

CSE 564

VISUALIZATION & VISUAL ANALYTICS

TIME VARYING AND STREAMING DATA

KLAUS MUELLER

COMPUTER SCIENCE DEPARTMENT
STONY BROOK UNIVERSITY

Lecture	Topic	Projects
1	Intro, schedule, and logistics	
2	Applications of visual analytics, basic tasks, data types	
3	Introduction to D3, basic vis techniques for non-spatial data	Project #1 out
4	Data assimilation and preparation	
5	Data reduction and notion of similarity and distance	
6	Visual perception and cognition	
7	Visual design and aesthetics	Project #1 due
8	Dimension reduction	Project #2 out
9	Data mining techniques: clusters, text, patterns	
10	Cluster analysis: numerical data	
11	Cluster analysis: categorical data	
12	Spatial data origins: medical imaging, scientific simulation	
13	Techniques to visualize spatial data: volume visualization	
14	Intro to GPU programming	
15	Techniques to visualize spatial data: flow visualization	Project #3 out
16	Midterm #1	Project #2 due
17	Illustrative rendering	
18	High-dimensional data	Project #3 due
19	High-dimensional data	
20	Principles of interaction	Final project proposal due
21	Hierarchies	
22	Midterm 1 discussion	
23	Visual analytics and sense making	
24	Causality and causal structures	Final Project preliminary report due
25	Time-varying and streaming data	
26	Midterm #2	
27	Maps, evaluation and user studies	
	Final project presentations	Final Project slides and final report due

Example (from Kuchar et al. IEEE CG&A May/June 2006)

Wednesday 28 April 1999; Posted: 11:33 p.m. EDT (03:33 GMT): Another robbery occurred in southwestern Ontario today, making this the fourth robbery in the past few months. Delaware Bank in Brantford was robbed by three masked individuals who stole \$150,000 in currency and several unknown items from the bank's vault. The bank robbery occurred at 2:30, lasting all of five minutes and injuring eight people. All injured parties were taken to the local hospital where one died on arrival. Two people were released and the remaining people are in intensive care. This robbery is similar to a crime spree that started on the Chinese New Year. The first robbery occurred in the morning at Allegiant Bank in Richmond Hill, with the robbers taking more than \$100,000 in currency. The second robbery occurred about two weeks later at Banner Bank in Ajax and was caught on tape. The robbers arrived just as the bank opened, riding in a white van and wearing black ski masks and black outfits, and carrying automatic machine guns. During the first three minutes, the robbers instructed all of the patrons to face the wall and place their hands on their heads. While two of the robbers watched the patrons, the other robber took the bank manager and instructed him to open the vault. Other video captured the movement in the vault. The vault was opened about five minutes after the robbers arrived. The safe was blown two minutes later, and the robber removed only two safety deposit boxes and placed them in a bag. He then continued to club the manager and was back upstairs, yelling instructions to his buddies as they left the bank, not more than 20 minutes after they arrived. What was unusual was that no alarms were sounded. The third robbery happened at night at Carter Bank in Brampton, with only nearby homeowners mentioning that they do not remember hearing the bank alarm, only dogs barking for a while....

Question

Why are temporal relationships difficult to discern?

Time Series Data Are Everywhere

Temporal relationships can be difficult to discern because

- temporal ordering can be hard to determine
- event may occur in spatially disjoint locations
- what came before what – cause and effect
- what time shifts are acceptable/plausible?

To understand temporal relationships, an analyst:

- might need to reread the paragraph many times
- needs to cognitively make inferences between pieces of information

Visualization is key to externalize these relationships

- put it all out on “paper” and reason with it

Time is Special

What actually is time?

- how can one work with the metaphor of time's flow?
- what is the proper formalism that captures the time's special role in reasoning

The time variable is different than the other variables

- people consider it as an independent quantity
- can't go back in time
- our perception is that we have no control over it
- time is an ever-present thread that can help tie events together

Time as a Reference Frame

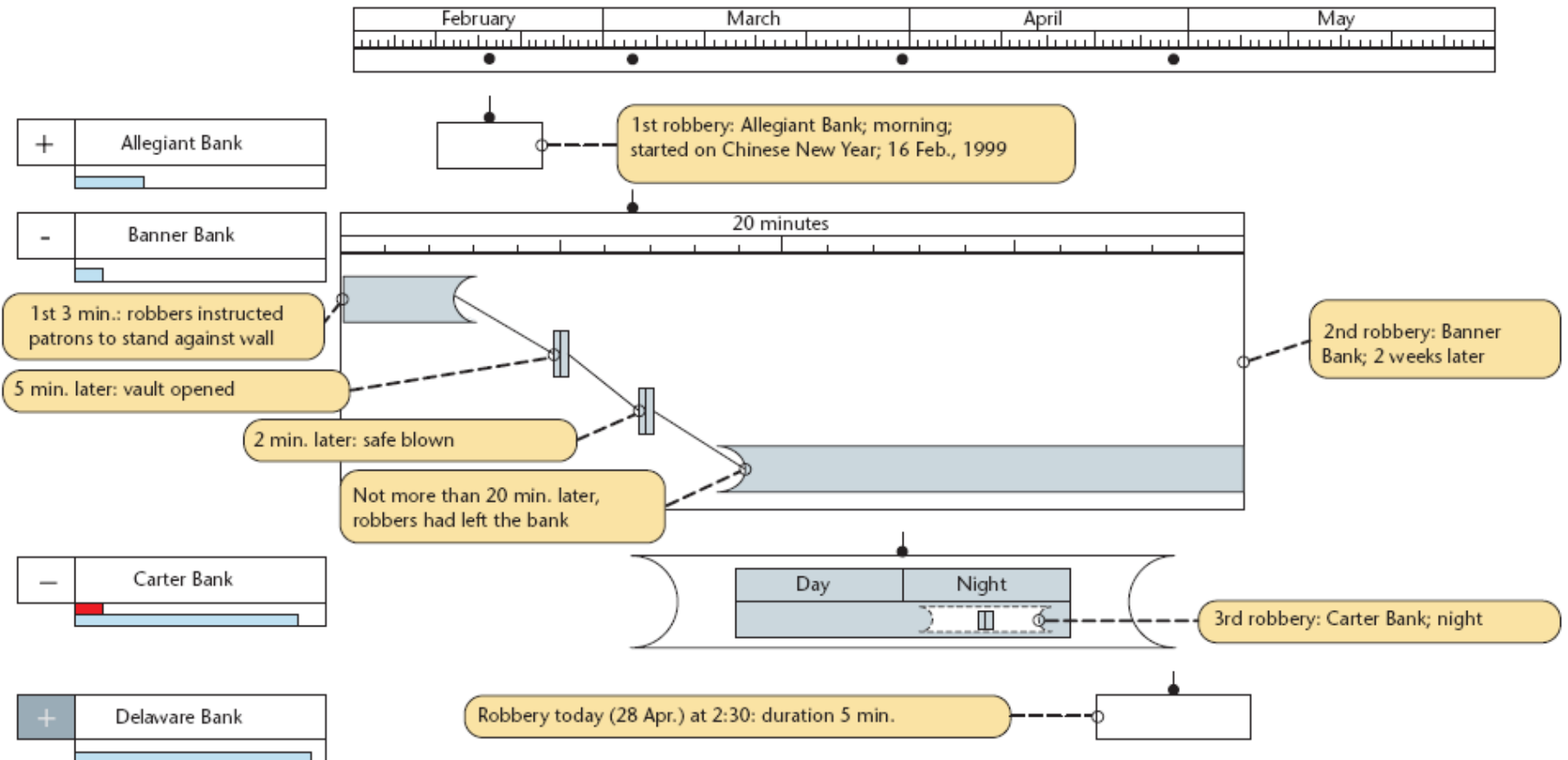
Calendars and time have reference frames

- Gregorian, Greenwich, EDT

Time is also often used in relative terms:

- “today”, “yesterday”, “fortnight”, “before Tuesday”, ...
- must normalize different reference systems into a common framework
- but it might be unknown what reference system was used individually
- “This robbery is similar to a crime spree that started on the Chinese New Year ...” – when is Chinese New Year?
- causes ambiguities, uncertainties, biases, conflicts

Back to Example (from Kuchar et al. IEEE CG&A May/June 2006)



Time: Tasks and Taxonomies

Often asked questions:

- when was something greatest/least?
- is there a pattern?
- are two series similar?
- does a data element exist at time t , and when?
- how long does a data element exist and how often?
- how fast are data elements changing
- in what order do they appear?
- do data elements exist together?

Different types of time series data:

- discrete vs. interval
- linear vs. cyclic
- ordinal vs. continuous
- ordered vs. branching vs. time with multiple perspectives

Traditional Time Series Visualizations

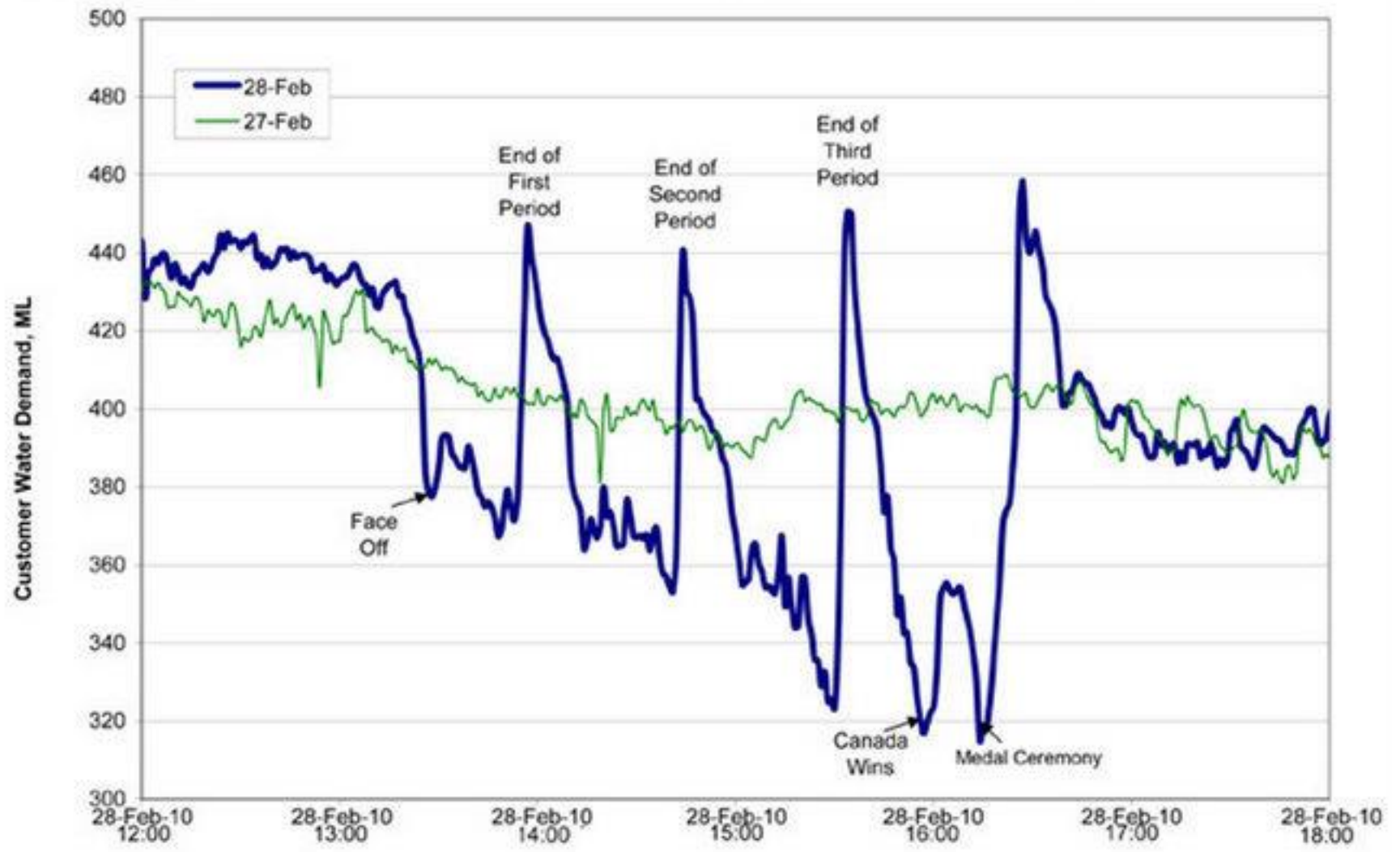
NVIDIA stock vs. NASDAQ (from yahoo! finance)



