

Carnegie Mellon

# Mining graphs and time series: patterns, anomalies, and fraud detection

## Part 2: Time Series

### Indexing, Fourier, Wavelets

Christos Faloutsos

CMU SCS

<https://www.cs.cmu.edu/~christos/TALKS/19-Gol>



1

---

---

---

---

---

---


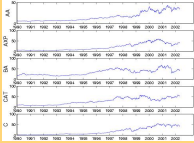
---

---

Carnegie Mellon

## Problem:

Q: mine/forecast (one, or more) time sequences

Gov. of India Copyright (C) 2019 C. Faloutsos

2

---

---

---

---

---


---



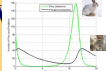
---

---

Carnegie Mellon

## Answers



- Similarity search: **Euclidean**/time-warping; **feature extraction** and **SAMs**
- Periodicities**: **DFT/DWT**
- Linear Forecasting: **AR** (Box-Jenkins)
- Non-linear forecasting: **lag-plots**
- Gray-box model:  **Lotka-V**  

Gov. of India Copyright (C) 2019 C. Faloutsos

3

---

---

---

---

---


---

---

---

Carnegie Mellon

## Outline



- Introduction
- Part#1: Graphs and Tensors
- ➔ • Part#2: Time series
- Part#3: extras (visualization, etc)
- Conclusions

Gov. of India
Copyright (C) 2019 C. Faloutsos
4

---

---

---

---

---

---


---

---

4

Carnegie Mellon

## Detailed Outline



- ➔ • Motivation
- Similarity Search and Indexing
- DSP (Digital Signal Processing)
- Linear Forecasting
- Non-linear forecasting
- Tensors
- Conclusions

Gov. of India
Copyright (C) 2019 C. Faloutsos
5

---

---

---

---

---

---

---

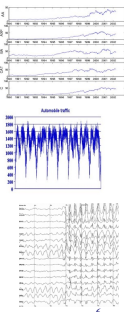
---

5

Carnegie Mellon

## Problem definition

- Given: one or more sequences  
 $x_1, x_2, \dots, x_t, \dots$   
 $(y_1, y_2, \dots, y_b, \dots)$   
 $\dots$
- Find
  - **Forecast**; similar sequences
  - patterns; clusters; outliers



Gov. of India
Copyright (C) 2019 C. Faloutsos
6

---

---

---

---

---

---

---

---

6

## Motivation - Applications

- Financial, sales, economic series
- Medical
  - reactions to new drugs
  - elderly care

7

ECG - [physionet.org](https://physionet.org)



8

## EEG - epilepsy



9

Carnegie Mellon

# Motivation - Applications (cont'd)

- ‘Smart house’
  - sensors monitor temperature, humidity, air quality
- video surveillance

Gov. of India
Copyright (C) 2019 C. Faloutsos
10

---

---

---

---

---

---

---


---

10

Carnegie Mellon

# Motivation - Applications (cont'd)

- civil/automobile infrastructure
  - bridge vibrations [Oppenheim+02]
  - road conditions / traffic monitoring



Gov. of India
Copyright (C) 2019 C. Faloutsos
11

---

---

---

---

---

---

---

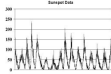
---

11

Carnegie Mellon

# Motivation - Applications (cont'd)

- Weather, environment/anti-pollution
  - volcano monitoring
  - air/water pollutant monitoring



Gov. of India
Copyright (C) 2019 C. Faloutsos
12

---

---

---

---

---

---

---

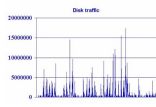
---

12

Carnegie Mellon

## Motivation - Applications (cont'd)

- Computer systems
  - ‘Active Disks’ (buffering, prefetching)
  - web servers (ditto)
  - network traffic monitoring
  - ...



Gov. of India

Copyright (C) 2019 C. Faloutsos

13

---

---

---

---

---

---

---

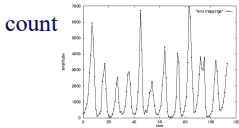
---

13

Carnegie Mellon

## Problem #1:

Goal: given a signal (eg., #packets over time)  
Find: patterns, periodicities, and/or compress



lynx caught per year  
(packets per day;  
temperature per day)

year

Gov. of India

Copyright (C) 2019 C. Faloutsos

14

---

---

---

---

---

---

---

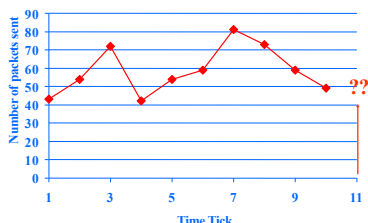
---

14

Carnegie Mellon

## Problem#2: Forecast

Given  $x_t, x_{t-1}, \dots$ , forecast  $x_{t+1}$



Gov. of India

Copyright (C) 2019 C. Faloutsos

15

---

---

---

---

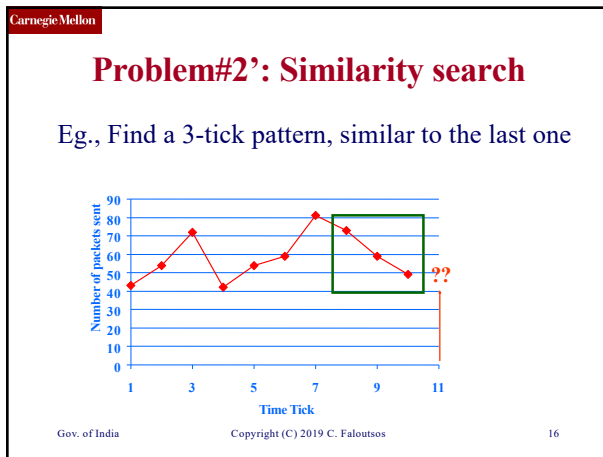
---

---

---

---

15



16

---

---

---

---

---

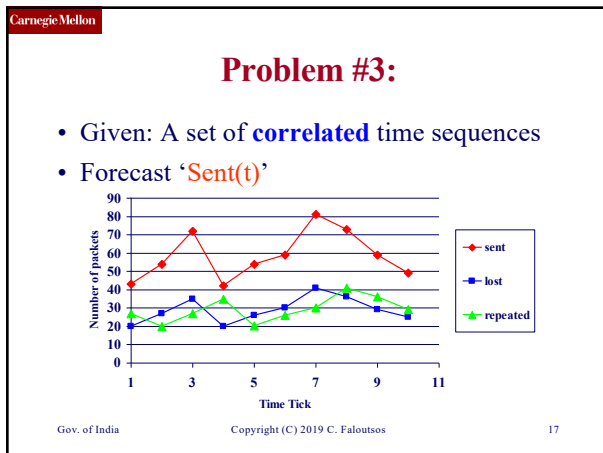
---

---

---

---

---



17

---

---

---

---

---

---

---

---

---

---

Carnegie Mellon

### Important observations

Patterns, rules, forecasting and similarity indexing are closely related:

- To do **forecasting**, we need
  - to find **patterns/rules**
  - compress
  - to find **similar** settings in the past
- to find outliers, we need to have forecasts
  - (outlier = too far away from our forecast)

Gov. of India Copyright (C) 2019 C. Faloutsos 18

18

---

---

---

---

---

---

---


---

---

---

Carnegie Mellon

# Outline



- Motivation
- Similarity Search and Indexing
- DSP
  - Linear Forecasting
  - Non-linear forecasting
  - Tensors
  - Conclusions

Gov. of India
Copyright (C) 2019 C. Faloutsos
19

---

---

---

---

---

---


---

---

19

Carnegie Mellon

# Outline



- Motivation
- Similarity Search and Indexing
  - distance functions: Euclidean; Time-warping
  - indexing
  - feature extraction
- DSP
- ...

Gov. of India
Copyright (C) 2019 C. Faloutsos
20

---

---

---

---

---

---


---

---

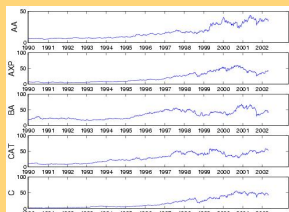
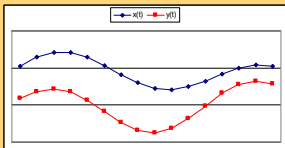
20

Carnegie Mellon

# Problem:



Q: How similar are two sequences?

