Carnegie Mellon

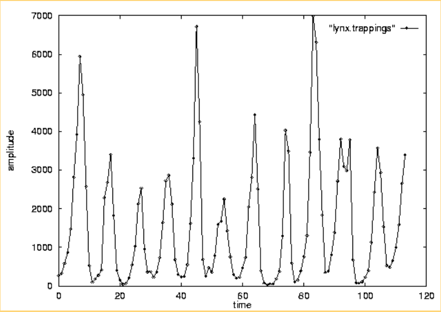
15-826: Multimedia (Databases) and Data Mining

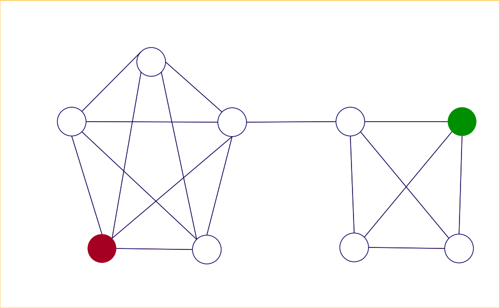
Lecture #31: Conclusions

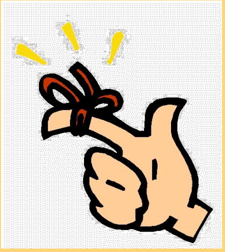
C. Faloutsos

Problem

- Given a large dataset (points; text doc’s; time series; images; nodes in a graph)
- Find similar/interesting things







Summary

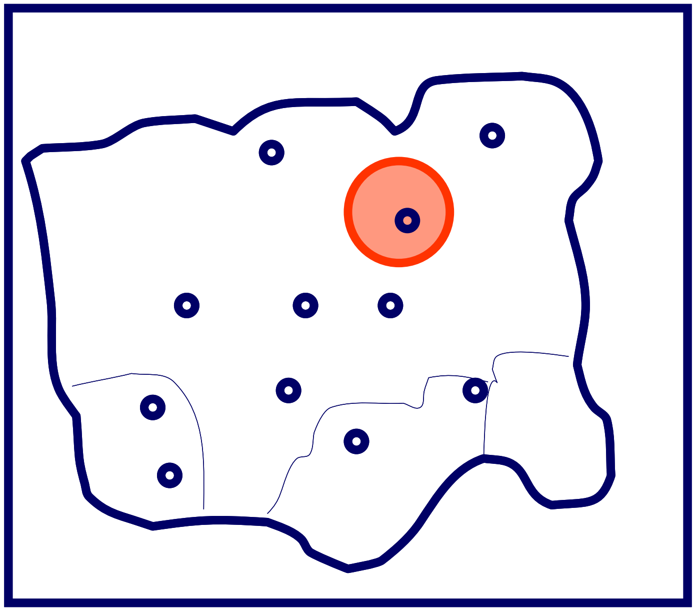
- **T1: fractals / power laws** lead to startling discoveries
  - ‘the mean may be meaningless’
  - Don’t assume Gaussian (average, k-means, etc)
- **T2: SVD**: behind PageRank/HITS/tensors/…
- **T3: Wavelets**: Nature seems to prefer them
- ~~**T4: RLS**: matrix inversion, without inverting~~

Outline

Goal: ‘Find similar / interesting things’

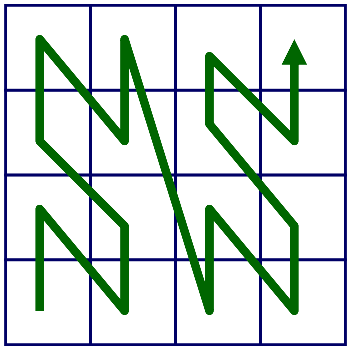
- Intro to DB
- Indexing - similarity search
  - Points
  - Text
  - Time sequences; images etc
  - Graphs

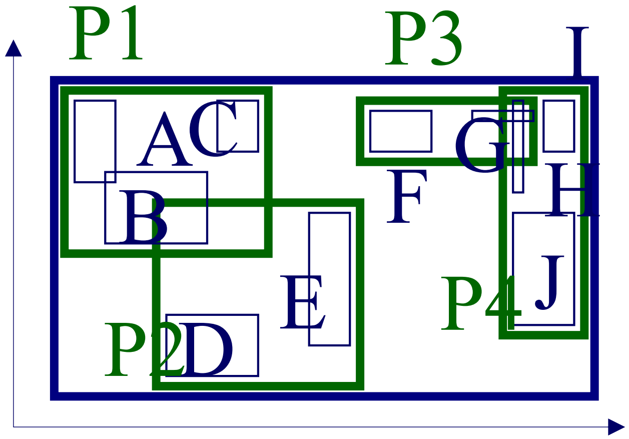
Indexing - similarity search



Indexing - similarity search

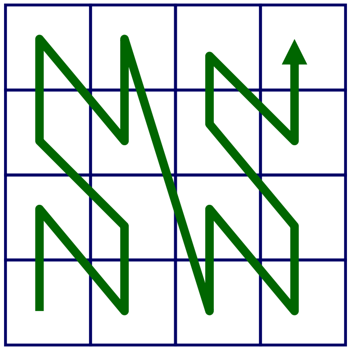
- R-trees
- z-ordering / hilbert curves
- M-trees
- (DON’ T FORGET …)

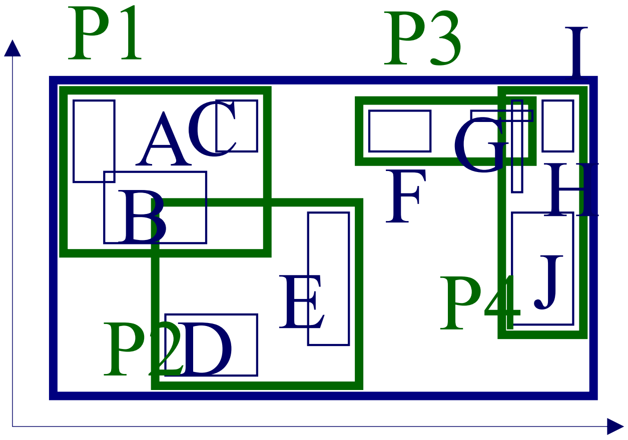




Indexing - similarity search

- R-trees
- z-ordering / hilbert curves
- M-trees
- **beware of high intrinsic dimensionality**





