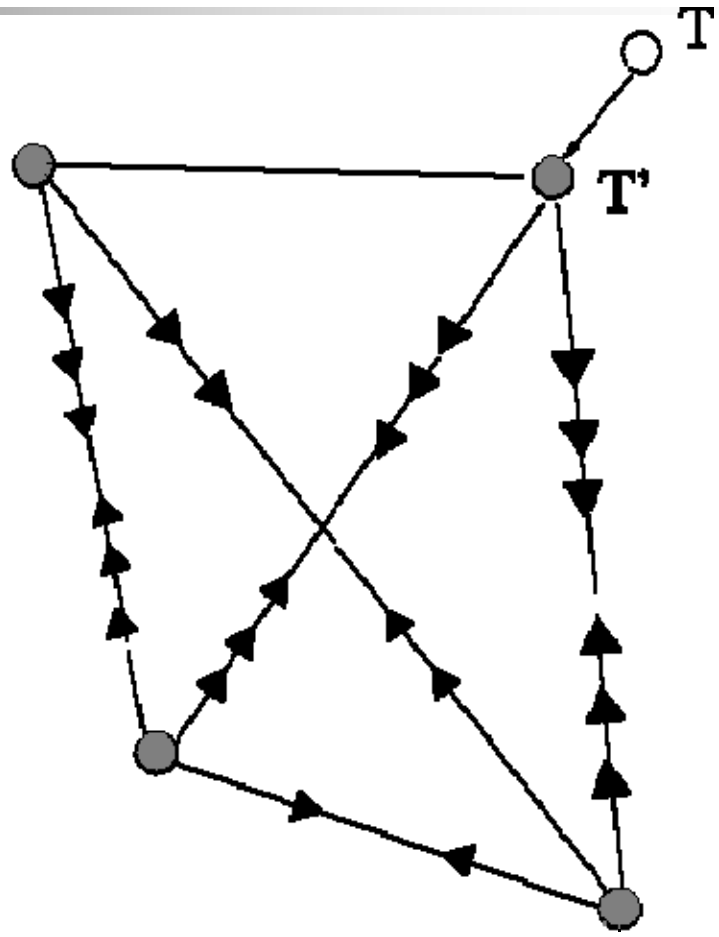


FastMap

- Multi-dimensional scaling (MDS) can do that, but in O(N**2) time

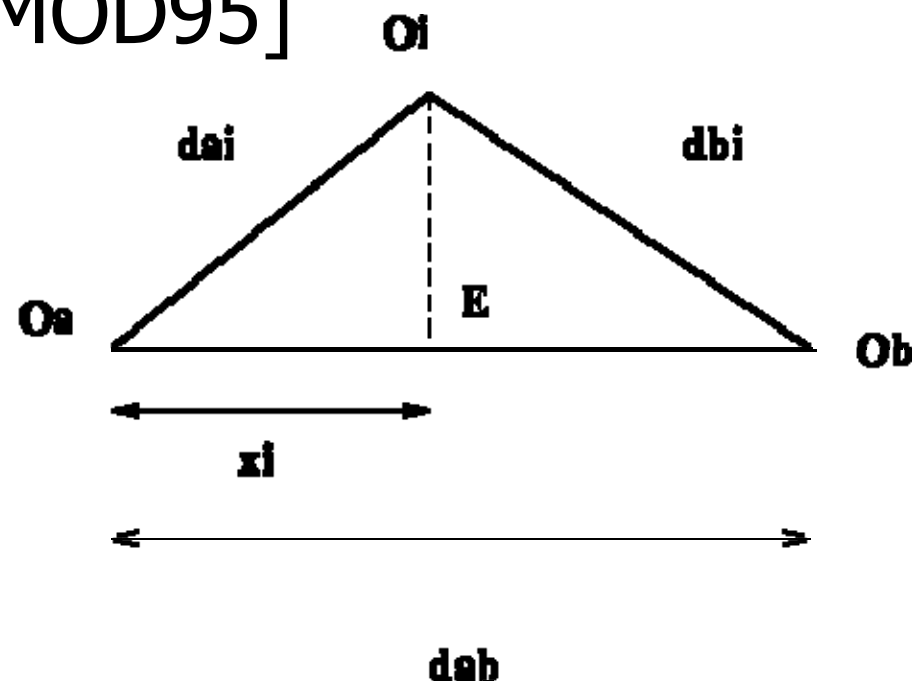
