

CMU SCS

# 15-826: Multimedia Databases and Data Mining

*Multimedia indexing*  
C. Faloutsos

---

---

---

---

---

---

---

---

CMU SCS

## Outline

Goal: 'Find similar / interesting things'

- Intro to DB
- ➔ • Indexing - similarity search
- Data Mining

15-826 Copyright: C. Faloutsos (2007) #2

---

---

---

---

---

---

---

---

CMU SCS

## Indexing - Detailed outline

- primary key indexing
- secondary key / multi-key indexing
- spatial access methods
- fractals
- text
- Singular Value Decomposition (SVD)
- ➔ • multimedia
- ...

15-826 Copyright: C. Faloutsos (2007) #3

---

---

---

---

---

---

---

---

CMU SCS

## Multimedia - Detailed outline

- multimedia
  - ➔ – Motivation / problem definition
  - Main idea / time sequences
  - images
  - sub-pattern matching
  - automatic feature extraction / FastMap

15-826 Copyright: C. Faloutsos (2007) #4

---

---

---

---

---

---

---

---

CMU SCS

## Problem

Given a large collection of (multimedia) records (eg. stocks)  
Allow fast, similarity queries

15-826 Copyright: C. Faloutsos (2007) #5

---

---

---

---

---

---

---

---

CMU SCS

## Applications

- time series: financial, marketing (click-streams!), ECGs, sound;
- images: medicine, digital libraries, education, art
- higher-d signals: scientific db (eg., astrophysics), medicine (MRI scans), entertainment (video)

15-826 Copyright: C. Faloutsos (2007) #6

---

---

---

---

---

---

---

---

CMU SCS

## Sample queries

- find medical cases similar to Smith's
- Find pairs of stocks that move in sync
- Find pairs of documents that are similar (plagiarism?)
- find faces similar to 'Tiger Woods'

15-826 Copyright: C. Faloutsos (2007) #7

---

---

---

---

---

---

---

---

CMU SCS

## Detailed problem defn.:

Problem:

- given a set of multimedia objects,
- find the ones similar to a desirable query object

• for example:

15-826 Copyright: C. Faloutsos (2007) #8

---

---

---

---

---

---

---

---

CMU SCS

Price

1 365 day

Price

1 365 day

Price

1 365 day

distance function: by expert  
(eg, Euclidean distance)

15-826 Copyright: C. Faloutsos (2007) #9

---

---

---

---

---

---

---

---

CMU SCS

## Types of queries

- whole match vs sub-pattern match
- range query vs nearest neighbors
- all-pairs query

15-826 Copyright: C. Faloutsos (2007) #10

---

---

---

---

---

---

---

---

CMU SCS

## Design goals

- Fast (faster than seq. scan)
- ‘correct’ (ie., no false alarms; no false dismissals)

15-826 Copyright: C. Faloutsos (2007) #11

---

---

---

---

---

---

---

---

CMU SCS

## Multimedia - Detailed outline

- multimedia
  - Motivation / problem definition
  - ➔ – Main idea / time sequences
  - images
  - sub-pattern matching
  - automatic feature extraction / FastMap

15-826 Copyright: C. Faloutsos (2007) #12

---

---

---

---

---

---

---

---

CMU SCS

### Main idea

- Eg., time sequences, 'whole matching', range queries, Euclidean distance

Price vs. day (1 to 365) graphs illustrating time sequences.

15-826 Copyright: C. Faloutsos (2007) #13

---

---

---

---

---

---

---

---

CMU SCS

### Main idea

- Seq. scanning works - how to do faster?