Outline

Goal: ‘Find similar / interesting things’

- Intro to DB
- Indexing - similarity search
  - Points
  - ➡ Text
  - Time sequences; images etc
  - Graphs

Text searching

- ‘find all documents with word *bla*’

Text searching

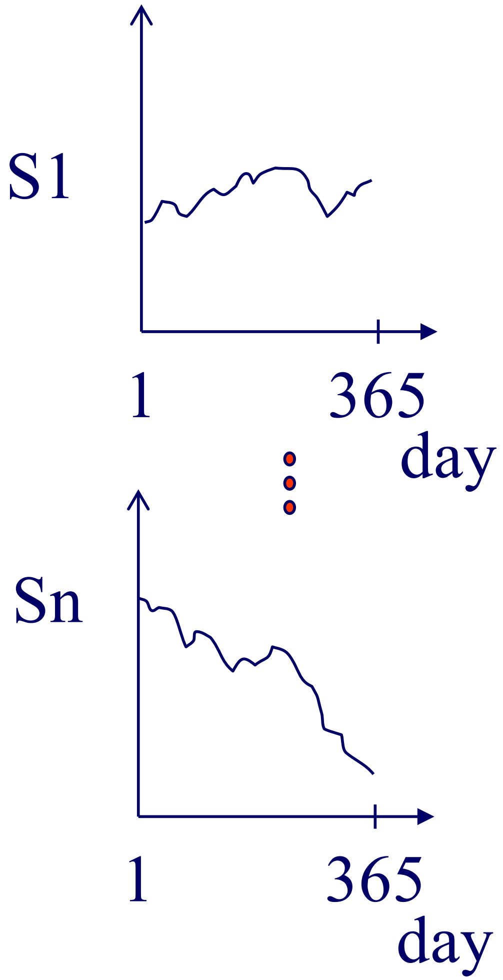
- Full text scanning ('grep')
- Inversion (B-tree or hash index)
- signature files – Bloom filters
- Vector space model
  - Ranked output
  - Relevance feedback
- String editing distance (-> dynamic prog.)

Outline

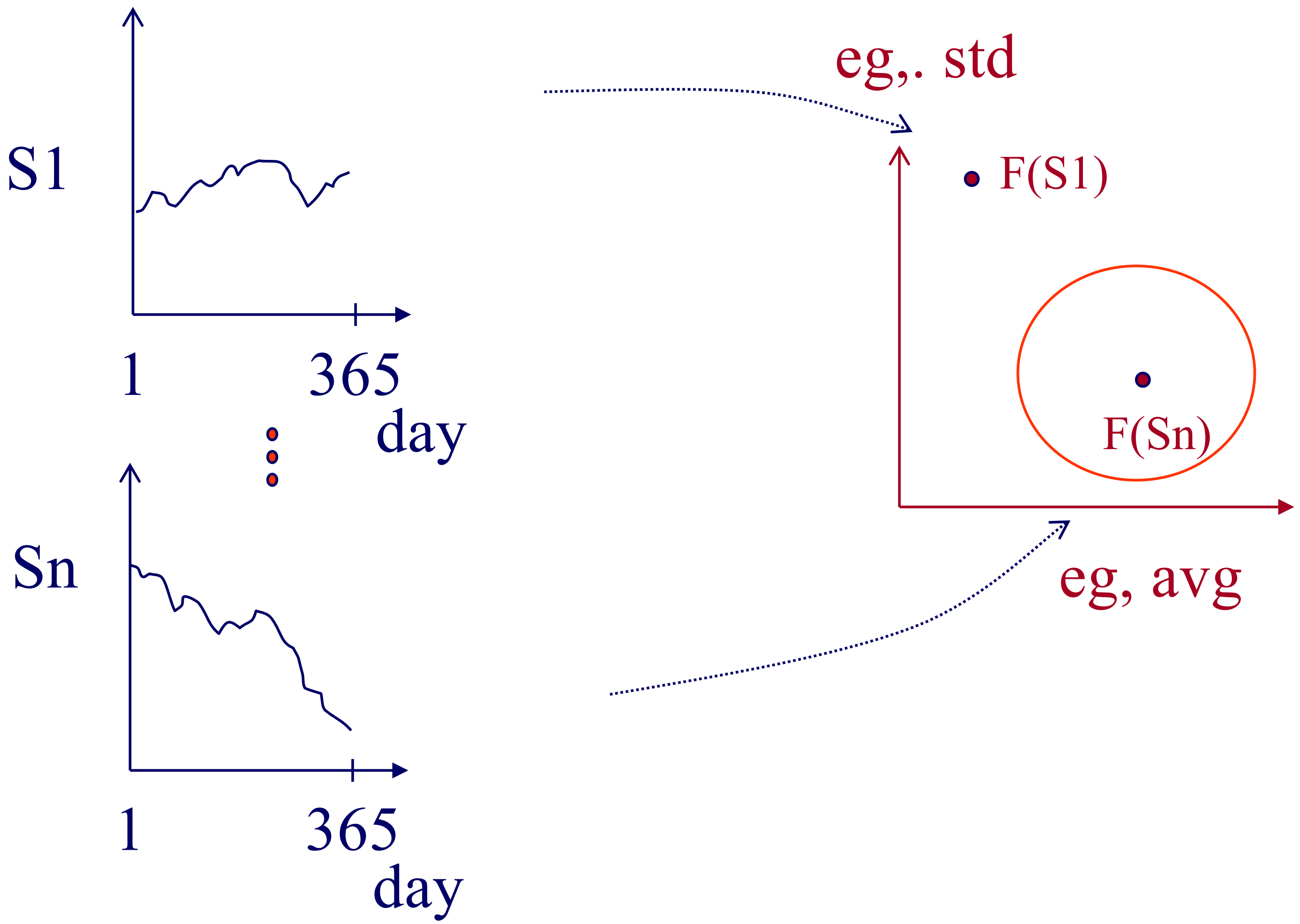
Goal: ‘Find similar / interesting things’

- Intro to DB
- Indexing - similarity search
  - Points
  - Text
  - ➡ Time sequences; images etc
  - Graphs