Gov. of India
Copyright (C) 2019 C. Faloutsos
21

---

---

---

---

---

---

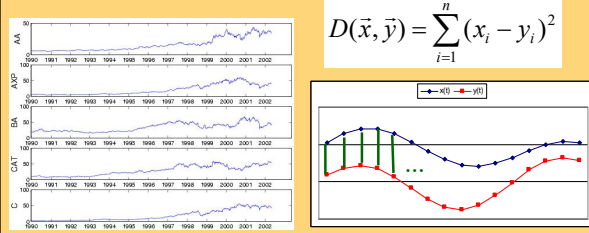
---

---

21

**Answer:**

Q: How similar are two sequences?  
A: Euclidean distance (<-> cosine similarity)



$$D(\vec{x}, \vec{y}) = \sum_{i=1}^n (x_i - y_i)^2$$

Gov. of India Copyright (C) 2019 C. Faloutsos 22

22

---

---

---

---

---

---

---

---

**Importance of distance functions**

Subtle, but **absolutely necessary**:

- A ‘must’ for similarity indexing (-> forecasting)
- A ‘must’ for clustering

Two major families

- Euclidean and Lp norms
- Time warping and variations

Gov. of India Copyright (C) 2019 C. Faloutsos 23

23

---

---

---

---

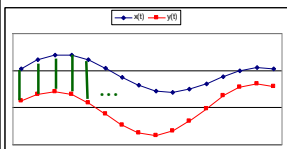
---

---

---

---

**Euclidean and Lp**



$$D(\vec{x}, \vec{y}) = \sum_{i=1}^n (x_i - y_i)^2$$

$$L_p(\vec{x}, \vec{y}) = \sum_{i=1}^n |x_i - y_i|^p$$

- $L_1$ : city-block = Manhattan
- $L_2$  = Euclidean
- $L_{\infty}$

Gov. of India Copyright (C) 2019 C. Faloutsos 24

24

---

---

---

---

---

---

---

---



Carnegie Mellon

## Observation #1

- Time sequence  $\rightarrow$  n-d vector

Gov. of India Copyright (C) 2019 C. Faloutsos 25

25

---

---

---

---

---

---

---

---

Carnegie Mellon

## Observation #2

Euclidean distance is closely related to

- cosine similarity
- dot product
- ‘cross-correlation’ function

Gov. of India Copyright (C) 2019 C. Faloutsos 26

26

---

---

---

---

---

---

---

---

Carnegie Mellon

## Time Warping

- allow accelerations - decelerations
  - (with or w/o penalty)
- THEN compute the (Euclidean) distance (+ penalty)
- related to the string-editing distance

Gov. of India Copyright (C) 2019 C. Faloutsos 27

27

---

---

---

---

---

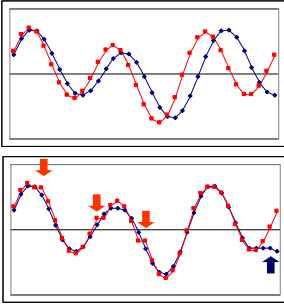
---

---

---

Carnegie Mellon

## Time Warping



Gov. of India Copyright (C) 2019 C. Faloutsos 28

28

---

---

---

---

---

---

---

---

Carnegie Mellon

## Other Distance functions

- piece-wise linear/flat approx.; compare pieces [Keogh+01] [Faloutsos+97]
- ‘cepstrum’ (for voice [Rabiner+Juang])
  - do DFT; take log of amplitude; do DFT again!
- Allow for small gaps [Agrawal+95]

Gov. of India Copyright (C) 2019 C. Faloutsos 29

29

---

---

---

---

---

---

---

---

Carnegie Mellon

## More distance functions.

- Chen + Ng [vlldb’04]: ERP ‘Edit distance with Real Penalty’: give a penalty to stutters
- Keogh+ [kdd’04]: VERY NICE, based on information theory: compress each sequence (quantize + Lempel-Ziv), using the **other** sequences’ LZ tables

*On The Marriage of Lp-norms and Edit Distance, [Lei Chen, Raymond T. Ng](#), VLDB’04*

*Towards Parameter-Free Data Mining, E. Keogh, S. Lonardi, C.A. Ratanamahatana, KDD’04*

30

---

---

---

---

---

---

---

---

Carnegie Mellon

## Conclusions

Prevailing distances:

- Euclidean and
- time-warping

Gov. of India

Copyright (C) 2019 C. Faloutsos

31

---

---

---

---

---

---

---

---

31

Carnegie Mellon

## Outline

- Motivation
- Similarity Search and Indexing
  - distance functions
  - indexing
- ➡ • feature extraction
- DSP
- ...

Gov. of India

Copyright (C) 2019 C. Faloutsos

32

---

---

---

---

---

---

---

---

32

Carnegie Mellon

## Important observations

Patterns, rules, forecasting and similarity indexing are closely related:

- To do forecasting, we need
  - to find patterns/rules
  - compress
  - **to find similar settings in the past**
- to find outliers, we need to have forecasts
  - (outlier = too far away from our forecast)

Gov. of India

Copyright (C) 2019 C. Faloutsos

33

---

---

---

---

---

---


---

---

33

Carnegie Mellon

# Outline



- Motivation
- Similarity Search and Indexing
  - distance functions
  - indexing
  - ➡ – feature extraction
- DSP
- ...

Gov. of India
Copyright (C) 2019 C. Faloutsos
34

34

---

---

---

---

---


---

---

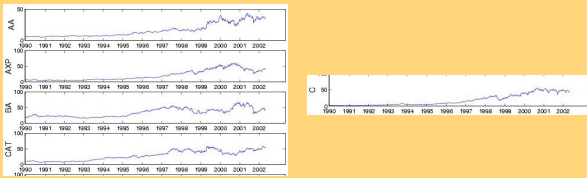
---

Carnegie Mellon

# Problem:



Q: find quickly stocks like 'C' (or customers like 'smith')



Gov. of India
Copyright (C) 2019 C. Faloutsos
35

35

---

---

---

---

---


---

---

---

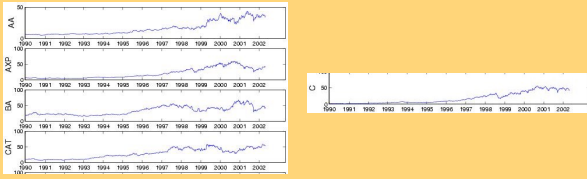
Carnegie Mellon

# Answer:



Q: find quickly stocks like 'C' (or customers like 'smith')

A: summarize seq. to a few numbers/features (eg., avg, stdv, Fourier coeff.)



Gov. of India
Copyright (C) 2019 C. Faloutsos
36

36

---

---

---

---

---

---

---

---

Carnegie Mellon

Indexing

Problem:

- given a set of time sequences,
- find the ones similar to a desirable query sequence

Gov. of India

Copyright (C) 2019 C. Faloutsos

37

---

---

---

---

---

---

---

---

37

Carnegie Mellon

distance function: by expert

Gov. of India

Copyright (C) 2019 C. Faloutsos

38

---

---

---

---

---

---

---

---

38

Carnegie Mellon

Idea: 'GEMINI'

Eg., 'find stocks similar to MSFT'

Seq. scanning: too slow

How to accelerate the search?

[Faloutsos96]

Gov. of India

Copyright (C) 2019 C. Faloutsos

39

---

---

---

---

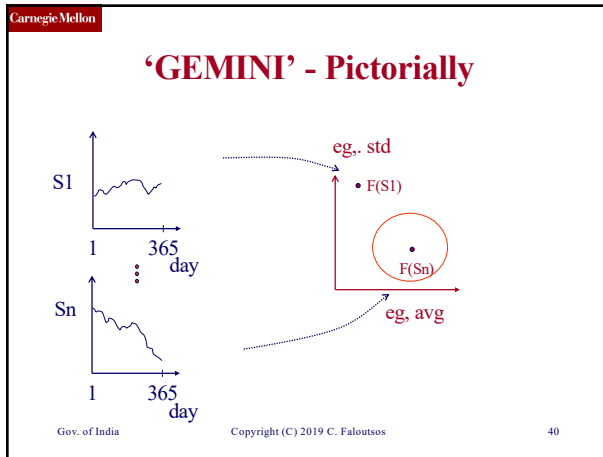
---

---

---

---

39



40

---

---

---

---

---

---

---

---

Carnegie Mellon

### 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

Gov. of India Copyright (C) 2019 C. Faloutsos 41

41

---

---

---

---

---

---

---

---

Carnegie Mellon

### Examples of GEMINI

- Time sequences: DFT (up to 100 times faster) [SIGMOD94];
- [Kanellakis+], [Mendelzon+]

Gov. of India Copyright (C) 2019 C. Faloutsos 42

42

---

---

---

---

---

---

---

---

Carnegie Mellon

Conclusions

- Fast indexing: through GEMINI
  - feature extraction and
  - (off the shelf) Spatial Access Methods [Gaede+98]

Gov. of India

Copyright (C) 2019 C. Faloutsos

43

43

---

---

---

---

---


---

---

---

Carnegie Mellon

Outline



- Motivation
- Similarity Search and Indexing
  - distance functions
  - indexing
  - feature extraction
- ➔ • DSP
- ...

Gov. of India

Copyright (C) 2019 C. Faloutsos

44

44

---

---

---

---

---


---

---

---

Carnegie Mellon

Outline



- Motivation
- Similarity Search and Indexing
  - distance functions
  - indexing
  - feature extraction
    - DFT, DWT, DCT (data independent)
    - SVD, etc (data dependent)

Gov. of India

Copyright (C) 2019 C. Faloutsos

45

45

---

---

---

---


---

---


---

---

Carnegie Mellon






## Important observations



Patterns, rules, forecasting and similarity indexing are closely related:

- To do forecasting, we need
  - to find patterns/rules
  - compress**
  - to find similar settings in the past
- to find outliers, we need to have forecasts
  - (outlier = too far away from our forecast)

Gov. of India
Copyright (C) 2019 C. Faloutsos
46

---

---

---

---

---

---


---

---


46

Carnegie Mellon

## Outline



- Motivation
- Similarity Search and Indexing
  - distance functions
  - indexing
  - feature extraction
    - DFT, DWT, DCT (data independent)
    - SVD etc (data dependent)



Gov. of India
Copyright (C) 2019 C. Faloutsos
47

---

---

---

---

---

---


---

---

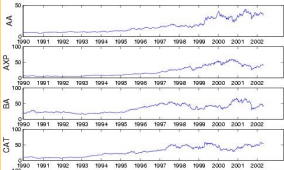
47

Carnegie Mellon

## Problem:



Q: how to extract features (commonalities)? (given the data)



Gov. of India
Copyright (C) 2019 C. Faloutsos
48

---

---

---

---

---

---

---

---

48



Carnegie Mellon

Answer

Q: how to extract features (commonalities)? (given the data)  
A: SVD, ICA

AA

1990 1991 1992 1993 1994 1995 1996 1997 1998 1999 2000 2001 2002

AXP

1990 1991 1992 1993 1994 1995 1996 1997 1998 1999 2000 2001 2002

BA

1990 1991 1992 1993 1994 1995 1996 1997 1998 1999 2000 2001 2002

CAT

1990 1991 1992 1993 1994 1995 1996 1997 1998 1999 2000 2001 2002

C

1990 1991 1992 1993 1994 1995 1996 1997 1998 1999 2000 2001 2002

Time (Yr)

Time (Yr)

Gov. of India

Copyright (C) 2019 C. Faloutsos

49

49

---

---

---

---

---

---

---

---

Carnegie Mellon

SVD

• THE optimal method for dimensionality reduction

– (under the Euclidean metric)

Gov. of India

Copyright (C) 2019 C. Faloutsos

50

50

---

---

---

---

---

---

---

---

Carnegie Mellon

Motivation:  
Find hidden variables

Alcoa

American Express

Boeing

Caterpillar

Citi Group

Dow Jones Industrial Average

Find common hidden variables, and weights.