15-826 Copyright: C. Faloutsos (2007) #14

---

---

---

---

---

---

---

---

CMU SCS

### Idea: 'GEMINI'

(GENeric Multimedia INDEXing)  
Extract a few numerical features, for a 'quick and dirty' test

15-826 Copyright: C. Faloutsos (2007) #15

---

---

---

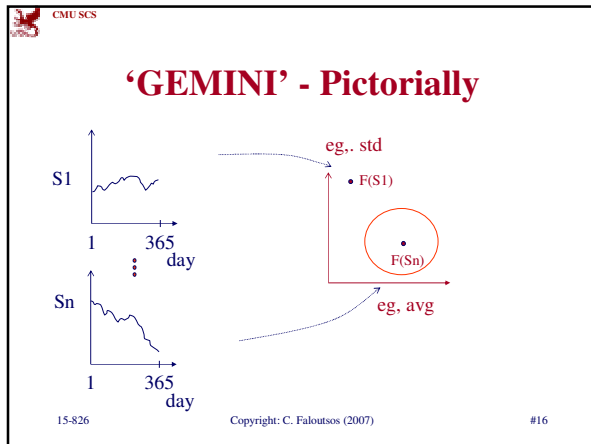
---

---

---

---

---




---

---

---

---

---

---

---

---

**GEMINI**

Solution: 'Quick-and-dirty' filter:

- extract  $n$  features (numbers, eg., avg., etc.)
- map into a point in  $n$ -d feature space
- organize points with off-the-shelf spatial access method ('SAM')
- discard false alarms

15-826 Copyright: C. Faloutsos (2007) #17

---

---

---

---

---

---

---

---

**GEMINI**

Important: Q: how to guarantee no false dismissals?

A1: preserve distances (but: difficult/impossible)

A2: Lower-bounding lemma: if the mapping 'makes things look closer', then there are no false dismissals

15-826 Copyright: C. Faloutsos (2007) #18

---

---

---


---

---

---

---

---

 CMU SCS

## GEMINI

Important:  
Q: how to extract features?  
A: *“if I have only one number to describe my object, what should this be?”*

15-826 Copyright: C. Faloutsos (2007) #19

---

---

---


---

---

---

---

---

 CMU SCS

## Time sequences

Q: what features?

15-826 Copyright: C. Faloutsos (2007) #20

---

---

---


---

---

---

---

---

 CMU SCS

## Time sequences

Q: what features?  
A: Fourier coefficients (we'll see them in detail soon)

15-826 Copyright: C. Faloutsos (2007) #21

---

---

---

---

---

---


---

---

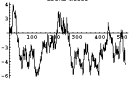
CMU SCS

## Time sequences

white noise brown noise

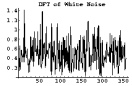


White Noise

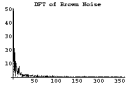


Brown Noise

Fourier spectrum

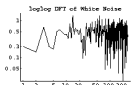


DFT of White Noise

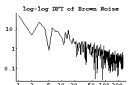


DFT of Brown Noise

... in log-log



log-log DFT of White Noise



log-log DFT of Brown Noise

15-826 #22

---

---

---

---

---

---

---

---

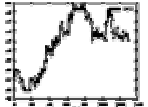
---

---


CMU SCS

## Time sequences

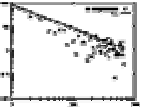
- Eg.:



(a) IBM stock



(b) spectrum  
(linear scale)



(c) spectrum  
(log scale)

15-826 Copyright: C. Faloutsos (2007) #23

---

---

---

---

---

---

---

---

---

---

CMU SCS

## Time sequences

- conclusion: colored noises are well approximated by their first few Fourier coefficients
- colored noises appear in nature:

15-826 Copyright: C. Faloutsos (2007) #24

---

---

---

---

---

---

---

---

---

---



CMU SCS

## Time sequences

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

15-826 Copyright: C. Faloutsos (2007) #25

---

---

---

---

---

---

---

---

CMU SCS

## Time sequences - results

- keep the first 2-3 Fourier coefficients
- faster than seq. scan
- NO false dismissals (see book)

The graph plots 'time' on the y-axis against '# coeff. kept' on the x-axis. Three lines are shown: 'total' (top), 'cleanup-time' (middle), and 'r-tree time' (bottom). All lines show a decreasing trend as the number of coefficients kept increases.