MDS

Multi Dimensional Scaling

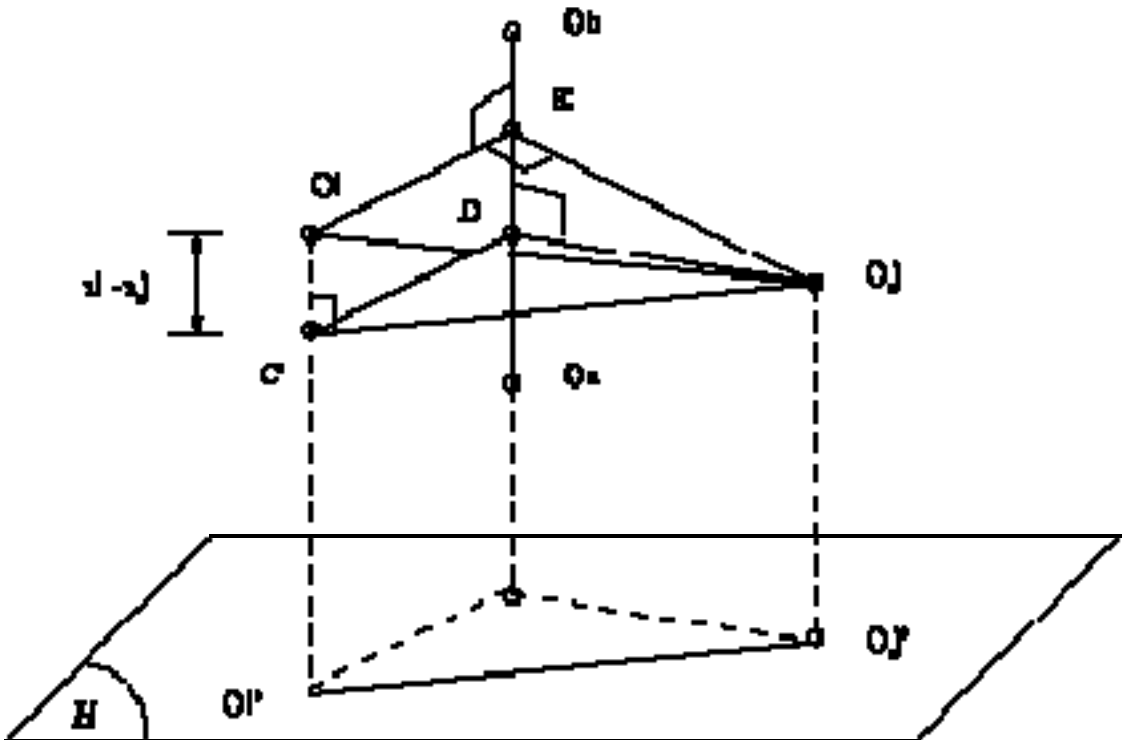


Main idea: projections

We want a linear algorithm: FastMap [SIGMOD95]

