CSE 564
Visualization and Visual Analytics

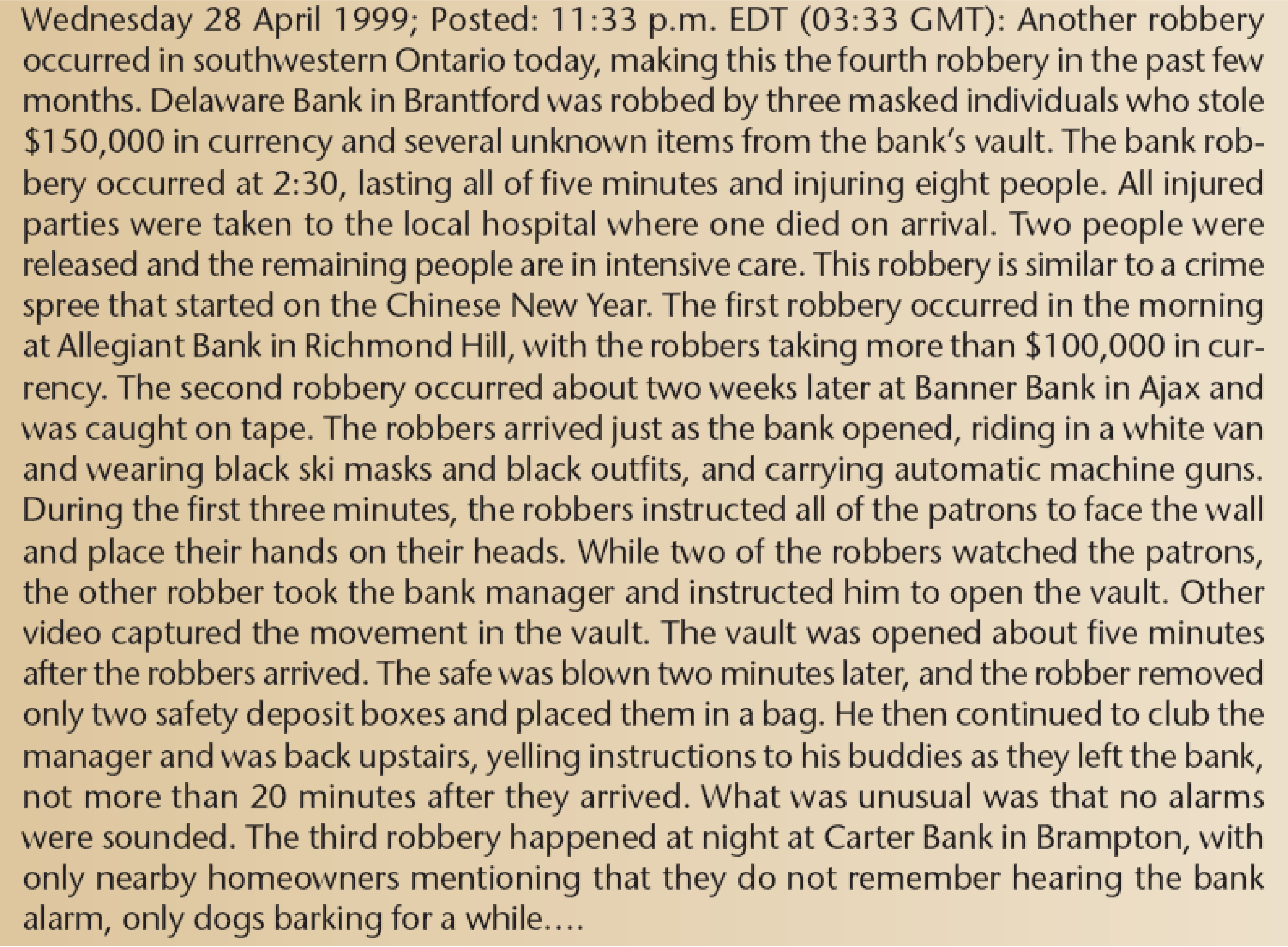
Time Varying and Streaming Data

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	
4	Data assimilation and preparation	Project #1 out
5	Data assimilation and preparation	
6	Bias in visualization	
7	Data reduction and dimension reduction	
8	Visual perception	Project #2(a) out
9	Visual cognition	
10	Visual design and aesthetics	
11	Cluster analysis: numerical data	
12	Cluster analysis: categorical data	Project #2(b) out
13	High-dimensional data visualization	
14	Dimensionality reduction and embedding methods	
15	Principles of interaction	
16	Midterm #1	
17	Visual analytics	Final project proposal call out
18	The visual sense making process	
19	Maps	
20	Visualization of hierarchies	
21	Visualization of time-varying and time-series data	Fnal project proposal due
22	Foundations of scientific and medical visualization	
23	Volume rendering	Project 3 out
24	Scientific and medical visualization	
25	Visual analytics system design and evaluation	Final Project preliminary report due
26	Memorable visualization and embellishments	
27	Infographics design	
28	Midterm #2	

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

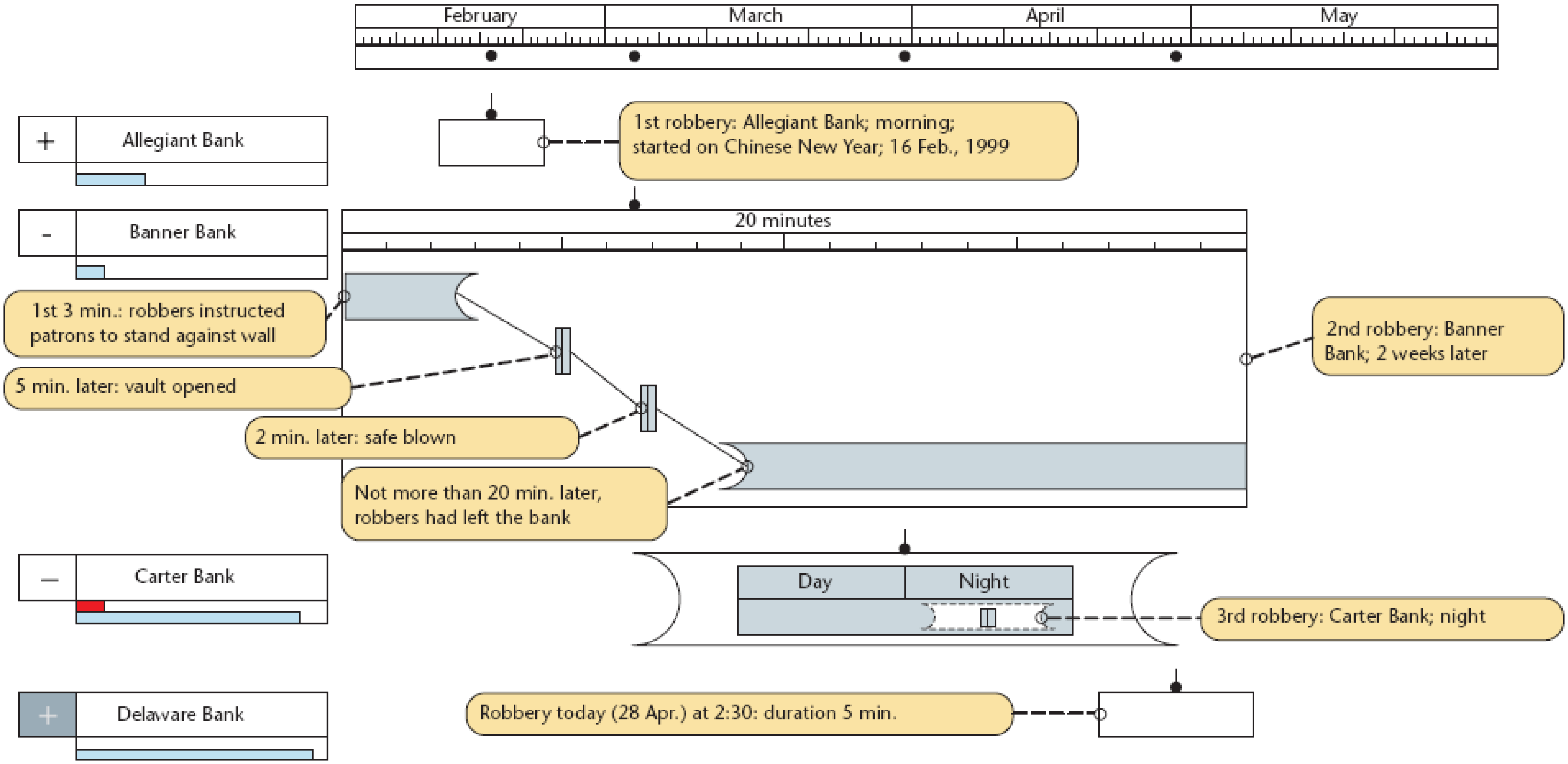
Calendars and time have reference frames

- 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)



1st robbery: Allegiant Bank; morning; started on Chinese New Year; 16 Feb., 1999

1st 3 min.: robbers instructed patrons to stand against wall

5 min. later: vault opened

2 min. later: safe blown

Not more than 20 min. later, robbers had left the bank

2nd robbery: Banner Bank; 2 weeks later

3rd robbery: Carter Bank; night

Robbery today (28 Apr.) at 2:30: duration 5 min.

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?