Multimedia indexing

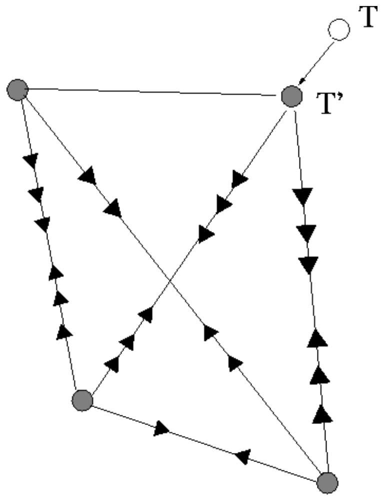


‘GEMINI’ - Pictorially



Multimedia indexing

- Feature extraction for indexing (GEMINI)
  - – Lower-bounding lemma, to guarantee no false alarms
- MDS/FastMap
- tSNE/UMap



Outline

Goal: ‘Find similar / interesting things’

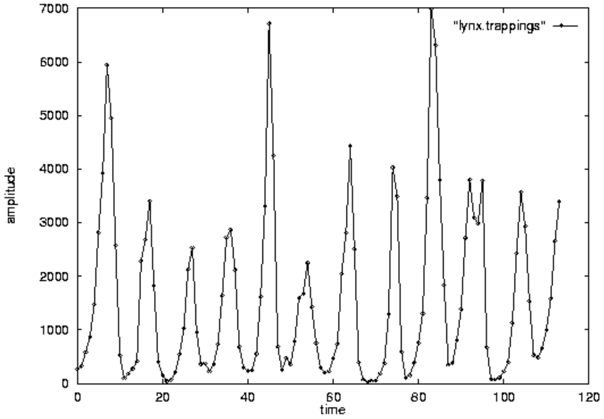
- Intro to DB
- Indexing - similarity search
  - Points
  - Text
  - ➡ Time sequences; images etc – **DFT/DWT**
  - Graphs

Time series ~~& forecasting~~

Goal: given a signal (eg., sales over time and/or space)

Find: patterns and/or compress

count

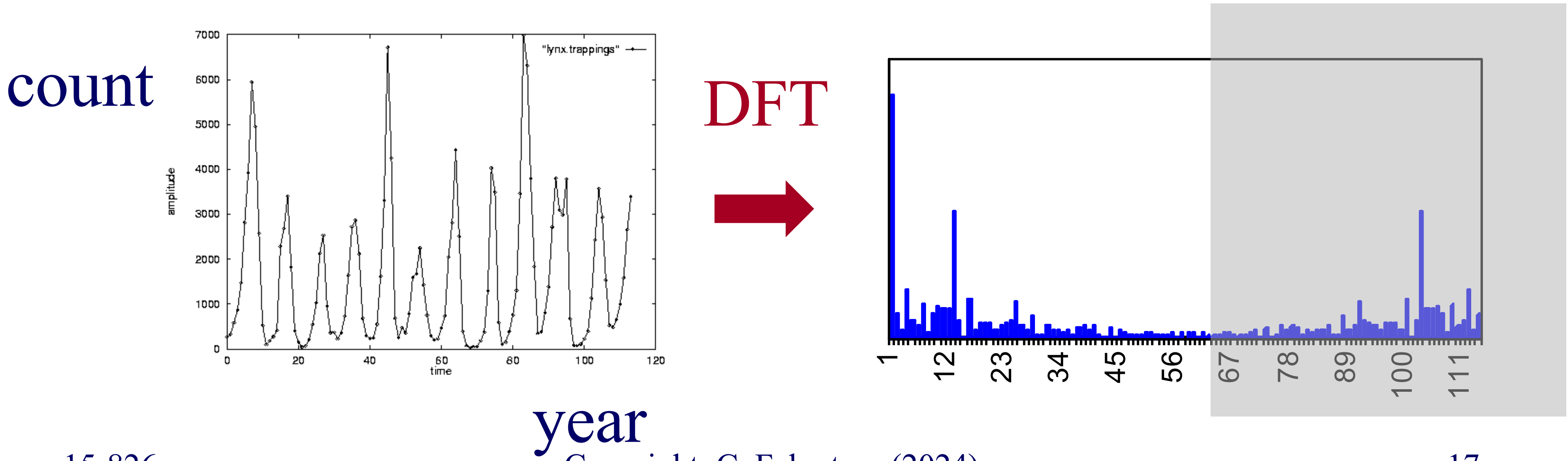


year

Time series ~~& forecasting~~

Goal: given a signal (eg., sales over time and/or space)