AA

1990 1991 1992 1993 1994 1995 1996 1997 1998 1999 2000 2001 2002

AXP

1990 1991 1992 1993 1994 1995 1996 1997 1998 1999 2000 2001 2002

BA

1990 1991 1992 1993 1994 1995 1996 1997 1998 1999 2000 2001 2002

CAT

1990 1991 1992 1993 1994 1995 1996 1997 1998 1999 2000 2001 2002

C

1990 1991 1992 1993 1994 1995 1996 1997 1998 1999 2000 2001 2002

Time (Yr)

Gov. of India

Copyright (C) 2019 C. Faloutsos

51

51

---

---

---

---

---

---

---

---

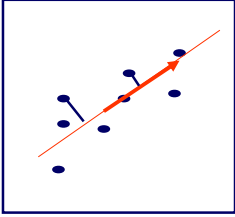
17

Carnegie Mellon

## Singular Value Decomposition (SVD)

- SVD ( $\sim$ LSI  $\sim$  KL  $\sim$  PCA  $\sim$  spectral analysis...)

day2



day1

LSI: S. Dumais; M. Berry  
 KL: eg, Duda+Hart  
 PCA: eg, Jolliffe  
 Details: [Press+], [Faloutsos96]

Gov. of India

Copyright (C) 2019 C. Faloutsos

52

---

---

---

---

---

---

---

---

52

Carnegie Mellon

## SVD

- Extremely** useful tool
  - (also behind PageRank/google and Kleinberg's algorithm for hubs and authorities)

Gov. of India

Copyright (C) 2019 C. Faloutsos

53

---

---

---

---

---

---

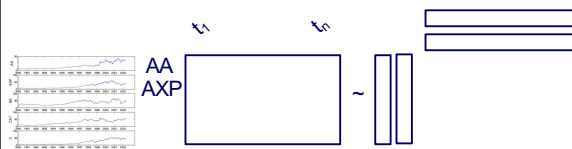
---

---

53

Carnegie Mellon

## SVD



AA  
AXP

~

U   Σ   V<sup>T</sup>

```
# svd code
import numpy as np
u, s, vh = np.linalg.svd(b)
```

Gov. of India

Copyright (C) 2019 C. Faloutsos

54

---

---

---

---

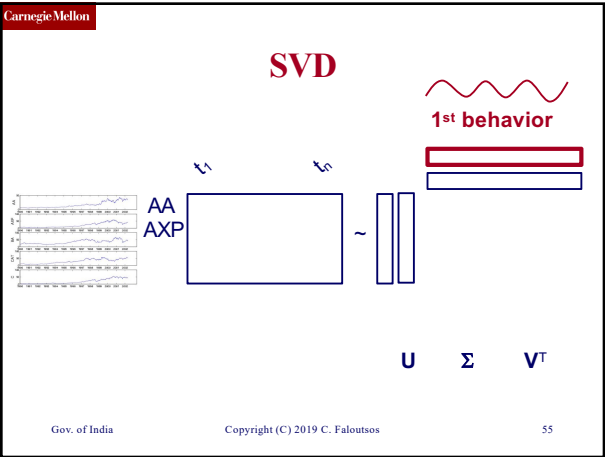
---

---

---

---

54



55

---

---

---

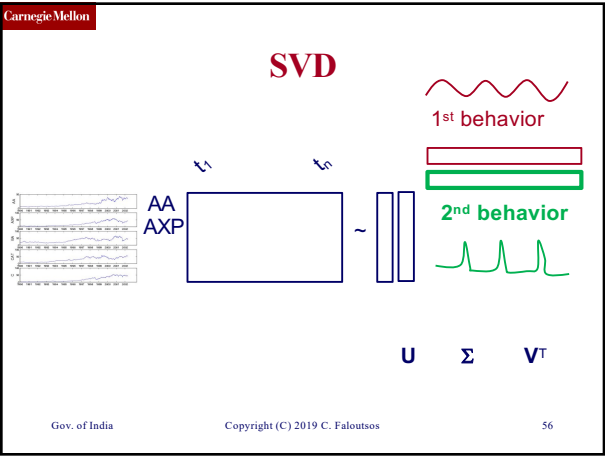
---

---

---

---

---



56

---

---

---

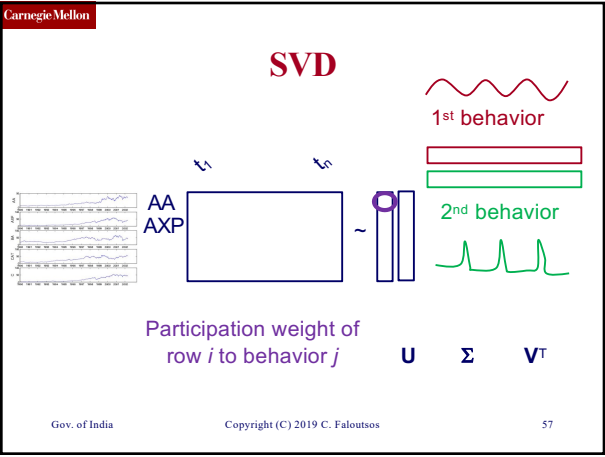
---

---

---

---

---



57

---

---

---

---

---

---

---

---

Carnegie Mellon

Outline

- Motivation
- Similarity Search and Indexing
  - distance functions
  - indexing
  - feature extraction
    - DFT, DWT, DCT (data independent)
    - SVD etc (data dependent), ICA

Gov. of India

Copyright (C) 2019 C. Faloutsos

58

---

---

---

---

---

---

---

---

58

Carnegie Mellon

Citation

- *AutoSplit: Fast and Scalable Discovery of Hidden Variables in Stream and Multimedia Databases*, **Jia-Yu Pan**, Hiroyuki Kitagawa, Christos Faloutsos and Masafumi Hamamoto

PAKDD 2004, Sydney, Australia

Gov. of India

Copyright (C) 2019 C. Faloutsos

59

---

---

---

---

---

---

---

---

59

Carnegie Mellon

ICA = BSS

- Independent Component Analysis =
- Blind Source Separation =
- ‘cocktail party problem’