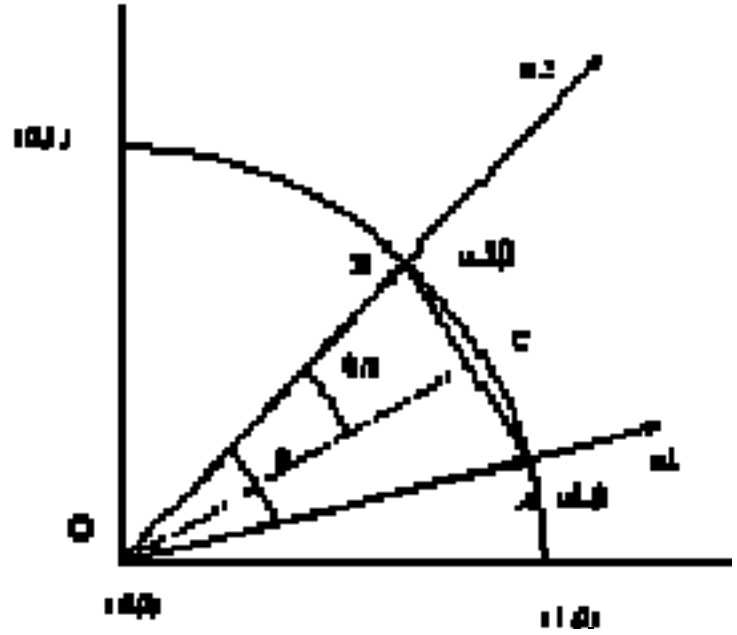


FastMap - next iteration



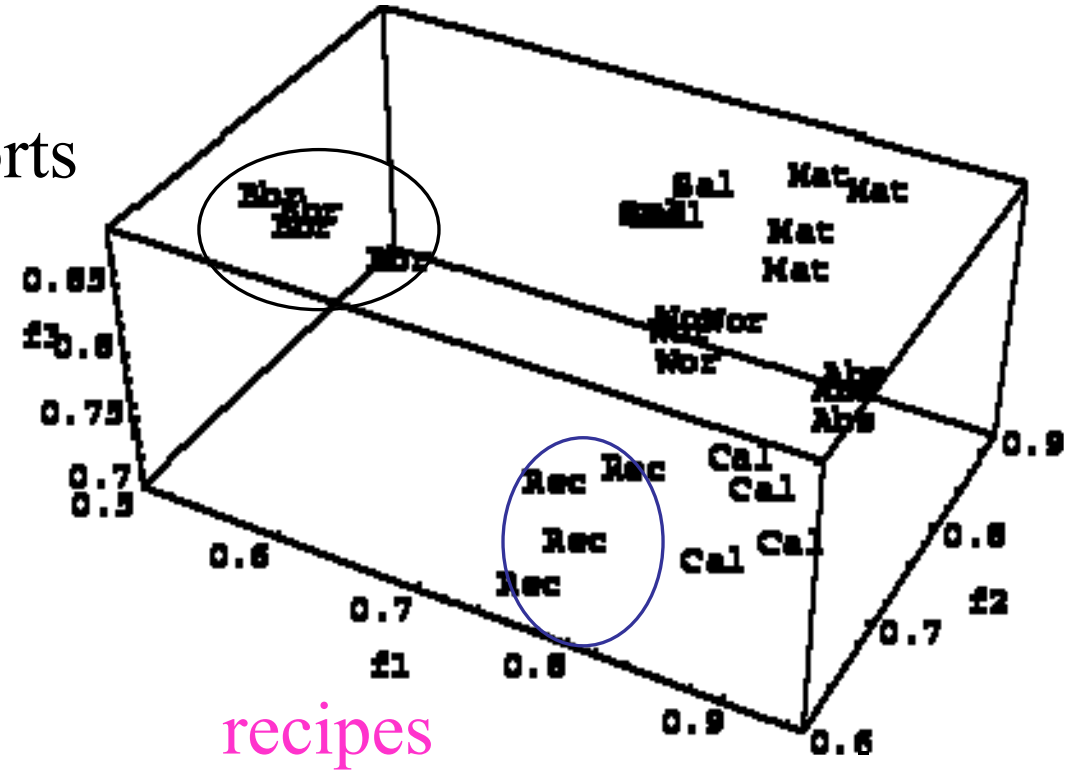
Results

Documents /cosine similarity ->
Euclidean distance (how?)