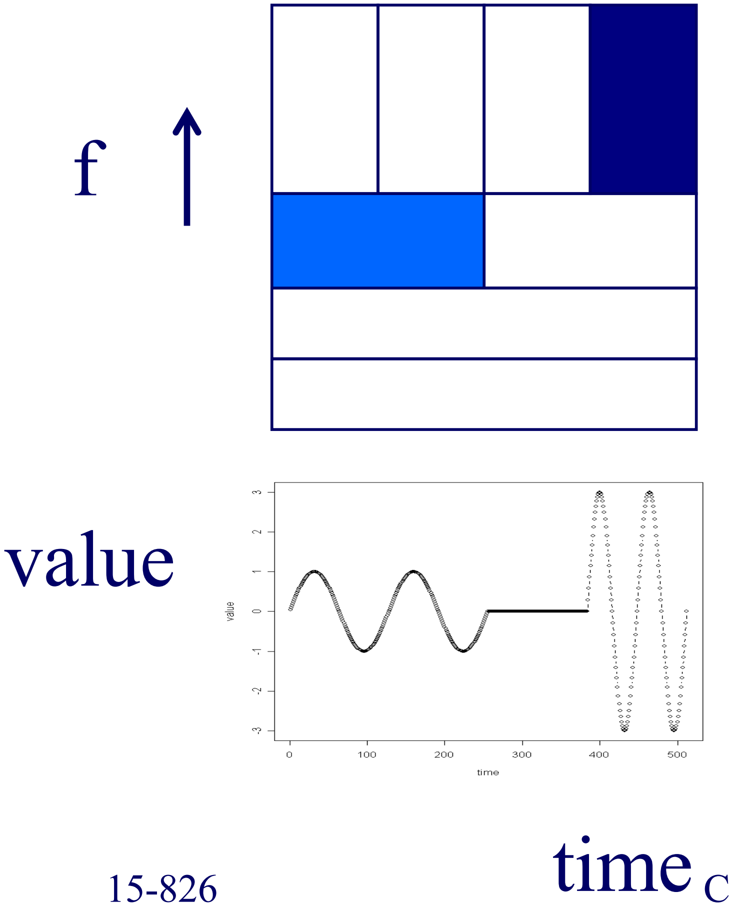
Find: patterns and/or compress

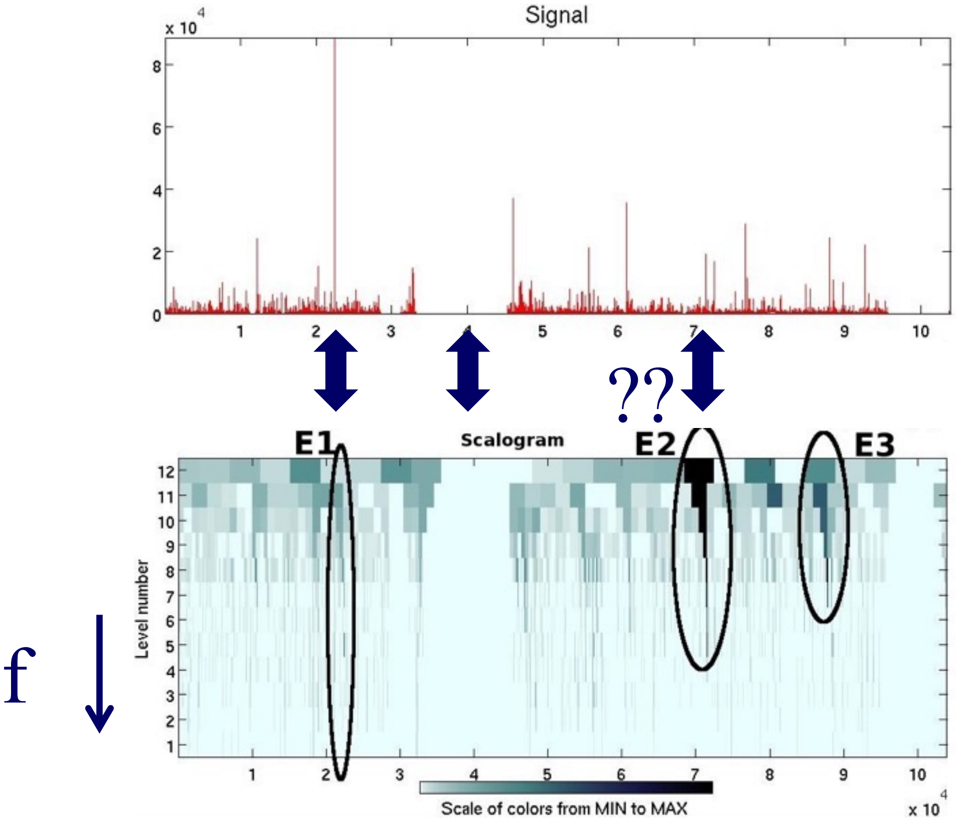


Copyright: C. Faloutsos (2024)

Wavelets

- Q: baritone/silence/soprano - DWT?

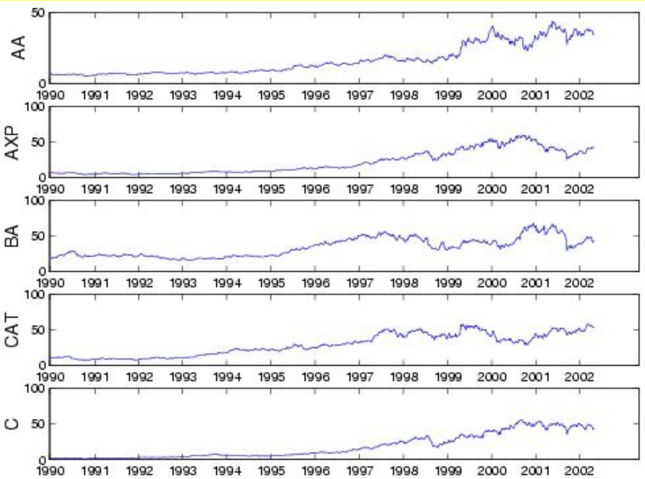




Not in the final exam

Problem:

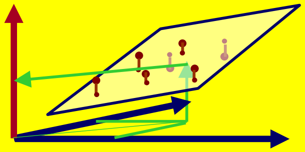
Q: mine/forecast (one, or more) time sequences

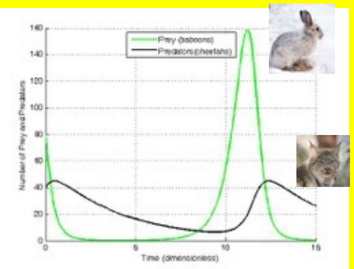


Not in the final exam

Answers

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





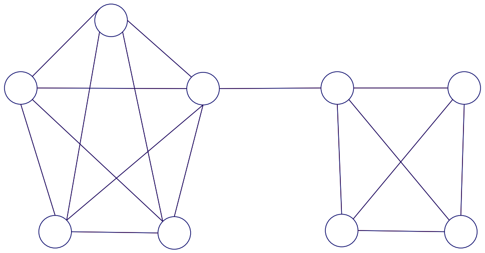
Outline

Goal: ‘Find similar / interesting things’

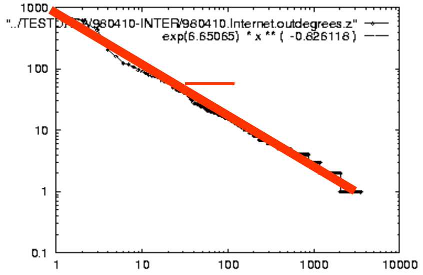
- Intro to DB
- Indexing - similarity search
  - Points
  - Text
  - Time sequences; images etc
  - ➡ Graphs

Graphs

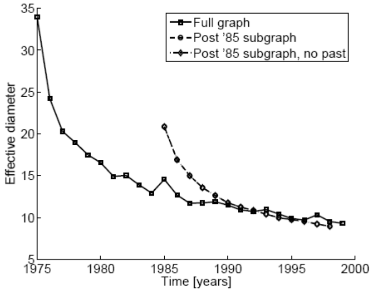
- Real graphs: surprising patterns
  - ??