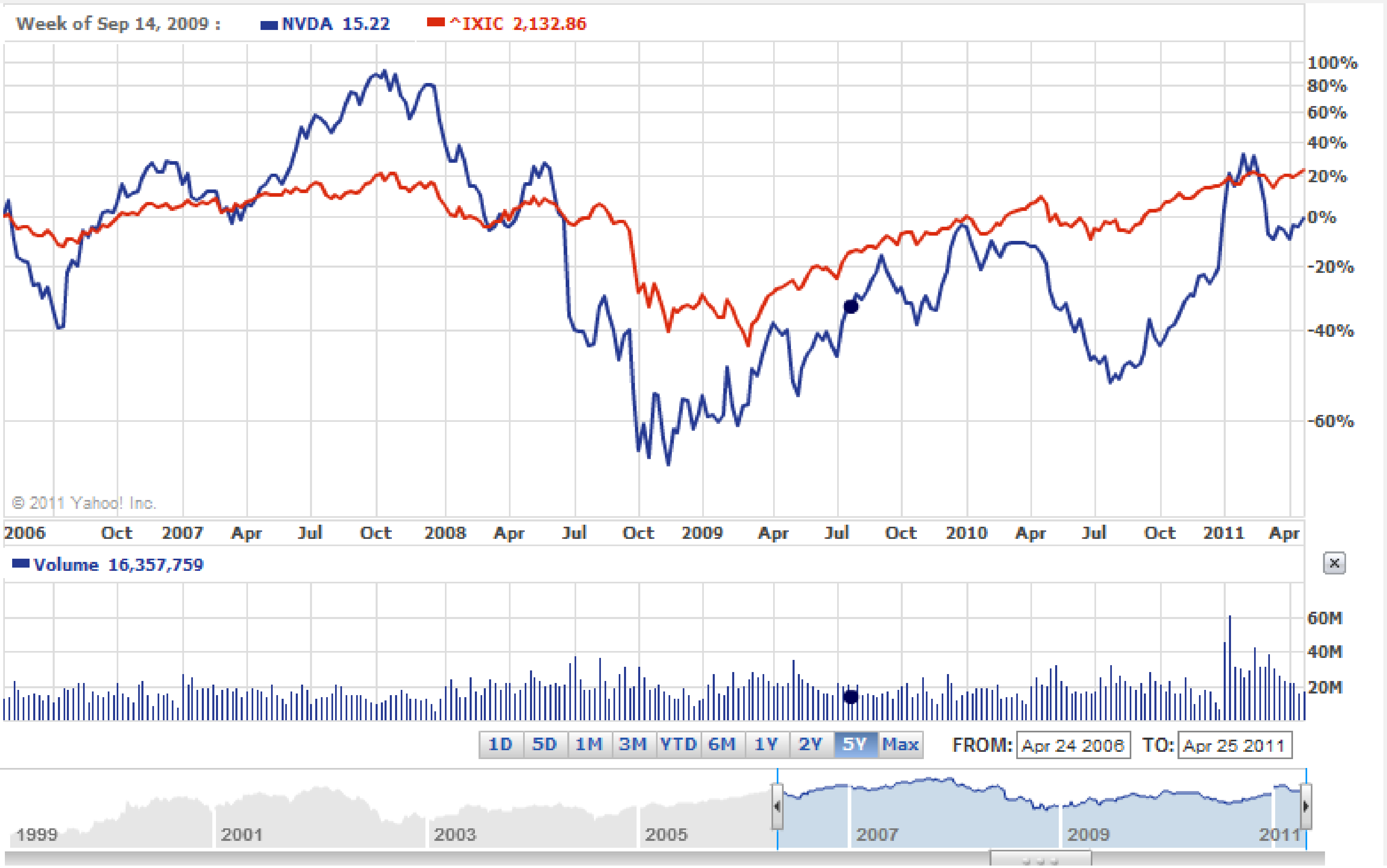
Different types of time series data:

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

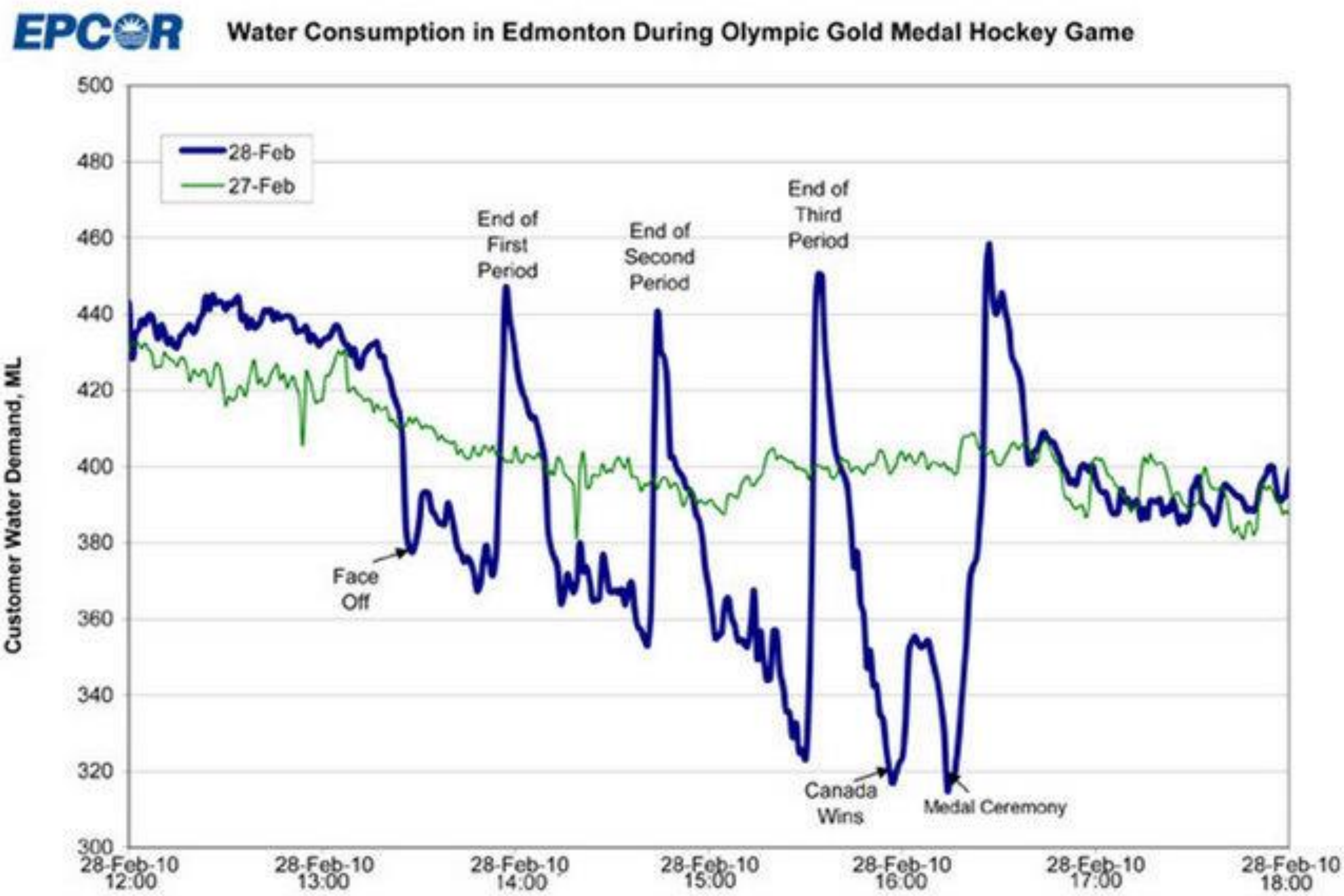
W. Mueller and H. Schumann ‘03

Traditional Time Series Visualizations

NVIDIA stock vs. NASDAQ (from yahoo! finance)



Fun one… (found in J. Stasko lecture)



Next…

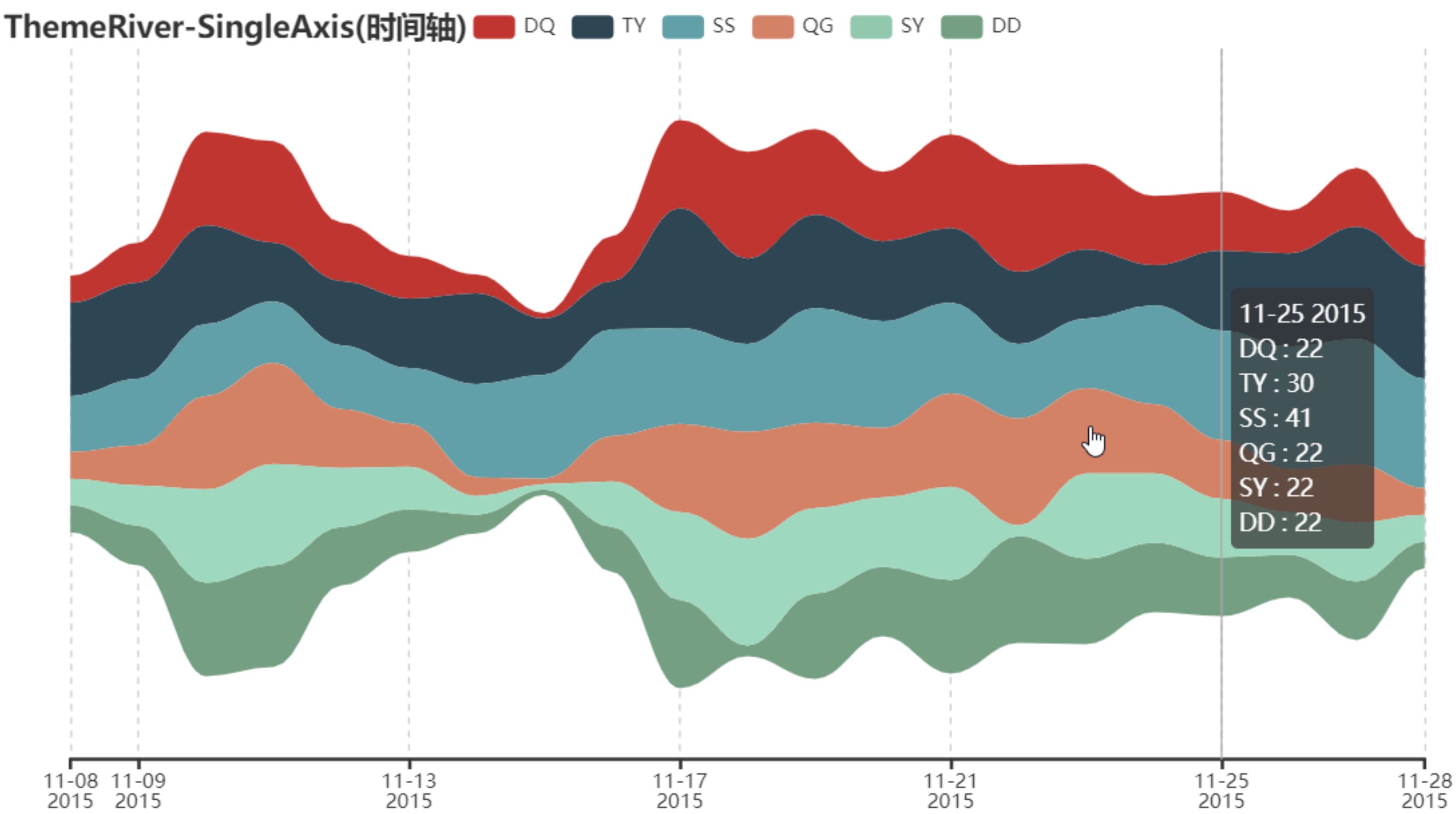
A few good visualization metaphors for time