15-826 Copyright: C. Faloutsos (2007) #26

---

---

---

---

---

---

---

---

CMU SCS

## Time sequences - improvements:

- improvements/variations: [Kanellakis+Goldin], [Mendelzon+Rafiei]
- could use Wavelets, or DCT
- could use segment averages [Yi+2000]

15-826 Copyright: C. Faloutsos (2007) #27

---

---

---

---

---

---

---

---

CMU SCS

## Multimedia - Detailed outline

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

15-826 Copyright: C. Faloutsos (2007) #28

---

---

---

---

---

---

---

---

CMU SCS

## Images - color

what is an image?  
A: 2-d array

COLOR IMAGE, eg. 256x256

1-ch pixel:  
(ci, gj, bi)

15-826 Copyright: C. Faloutsos (2007) #29

---

---

---

---

---

---

---

---

CMU SCS

## Images - color

Color histograms, and distance function

15-826 Copyright: C. Faloutsos (2007) #30

---

---

---

---

---

---

---

---

CMU SCS

## Images - color

Mathematically, the distance function is:

$$\text{distance}(\vec{x}, \vec{q}) = (\vec{x} - \vec{q}) \begin{bmatrix} 0.333 & 0.333 & \dots \\ 0.333 & 0.333 & \dots \\ \dots & \dots & \dots \end{bmatrix} (\vec{x} - \vec{q})^T$$

$$\dots = (\vec{x} - \vec{q})^T (\vec{x} - \vec{q})$$

15-826 Copyright: C. Faloutsos (2007) #31

---

---

---

---

---

---

---

---

CMU SCS

## Images - color

Problem: 'cross-talk':

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

• Q: what to do?

• A: feature-extraction question

15-826 Copyright: C. Faloutsos (2007) #32

---

---

---

---

---

---

---

---

CMU SCS

## 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

15-826 Copyright: C. Faloutsos (2007) #33

---

---

---

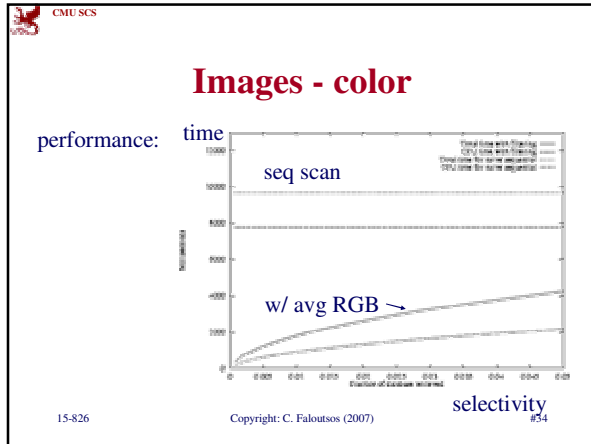
---

---

---

---

---




---

---

---

---

---

---

---

---

- ### Multimedia - Detailed outline
- multimedia
    - Motivation / problem definition
    - Main idea / time sequences
    - ➔ - images (color; shape)
    - sub-pattern matching
    - automatic feature extraction / FastMap
- 15-826 Copyright: C. Faloutsos (2007) #35

---

---

---

---

---

---

---

---

- ### Images - shapes
- distance function: Euclidean, on the area, perimeter, and 20 'moments'
  - (Q: how to normalize them?)
- 15-826 Copyright: C. Faloutsos (2007) #36