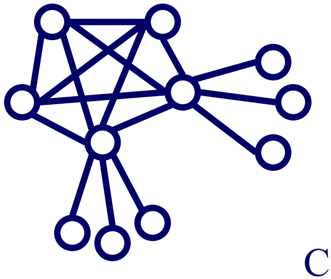


Graphs



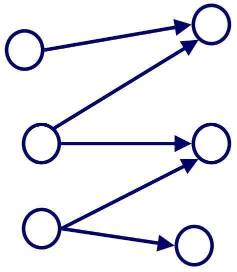
- Real graphs: surprising patterns
  - ‘**six degrees**’
  - **Skewed** degree distribution (‘rich get richer’)
  - Super-linearities (2x nodes -> 3x edges)
  - Diameter: **shrinks** (!)
  - Might have **no** good cuts

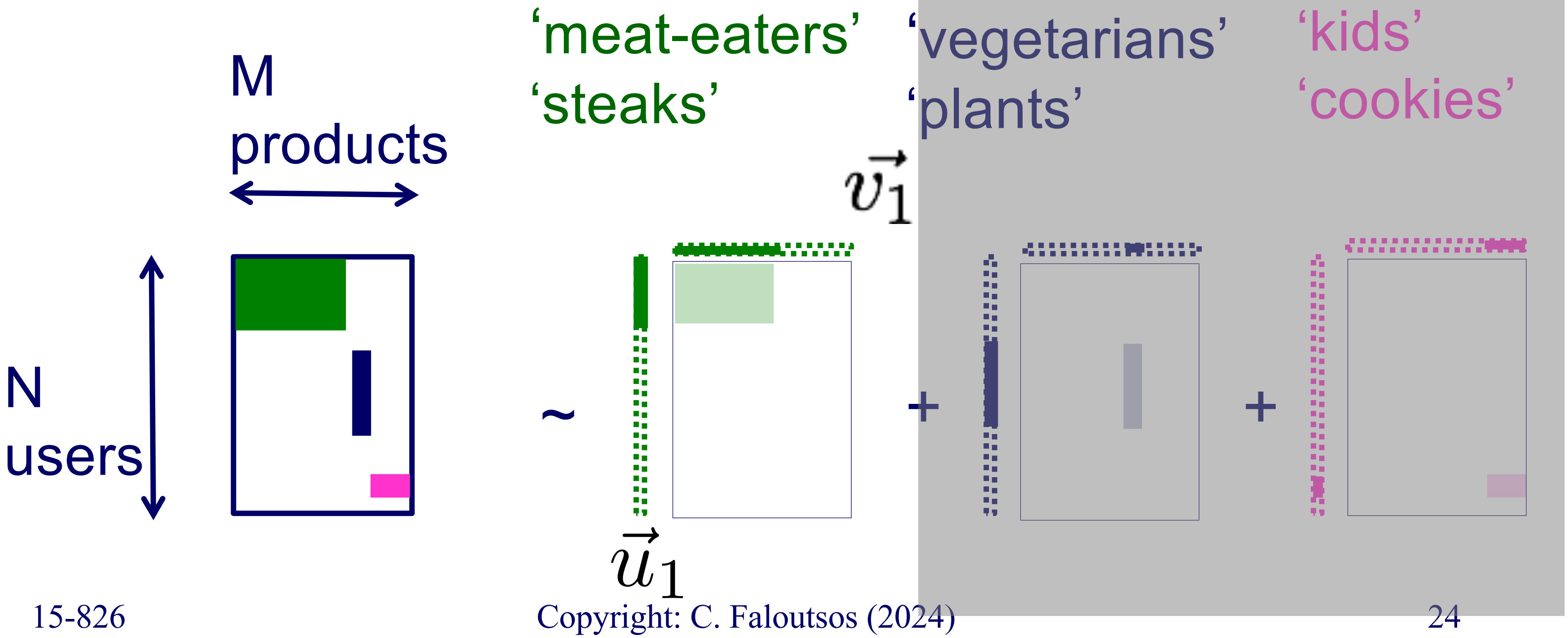




Graphs - SVD

- Hubs/Authorities (SVD on adjacency matrix)



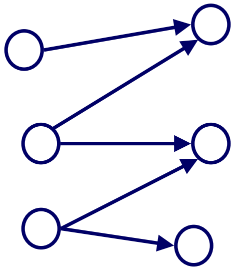


‘meat-eaters’
‘steaks’

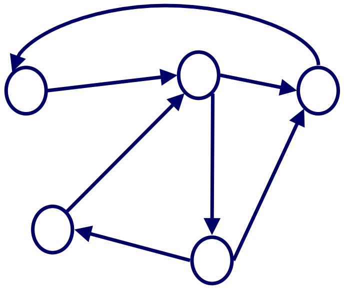
‘vegetarians’
‘plants’

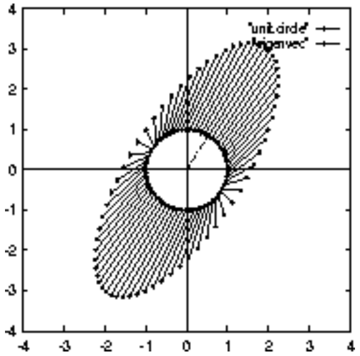
‘kids’
‘cookies’

Graphs - PageRank



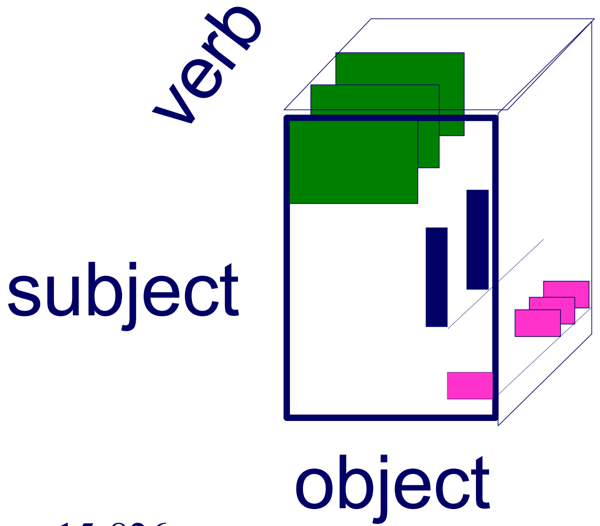
- Hubs/Authorities (SVD on adjacency matrix)
- PageRank (fixed point -> eigenvector)