Gov. of India

Copyright (C) 2019 C. Faloutsos

60

---

---

---

---

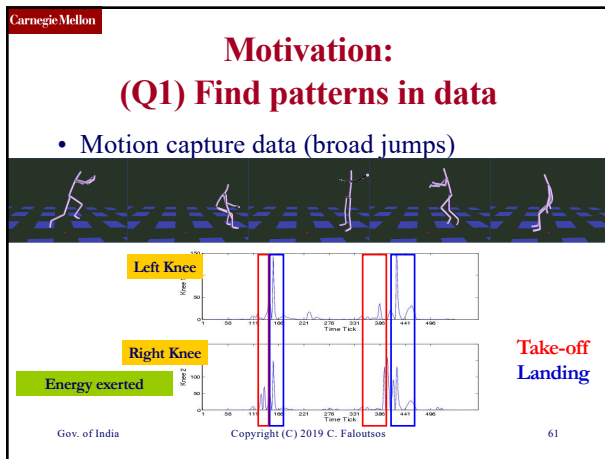
---

---

---

---

60



61

---

---

---

---

---

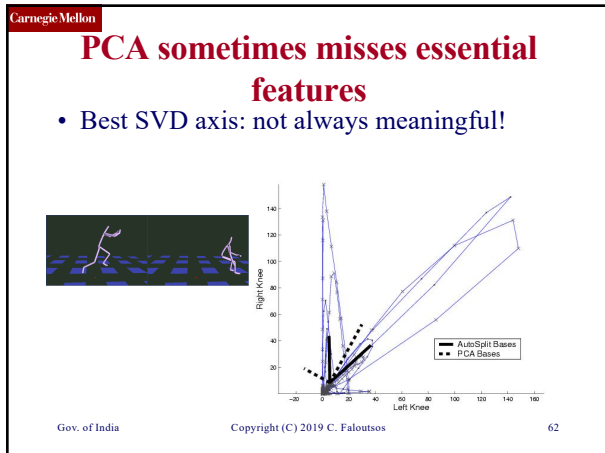
---

---

---

---

---



62

---

---

---

---

---

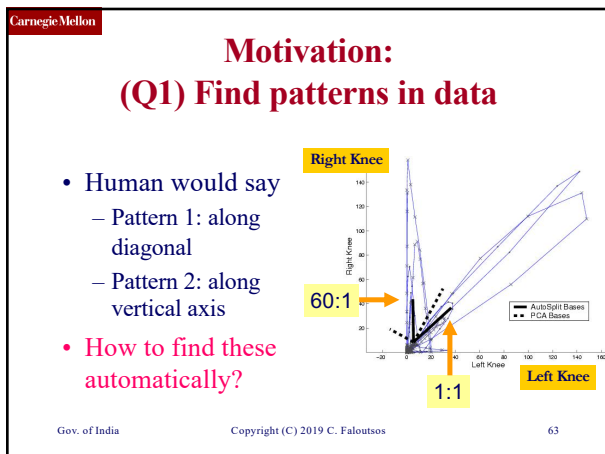
---

---

---

---

---



63

---

---

---

---

---

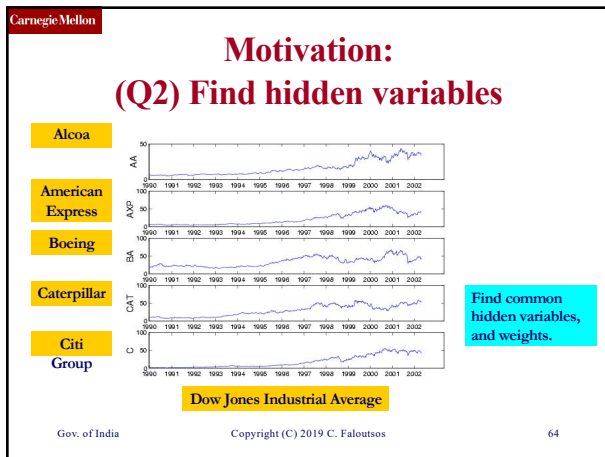
---

---

---

---

---



64

---

---

---

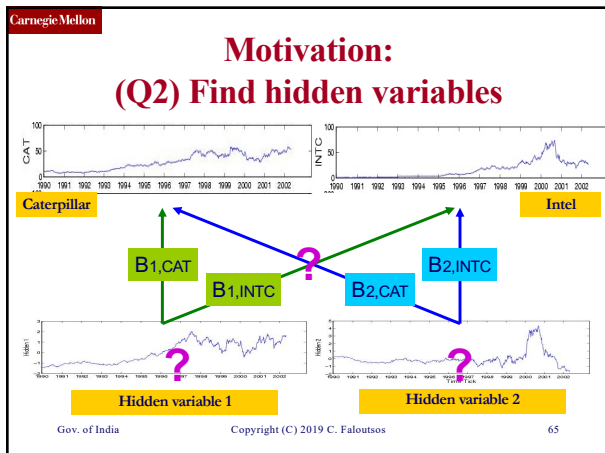
---

---

---

---

---



65

---

---

---

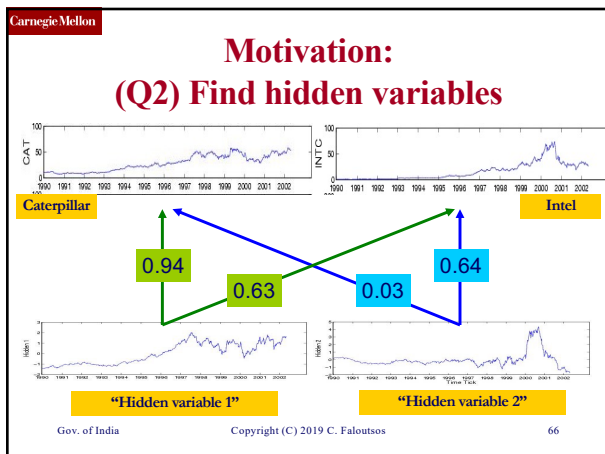
---

---

---

---

---



66

---

---

---

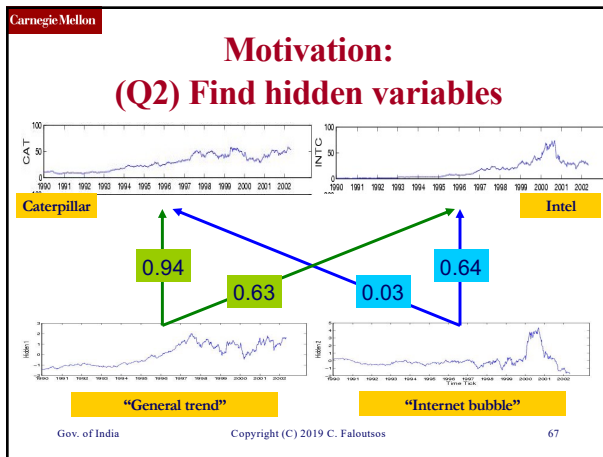
---

---

---

---

---



67

---

---

---

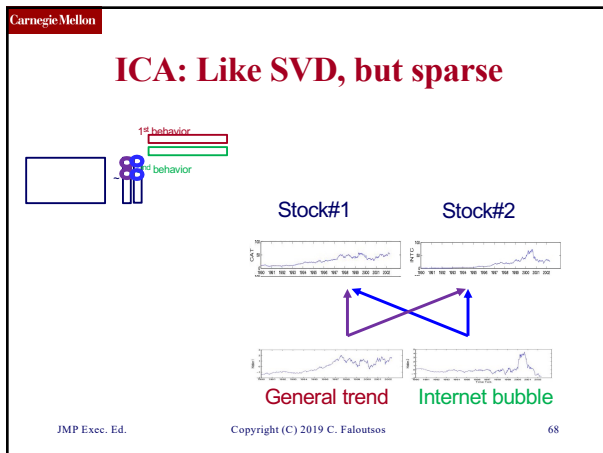
---

---

---

---

---



68

---

---

---

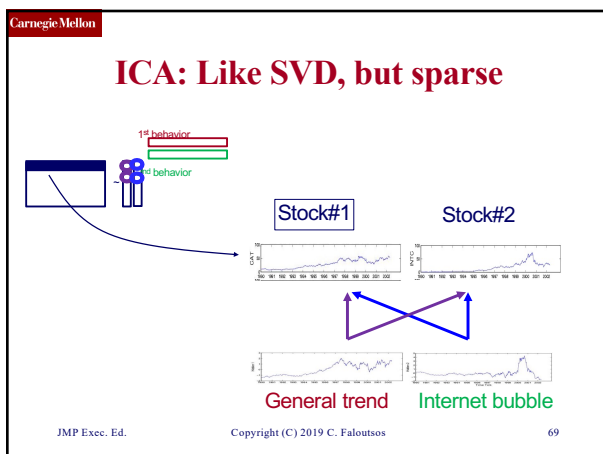
---

---

---

---

---



69

---

---

---

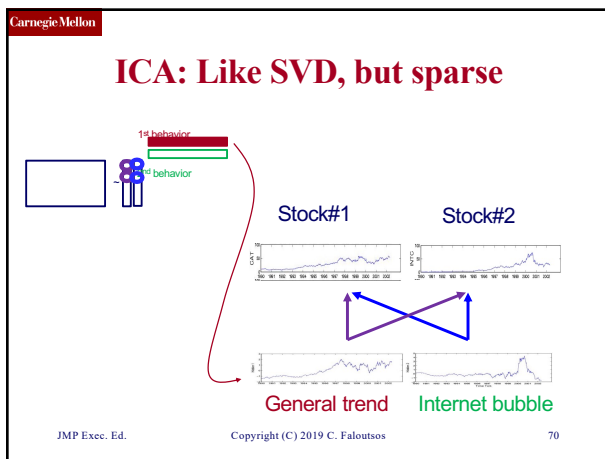
---

---

---

---

---



70

---

---

---

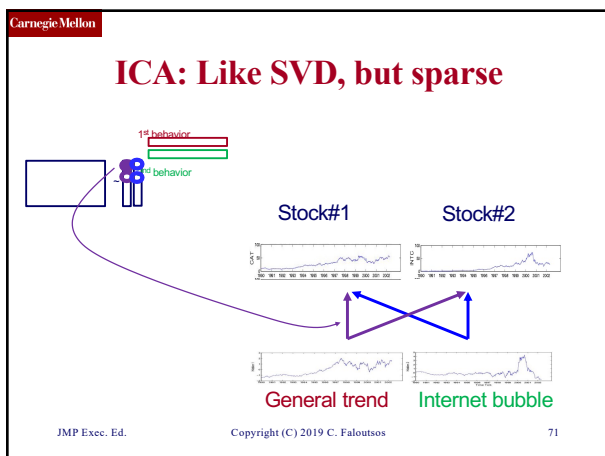
---

---

---

---

---



71

---

---

---

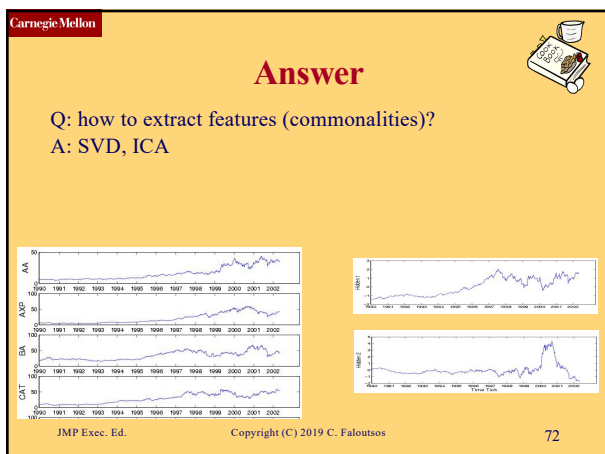
---

---

---

---

---



72

---

---

---

---

---

---

---

---



Carnegie Mellon

Books

- William H. Press, Saul A. Teukolsky, William T. Vetterling and Brian P. Flannery: *Numerical Recipes in C*, Cambridge University Press, 1992, 2nd Edition. (Great description, intuition and code for SVD)
- C. Faloutsos: *Searching Multimedia Databases by Content*, Kluwer Academic Press, 1996 (introduction to SVD, and GEMINI)