---

---

---


---

---

---

---

---

 CMU SCS

### Images - shapes

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

15-826 Copyright: C. Faloutsos (2007) #37

---

---

---


---

---

---

---

---

 CMU SCS

### Images - shapes

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

15-826 Copyright: C. Faloutsos (2007) #38

---

---

---


---

---

---

---

---

 CMU SCS

### Images - shapes

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

15-826 Copyright: C. Faloutsos (2007) #39

---

---

---

---

---

---

---

---

CMU SCS

## Images - shapes

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

15-826 Copyright: C. Faloutsos (2007) #40

---

---

---

---

---

---

---

---

CMU SCS

## Images - shapes

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

15-826 Copyright: C. Faloutsos (2007) #41

---

---

---

---

---

---

---

---

CMU SCS

## Images - shapes

- Performance: ~10x faster

log(# of I/Os)

← all kept

# of features kept

15-826 Copyright: C. Faloutsos (2007) #42

---

---

---

---

---

---

---

---

CMU SCS

## Other shape features?

15-826 Copyright: C. Faloutsos (2007) #43

---

---

---

---

---

---

---


---

CMU SCS

## Other shape features

- Morphology (dilations, erosions, openings, closings) [Korn+, VLDB96]

shape



“structuring element”

R=1 ●

15-826 Copyright: C. Faloutsos (2007) #44

---

---

---

---

---

---

---


---

CMU SCS

## Other shape features

- Morphology (dilations, erosions, openings, closings) [Korn+, VLDB96]

shape



“structuring element”

R=0.5 ●

R=1 ●

R=2 ●

15-826 Copyright: C. Faloutsos (2007) #45

---

---

---

---

---

---

---


---

CMU SCS


### Other shape features


- Morphology (dilations, erosions, openings, closings) [Korn+, VLDB96]


shape



“structuring element”

R=0.5 

R=1 

R=2 

15-826 Copyright: C. Faloutsos (2007) #46

---

---

---

---

---

---

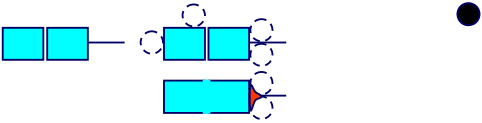
---

---

CMU SCS

### Morphology: closing

- fill in small gaps
- very similar to ‘alpha contours’



15-826 Copyright: C. Faloutsos (2007) #47

---

---

---

---

---

---


---

---

CMU SCS

### Morphology: closing

- fill in small gaps



‘closing’,  
with R=1

15-826 Copyright: C. Faloutsos (2007) #48

---

---

---

---

---

---

---

---



CMU SCS

### Morphology: opening

- ‘closing’, for the complement =
- trim small extremities

15-826 Copyright: C. Faloutsos (2007) #49

---

---

---

---

---

---

---

---

CMU SCS

### Morphology: opening

- ‘closing’, for the complement =
- trim small extremities

‘opening’ with  $R=1$

15-826 Copyright: C. Faloutsos (2007) #50

---

---

---

---

---




---

---

---

CMU SCS

### Morphology

- Closing: fills in gaps 
- Opening: trims extremities 
- All wrt a structuring element: 