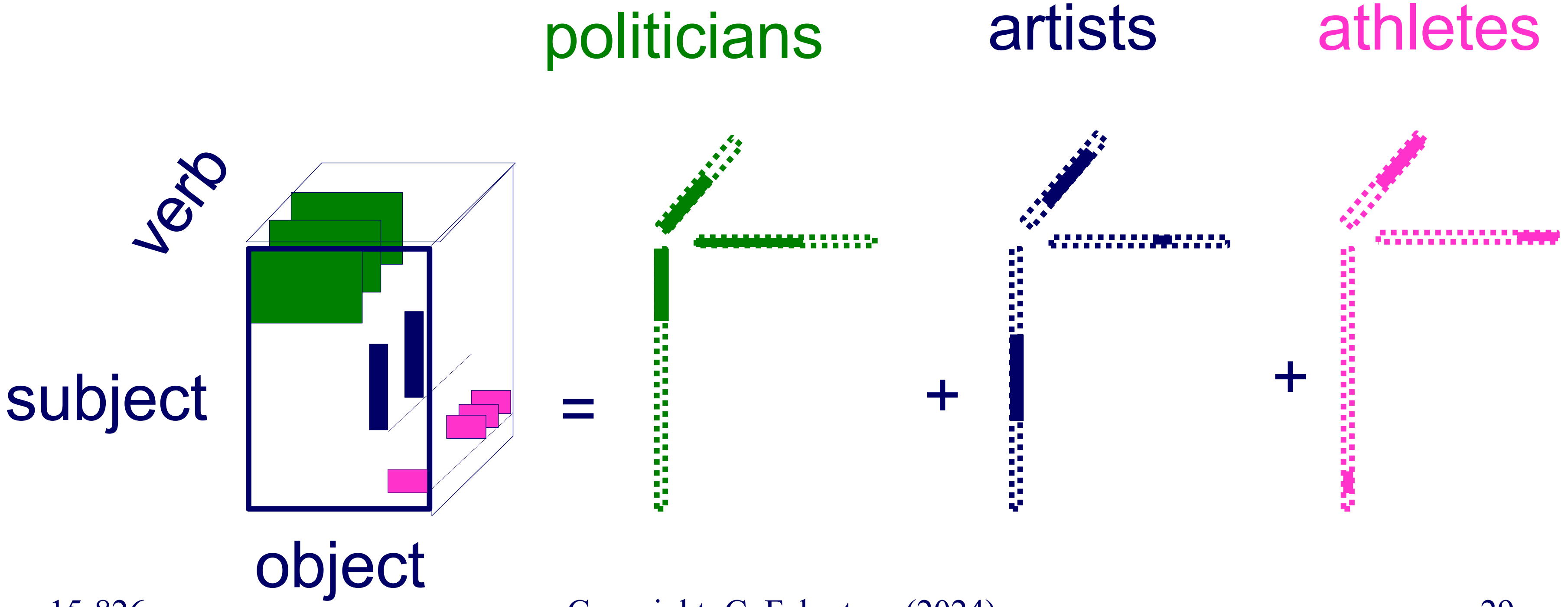
Tensors

- Eg., time evolving graphs; Subject-verb-object triplets; etc



Tensors

- Eg., time evolving graphs; Subject-verb-object triplets; etc

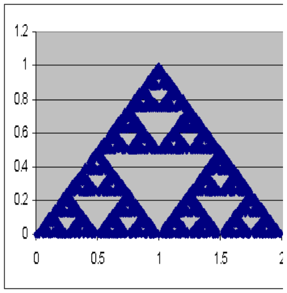


Taking a step back:

We saw some fundamental, recurring concepts and tools:

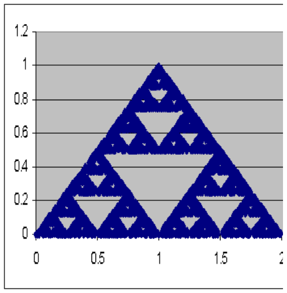
T1: Powerful, recurring tools

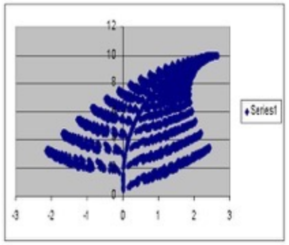
- Fractals/ self similarity

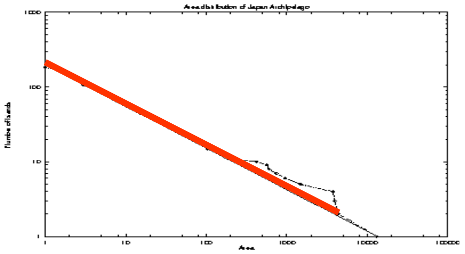


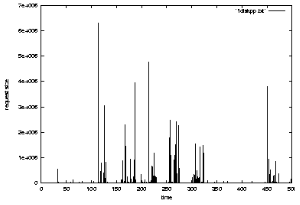
T1: Powerful, recurring tools

- Fractals/ self similarity <-> Power laws
  - Zipf, Korcak, Pareto's laws
  - intrinsic dimension (Sierpinski triangle)
  - correlation integral
  - Barnsley's IFS compression
  - ~~Kronecker graphs~~