- there are quite a few of them…

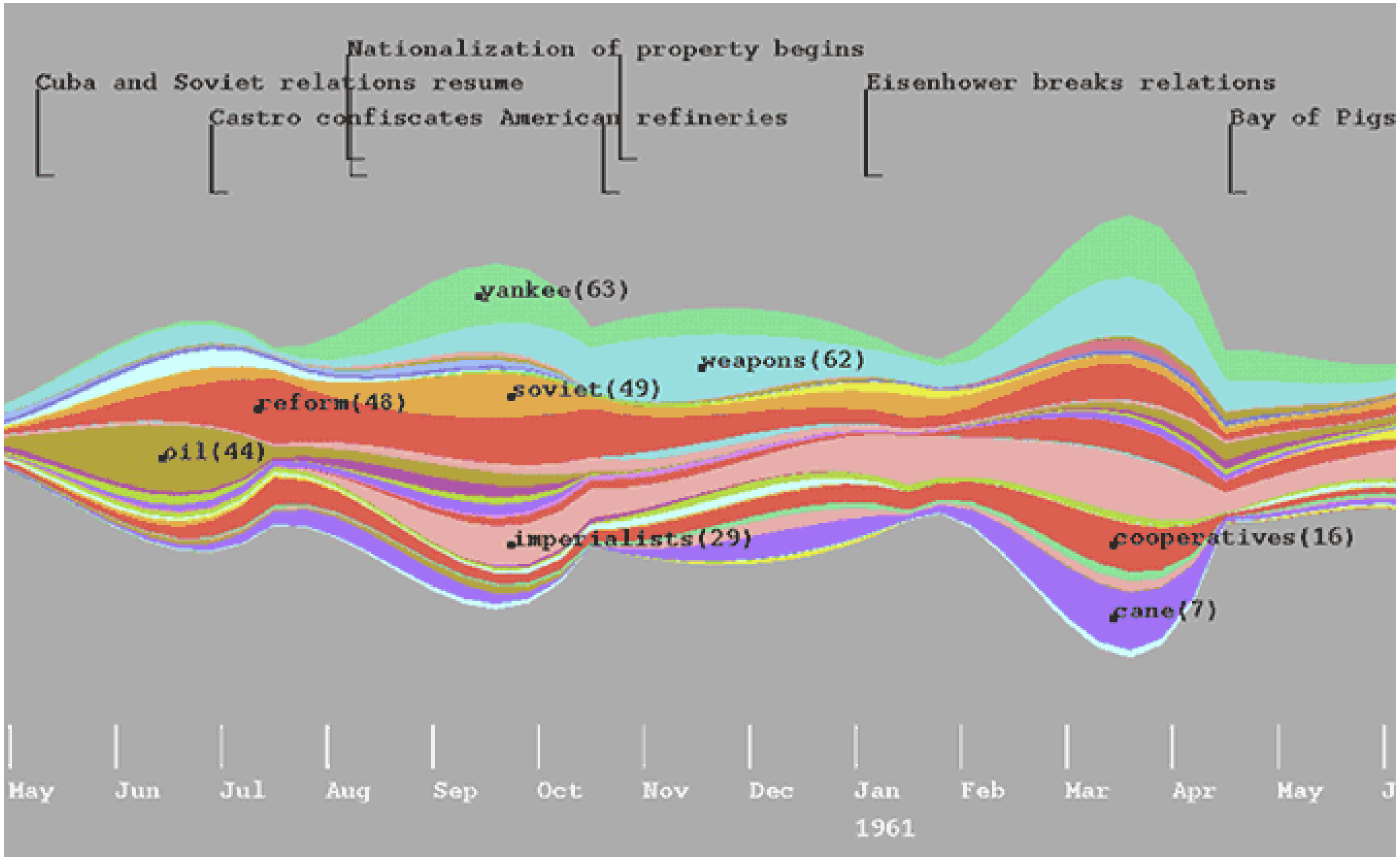
ThemeRiver (Havre et al., 2002)

Each vertical line is a stacked bar chart for data at time *t*

- assembled into a “river” by ordering the bar charts along *t* and centering them on *y* for at each *t*



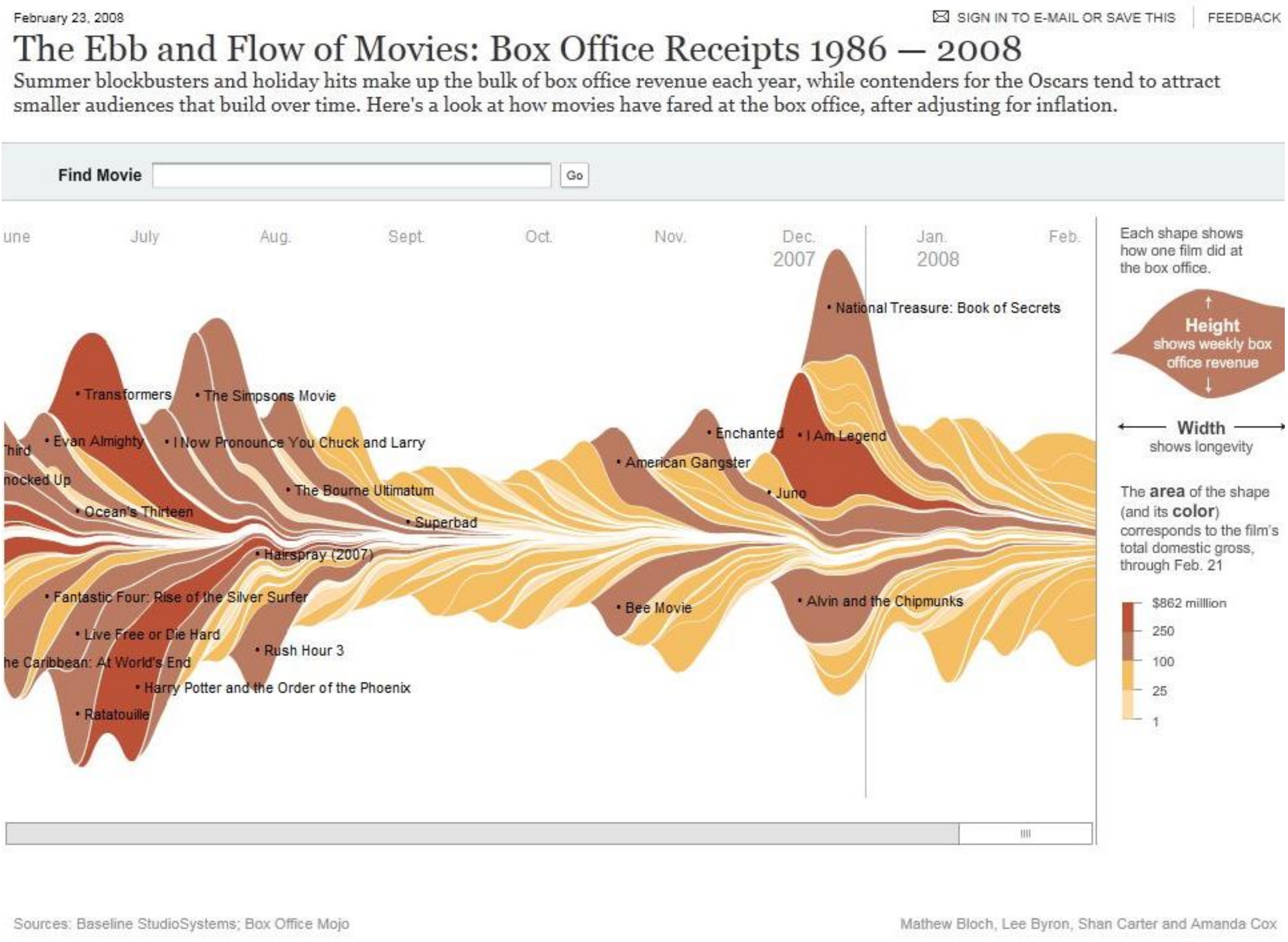
ThemeRiver (Havre et al., 2002)



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.

Find Movie [ ] Go

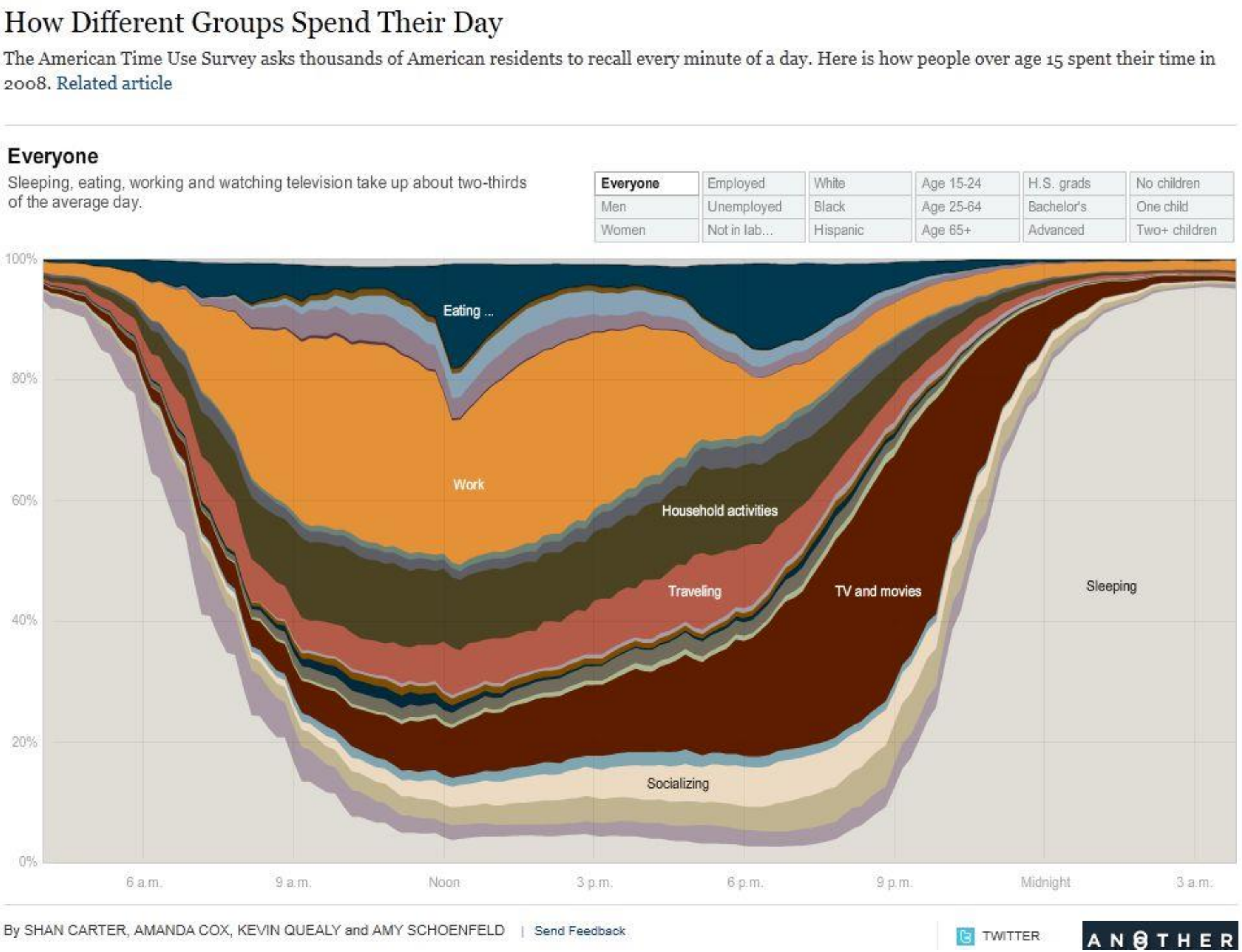
Each shape shows how one film did at the box office.