15-826 Copyright: C. Faloutsos (2007) #51

---

---

---

---

---

---

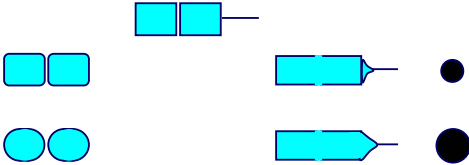
---

---

CMU SCS

## Morphology

- Features: areas of openings ( $R=1, 2, \dots$ ) and closings



15-826 Copyright: C. Faloutsos (2007) #52

---

---

---

---

---

---

---

---

CMU SCS

## Multimedia - Detailed outline

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

15-826 Copyright: C. Faloutsos (2007) #53

---

---

---

---

---

---

---

---

CMU SCS

## Sub-pattern matching

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

15-826 Copyright: C. Faloutsos (2007) #54

---

---

---

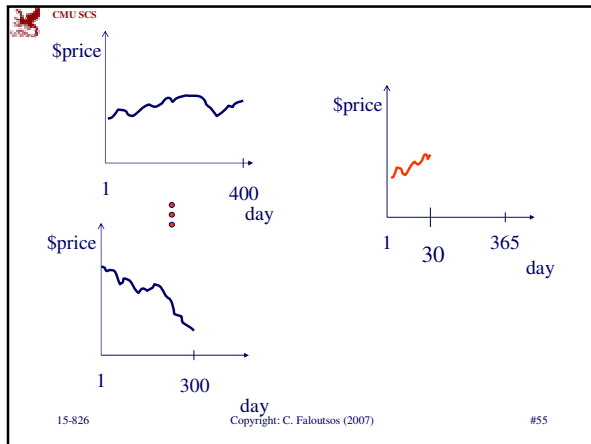
---

---

---

---

---



---

---

---

---

---

---

---

---

**Sub-pattern matching**

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

15-826 Copyright: C. Faloutsos (2007) #56

---

---

---

---

---

---

---

---

**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?)

15-826 Copyright: C. Faloutsos (2007) #57

---

---

---

---

---

---

---

---

CMU SCS

## 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:

15-826 Copyright: C. Faloutsos (2007) #58

---

---

---

---

---

---

---

---

CMU SCS

## Sub-pattern matching

Figure 1: Original data sequence

Figure 2: Sub-pattern (from a window of size  $w$ )

15-826 Copyright: C. Faloutsos (2007) #59

---

---

---

---

---

---

---

---

CMU SCS

## Sub-pattern matching

sequences  $\rightarrow$  trails  $\rightarrow$  MBRs in feature space

15-826 Copyright: C. Faloutsos (2007) #60

---

---

---

---

---

---

---

---

CMU SCS

### Sub-pattern matching

Q: do we store all points? why not?

15-826 Copyright: C. Faloutsos (2007) #61

---

---

---

---

---

---

---

---

CMU SCS

### Sub-pattern matching

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

15-826 Copyright: C. Faloutsos (2007) #62

---

---

---

---

---

---

---

---

CMU SCS

### Sub-pattern matching

(improvement [Moon+2001])

- use non-overlapping windows, for data

15-826 Copyright: C. Faloutsos (2007) #63

---

---

---

---

---

---

---

---

CMU SCS

## 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?)

15-826 Copyright: C. Faloutsos (2007) #64

---

---

---

---

---

---

---

---

CMU SCS

## Multimedia - Detailed outline

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

15-826 Copyright: C. Faloutsos (2007) #65

---

---

---

---

---

---

---

---

CMU SCS

## 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)

15-826 Copyright: C. Faloutsos (2007) #66

---

---

---

---

---

---

---

---

CMU SCS

## 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

15-826 Copyright: C. Faloutsos (2007) #67

---

---

---

---

---

---

---

---

CMU SCS

## FastMap

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

15-826 Copyright: C. Faloutsos (2007) #68

---

---

---

---

---

---

---

---

CMU SCS

## MDS

Multi Dimensional Scaling

15-826 Copyright: C. Faloutsos (2007) #69

---

---

---

---

---

---

---

---

CMU SCS

## Main idea: projections

We want a **linear** algorithm: FastMap  
[SIGMOD95]