Gov. of IndiaCopyright (C) 2019 C. Faloutsos73

73

---

---

---

---

---

---

---

---

Carnegie Mellon

References

- Agrawal, R., K.-I. Lin, et al. (Sept. 1995). Fast Similarity Search in the Presence of Noise, Scaling and Translation in Time-Series Databases. Proc. of VLDB, Zurich, Switzerland.
- Babu, S. and J. Widom (2001). "Continuous Queries over Data Streams." SIGMOD Record 30(3): 109-120.
- Breunig, M. M., H.-P. Kriegel, et al. (2000). LOF: Identifying Density-Based Local Outliers. SIGMOD Conference, Dallas, TX.
- Berry, Michael: <http://www.cs.utk.edu/~lsi/>

Gov. of IndiaCopyright (C) 2019 C. Faloutsos74

74

---

---

---

---

---

---

---

---

Carnegie Mellon

References

- Ciaccia, P., M. Patella, et al. (1997). M-tree: An Efficient Access Method for Similarity Search in Metric Spaces. VLDB.
- Foltz, P. W. and S. T. Dumais (Dec. 1992). "Personalized Information Delivery: An Analysis of Information Filtering Methods." Comm. of ACM (CACM) 35(12): 51-60.
- Guttman, A. (June 1984). R-Trees: A Dynamic Index Structure for Spatial Searching. Proc. ACM SIGMOD, Boston, Mass.

Gov. of IndiaCopyright (C) 2019 C. Faloutsos75

75

---

---

---

---

---

---

---

---

Carnegie Mellon

References

- Gaede, V. and O. Guenther (1998). “Multidimensional Access Methods.” Computing Surveys 30(2): 170-231.
- Gehrke, J. E., F. Korn, et al. (May 2001). On Computing Correlated Aggregates Over Continual Data Streams. ACM Sigmod, Santa Barbara, California.

Gov. of India      Copyright (C) 2019 C. Faloutsos      76

76

---

---

---

---

---

---

---

---

Carnegie Mellon

References

- Gunopulos, D. and G. Das (2001). Time Series Similarity Measures and Time Series Indexing. SIGMOD Conference, Santa Barbara, CA.
- Eamonn J. Keogh, [Themis Palpanas](#), [Victor B. Zordan](#), [Dimitrios Gunopulos](#), [Marc Cardle](#): Indexing Large Human-Motion Databases. [VLDB 2004](#): 780-791

Gov. of India      Copyright (C) 2019 C. Faloutsos      Part2.1 #77

77

---

---

---

---

---

---

---

---

Carnegie Mellon

References

- Hatonen, K., M. Klemettinen, et al. (1996). Knowledge Discovery from Telecommunication Network Alarm Databases. ICDE, New Orleans, Louisiana.
- Jolliffe, I. T. (1986). Principal Component Analysis, Springer Verlag.

Gov. of India      Copyright (C) 2019 C. Faloutsos      78

78

---

---

---

---

---

---

---

---

Carnegie Mellon

References

- Keogh, E. J., K. Chakrabarti, et al. (2001). Locally Adaptive Dimensionality Reduction for Indexing Large Time Series Databases. SIGMOD Conference, Santa Barbara, CA.
- Kobla, V., D. S. Doermann, et al. (Nov. 1997). VideoTrails: Representing and Visualizing Structure in Video Sequences. ACM Multimedia 97, Seattle, WA.

Gov. of India

Copyright (C) 2019 C. Faloutsos

79

---

---

---

---

---

---

---

---

79

Carnegie Mellon

References

- Oppenheim, I. J., A. Jain, et al. (March 2002). A MEMS Ultrasonic Transducer for Resident Monitoring of Steel Structures. SPIE Smart Structures Conference SS05, San Diego.
- Papadimitriou, C. H., P. Raghavan, et al. (1998). Latent Semantic Indexing: A Probabilistic Analysis. PODS, Seattle, WA.
- Rabiner, L. and B.-H. Juang (1993). Fundamentals of Speech Recognition, Prentice Hall.

Gov. of India

Copyright (C) 2019 C. Faloutsos

80

---

---

---

---

---

---

---

---

80

Carnegie Mellon

References

- Traina, C., A. Traina, et al. (October 2000). Fast feature selection using the fractal dimension,. XV Brazilian Symposium on Databases (SBB D), Paraiba, Brazil.

Gov. of India

Copyright (C) 2019 C. Faloutsos

81

---

---

---

---

---

---

---

---

81

Carnegie Mellon

References

- Dennis Shasha and Yunyue Zhu *High Performance Discovery in Time Series: Techniques and Case Studies* Springer 2004
- Yunyue Zhu, Dennis Shasha ``StatStream: Statistical Monitoring of Thousands of Data Streams in Real Time'' VLDB, August, 2002. pp. 358-369.
- Samuel R. Madden, Michael J. Franklin, Joseph M. Hellerstein, and Wei Hong. *The Design of an Acquisitional Query Processor for Sensor Networks*. SIGMOD, June 2003, San Diego, CA.

Gov. of India

Copyright (C) 2019 C. Faloutsos

82

82

---

---

---

---

---

---

---

---

Carnegie Mellon

References

- Lawrence Saul & Sam Roweis. *An Introduction to Locally Linear Embedding* (draft)
- Sam Roweis & Lawrence Saul. *Nonlinear dimensionality reduction by locally linear embedding*. Science, v.290 [no.5500](#) , Dec.22, 2000. pp.2323--2326.
- B. Shaw and T. Jebara. "*Minimum Volume Embedding*". Artificial Intelligence and Statistics, AISTATS, March 2007.

Gov. of India

Copyright (C) 2019 C. Faloutsos

#83

83

---

---

---

---

---

---

---

---

Carnegie Mellon

References

- Josh Tenenbaum, Vin de Silva and John Langford. *A Global Geometric Framework for Nonlinear dimensionality Reduction*. Science 290, pp. 2319-2323, 2000.

Gov. of India

Copyright (C) 2019 C. Faloutsos

#84

84

---

---

---

---

---

---

---

---

Carnegie Mellon

# Part 2.2: DSP (Digital Signal Processing)

Gov. of India

Copyright (C) 2019 C. Faloutsos

85

85

---

---

---

---

---


---

---

---

Carnegie Mellon

## Outline



- Motivation
- Similarity Search and Indexing
- ➔ • DSP (DFT, DWT)
- Linear Forecasting
- Non-linear forecasting
- Tensors
- Conclusions

Gov. of India

Copyright (C) 2019 C. Faloutsos

86

86

---

---

---

---

---


---

---

---

Carnegie Mellon

## Outline



- ➔ • DFT
  - Definition of DFT and properties
  - how to read the DFT spectrum
- DWT
  - Definition of DWT and properties
  - how to read the DWT scalogram

Gov. of India

Copyright (C) 2019 C. Faloutsos

87

87

---

---

---

---


---

---

---

---




Carnegie Mellon



## Important observations

Patterns, rules, forecasting and similarity indexing are closely related:

- To do forecasting, we need
  - to find patterns/rules
  - compress**
  - to find similar settings in the past
- to find outliers, we need to have forecasts
  - (outlier = too far away from our forecast)

Gov. of India Copyright (C) 2019 C. Faloutsos 88

88

---

---

---

---

---

---

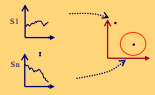
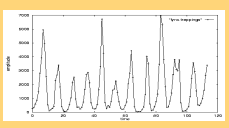

---

---

Carnegie Mellon

## Problem

Q: How to summarize / extract few features

JMP Exec. Ed. Copyright (C) 2019 C. Faloutsos 89

89

---

---

---

---

---

---

---

---

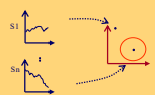
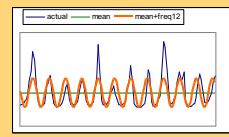

Carnegie Mellon

## Answer:

Q: How to summarize / extract few features

A1: Data dep.: SVD, ICA

A2: Data indep.: Fourier; Wavelets

JMP Exec. Ed. Copyright (C) 2019 C. Faloutsos 90

90

---

---

---

---

---

---

---

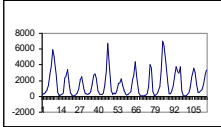
---

Carnegie Mellon


## Introduction - Problem#1

Goal: given a signal (eg., packets over time)  
Find: patterns and/or compress

count



lynx caught per year  
(packets per day;  
automobiles per hour)