The **area** of the shape (and its **color**) corresponds to the film's total domestic gross, through Feb. 21

Sources: Baseline StudioSystems; Box Office Mojo

Mathew Bloch, Lee Byron, Shan Carter and Amanda Cox

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

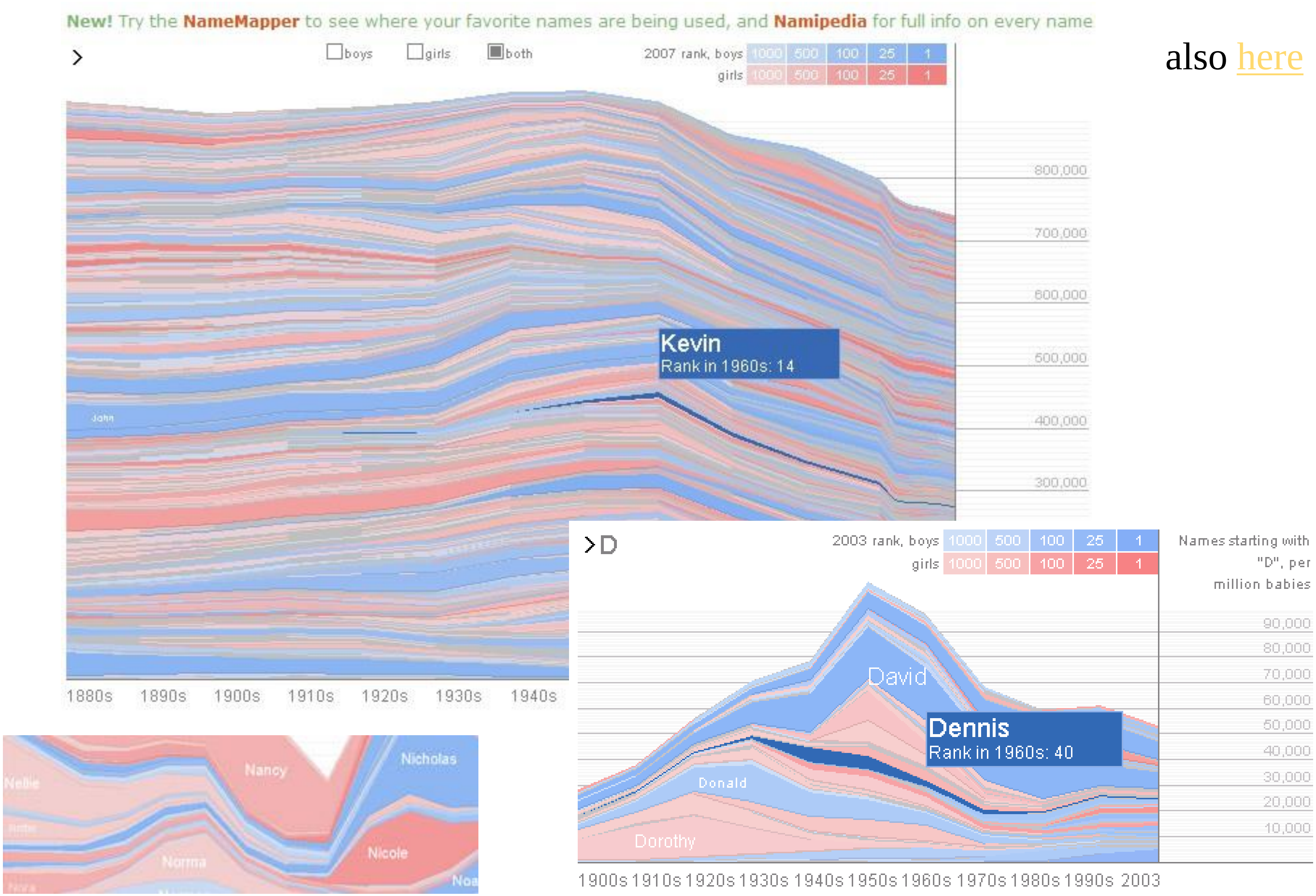
Everyone

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

By SHAN CARTER, AMANDA COX, KEVIN QUEALY and AMY SCHOENFELD | Send Feedback

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

also here

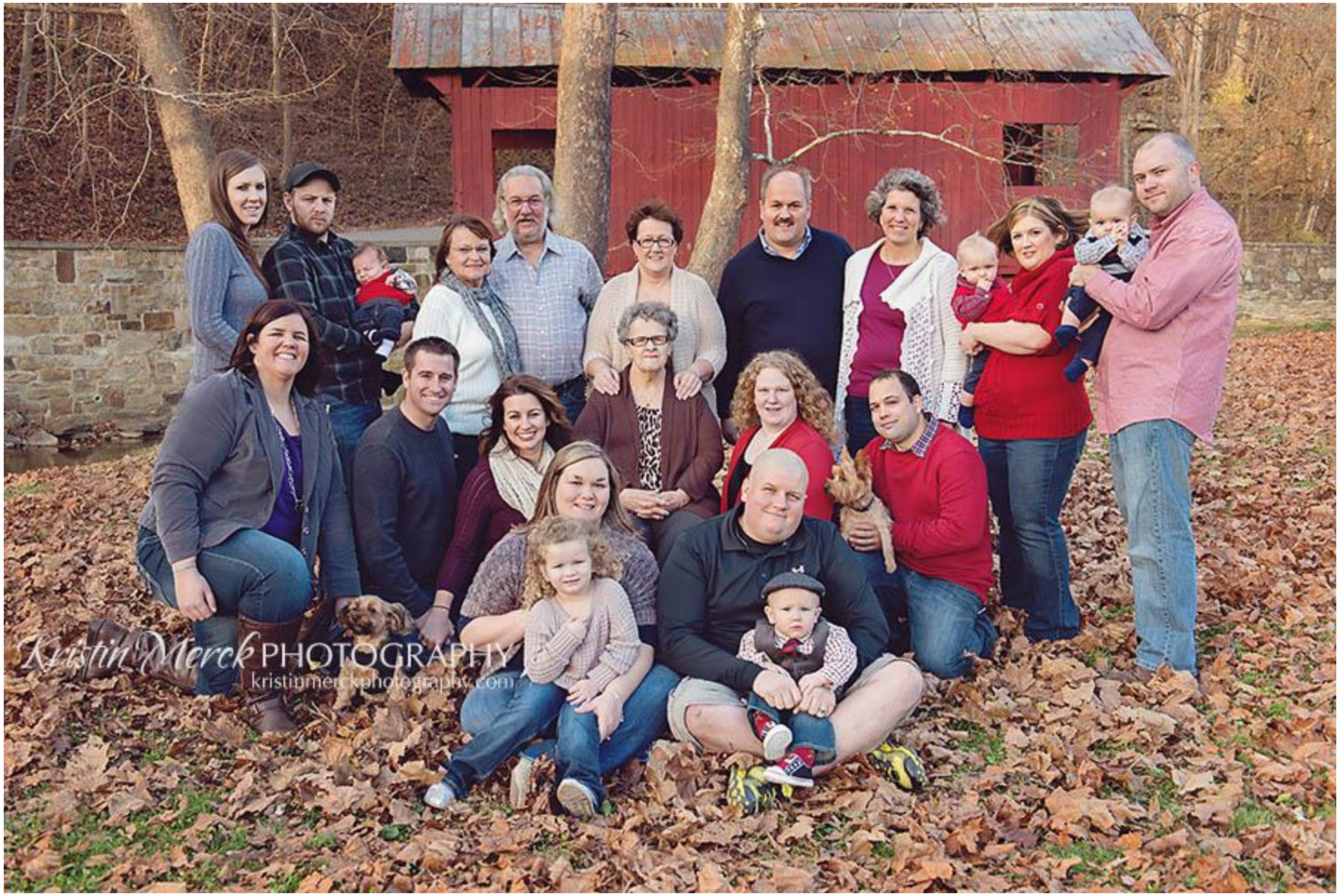


New! Try the NameMapper to see where your favorite names are being used, and Namipedia for full info on every name

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

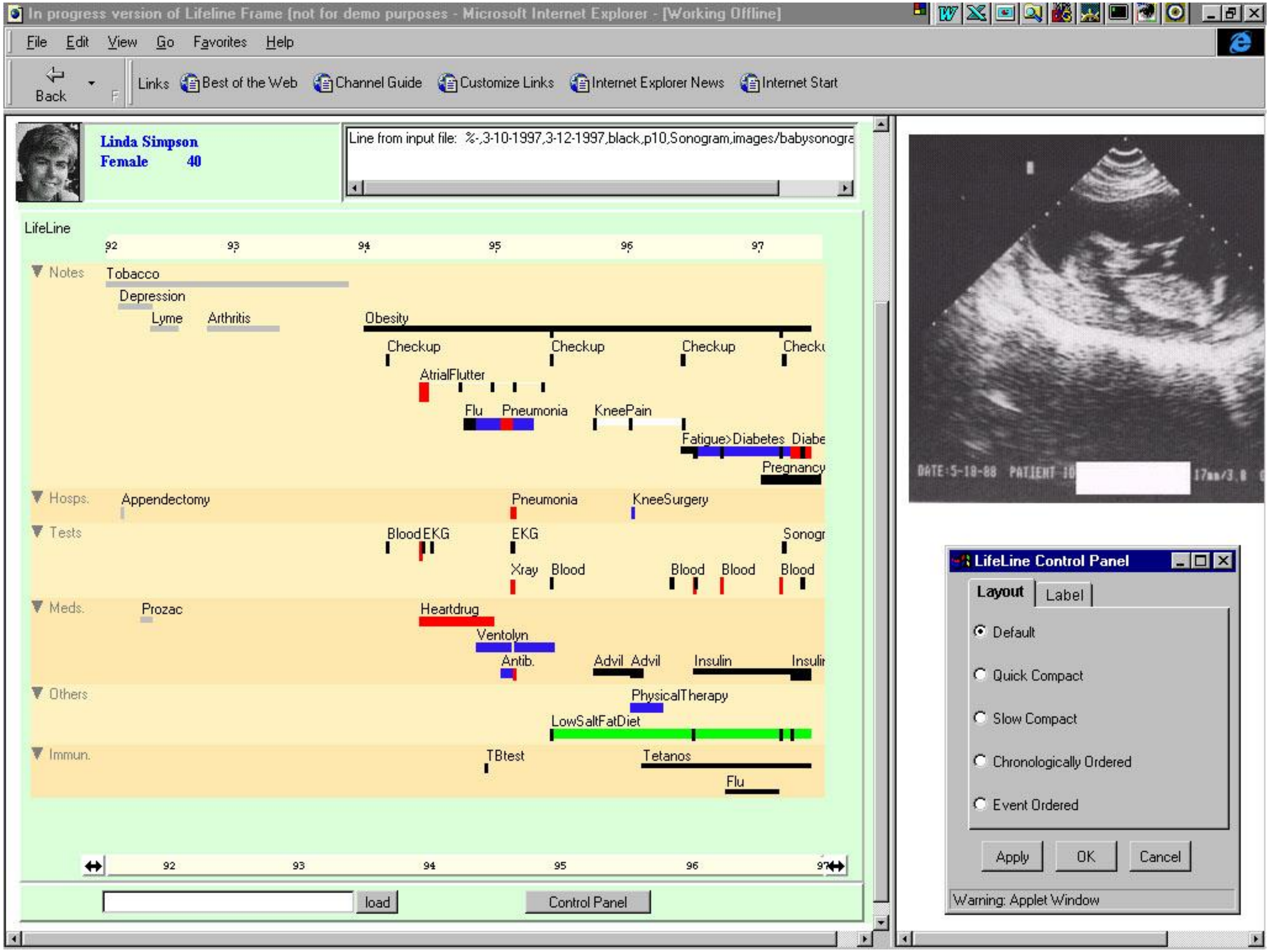
Medical data are often displayed along time