15-826 Copyright: C. Faloutsos (2007) #70

---

---

---

---

---

---

---

---

CMU SCS

## FastMap - next iteration

15-826 Copyright: C. Faloutsos (2007) #71

---

---

---

---

---

---

---

---

CMU SCS

## Results

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

15-826 Copyright: C. Faloutsos (2007) #72

---

---

---

---

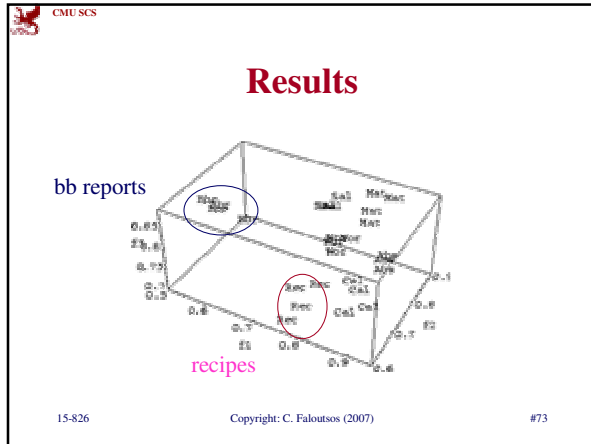
---

---

---

---






---

---

---

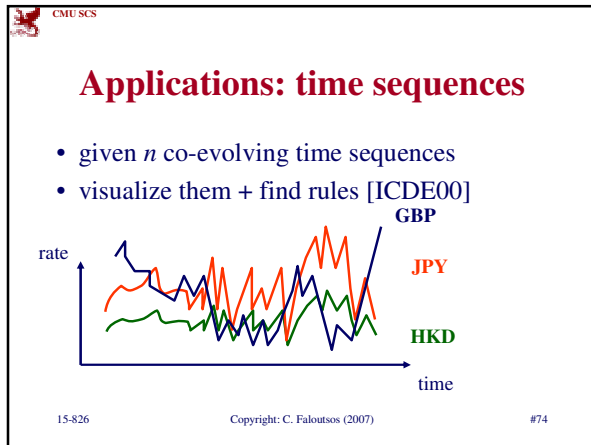
---

---

---

---

---




---

---

---

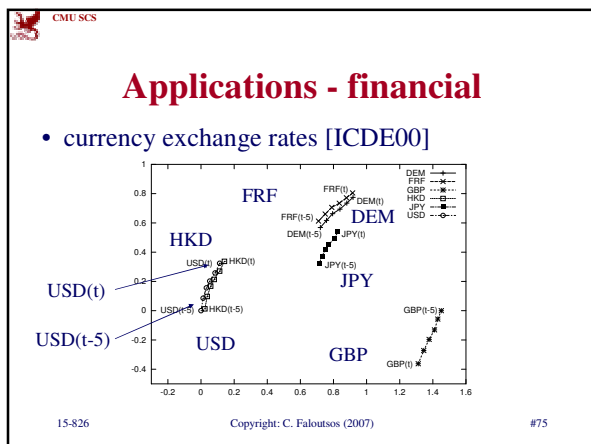
---

---

---

---

---




---

---

---

---

---

---

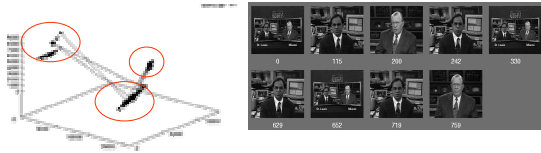
---

---



## Video Trails

[ACM MM97]



15-826

Copyright: C. Faloutsos (2007)

#76

---

---

---

---

---

---

---

---

---

---



## Conclusions

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

15-826

Copyright: C. Faloutsos (2007)

#77

---

---

---

---

---

---

---

---

---

---



## 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.

15-826

Copyright: C. Faloutsos (2007)

#78

---

---

---

---

---

---

---

---

---

---

CMU SCS

## References

- 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.
- Flip Korn, Nikolaos Sidiropoulos, Christos Faloutsos, Eliot Siegel, Zenon Protopapas: *Fast Nearest Neighbor Search in Medical Image Databases*. VLDB 1996: 215-226

15-826 Copyright: C. Faloutsos (2007) #79

---

---

---

---

---

---

---

---

CMU SCS

## 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.

15-826 Copyright: C. Faloutsos (2007) #80

---

---

---

---

---

---

---

---

CMU SCS

## References

- 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, Cairo, Egypt.

15-826 Copyright: C. Faloutsos (2007) #81

---

---

---

---

---

---

---

---