Fun one... (found in J. Stasko lecture)



Water Consumption in Edmonton During Olympic Gold Medal Hockey Game

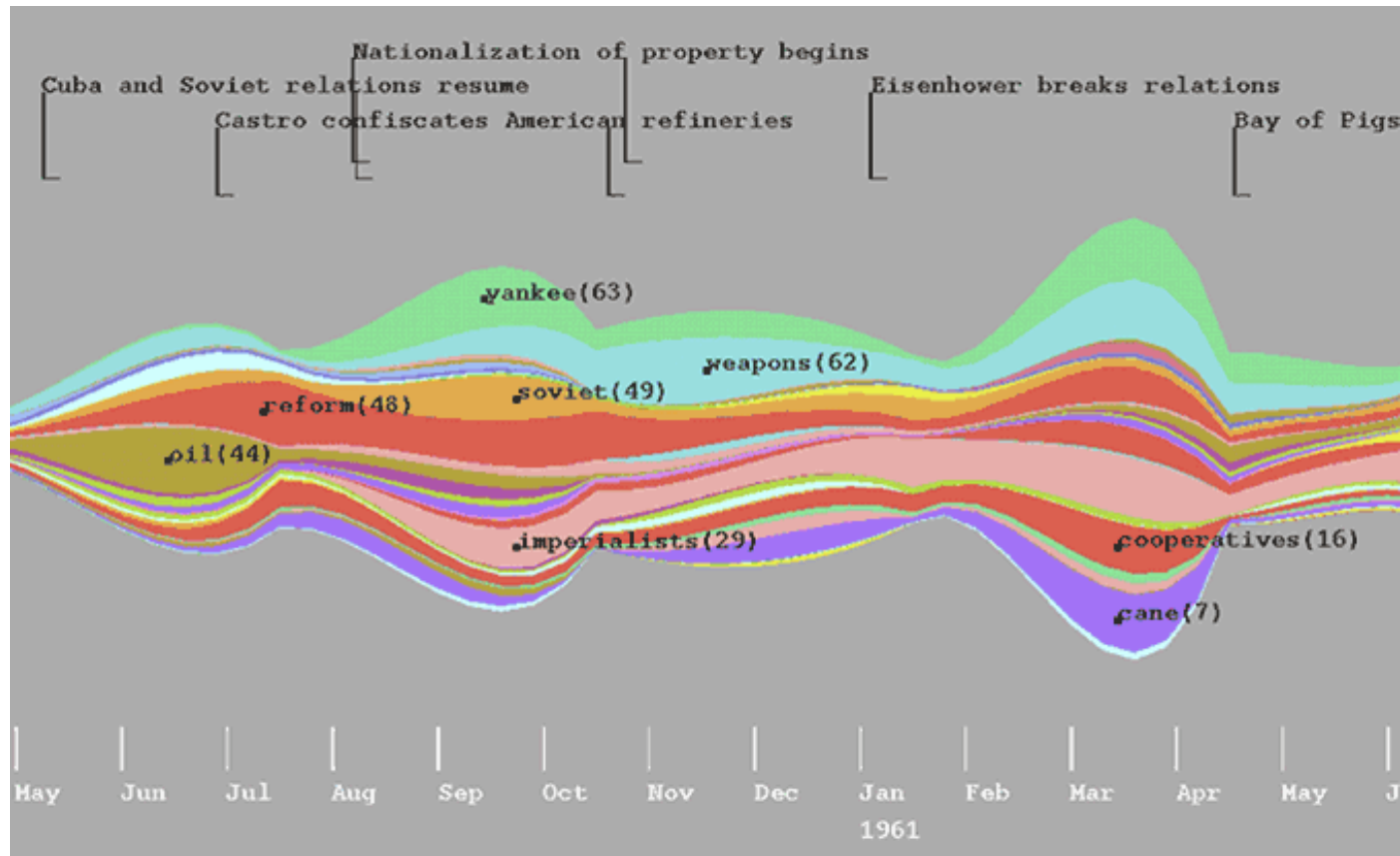


Next...

A few good visualization metaphors for time

- there are quite a few of them...

ThemeRiver (Havre et al., 2002)



River widens or narrows to depict changes in the collective strength of selected themes in the underlying documents. Individual themes are represented as colored "currents" flowing within the river.

Example shown here: newspaper themes around the Cuban Missile crisis

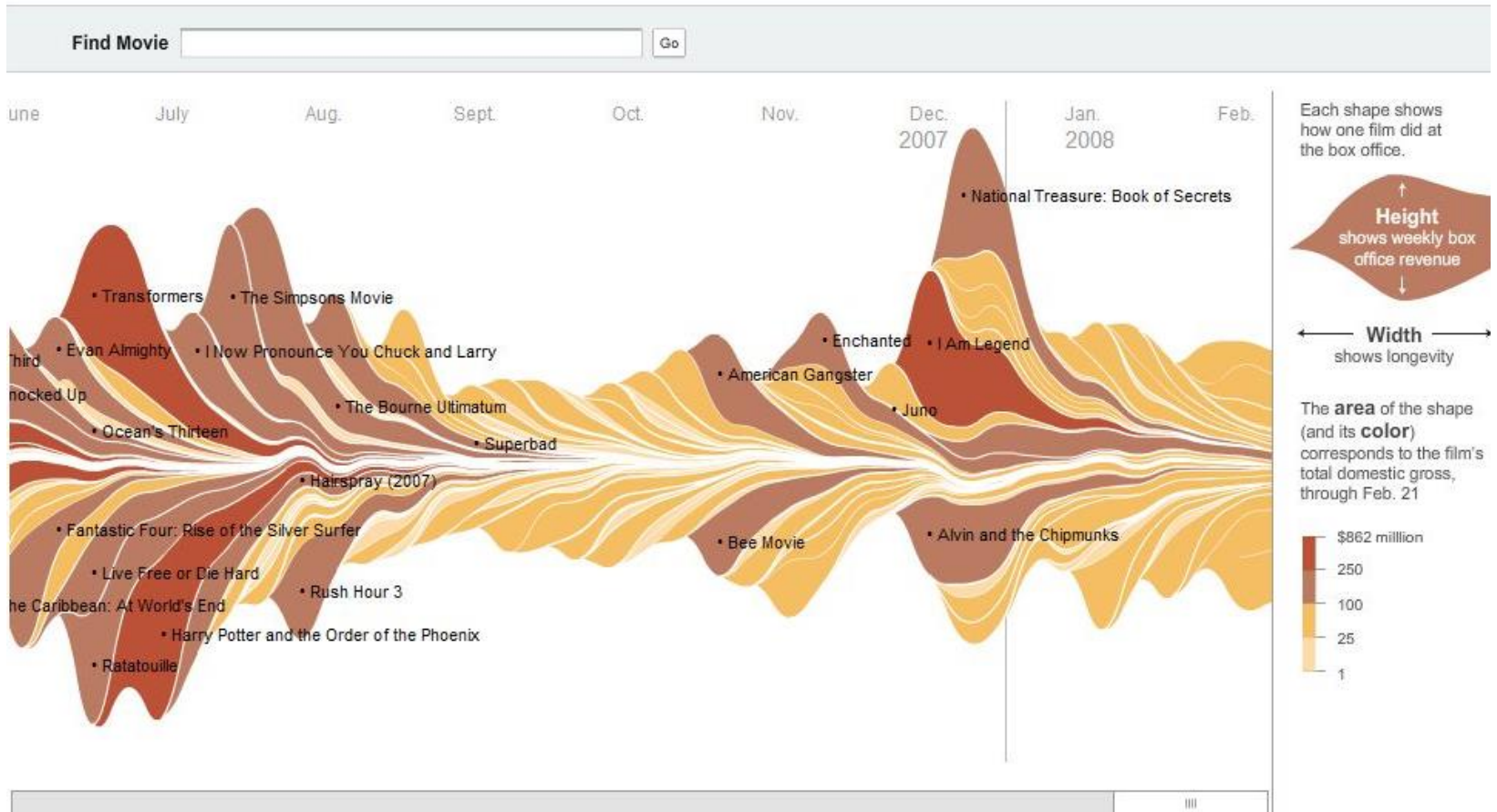
Stream Graphs

February 23, 2008

[SIGN IN TO E-MAIL OR SAVE THIS](#) | [FEEDBACK](#)

The Ebb and Flow of Movies: Box Office Receipts 1986 — 2008

Summer blockbusters and holiday hits make up the bulk of box office revenue each year, while contenders for the Oscars tend to attract smaller audiences that build over time. Here's a look at how movies have fared at the box office, after adjusting for inflation.