- 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



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

Layout:
- ● Default
- ○ Quick Compact
- ○ Slow Compact
- ○ Chronologically Ordered
- ○ Event Ordered

LifeLines2: Pattern-Centric

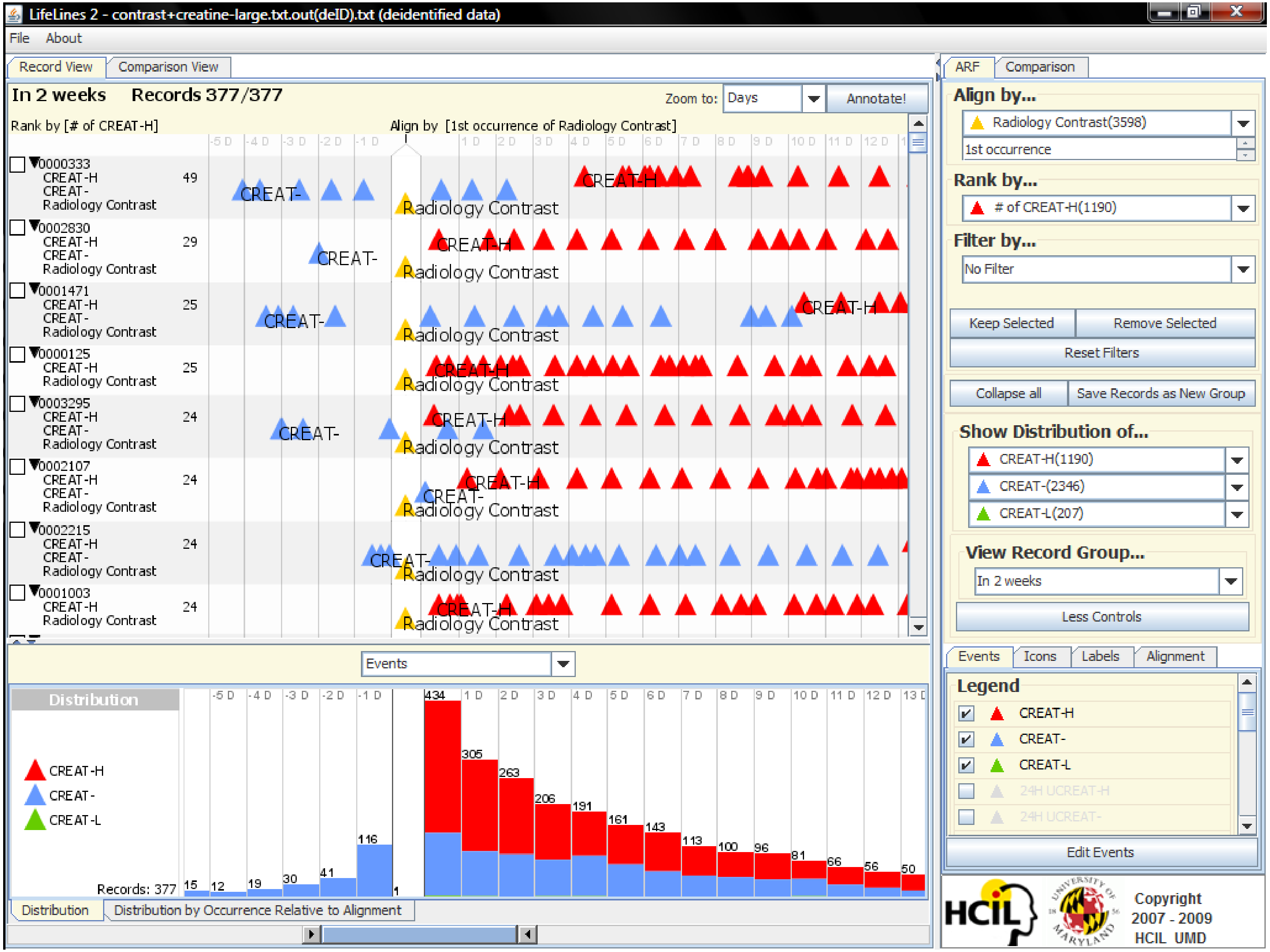
Goals:

- 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



LifeLines 2 - contrast+creatine-large.txt.out(deID).txt (deidentified data)

In 2 weeks Records 377/377

Zoom to: Days

Rank by [# of CREAT-H]

Align by [1st occurrence of Radiology Contrast]

□ 0000333 CREAT-H CREAT- Radiology Contrast 49

□ 0002830 CREAT-H CREAT- Radiology Contrast 29

□ 0001471 CREAT-H CREAT- Radiology Contrast 25

□ 0000125 CREAT-H Radiology Contrast 25

□ 0003295 CREAT-H CREAT- Radiology Contrast 24

□ 0002107 CREAT-H CREAT- Radiology Contrast 24

□ 0002215 CREAT-H CREAT- Radiology Contrast 24

□ 0001003 CREAT-H Radiology Contrast 24

Align by... Radiology Contrast(3598)

1st occurrence

Rank by... # of CREAT-H(1190)

Filter by... No Filter

Show Distribution of...

- CREAT-H(1190)
- CREAT-(2346)
- CREAT-L(207)

View Record Group... In 2 weeks

Legend

- ☑ CREAT-H
- ☑ CREAT-
- ☑ CREAT-L
- □ 24H UCREAT-H
- □ 24H UCREAT-

Distribution

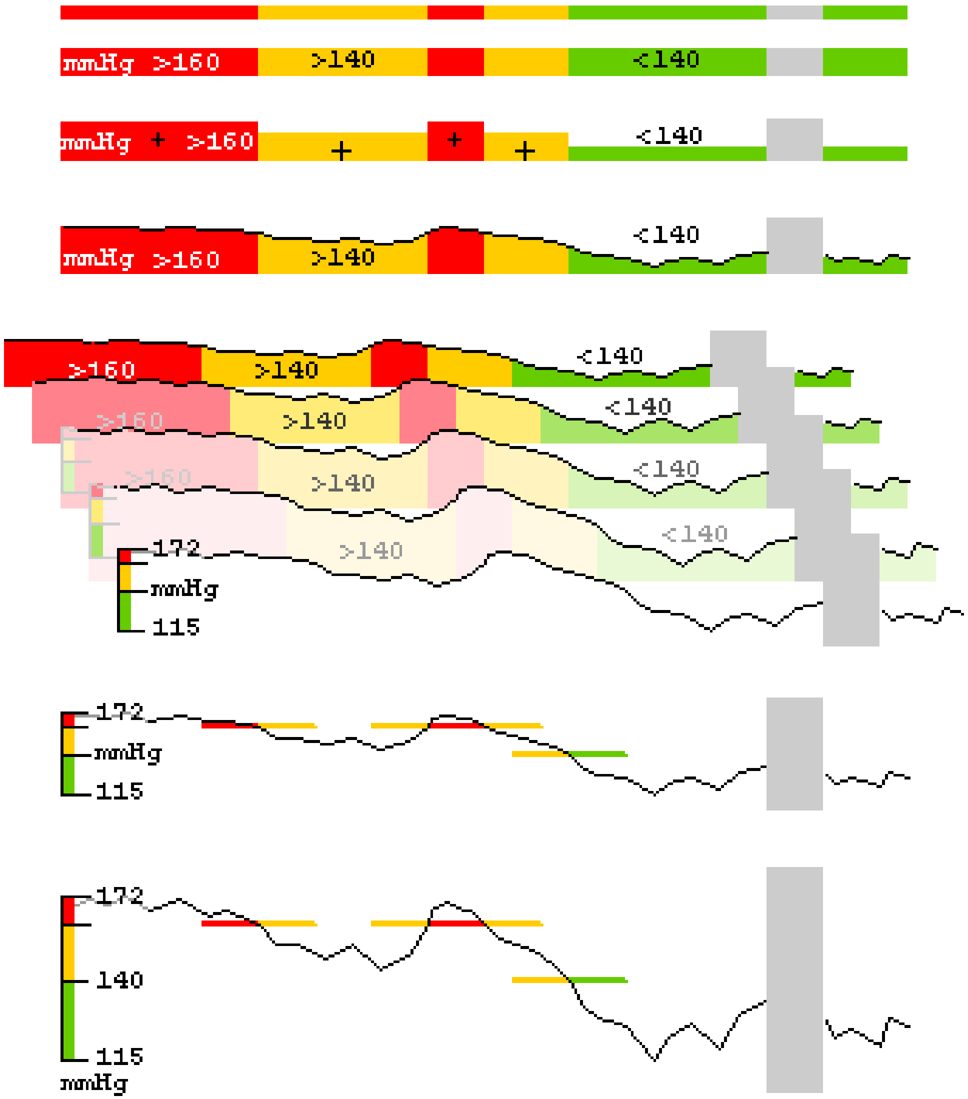
Records: 377

Copyright
2007 - 2009
HCIL UMD

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

Deal with different levels of detail

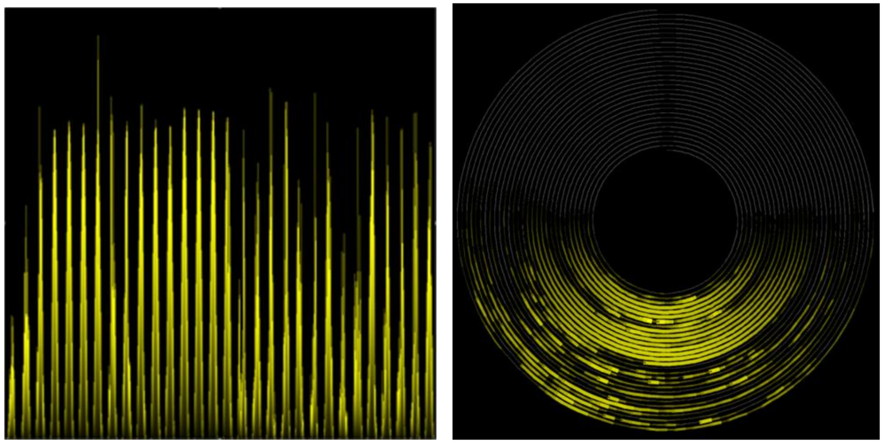
- illustrative abstraction
- overview + detail
- used here for medical data



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

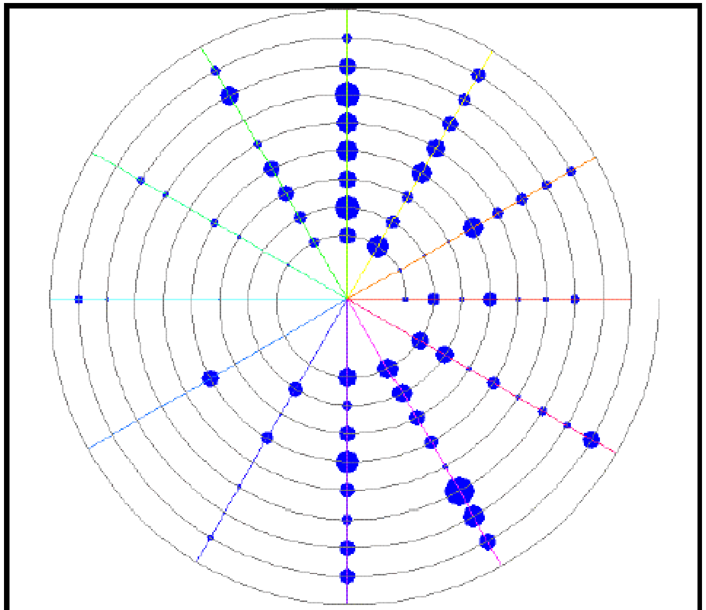


Figure 2. An indented spiral, with spokes, showing monthly consumption percentages for Baphia Capparidifolia during the period 1980 – 1988.