Results

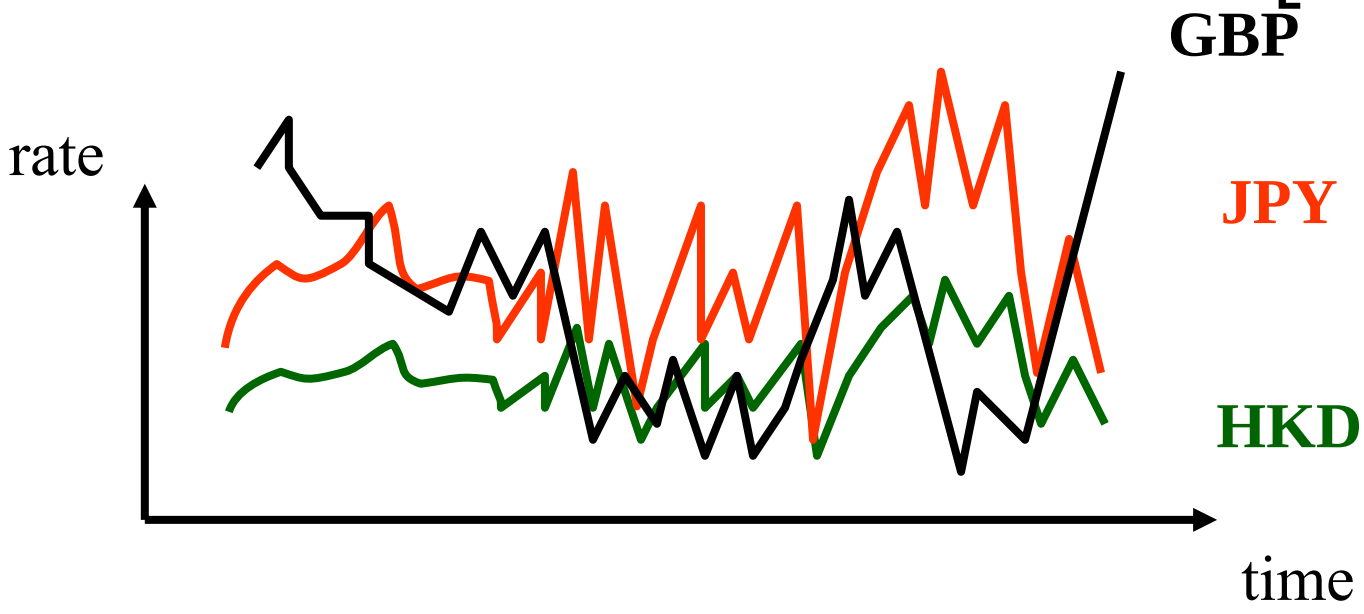
bb reports



recipes

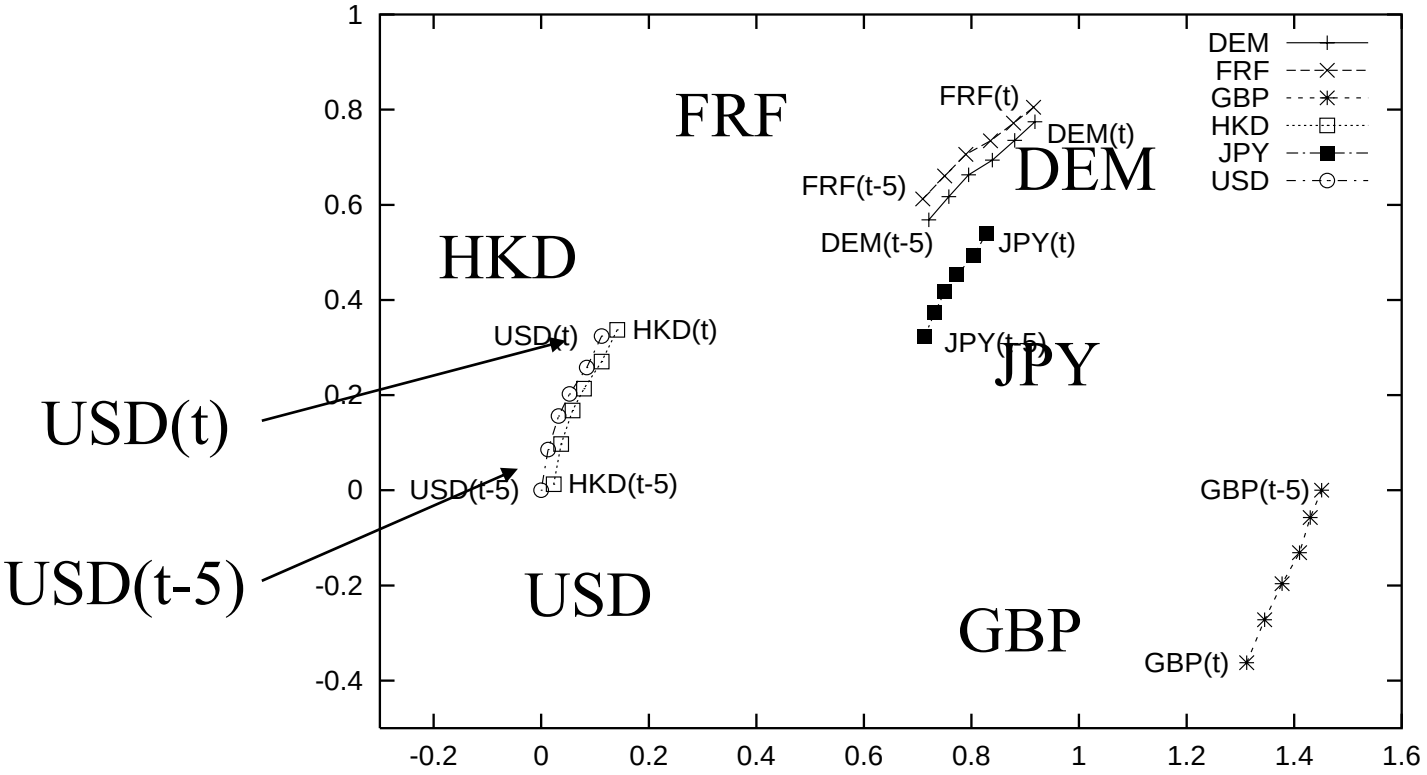
Applications: time sequences

- given *n* co-evolving time sequences
- visualize them + find rules [ICDE00]