Stacked Area Charts

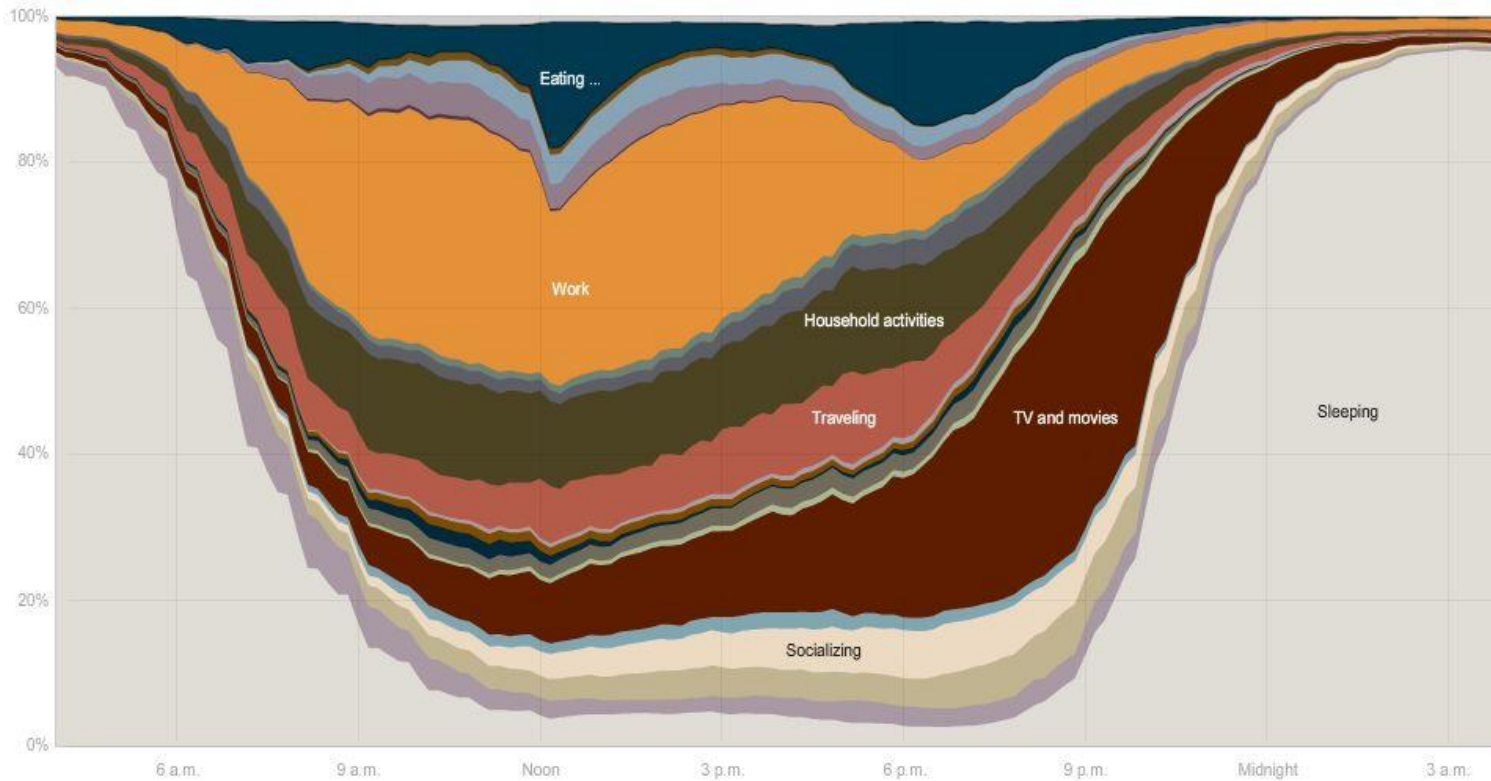
How Different Groups Spend Their Day

The American Time Use Survey asks thousands of American residents to recall every minute of a day. Here is how people over age 15 spent their time in 2008. [Related article](#)

Everyone

Sleeping, eating, working and watching television take up about two-thirds of the average day.

Everyone	Employed	White	Age 15-24	H.S. grads	No children
Men	Unemployed	Black	Age 25-64	Bachelor's	One child
Women	Not in lab...	Hispanic	Age 65+	Advanced	Two+ children



Name Voyager (<http://www.babynamewizard.com>)

New! Try the **NameMapper** to see where your favorite names are being used, and **Namipedia** for full info on every name

>

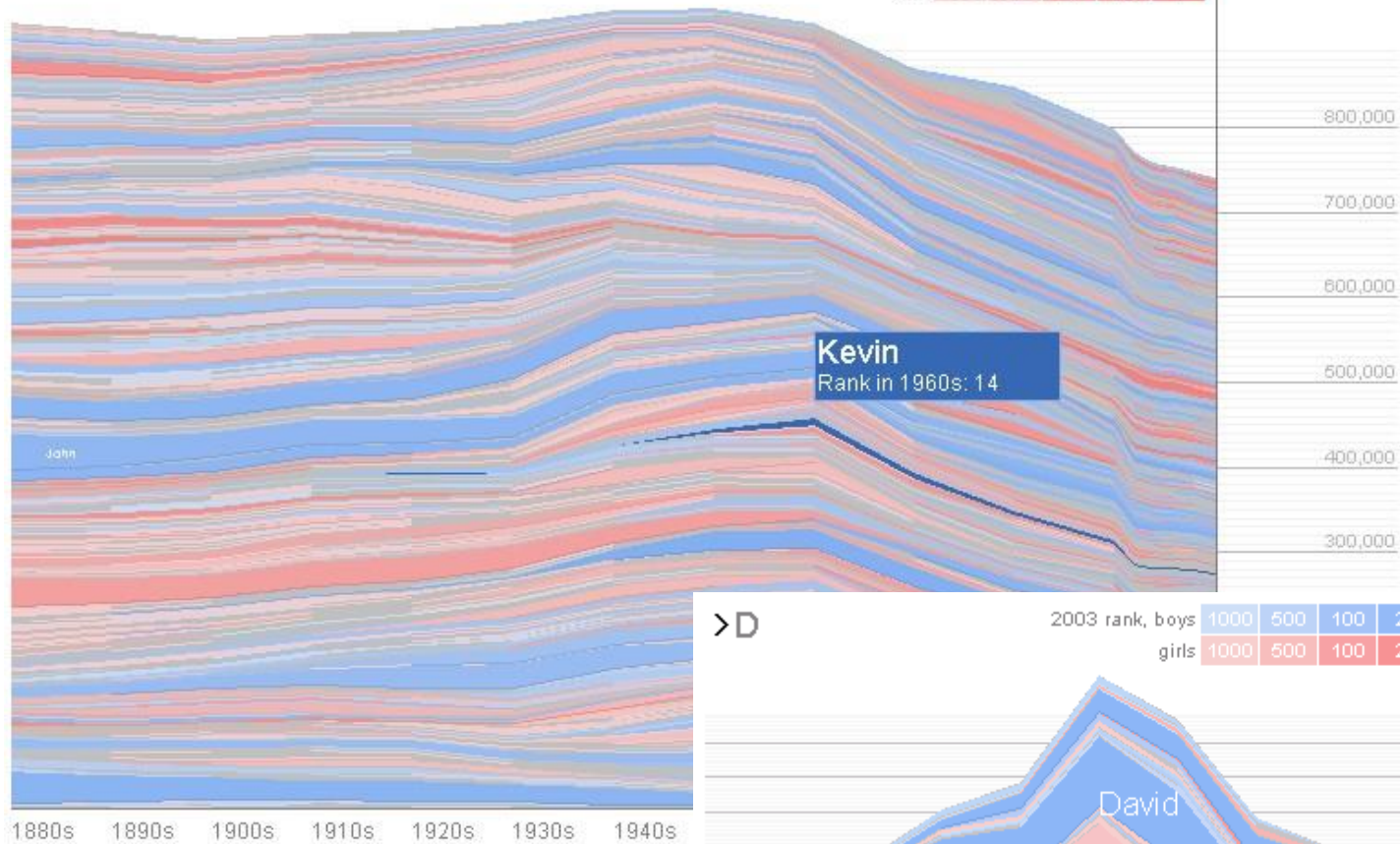
boys girls both

2007 rank, boys

1000	500	100	25	1
------	-----	-----	----	---

girls

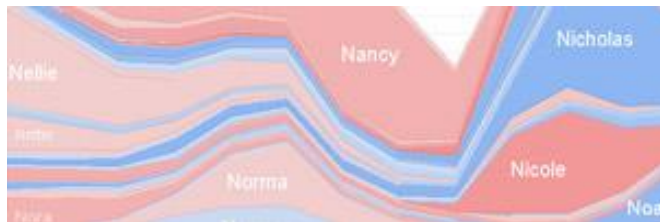
1000	500	100	25	1
------	-----	-----	----	---



Kevin
Rank in 1960s: 14

John

1880s 1890s 1900s 1910s 1920s 1930s 1940s



>D

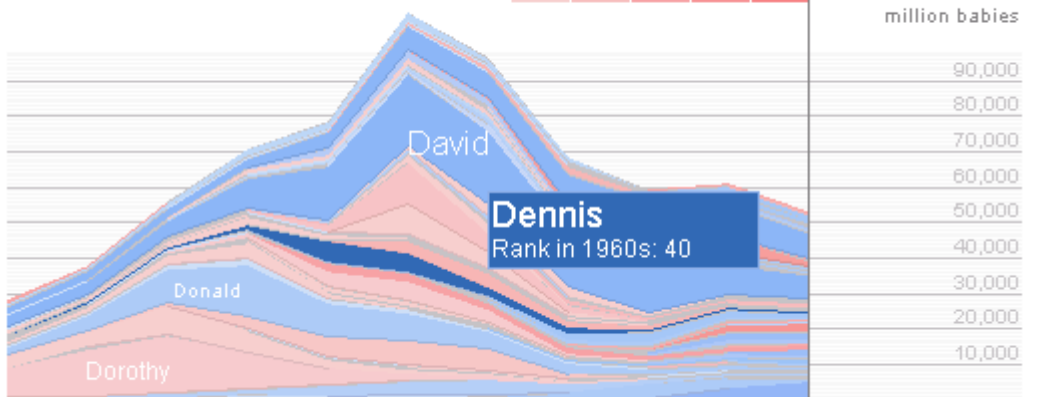
2003 rank, boys

1000	500	100	25	1
------	-----	-----	----	---

girls

1000	500	100	25	1
------	-----	-----	----	---

Names starting with "D", per million babies



David

Dennis
Rank in 1960s: 40

Donald

Dorothy

1900s 1910s 1920s 1930s 1940s 1950s 1960s 1970s 1980s 1990s 2003

Question

Can you tell me who is who?

- I tell you all the names there are and the age of each person
- can you assign them? (hint: use the Name Voyager)



Medical Data

Medical data are often displayed along time

- natural to humans
- progression of disease
- appearance of symptoms
- time course of treatment and outcome
- but also time signals (ECG, blood pressure, etc.)

A popular example is Lifelines and Lifelines2


- Shneiderman and Plaisant et al.
- <http://www.cs.umd.edu/hcil/lifelines/>

LifeLines: Patient-Centric

In progress version of LifeLine Frame [not for demo purposes - Microsoft Internet Explorer - [Working Offline]

File Edit View Go Favorites Help

Back Links Best of the Web Channel Guide Customize Links Internet Explorer News Internet Start

 **Linda Simpson**
Female 40

Line from input file: %:3-10-1997,3-12-1997_black,p10,Sonogram,images/babysonogra

LifeLine

92 93 94 95 96 97

Notes: Tobacco, Depression, Lyme, Arthritis, Obesity, AtrialFlutter, Flu, Pneumonia, KneePain, Fatigue>Diabetes, Diabe, Pregnancy