Weber et al., 2001

Cyclic Patterns

May have to play around to discover the cycles

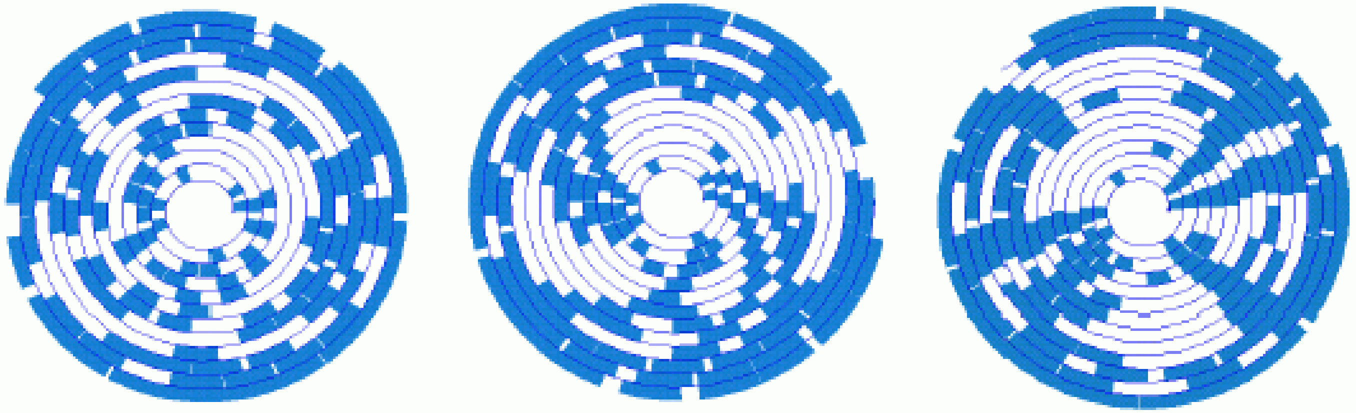


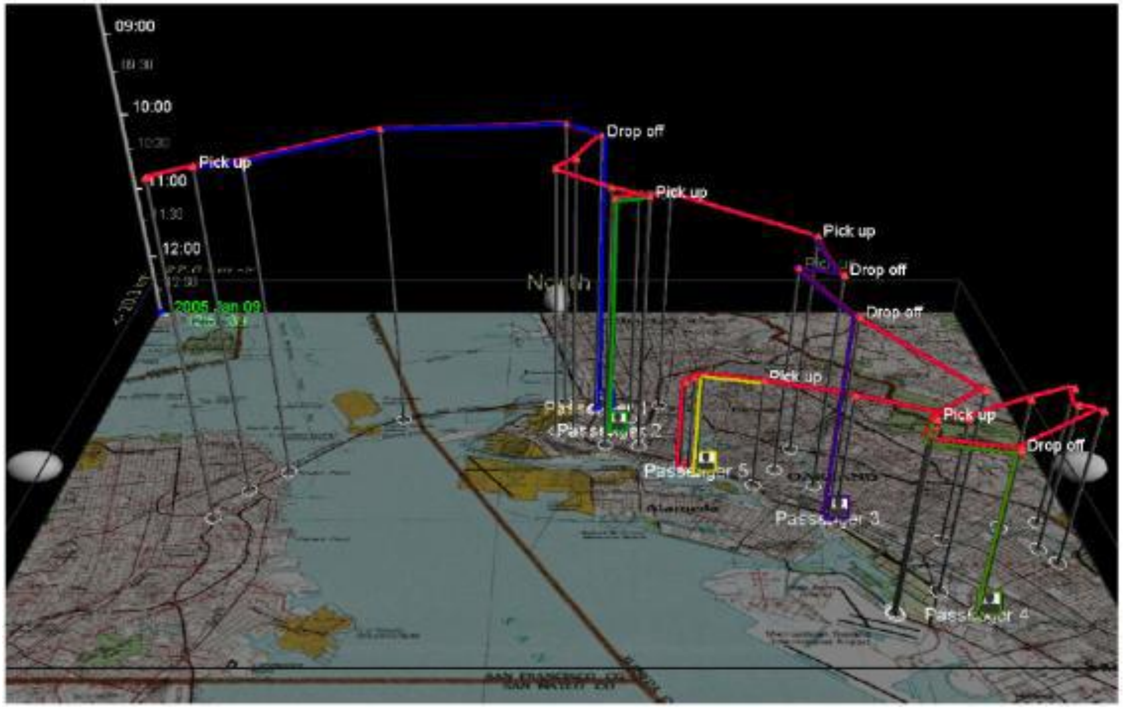
Figure 7. Tightening a spiral view of sound data from five instruments. From left to right the structure of the sound reveals itself.

from J. Stasko, lecture notes

Combine Space and Time

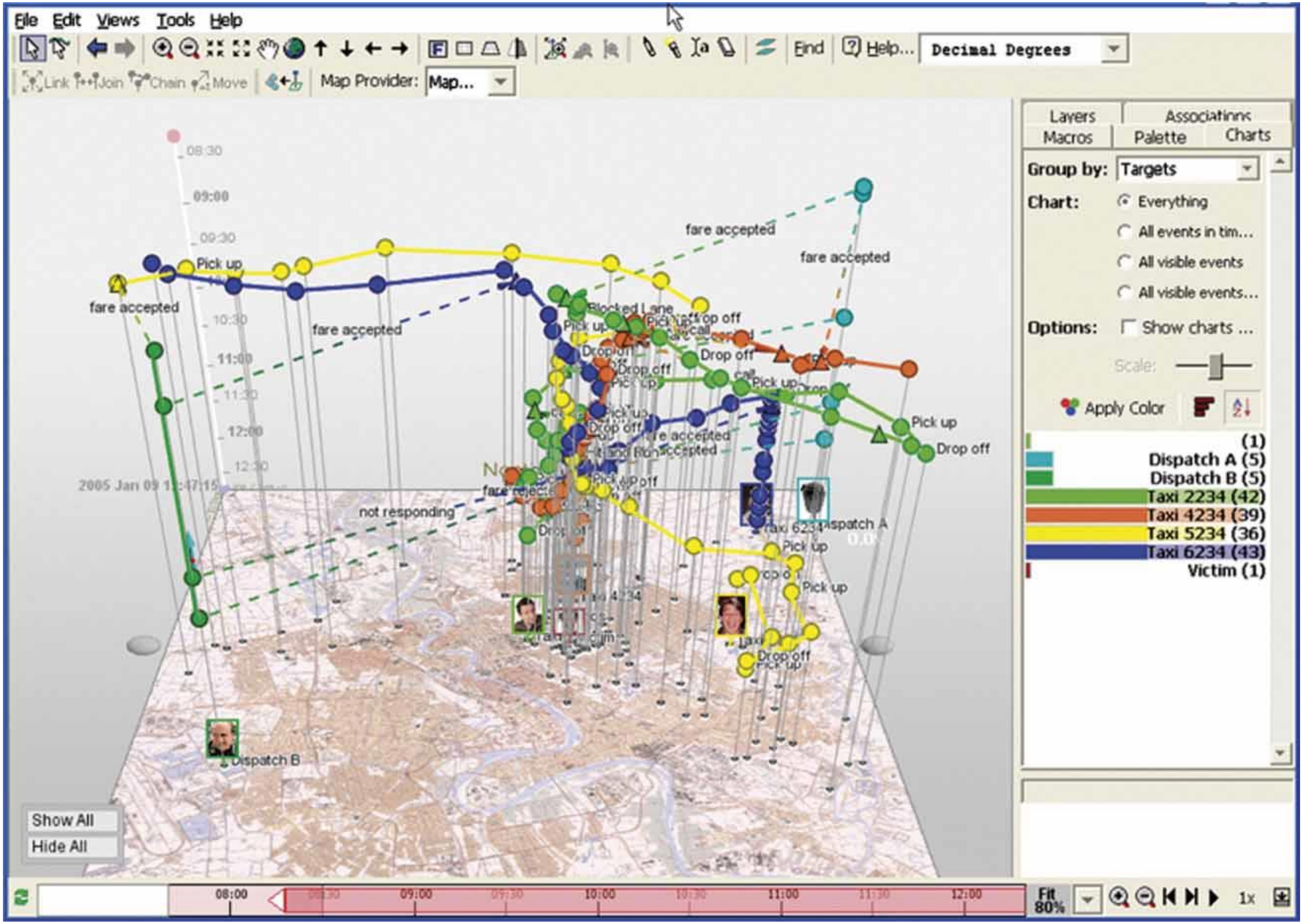
OculusInfo Geotime application

- 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.



http://www.oculusinfo.com/SoftwareProducts/GeoTime.html

Geotime



Map Provider: Map...

Group by: Targets

Chart:
- ● Everything
- ○ All events in tim...
- ○ All visible events
- ○ All visible events...

Options: □ Show charts ...

	(1)
🟦	Dispatch A (5)
🟩	Dispatch B (5)
🟩	Taxi 2234 (42)
🟧	Taxi 4234 (39)
🟨	Taxi 5234 (36)
🟦	Taxi 6234 (43)
🟥	Victim (1)