Gov. of India Copyright (C) 2019 C. Faloutsos 91

91

---

---

---

---

---

---

---

---

Carnegie Mellon

## What does DFT do?

A: highlights the periodicities

Gov. of India Copyright (C) 2019 C. Faloutsos 92

92

---

---

---

---

---

---

---

---

Carnegie Mellon

## DFT: definition

Skip

- For a sequence  $x_0, x_1, \dots, x_{n-1}$
- the (**n-point**) Discrete Fourier Transform is
- $X_0, X_1, \dots, X_{n-1}$ :

$$X_f = 1/\sqrt{n} \sum_{t=0}^{n-1} x_t * \exp(-j2\pi tf/n) \quad f = 0, \dots, n-1$$

( $j = \sqrt{-1}$ )

$$x_t = 1/\sqrt{n} \sum_{f=0}^{n-1} X_f * \exp(+j2\pi tf/n)$$

inverse DFT

Gov. of India Copyright (C) 2019 C. Faloutsos 93

93

---

---

---

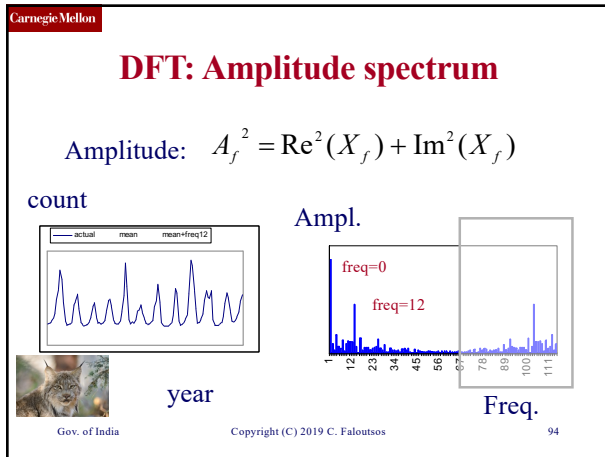
---

---

---

---

---



94

---

---

---

---

---

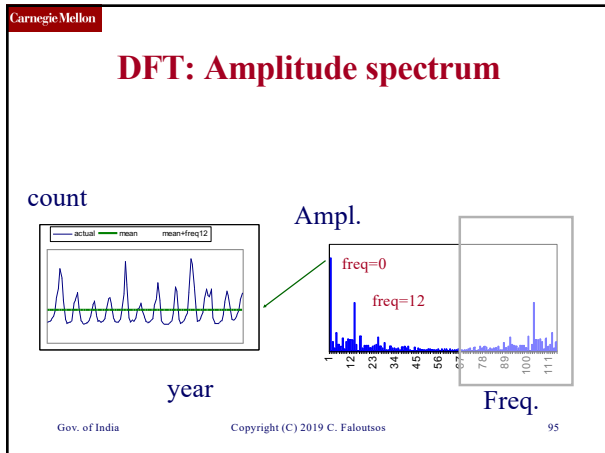
---

---

---

---

---



95

---

---

---

---

---

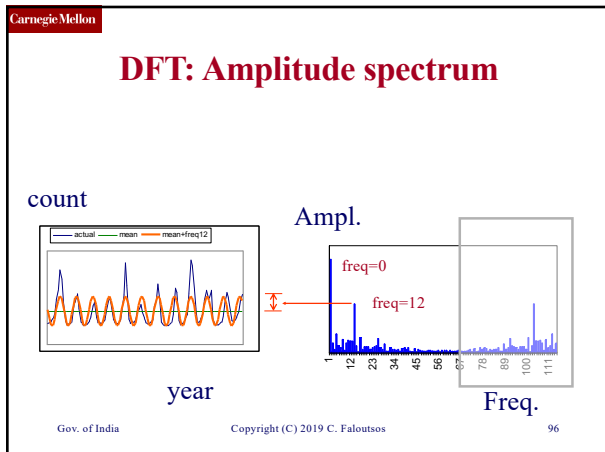
---

---

---

---

---



96

---

---

---

---

---

---

---

---

---

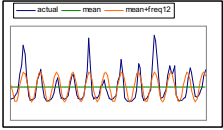
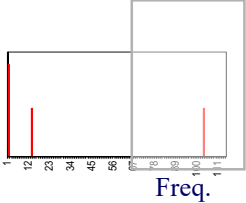
---



Carnegie Mellon

## DFT: Amplitude spectrum

- excellent approximation, with only 2 frequencies!
- so what?

Gov. of India Copyright (C) 2019 C. Faloutsos 97

---

---

---

---

---

---

---


---

97

Carnegie Mellon

## DFT: Amplitude spectrum

- excellent approximation, with only 2 frequencies!
- so what?
- A1: **(lossy) compression**
- A2: pattern discovery



Gov. of India Copyright (C) 2019 C. Faloutsos 98

---

---

---

---

---

---

---

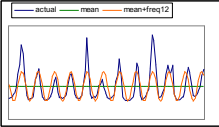
---

98

Carnegie Mellon

## DFT: Amplitude spectrum

- excellent approximation, with only 2 frequencies!
- so what?
- A1: (lossy) compression
- A2: **pattern discovery**



Gov. of India Copyright (C) 2019 C. Faloutsos 99

---

---

---

---

---

---

---

---

99

Carnegie Mellon

## DFT - Conclusions

- It spots periodicities (with the ‘**amplitude spectrum**’)
- can be quickly computed ( $O(n \log n)$ ), thanks to the FFT algorithm.
- **standard** tool in signal processing (speech, image etc signals)
- (closely related to DCT and JPEG)

Gov. of India      Copyright (C) 2019 C. Faloutsos      100

100

---

---

---

---

---


---

---


---

Carnegie Mellon

## Outline



- Motivation
- Similarity Search and Indexing
- DSP
  - DFT
  - DWT



- Definition of DWT and properties
- how to read the DWT scalogram

Gov. of India      Copyright (C) 2019 C. Faloutsos      101

101

---

---

---

---

---

---

---

---

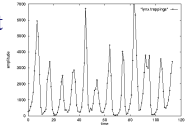
Carnegie Mellon

## Problem #1:

Goal: given a signal (eg., #packets over time)

Find: patterns, periodicities, and/or compress

count



year

lynx caught per year  
(packets per day;  
virus infections per month)

Gov. of India      Copyright (C) 2019 C. Faloutsos      102

102

---

---

---

---

---

---

---

---

Carnegie Mellon

## Wavelets - DWT

- DFT is great - but, how about compressing a spike?

value

time

Gov. of India

Copyright (C) 2019 C. Faloutsos

103

---

---

---

---

---

---

---

---

103

Carnegie Mellon

## Wavelets - DWT

- DFT is great - but, how about compressing a spike?
- A: Terrible - all DFT coefficients needed!

value

time

Ampl

Gov. of India

Copyright (C) 2019 C. Faloutsos

104

---

---

---

---

---

---

---

---

104

Carnegie Mellon

## Wavelets - DWT

- DFT is great - but, how about compressing a spike?
- A: Terrible - all DFT coefficients needed!

value

time

Gov. of India

Copyright (C) 2019 C. Faloutsos

105

---

---

---

---

---

---

---

---

105

Carnegie Mellon

## Wavelets - DWT

- Similarly, DFT suffers on short-duration waves (eg., baritone, silence, soprano)

Gov. of India Copyright (C) 2019 C. Faloutsos 106

106

---

---

---

---

---

---

---

---

Carnegie Mellon

## Wavelets - DWT

- Solution#1: Short window Fourier transform (SWFT)
- But: how short should be the window?

Gov. of India Copyright (C) 2019 C. Faloutsos 107

107

---

---

---

---

---

---

---

---

Carnegie Mellon

## Wavelets - DWT

- Answer: **multiple** window sizes! -> DWT

Time domain

Gov. of India Copyright (C) 2019 C. Faloutsos 108

108

---

---

---

---

---

---

---

---

Carnegie Mellon

## Haar Wavelets

- subtract sum of left half from right half
- repeat recursively for quarters, eight-ths, ...

Gov. of India

Copyright (C) 2019 C. Faloutsos

109

---

---

---

---

---

---

---

---

109

Carnegie Mellon

## Wavelets - construction

Skip

$x_0 \ x_1 \ x_2 \ x_3 \ x_4 \ x_5 \ x_6 \ x_7$

Gov. of India

Copyright (C) 2019 C. Faloutsos

110

---

---

---

---

---

---

---

---

110

Carnegie Mellon

## Wavelets - construction

Skip

level 1  $d_{1,0}$   $s_{1,0}$   $d_{1,1}$   $s_{1,1}$  .....

$-$   $+$

$x_0 \ x_1 \ x_2 \ x_3 \ x_4 \ x_5 \ x_6 \ x_7$

Gov. of India

Copyright (C) 2019 C. Faloutsos

111

---

---

---

---

---

---

---

---

111

Carnegie Mellon
Skip

## Wavelets - construction

level 2

Gov. of India      Copyright (C) 2019 C. Faloutsos      112

112

---

---

---

---

---

---

---

---

Carnegie Mellon
Skip

## Wavelets - construction

etc ...