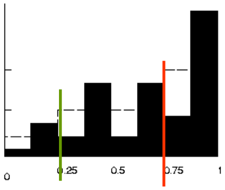








80/20



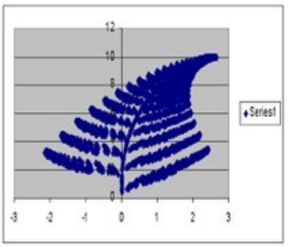
T1: Powerful, recurring tools

- Fractals/ self similarity
  - Zipf, Ko

pression

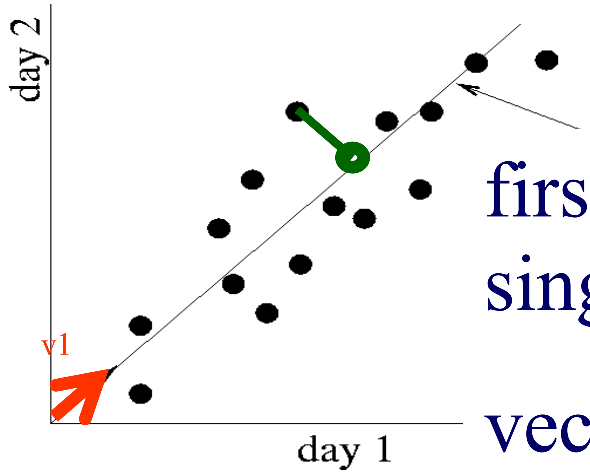
ecker graphs)

- 'Take logarithms'
  - mean-> meaningless
  - Gaussian trap



T2: Powerful, recurring tools

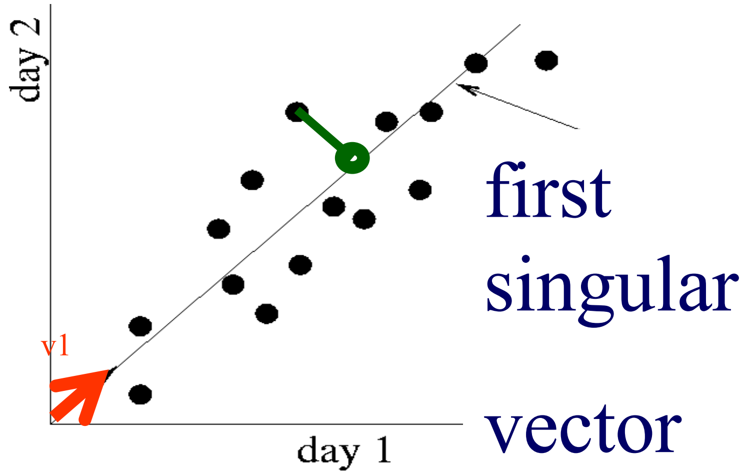
- SVD (optimal L2 approx)



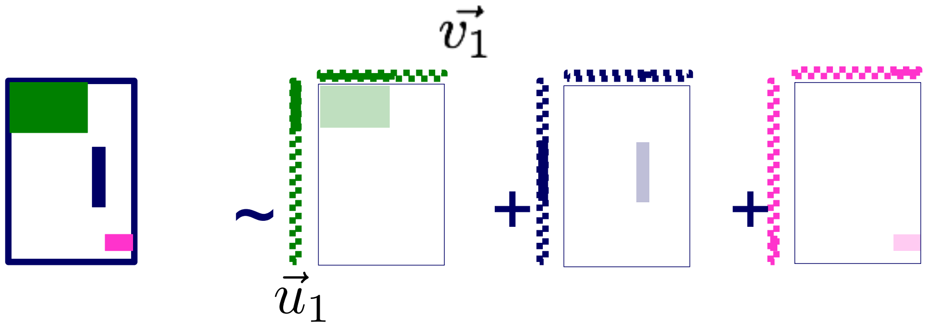
first singular vector

T2: Powerful, recurring tools

- Q: Cases we have a matrix as input?
- A: …

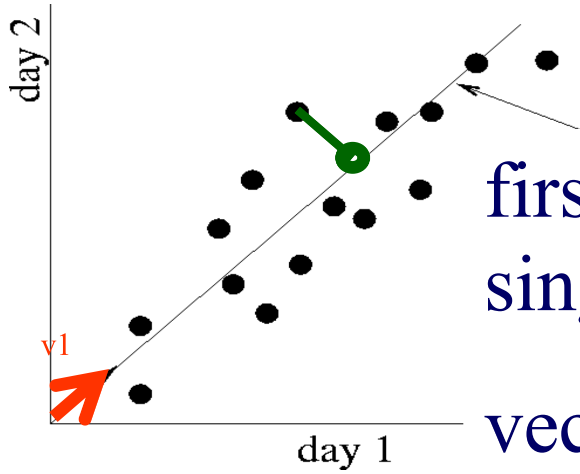


first singular vector

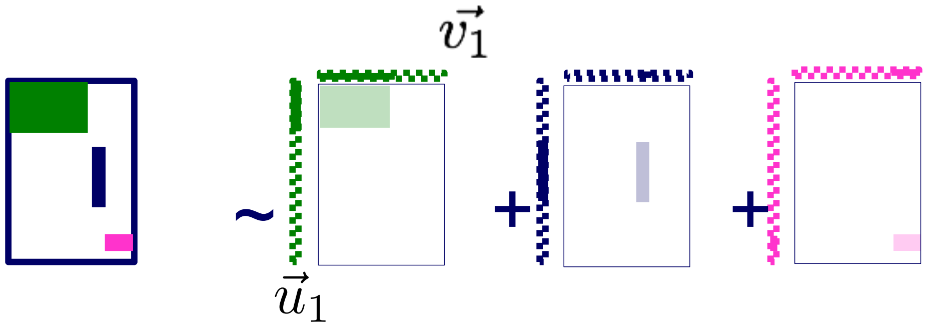


T2: Powerful, recurring tools

- Q: Cases we have a matrix as input?
- A1: graphs
- A2: co-evolving time sequences
- A3: entities in feature space