Hosp.: Appendectomy, Pneumonia, KneeSurgery

Tests: BloodEKG, EKG, Xray, Blood, Blood, Blood, Sonogr

Meds.: Prozac, Heartdrug, Ventolyn, Antib., Advil, Advil, Insulin, Insulir

Others: PhysicalTherapy, LowSaltFatDiet

Immun.: TBtest, Tetanos, Flu

DATE: 5-18-88 PATIENT ID: [redacted] 17aa/3.8

LifeLine Control Panel

Layout Label

Default

Quick Compact


Slow Compact

Chronologically Ordered

Event Ordered

Apply OK Cancel

Warning: Applet Window



LifeLines2: Pattern-Centric

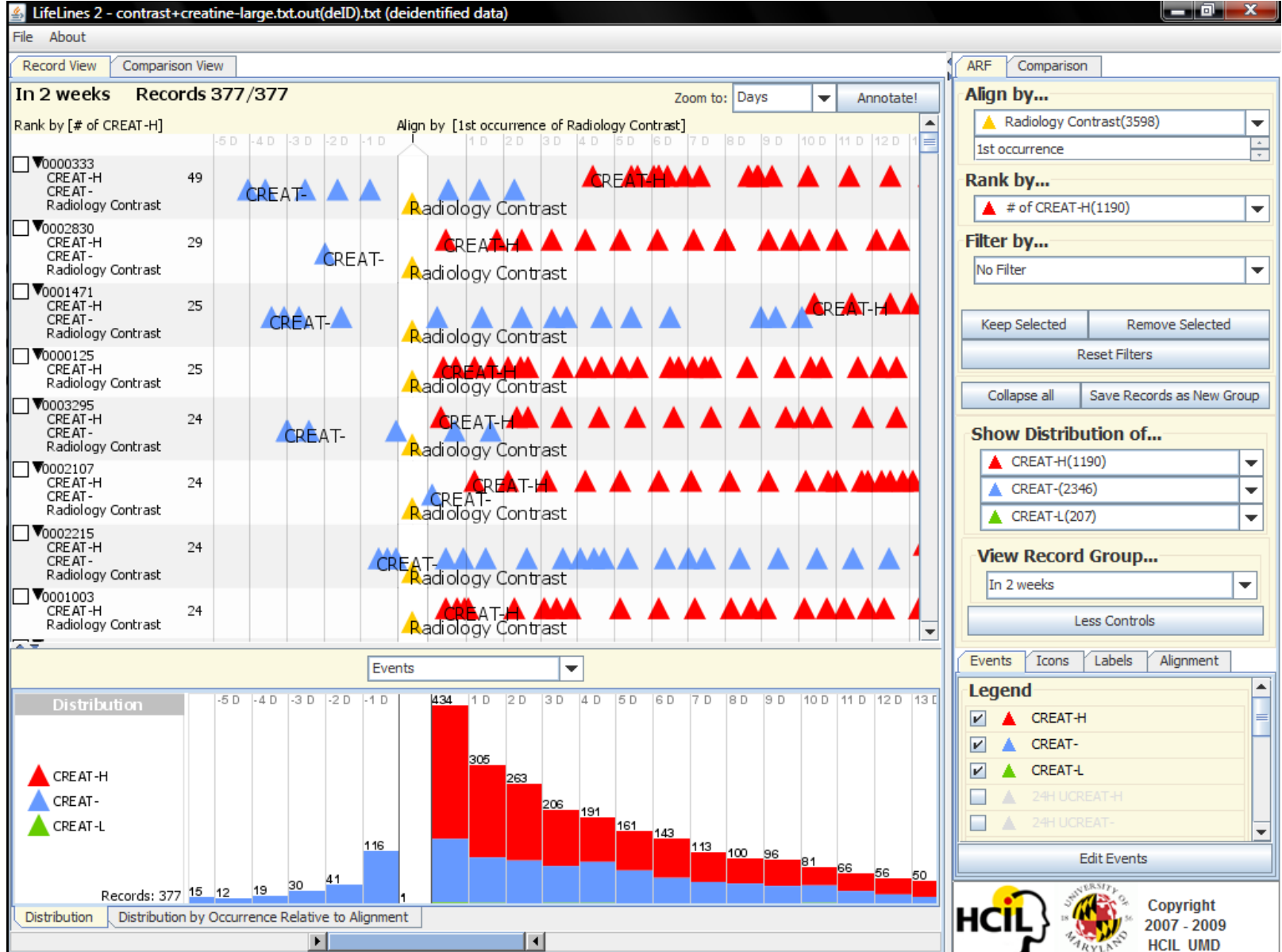
Goals:

- bring out temporal categorical patterns across multiple records
- categorical event data such as complaints, diagnoses, treatments
- play important roles in health providers decision making

Features

- allows users to manipulate multiple records simultaneously
- understand relative temporal relationships across records
- 3 operators: align, rank, filter
- temporal summaries allow multiple groups of records to be compared

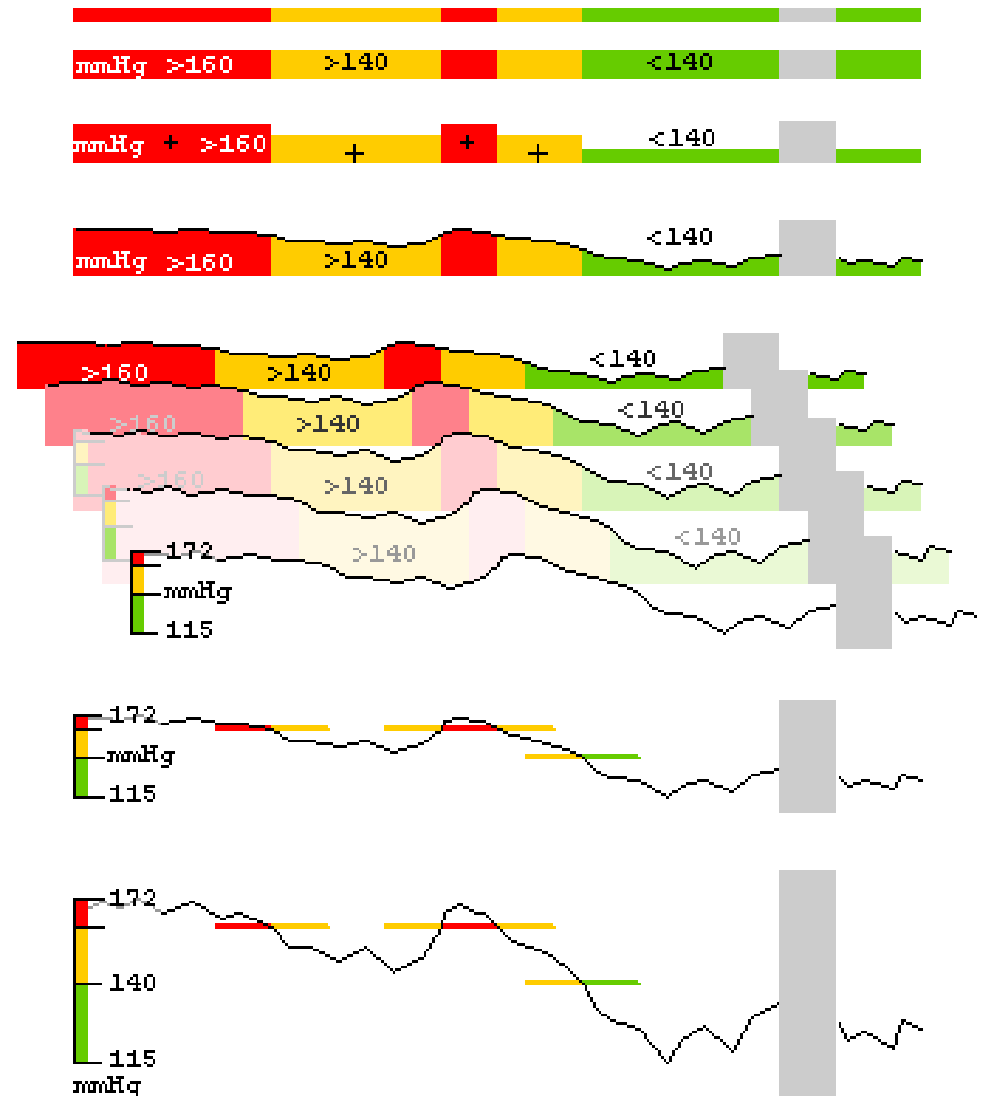
LifeLines2: Screenshot



Multi-Scale and Abstractions (Aigner et al., IEEE TVCG, 2008)

Deal with different levels of detail

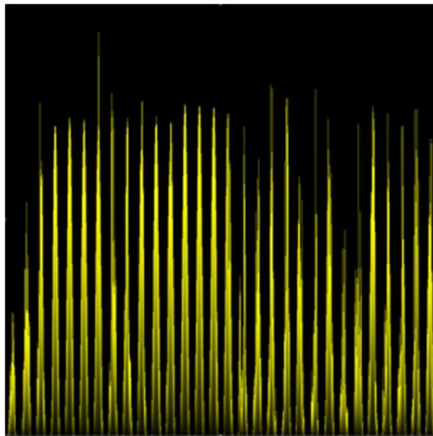
- illustrative abstraction
- overview + detail
- used here for medical data



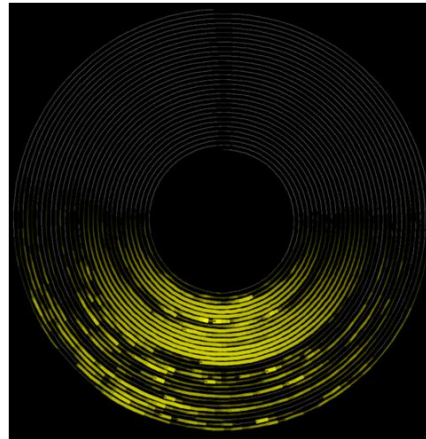
Cyclic Patterns

Time data are often cyclic

- spiral displays are good to bring out cyclic patterns
- one period per loop (for example, a year)



linear layout



radial layout

sunshine pattern

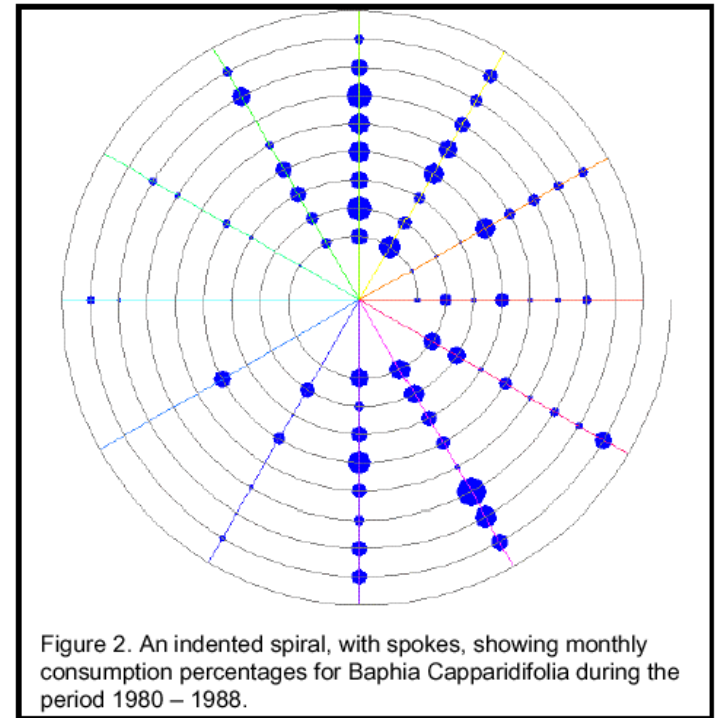
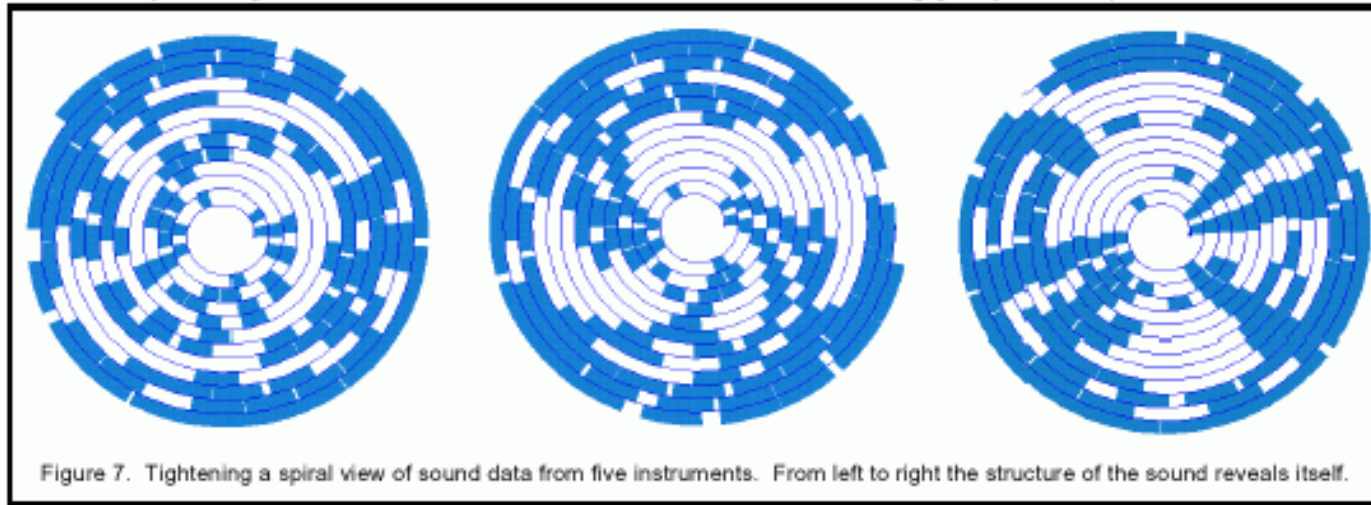


Figure 2. An indented spiral, with spokes, showing monthly consumption percentages for Baphia Capparidifolia during the period 1980 – 1988.