Gov. of India      Copyright (C) 2019 C. Faloutsos      113

113

---

---

---

---

---

---

---

---

Carnegie Mellon
Skip

## Wavelets - construction

Q: map each coefficient on the time-freq. plane


f

t

Gov. of India      Copyright (C) 2019 C. Faloutsos      114

114

---

---

---

---

---

---

---

---

Carnegie Mellon

## Wavelets - construction

Q: map each coefficient on the time-freq. plane

Gov. of India Copyright (C) 2019 C. Faloutsos 115

115

Carnegie Mellon

## Haar wavelets - code

```
#!/usr/bin/perl5
# expects a file with numbers
# and prints the dwt transform
# The number of time-ticks should be a power of 2
# USAGE:
#   haar.pl <fname>

my @vals=();
my @smooth; # the smooth component of the signal
my @diff;   # the high-freq. component

# collect the values into the array @vals
while(<>){
    @vals = ( @vals , split );
}

my $len = scalar(@vals);
my $half = int($len/2);
while($half >= 1 ){
    for(my $i=0; $i<$half; $i++){
        $diff[$i] = ($vals[2*$i] - $vals[2*$i+1]) / sqrt(2);
        print "d", $diff[$i];
        $smooth[$i] = ($vals[2*$i] + $vals[2*$i+1]) / sqrt(2);
    }
    print "a";
    @vals = @smooth;
    $half = int($half/2);
}
print "d", $vals[0], "a"; # the final, smooth component
```

Gov. of India Copyright (C) 2019 C. Faloutsos 116

116

Carnegie Mellon

## Wavelets - construction

Observation1:  
 '+' can be some weighted addition  
 '-' is the corresponding weighted difference  
 ('Quadrature mirror filters')

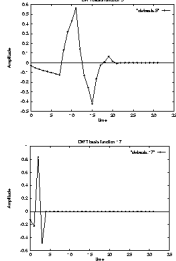
Observation2: unlike DFT/DCT,  
 there are \*many\* wavelet bases: Haar, Daubechies-4, Daubechies-6, Coifman, Morlet, Gabor, ...

Gov. of India Copyright (C) 2019 C. Faloutsos 117

117

Carnegie Mellon

Wavelets - how do they look like?



- E.g., Daubechies-4

Gov. of India

Copyright (C) 2019 C. Faloutsos

118

---

---

---

---

---

---

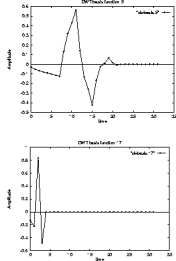
---

---


118

Carnegie Mellon

Wavelets - how do they look like?



- E.g., Daubechies-4



Gov. of India

Copyright (C) 2019 C. Faloutsos

119

---

---

---

---

---

---

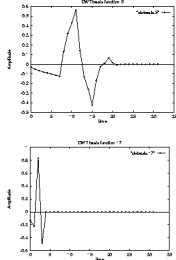
---

---


119

Carnegie Mellon

Wavelets - how do they look like?



- E.g., Daubechies-4



Gov. of India

Copyright (C) 2019 C. Faloutsos

120

---

---

---

---

---

---

---

---


120



Carnegie Mellon

Outline

- Motivation
- Similarity Search and Indexing
- DSP
  - DFT
  - DWT
    - Definition of DWT and properties
    - how to read the DWT scalogram



Gov. of IndiaCopyright (C) 2019 C. Faloutsos121

---

---

---

---

---

---

---

---

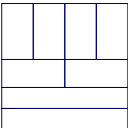
121

Carnegie Mellon

Wavelets - Drill#1:

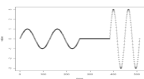
- Q: baritone/silence/soprano - DWT?

f



t

value



time

Gov. of IndiaCopyright (C) 2019 C. Faloutsos122

---

---

---

---

---

---

---

---

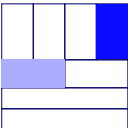
122

Carnegie Mellon

Wavelets - Drill#1:

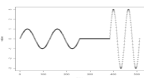
- Q: baritone/silence/soprano - DWT?

f



t

value



time

Gov. of IndiaCopyright (C) 2019 C. Faloutsos123

---

---

---

---

---

---

---

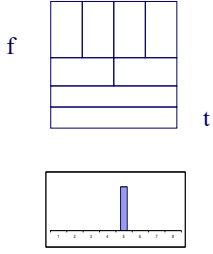
---

123

Carnegie Mellon

### Wavelets - Drill#2:

- Q: spike - DWT?



Gov. of India Copyright (C) 2019 C. Faloutsos 124

124

---

---

---

---

---

---

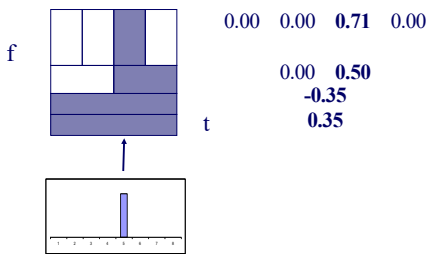
---

---

Carnegie Mellon

### Wavelets - Drill#2:

- Q: spike - DWT?



Gov. of India Copyright (C) 2019 C. Faloutsos 125

125

---

---

---

---

---

---

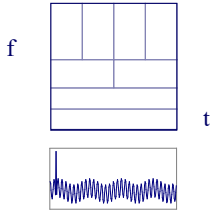
---

---

Carnegie Mellon

### Wavelets - Drill#3:

- Q: weekly + daily periodicity, + spike - DWT?



Gov. of India Copyright (C) 2019 C. Faloutsos 126

126

---

---

---

---

---

---

---

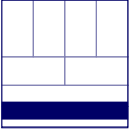
---

Carnegie Mellon

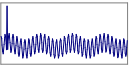
### Wavelets - Drill#3:

- Q: **weekly** + daily periodicity, + spike - DWT?

f



t



Gov. of India

Copyright (C) 2019 C. Faloutsos

127

---

---

---

---

---

---

---

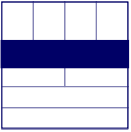
---

Carnegie Mellon

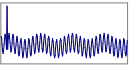
### Wavelets - Drill#3:

- Q: weekly + **daily** periodicity, + spike - DWT?

f



t



Gov. of India

Copyright (C) 2019 C. Faloutsos

128

---

---

---

---

---

---

---

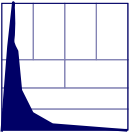
---

Carnegie Mellon

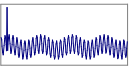
### Wavelets - Drill#3:

- Q: weekly + daily periodicity, + **spike** - DWT?

f



t



Gov. of India

Copyright (C) 2019 C. Faloutsos

129

---

---

---

---

---

---

---

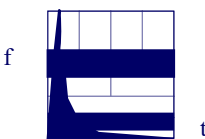
---

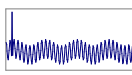
129

Carnegie Mellon

### Wavelets - Drill#3:

- Q: weekly + daily periodicity, + spike - DWT?