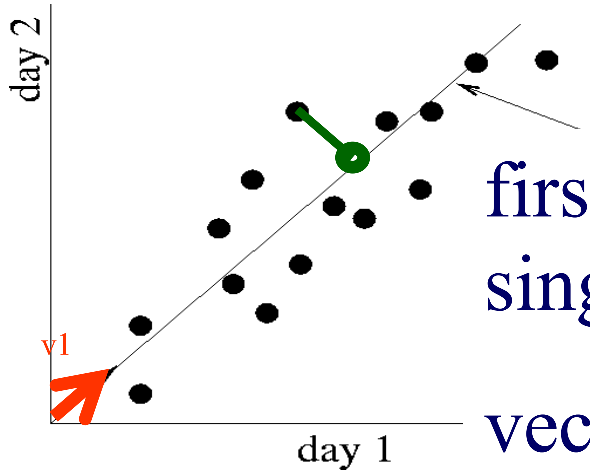


first singular vector



T2: Powerful, recurring tools

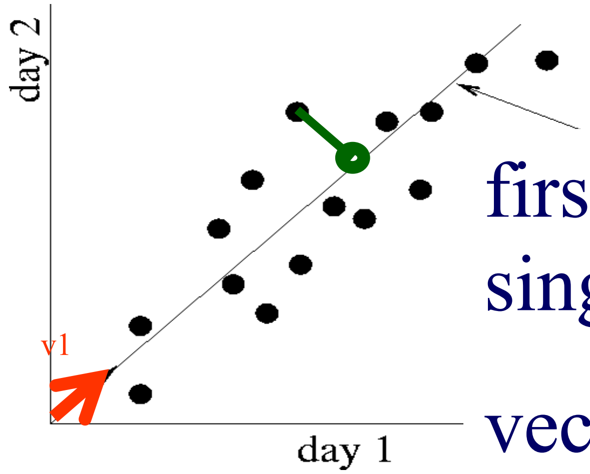
- SVD (optimal L2 approx)
- Algorithms in course, where SVD worked?



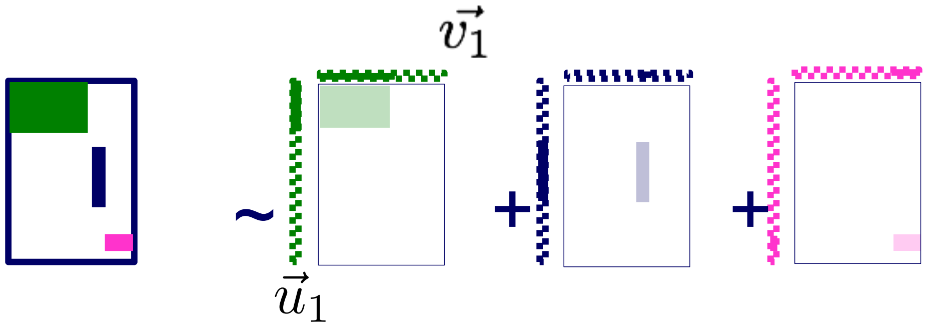
first singular vector

T2: Powerful, recurring tools

- SVD (optimal L2 approx)
  - LSI, KL, PCA, ‘eigenSpokes’, (& in ICA)
  - HITS (PageRank)



first singular vector

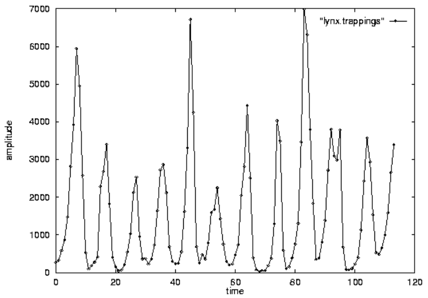


T3: powerful, recurring tools

DFT (Discrete Fourier Transform)

DWT (Discrete Wavelet Transform)

count



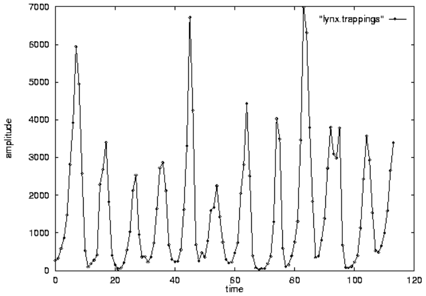
year

T3: powerful, recurring tools

DFT (Discrete Fourier Transform)

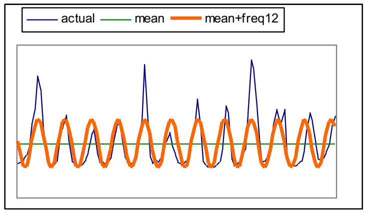
DWT (Discrete Wavelet Transform)

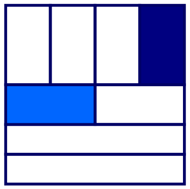
count



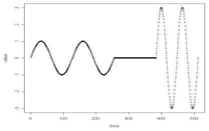
year

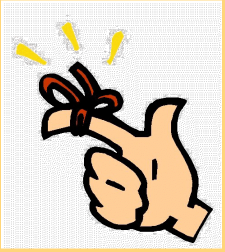
A1: Fourier (DFT)





A2: Wavelets (DWT)

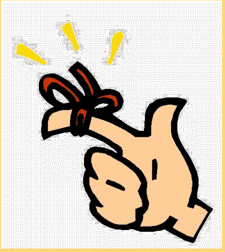




Summary of summary

- **T1: fractals / power laws** lead to startling discoveries
  - ‘the mean may be meaningless’
  - Don’t assume Gaussian (average, k-means, etc)
- **T2: SVD**: behind PageRank/HITS/tensors/…
- **T3: Wavelets**: Nature seems to prefer them
- ~~**T4: RLS**: matrix inversion, without inverting~~

Carnegie Mellon



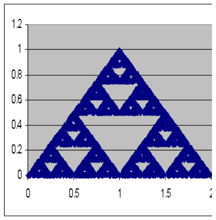
Summary of summary

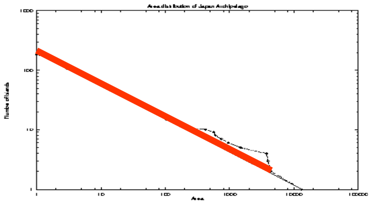
- **T1: fractals / power laws** lead to st... g discoveries
  - ‘the me...
  - ...ans, etc)
- T... ...geRank/HITS/tensors/…
- T... **...avelets**: Nature seems to prefer them
- ~~**T4: RLS**: matrix inversion, without inverting~~

- **‘Take logarithms’**
  - **mean -> meaningless**
  - **Gaussian trap**

Thank you!

- Feel free to contact me:
  - Cell#; christos@cs; GHC 7003
- Reminder: faculty course eval’s:
  - http://www.cmu.edu/hub/fce/
- Have a great holiday!





- ‘Take logarithms’
- mean -> meaningless
- Gaussian trap