Cyclic Patterns

May have to play around to discover the cycles



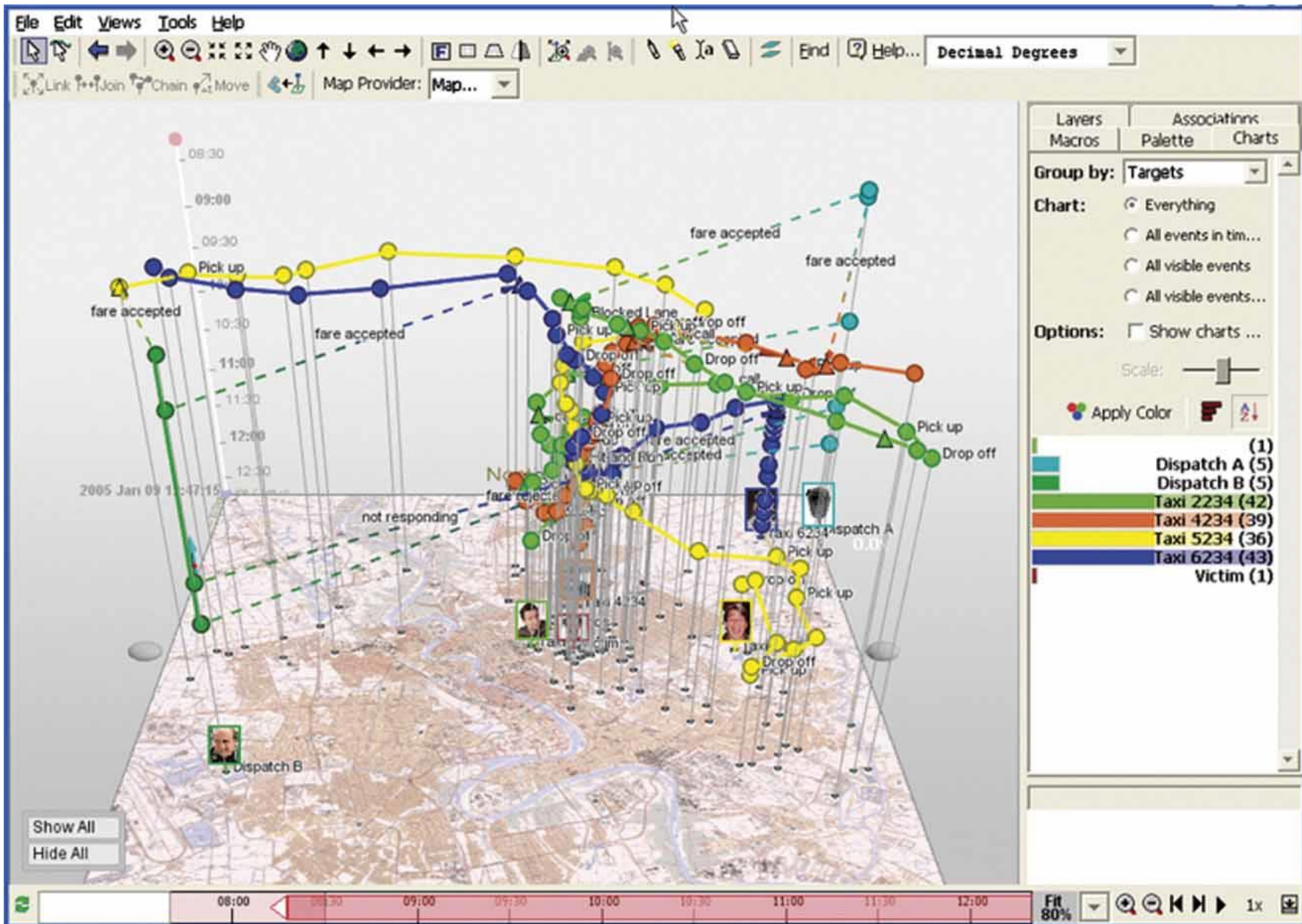
Combine Space and Time

OculusInfo Geotime application

- events are represented in an X,Y,T coordinate space
- the X,Y plane shows geography
- the vertical T axis represents time
- events animate in time vertically through the 3-D space as the time slider bar is moved.



Geotime



Interaction

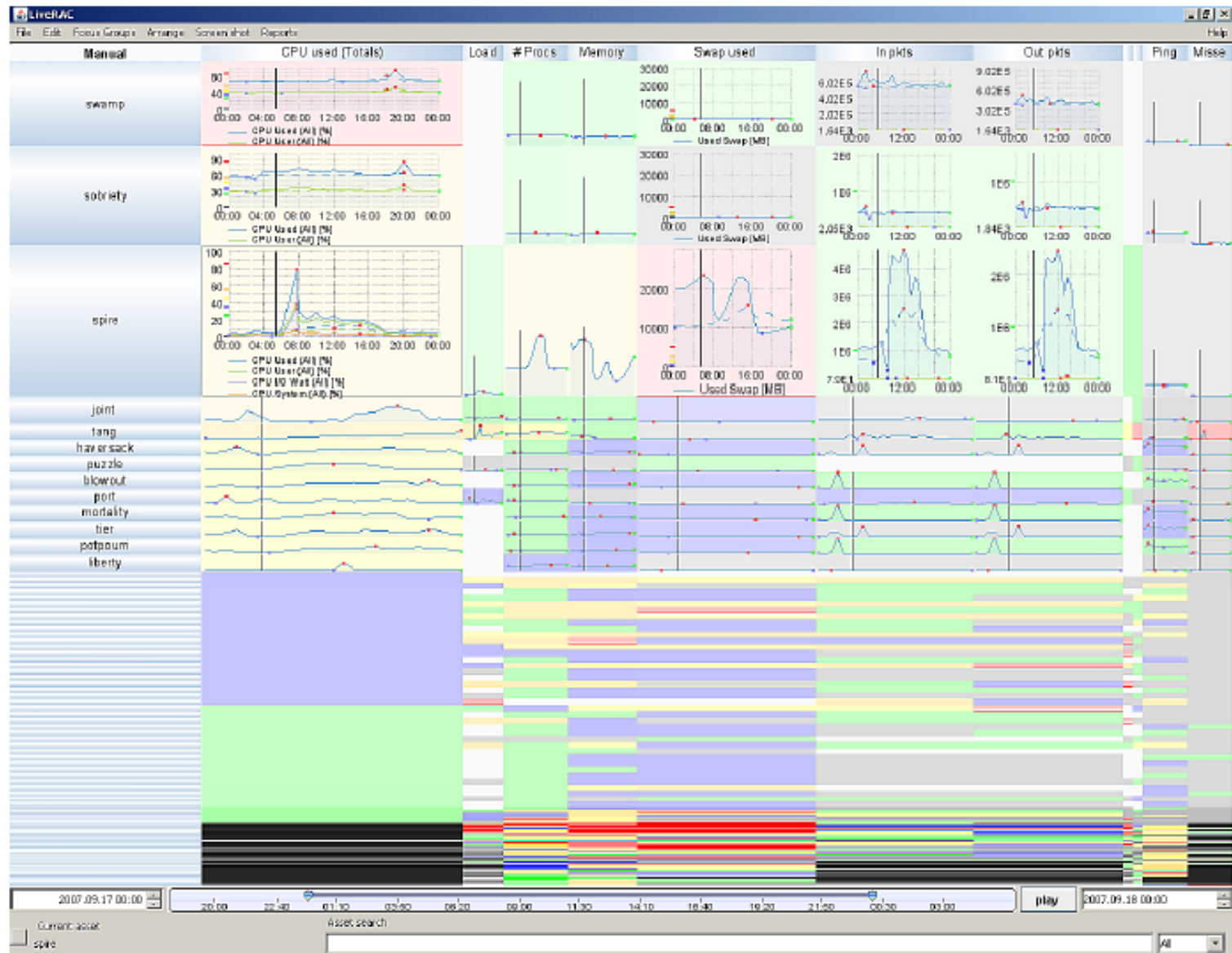
As complexity increases, interaction capabilities are key

- show more context of what else was going on at that time
- likely have to abstract some of the information
- allow several different levels of detail at once
- allow drill-down for details
- use dashboard design with many linked information displays

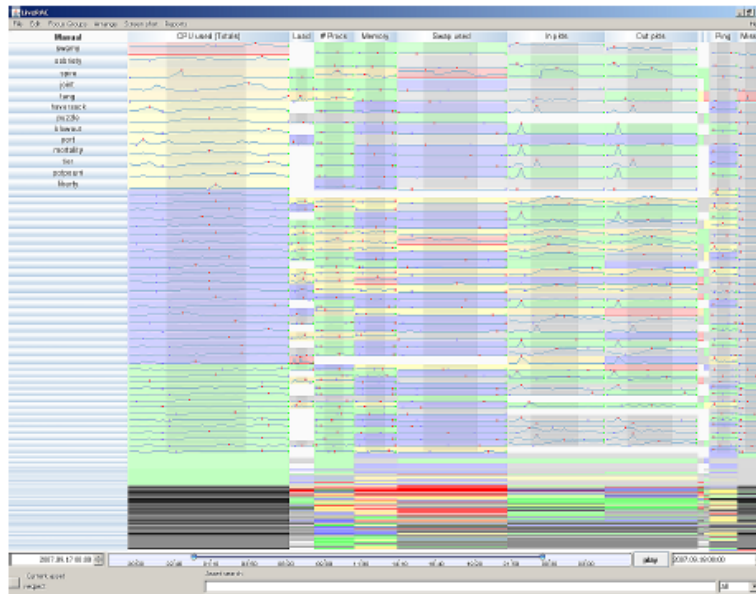
Example: Computer system management

- LiveRAC system (McLachlan et al.)
- next two slides

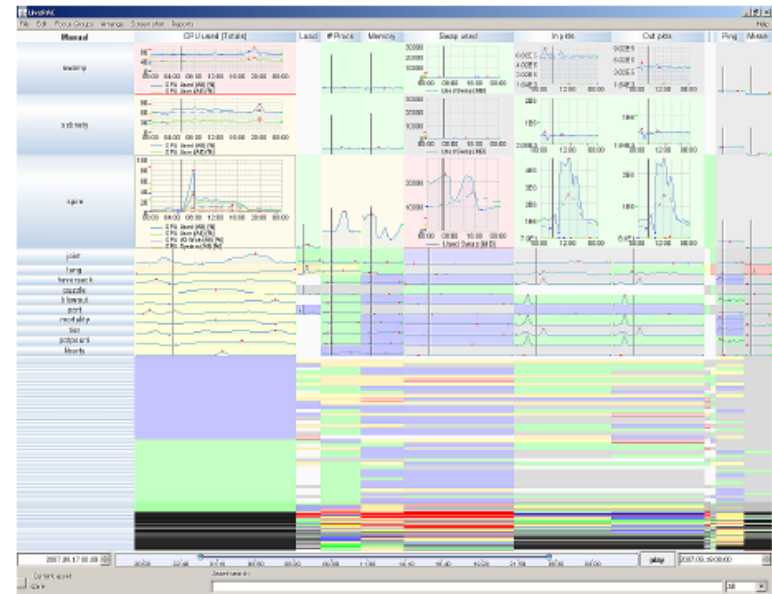
LiveRAC



LiveRAC



(a)