Gov. of India

Copyright (C) 2019 C. Faloutsos

130

---

---

---

---

---

---

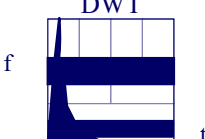
---


---

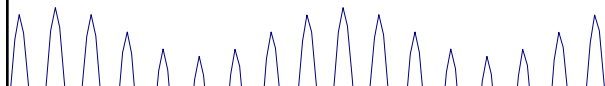
130

Carnegie Mellon

### Wavelets - Drill#3:







---

---

---

---

---

---

---

---

131

Carnegie Mellon

### Wavelets in action

Gov. of India

Copyright (C) 2019 C. Faloutsos

132

---

---

---

---

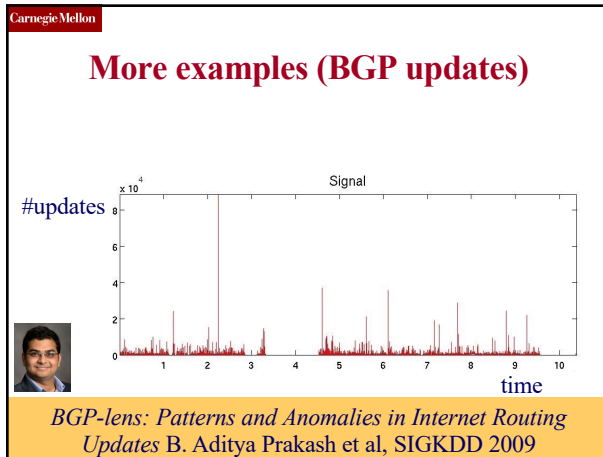
---

---

---

---

132



133

---

---

---

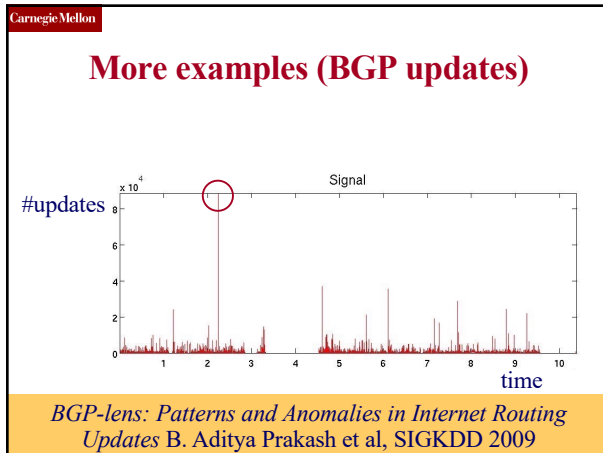
---

---

---

---

---



134

---

---

---

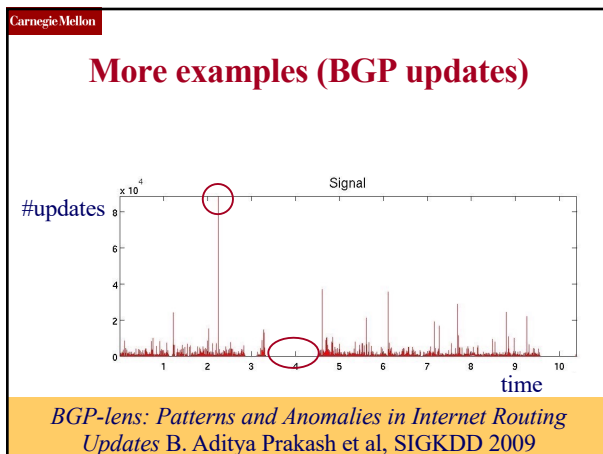
---

---

---

---

---



135

---

---

---

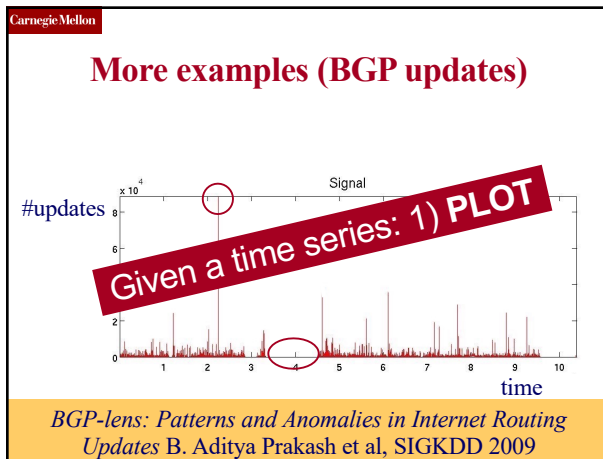
---

---

---

---

---



136

---

---

---

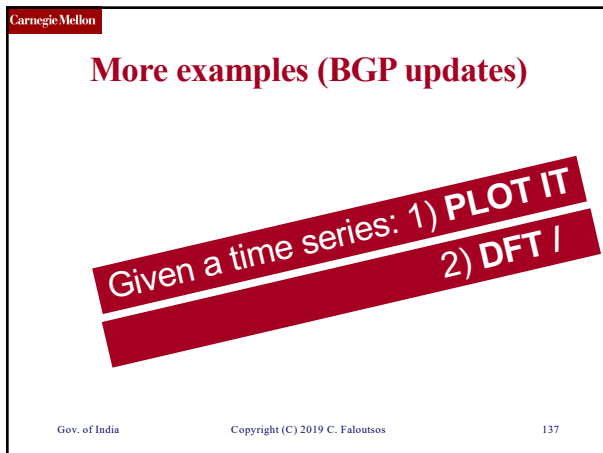
---

---

---

---

---



137

---

---

---

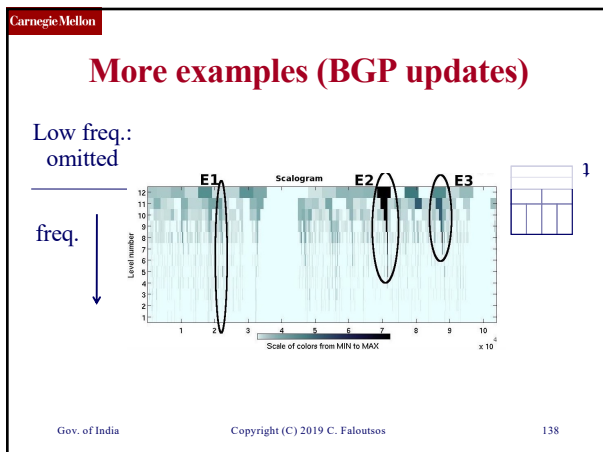
---

---

---

---

---



138

---

---

---

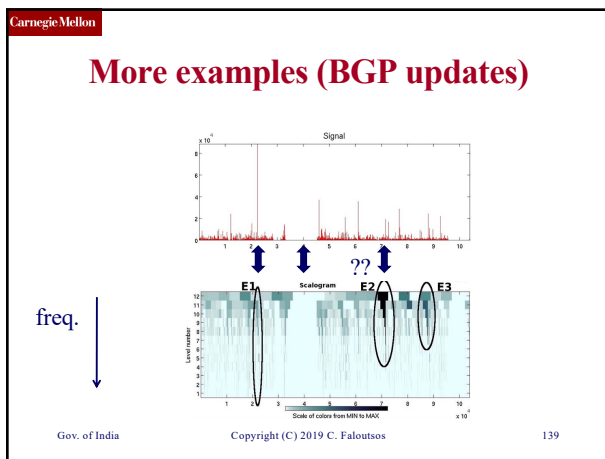
---

---

---

---

---



139

---

---

---

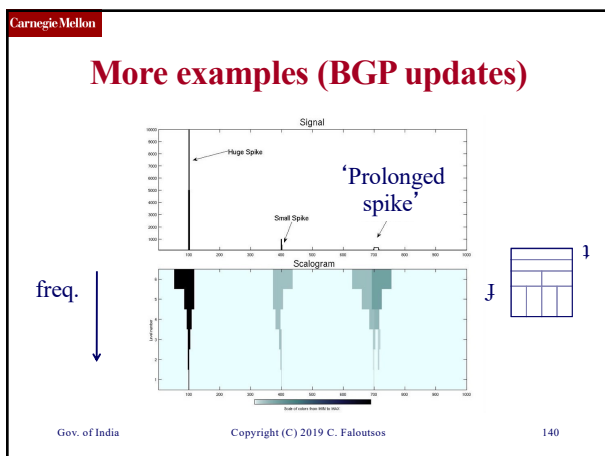
---

---

---

---

---



140

---

---

---

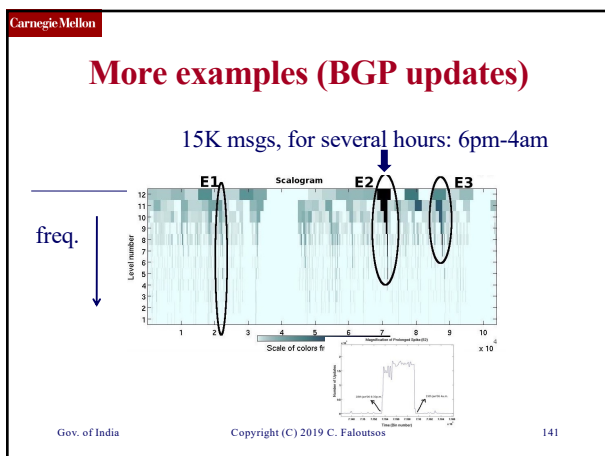
---

---

---

---

---



141

---

---

---

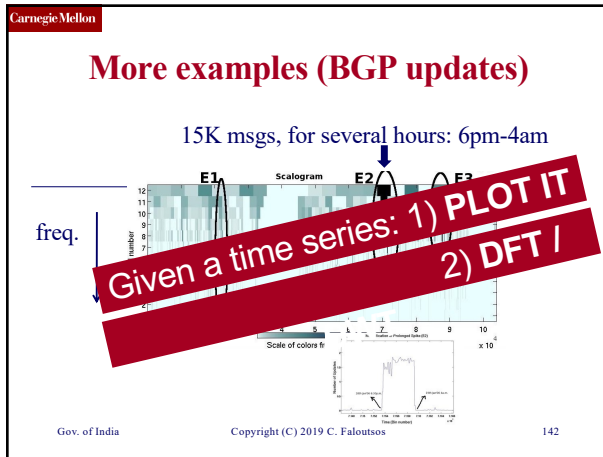
---

---

---

---

---



142

---

---

---

---

---

---

---

---

Carnegie Mellon

### Advantages of Wavelets

- Better compression (better RMSE with same number of coefficients - used in JPEG-2000)
- fast to compute (usually:  $O(n)$ !)
- good for 'spikes'
- good for de-noising
- mammalian eye and ear: Gabor wavelets

JMP Exec. Ed.

Copyright (C) 2019 C. Faloutsos

143

143

---

---

---

---

---

---

---

---

Carnegie Mellon

### Overall Conclusions

- DFT spots periodicities
- **DWT** : multi-resolution - matches processing of mammalian ear/eye better
- Both: powerful tools for **compression**, **pattern detection** in real signals

JMP Exec. Ed.

Copyright (C) 2019 C. Faloutsos

144

144

---

---

---

---

---

---

---

---



Carnegie Mellon

Resources: software and urls

- *xwpl*: open source wavelet package from Yale, with excellent GUI
- <http://monet.me.ic.ac.uk/people/gavin/java/waveletDemos.html> : wavelets and scalograms

Gov. of India

Copyright (C) 2019 C. Faloutsos

145

145

---

---

---

---

---

---

---

---

Carnegie Mellon

Books

- William H. Press, Saul A. Teukolsky, William T. Vetterling and Brian P. Flannery: *Numerical Recipes in C*, Cambridge University Press, 1992, 2nd Edition. (Great description, intuition and code for DFT, DWT)
- C. Faloutsos: *Searching Multimedia Databases by Content*, Kluwer Academic Press, 1996 (introduction to DFT, DWT)

Gov. of India

Copyright (C) 2019 C. Faloutsos

146

146

---

---

---

---

---

---

---

---

Carnegie Mellon

Additional Reading

- [Gilbert+01] Anna C. Gilbert, Yannis Kotidis and S. Muthukrishnan and Martin Strauss, *Surfing Wavelets on Streams: One-Pass Summaries for Approximate Aggregate Queries*, VLDB 2001

Gov. of India

Copyright (C) 2019 C. Faloutsos

147

147

---

---

---

---

---

---

---

---