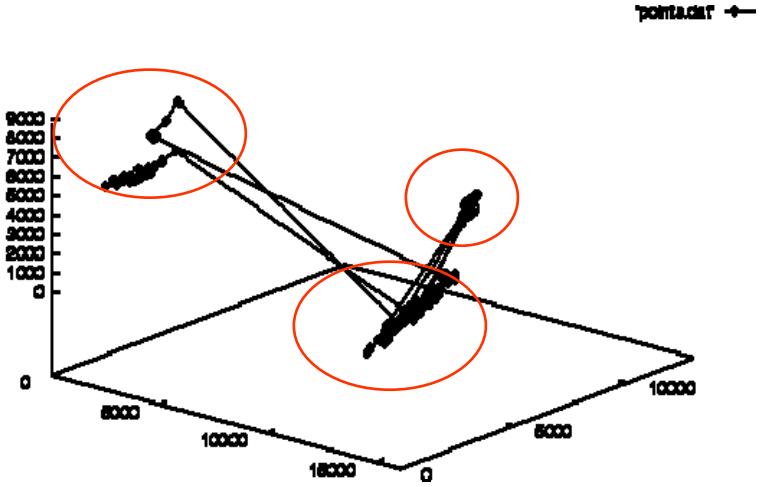
Applications - financial

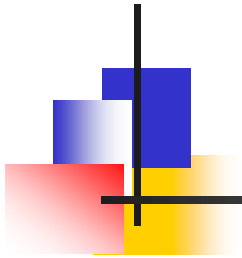
- currency exchange rates [ICDE00]



VideoTrails

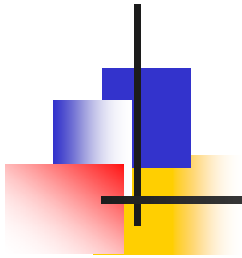
[ACM MM97]





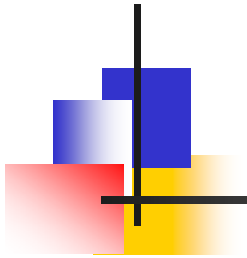
Conclusions

- GEMINI works for multiple settings
- FastMap can extract ‘features’ automatically (-> indexing, visual d.m.)



References

- Faloutsos, C., R. Barber, et al. (July 1994). "Efficient and Effective Querying by Image Content." J. of Intelligent Information Systems 3(3/4): 231-262.
- Faloutsos, C. and K.-I. D. Lin (May 1995). FastMap: A Fast Algorithm for Indexing, Data-Mining and Visualization of Traditional and Multimedia Datasets. Proc. of ACM-SIGMOD, San Jose, CA.
- Faloutsos, C., M. Ranganathan, et al. (May 25-27, 1994). Fast Subsequence Matching in Time-Series Databases. Proc. ACM SIGMOD, Minneapolis, MN.
- Flickner, M., H. Sawhney, et al. (Sept. 1995). "Query by Image and Video Content: The QBIC System." IEEE Computer 28(9): 23-32.
- Goldin, D. Q. and P. C. Kanellakis (Sept. 19-22, 1995). On Similarity Queries for Time-Series Data: Constraint Specification and Implementation. Int. Conf. on Principles and Practice of Constraint Programming (CP95), Cassis, France.



References

- Leland, W. E., M. S. Taqqu, et al. (Feb. 1994). "On the Self-Similar Nature of Ethernet Traffic." IEEE Transactions on Networking 2(1): 1-15.
- Moon, Y.-S., K.-Y. Whang, et al. (2001). Duality-Based Subsequence Matching in Time-Series Databases. ICDE, Heidelberg, Germany.
- Rafiei, D. and A. O. Mendelzon (1997). Similarity-Based Queries for Time Series Data. SIGMOD Conference, Tucson, AZ.
- Schroeder, M. (1991). Fractals, Chaos, Power Laws: Minutes from an Infinite Paradise. New York, W.H. Freeman and Company.
- Yi, B.-K. and C. Faloutsos (2000). Fast Time Sequence Indexing for Arbitrary Lp Norms. VLDB, Kairo, Egypt.