(b)

Figure 3. LiveRAC shows a full day of system management time-series data using a reorderable matrix of area-aware charts. Over 4000 devices are shown in rows, with 11 columns representing groups of monitored parameters. (a): The user has sorted by the maximum value in the *CPU* column. The first several dozen rows have been stretched to show sparklines for the devices, with the top 13 enlarged enough to display text labels. The time period of business hours has been selected, showing the increase in the *In pkts* parameter for many devices. (b): The top three rows have been further enlarged to show fully detailed charts in the *CPU* column and partially detailed ones in *Swap* and two other columns. The time marker (vertical black line on each chart) indicates the start of anomalous activity in several of *spire*'s parameters. Below the labeled rows, we see many blocks at the lowest semantic zoom level, and further below we see a compressed region of highly saturated blocks that aggregate information from many charts.

LiveRAC: Interactive Visual Exploration of System Management Time-Series Data

Next – Streaming Data

Time series data with no end...

Types of Streaming data

Transaction streams

- credit card, point-of-sale transaction
- at a supermarket, or online purchase of an item

Web click-streams

Social streams

- online social networks such as Twitter
- speed and volume of the stream typically scale super-linearly with the number of actors

Network streams

- communication networks contain large volumes of traffic streams
- often mined for intrusions, outliers, or other unusual activity

Challenges (1)

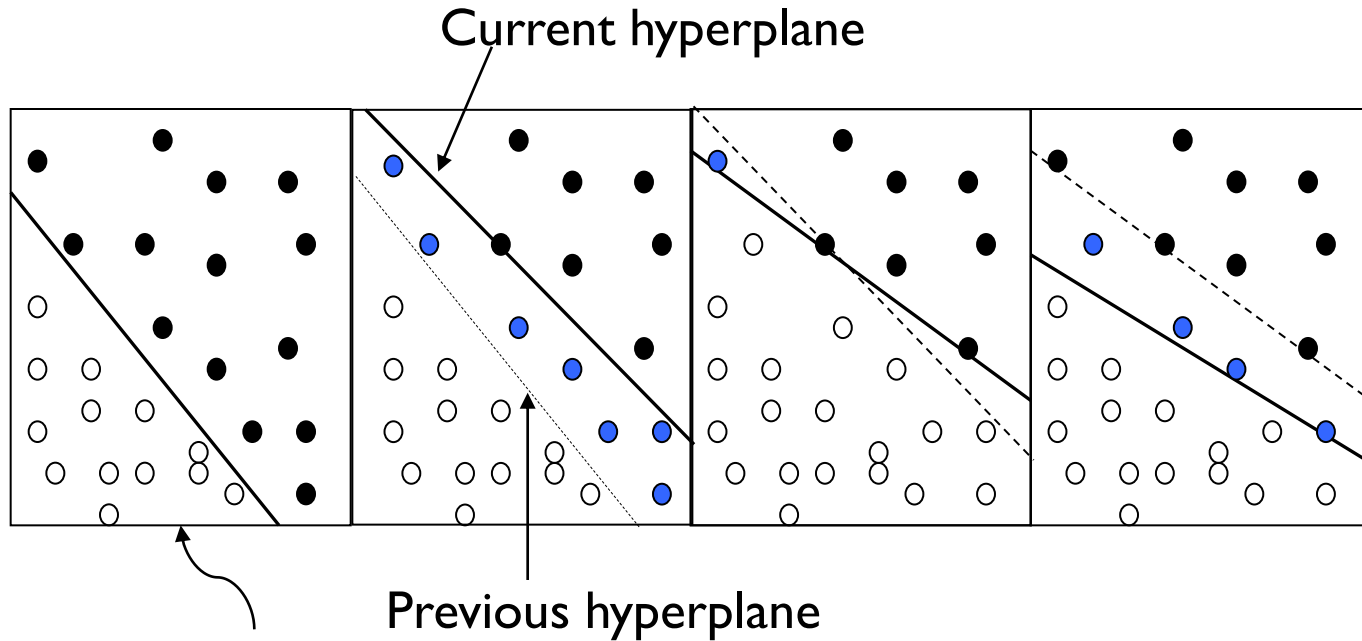
One-pass constraint

- data is generated continuously and rapidly
- it is assumed that the data can be processed only once
- archival for future processing is not possible
- prevents use of iterative mining or model building algorithms that require multiple passes over the data

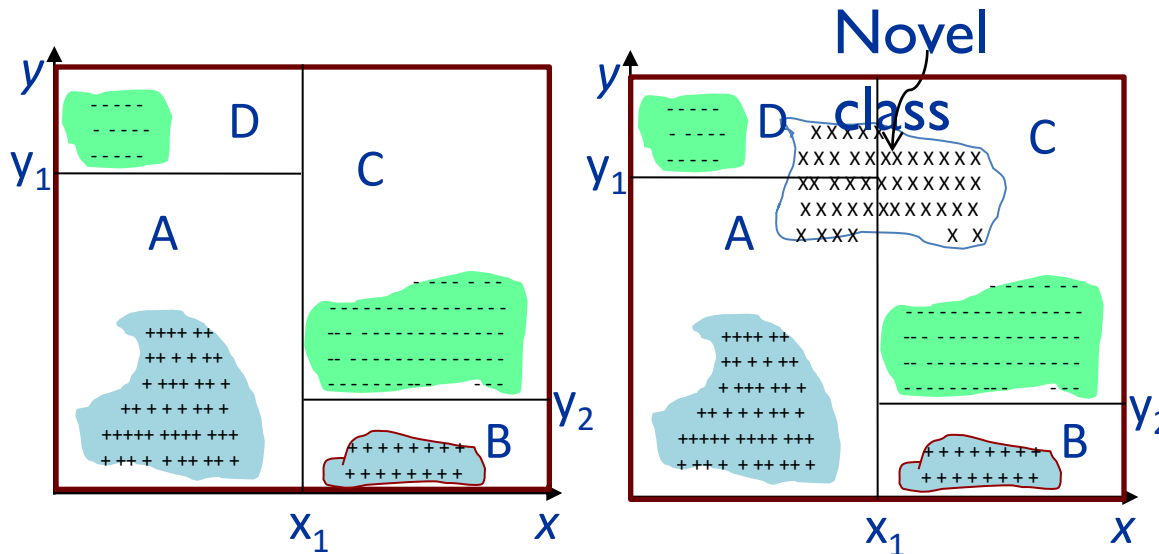
Concept drift, concept evolution, feature evolution

- data may evolve over time
- various statistical properties, such as correlations between attributes, correlations between attributes and class labels, and cluster distributions may change over time

Concept Drift



Concept Evolution



Classification rules:

R1. if $(x > x_1 \text{ and } y < y_2)$ or $(x < x_1 \text{ and } y < y_1)$ then class = +

R2. if $(x > x_1 \text{ and } y > y_2)$ or $(x < x_1 \text{ and } y > y_1)$ then class = -

Existing classification models misclassify novel class instances

Key Element – Online Synopsis

Virtually all streaming methods use an online synopsis (summary) construction approach in the mining process

- create an online synopsis that is then leveraged for mining

Many different kinds of synopsis approaches

- the nature of a synopsis highly influences the type of insights that can be mined from it.

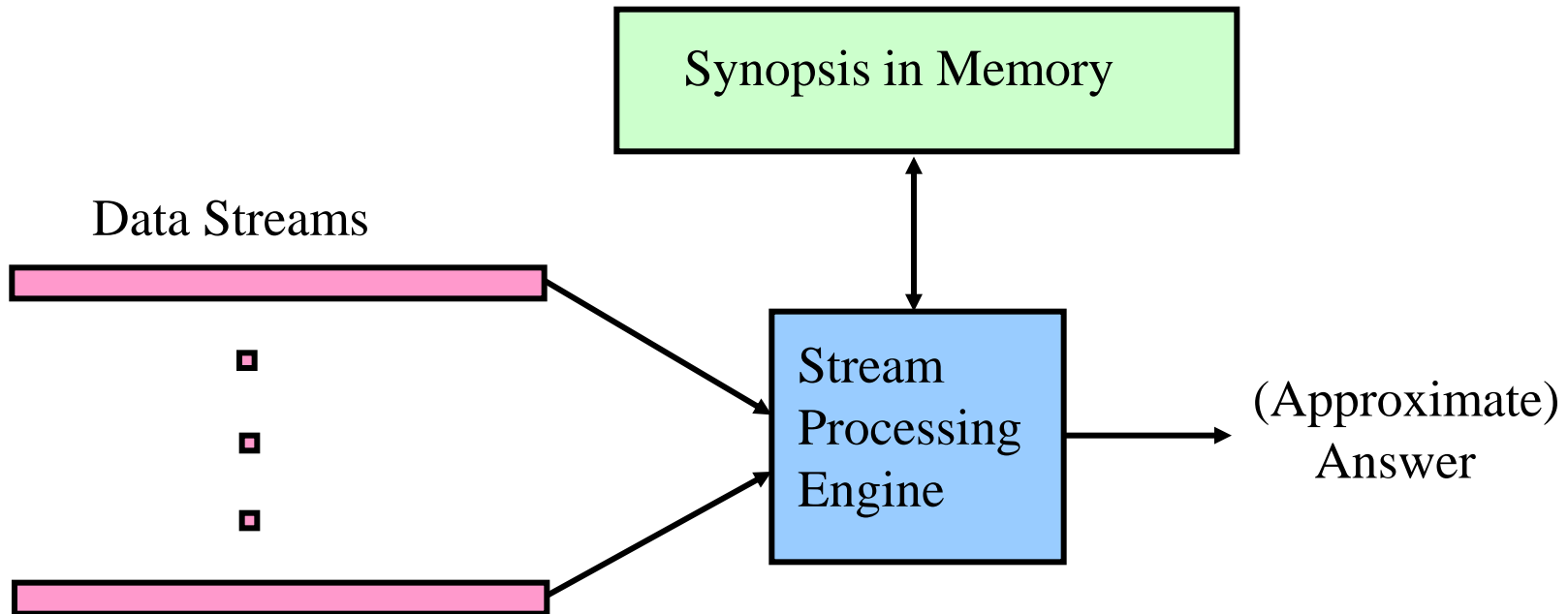
Some examples of synopsis structures:

- random samples
- bloom filters, sketches
- distinct element-counting data structures
- traditional data mining applications, such as clustering

Synopsis

Stream processing requirements

- single pass: each record is examined (sampled) at most once
- bounded storage: limited Memory (M) for storing synopsis
- real-time: per record processing time (to maintain synopsis) must be low



Samples

A data stream is a (massive) sequence of elements e_1, e_2, \dots, e_n

Idea:

- a small random sample S of the data often well represents all the data
- many different ways to obtain this sample

Data stream:

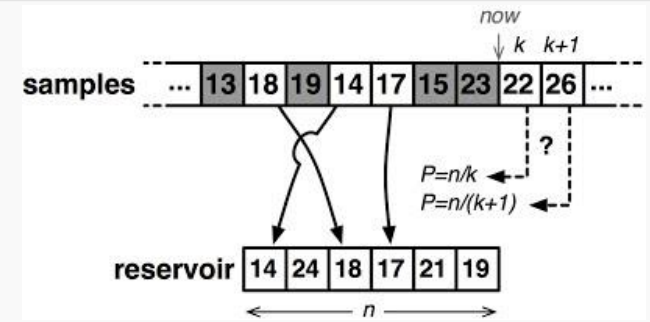
9	3	5	2	7	1	6	5	8	4	9	1
---	---	---	---	---	---	---	---	---	---	---	---

Sample S :

9	5	1	8
---	---	---	---

Reservoir Sampling

```
/*  
  S has items to sample, R will contain the result  
*/  
ReservoirSample(S[1..n], R[1..k])  
  // fill the reservoir array  
  for i = 1 to k  
    R[i] := S[i]  
  
  // replace elements with gradually decreasing probability  
  for i = k+1 to n  
    j := random(1, i) // important: inclusive range  
    if j <= k  
      R[j] := S[i]
```



Probabilities

- k/i for the i^{th} sample to go into the reservoir
- $1/k \cdot k/i = 1/i$ for the j^{th} reservoir element to be replaced
- k/n for all elements in the reservoir after n has been reached
- can be shown via induction

A good algorithm to use for streaming data when n is growing

Sliding Window Approach

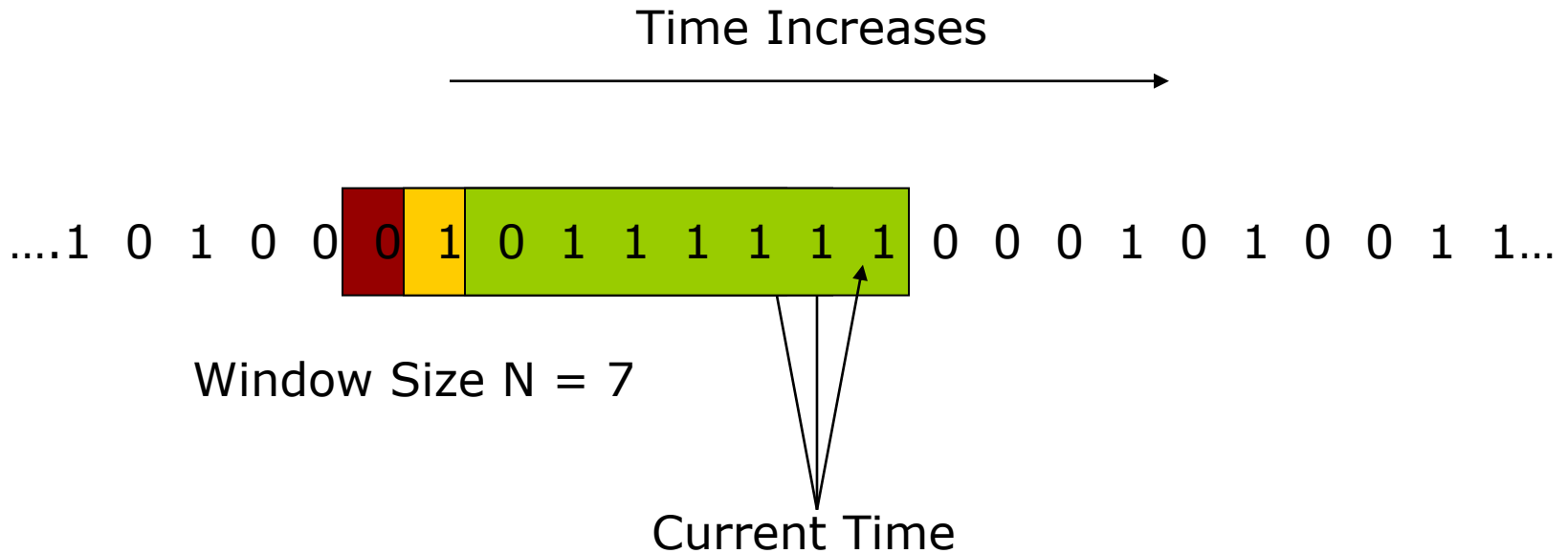
Background:

- some applications rely on ALL historical data
- but for most applications, OLD data is considered less relevant and could skew results from NEW trends or conditions
 - new processes/procedures
 - new hardware/sensors
 - new fashion trends

Sliding Windows Model

- only last “N” elements are considered
- incorporate examples as they arrive
- the record “expires” at time $t+N$ (N is the window length)

Sliding Window Approach



CluStream Clustering

The concept drift in an evolving data stream changes the clusters significantly over time

- need a clustering algorithm that can deal with this
- CluStream is such an algorithm

CluStream's online microclustering clustering stage

- processes the stream in real time to continuously maintain summarized but detailed (micro-)cluster statistics of the stream

CluStream's offline macroclustering stage

- further summarizes these detailed clusters
- provides the user with a more concise understanding of the clusters over different time horizons and levels of temporal granularity.

Microclustering Algorithm

There are k microclusters

- a new data point either needs to be absorbed by a microcluster, or it needs to be put in a cluster of its own

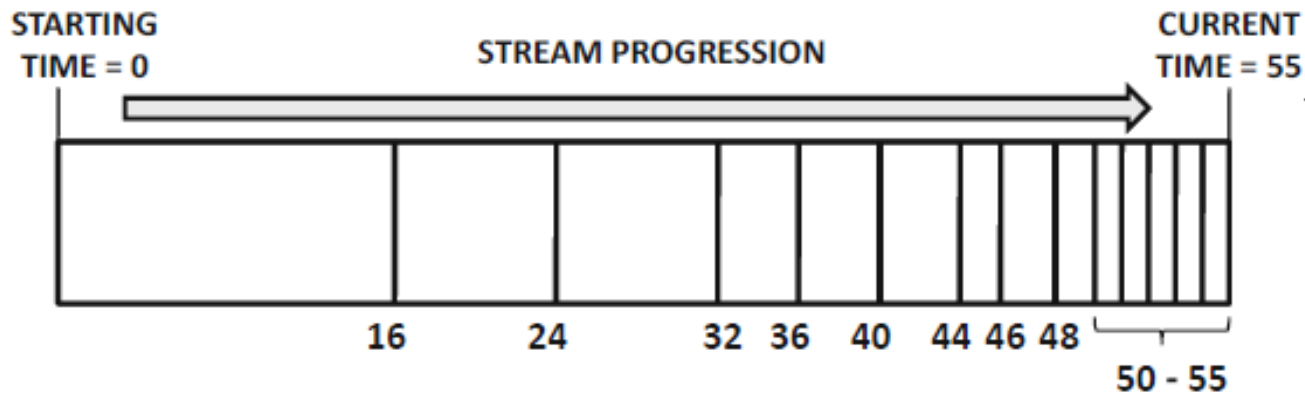
Algorithm

- determine distance of the new data point to all current microcluster centroids
- assign the point to the closest cluster and update the statistics
- if the point does not fall within the maximum boundary of any microcluster create a new microcluster
- to create this new microcluster, the number of other microclusters must be reduced by 1 to free memory availability
- achieve this by either deleting an old microcluster or merging two of the older clusters
- decide by examining the staleness (using the time stamp statistics) of the different clusters, and the number of points in them
- determine whether one of them is “sufficiently” stale to merit removal
- if no microcluster is stale, then a merging of the two microclusters is initiated

Pyramidal Time Frame

Store microclusters statistics periodically to enable time horizon-specific analysis of the clusters

- the microcluster snapshots are stored at varying levels of granularity depending on the recency of the snapshot



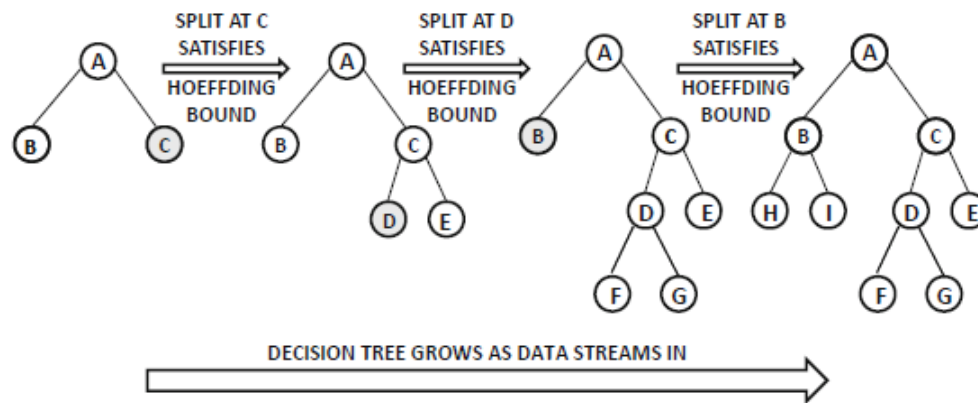
Other Stream Mining Issues

Streaming outlier (anomaly) detection

- use time windows and k-nearest neighbor scores
- new concepts or trends can manifest themselves as outliers in the onset

Streaming classifiers

- the *Hoeffding tree* is constructed incrementally by growing the tree simultaneously with stream arrival.



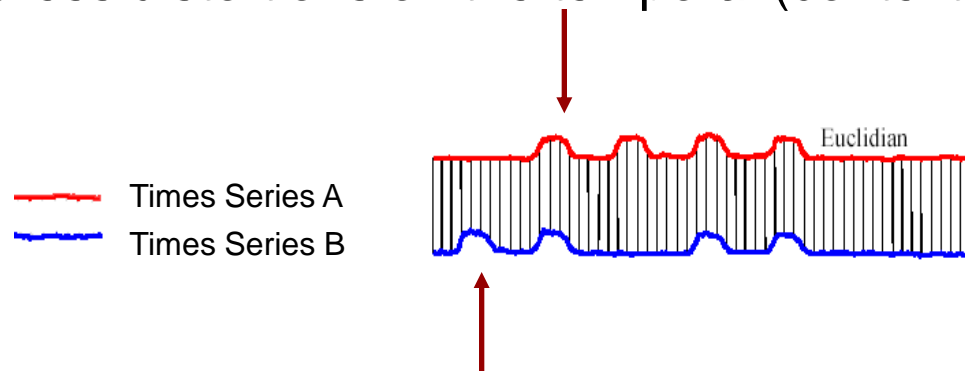
L_p Norm and its Shortcomings

Standard pairwise distance

$$Dist(\bar{X}, \bar{Y}) = \left(\sum_{i=1}^n |x_i - y_i|^p \right)^{1/p}$$

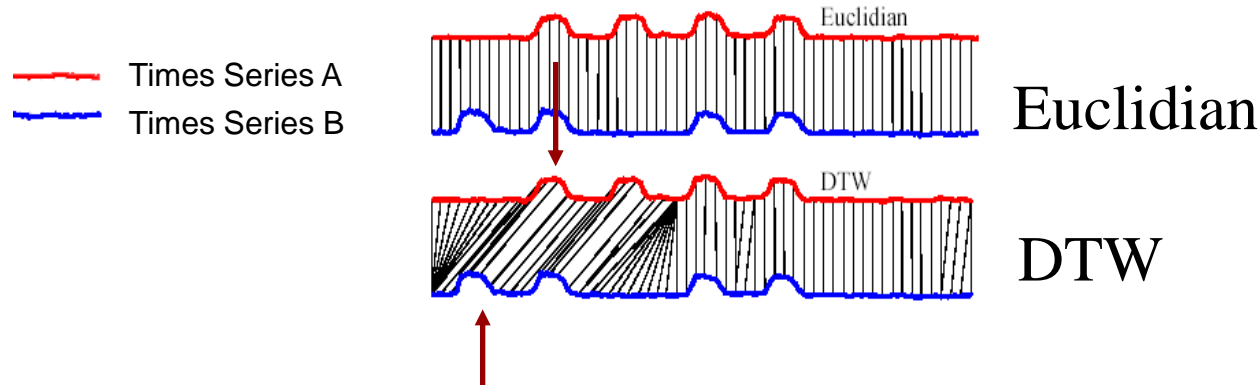
Shortcomings:

- designed for time series of equal length
- cannot address distortions on the temporal (contextual) attributes



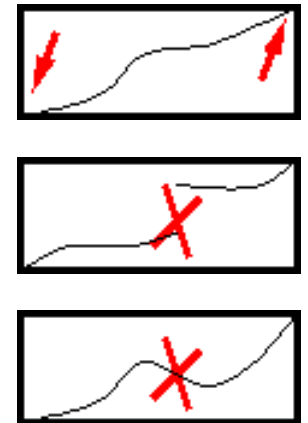
Dynamic Time Warping Distance

Can better accommodate local mismatches

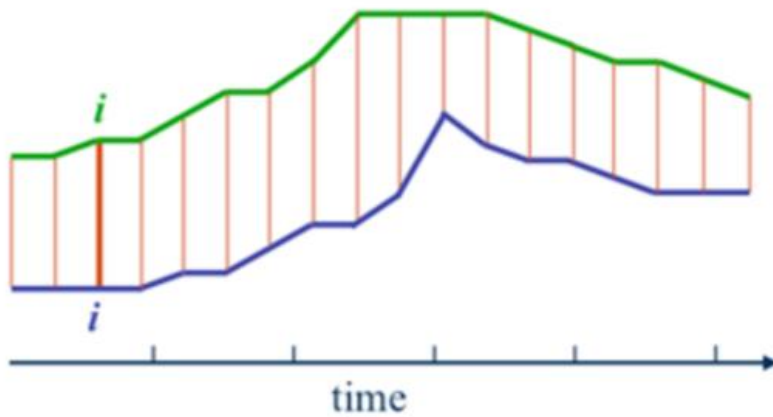


Three constraints

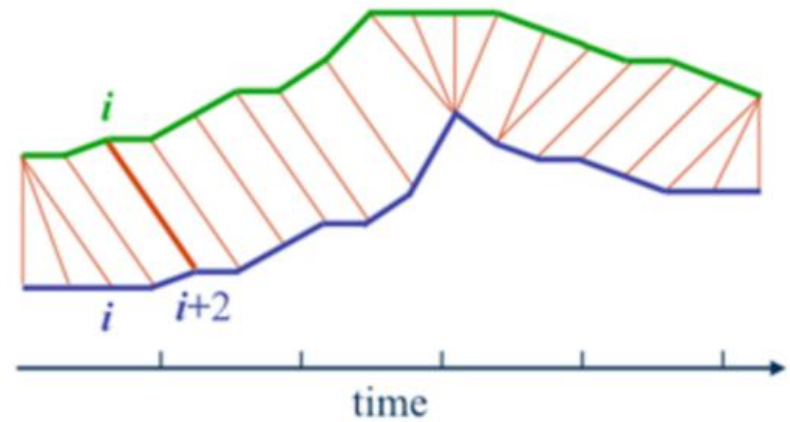
- no skipping of beginning or ends of either sequence
- continuity – no jumps
- monotonicity – can't go back in time



DTW – Find The Minimum Cost Path

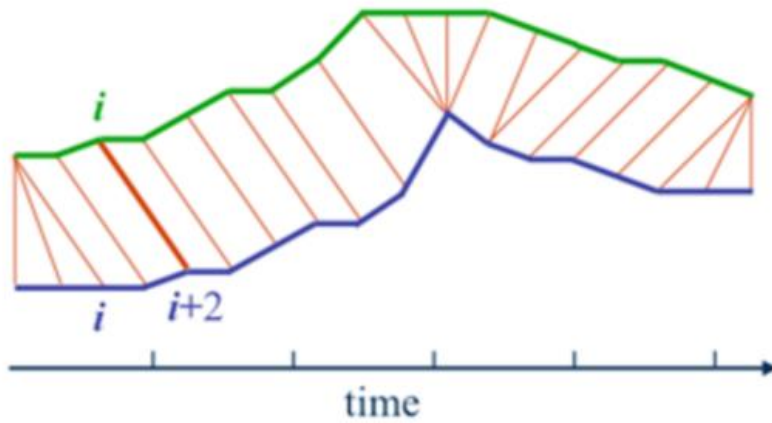


Euclidian



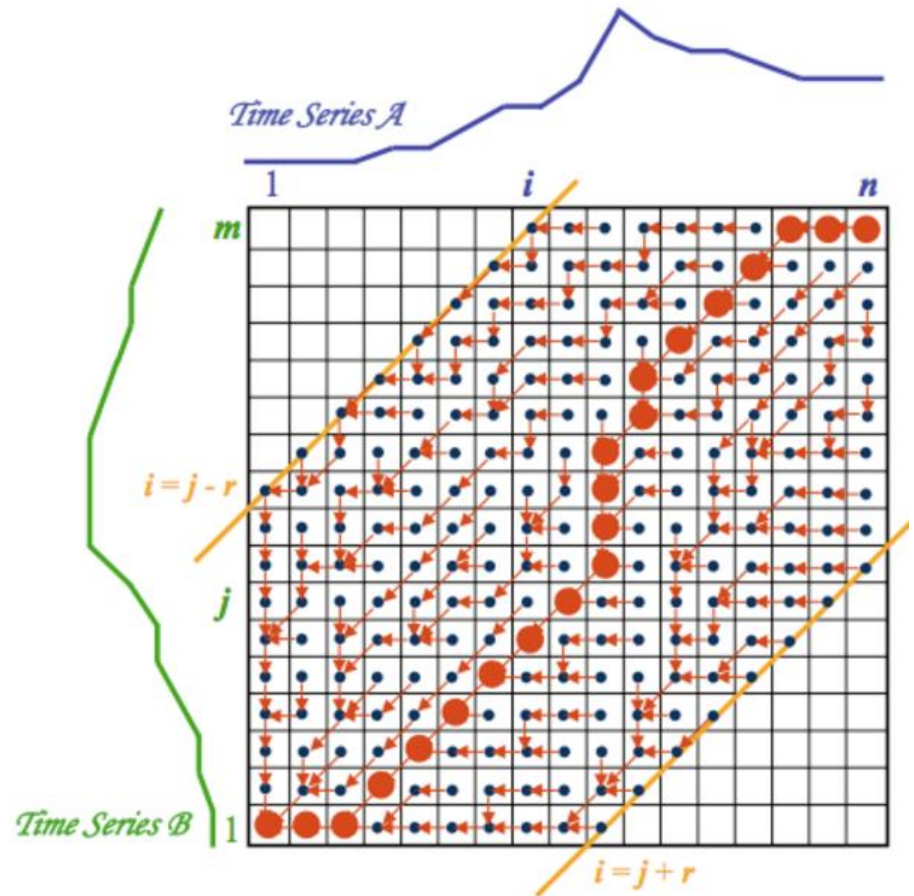
DTW

DTW – Find The Minimum Cost Path

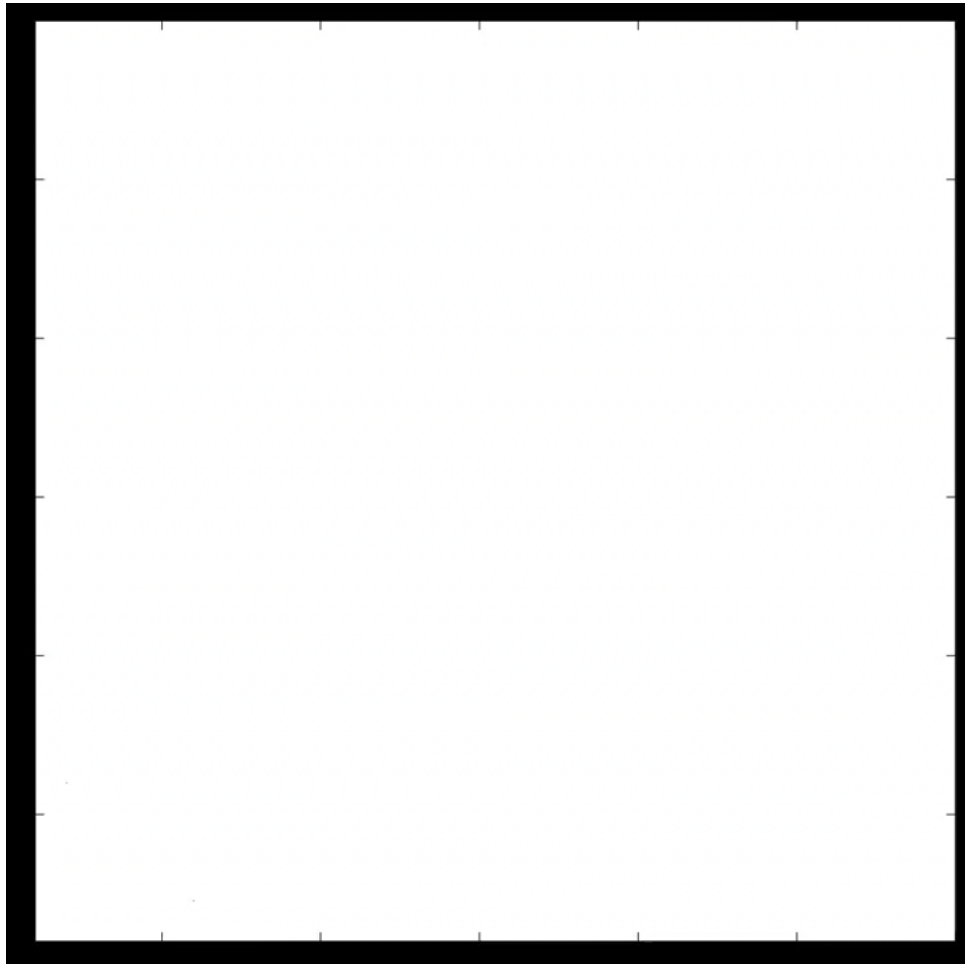


DTW

Compute using dynamic programming



DTW Video



[YouTube video](#)