Interaction

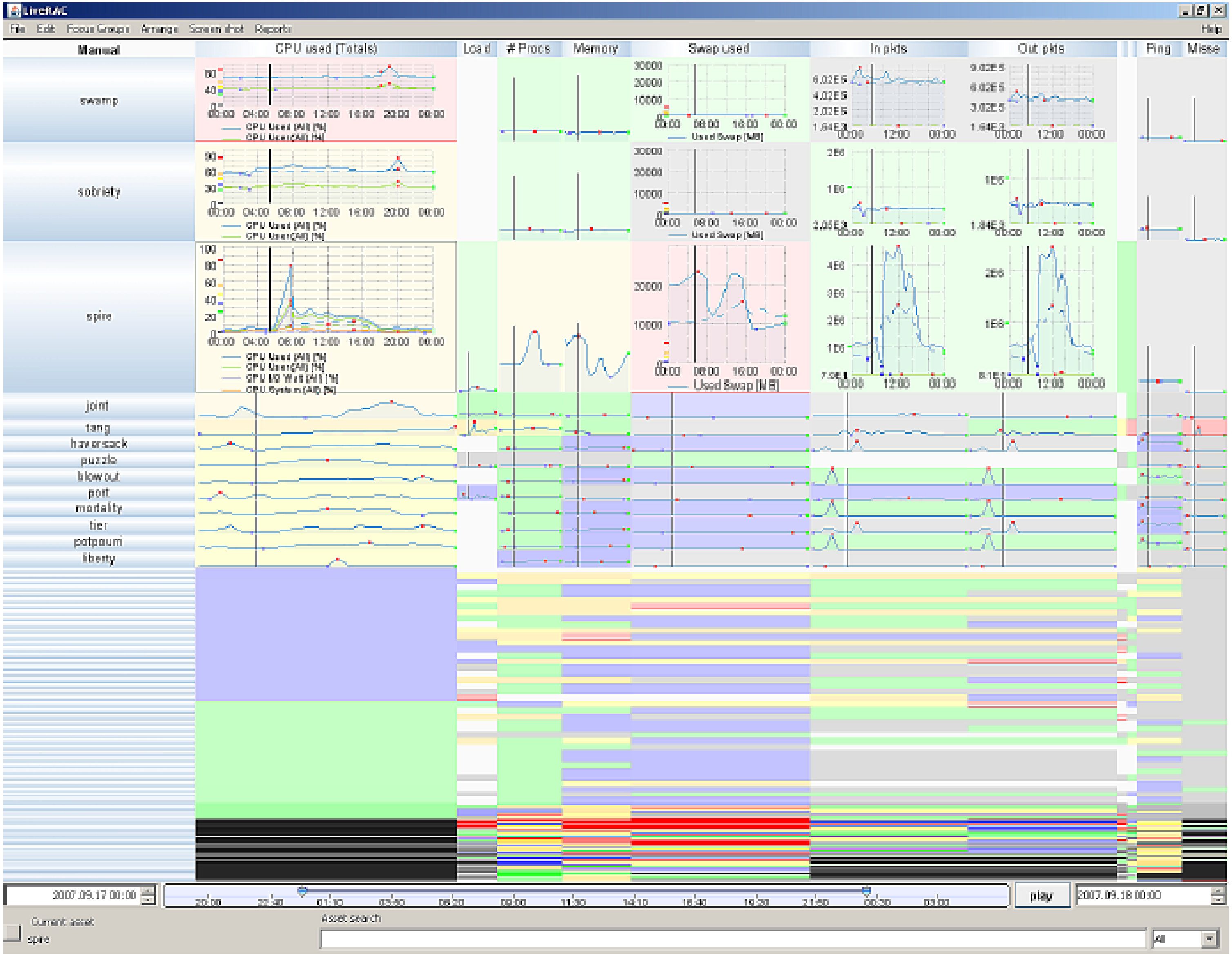
As complexity increases, interaction capabilities are key

- 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

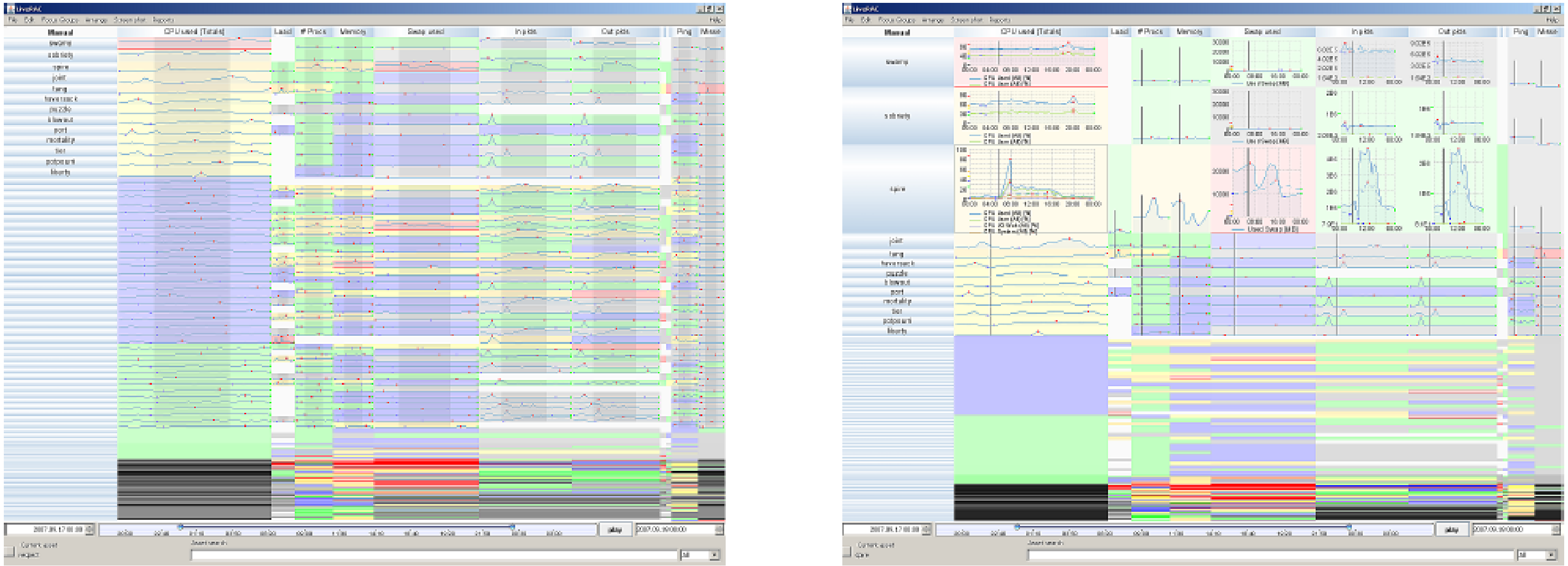


2007.09.17 00:00

2007.09.18 00:00

Current asset
spire

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 Video

video

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)

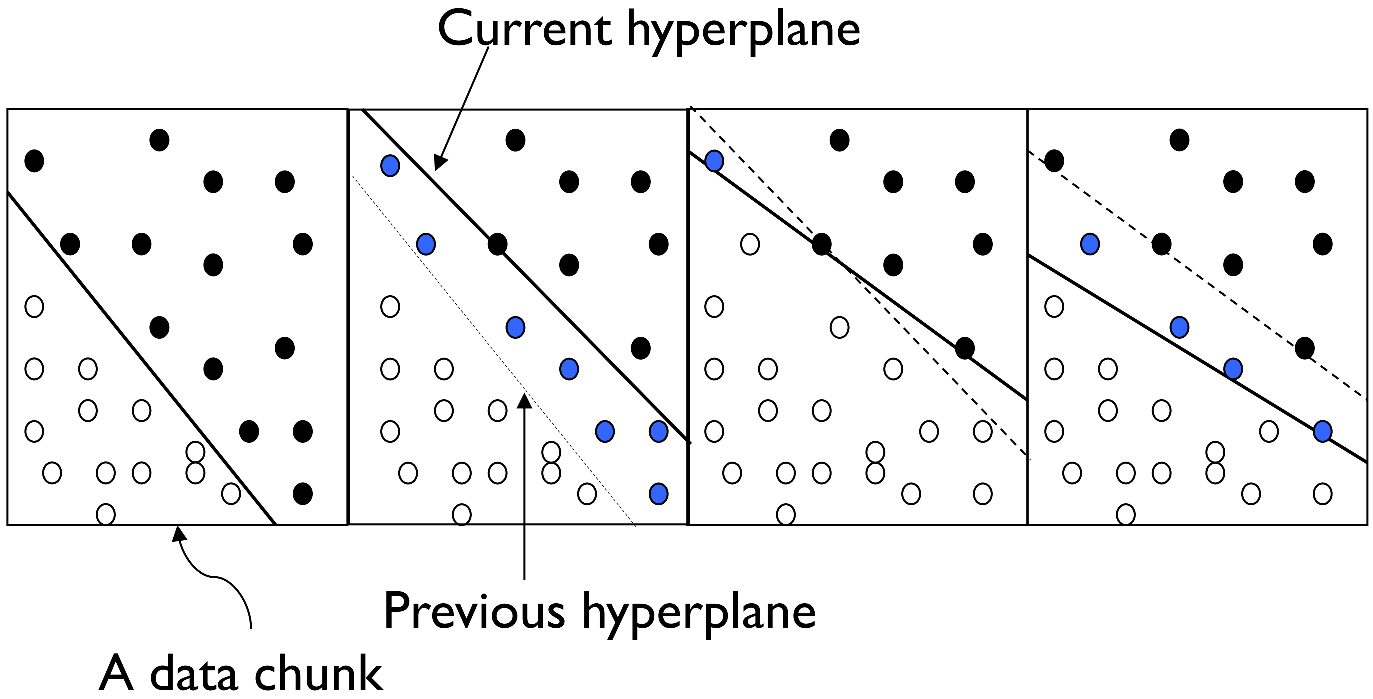
One-pass constraint

- 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



Current hyperplane

Previous hyperplane

A data chunk

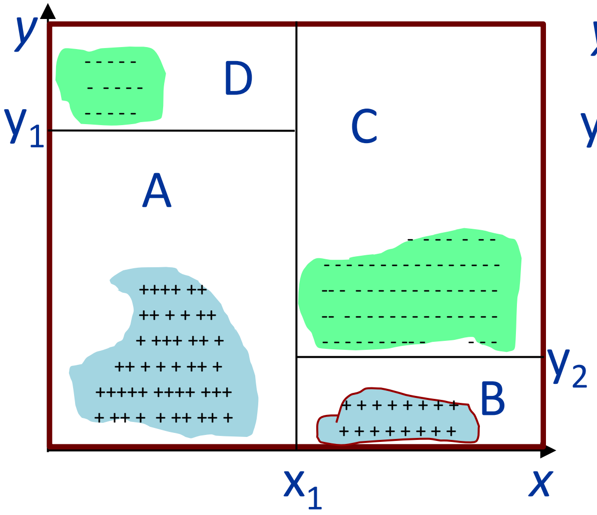
Negative instance ●

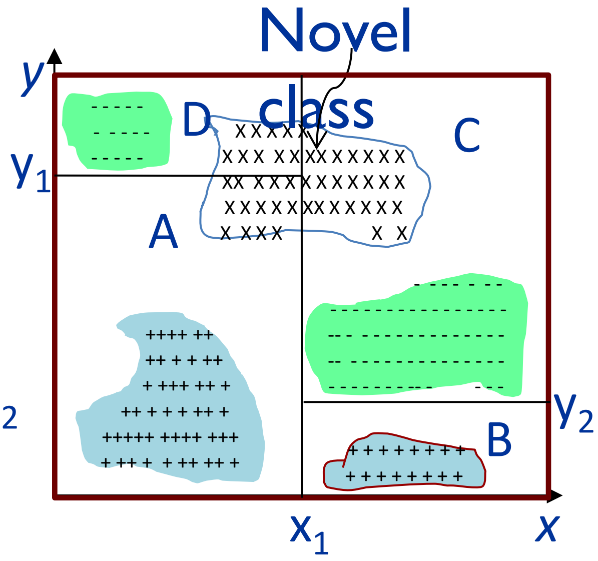
Positive instance ○

Instances victim of concept-drift ●

Latifur Khan, et al. IBM

Concept Evolution





Classification rules:

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

R2. if ($x > x_1$ and $y > y2$) or ($x < x_1$ and $y > y_1$) then class = -

Existing classification models misclassify novel class instances

Latifur Khan, et al. IBM

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

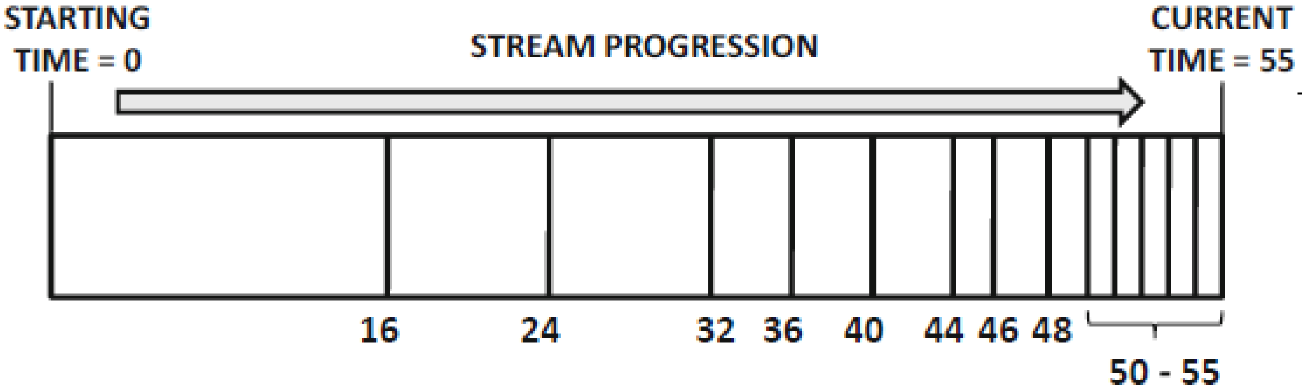
Algorithm

- 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



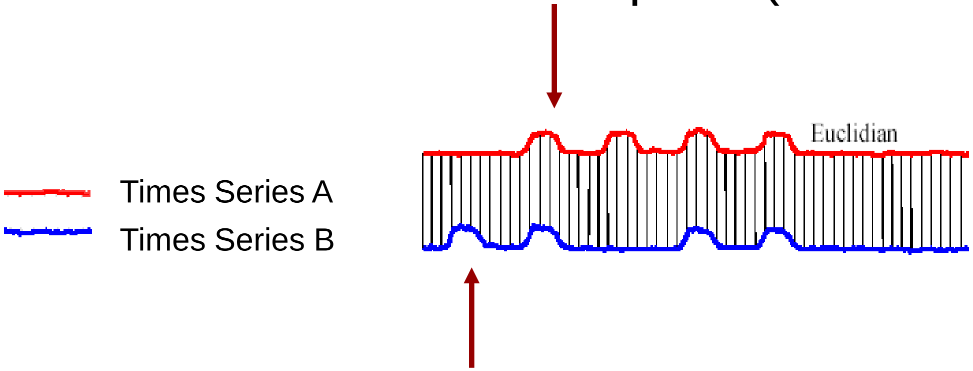
L_p Norm and its Shortcomings

Standard pairwise distance

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

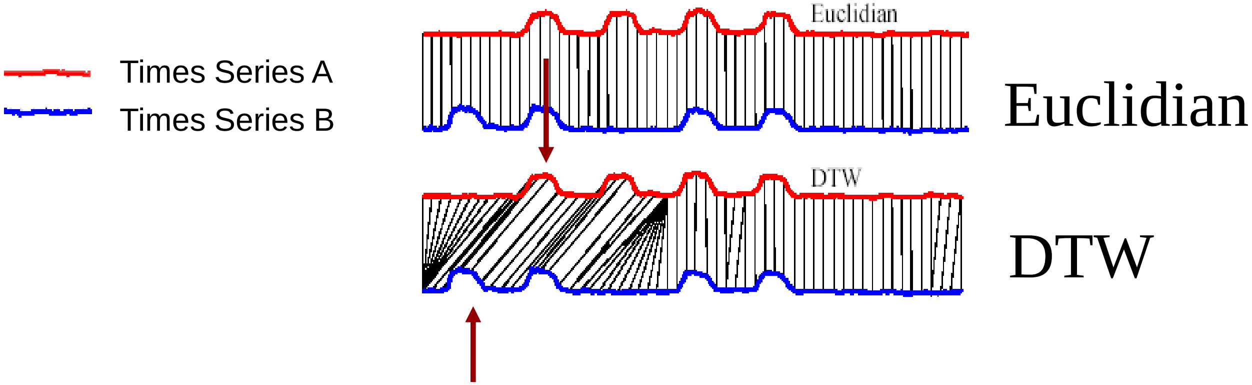
Shortcomings:

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



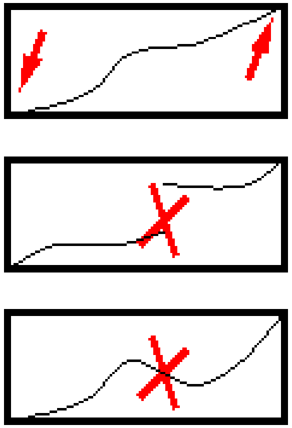
Dynamic Time Warping Distance

Can better accommodate local mismatches

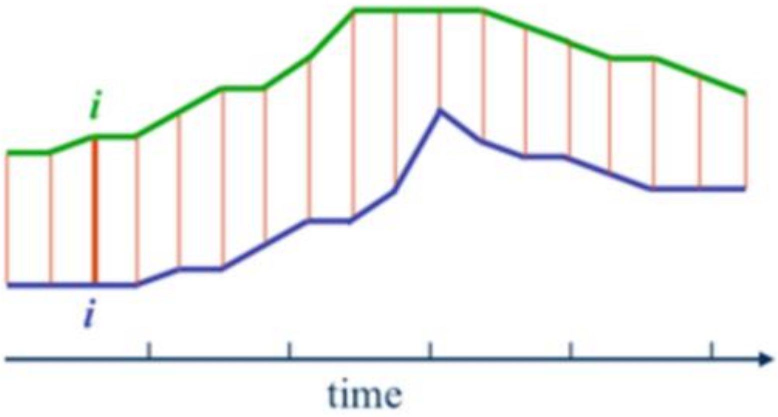


Three constraints

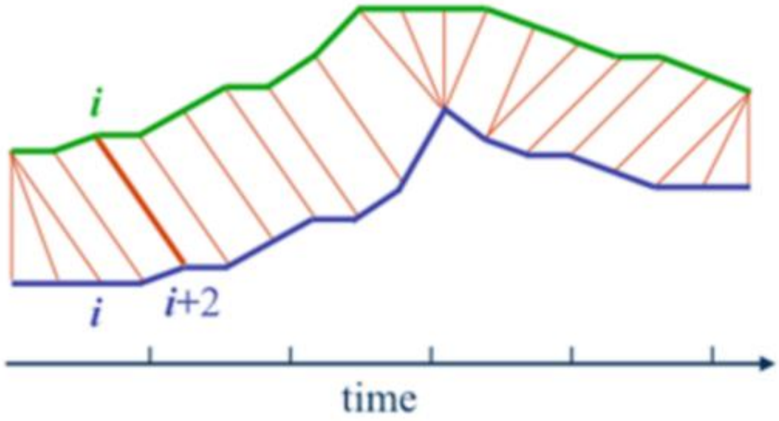
- no skipping of beginning or ends of either sequence
- continuity – no jumps
- monotonicity – can't go back in time



DTW – Find The Minimum Cost Path

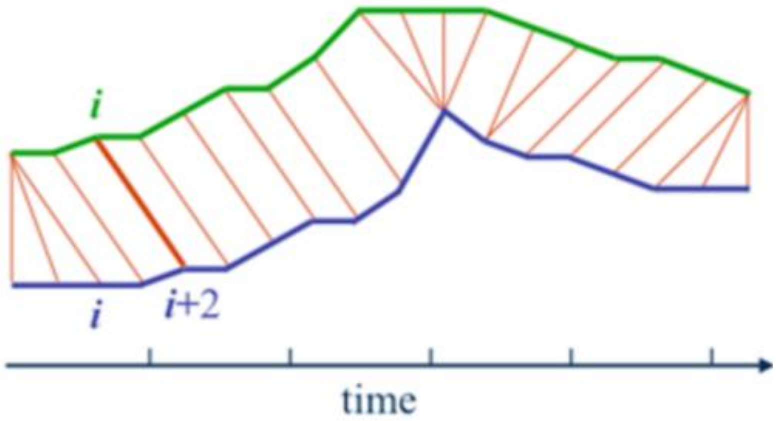


Euclidian

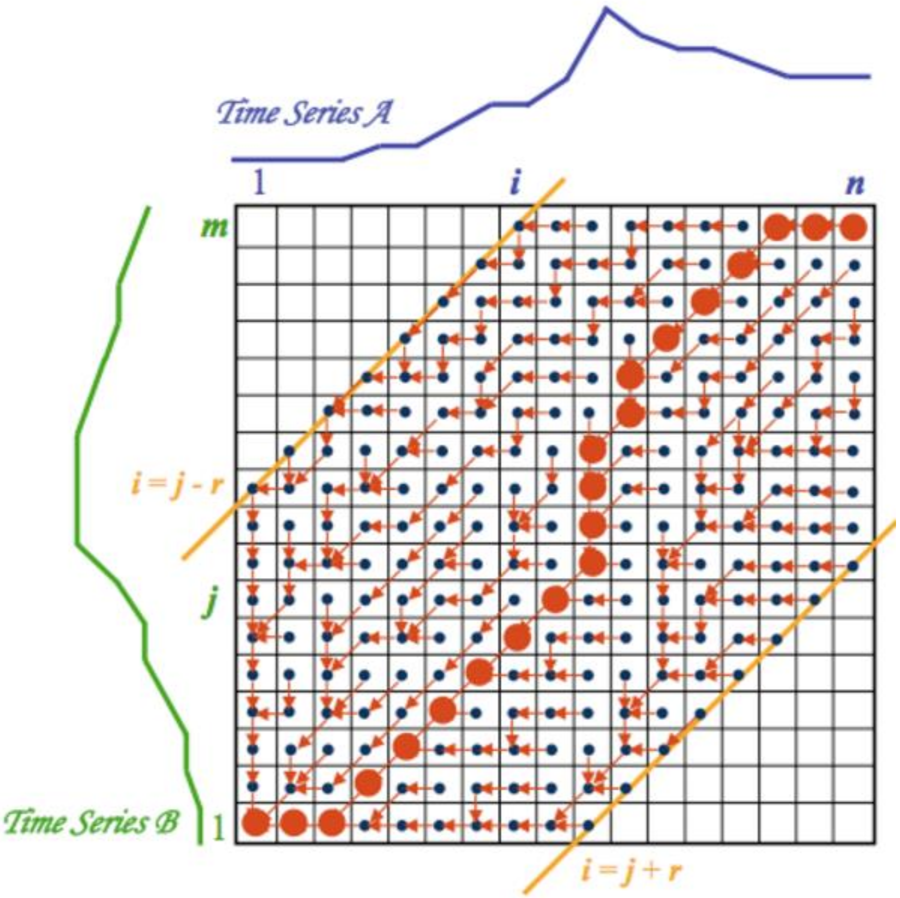


DTW

DTW – Find The Minimum Cost Path

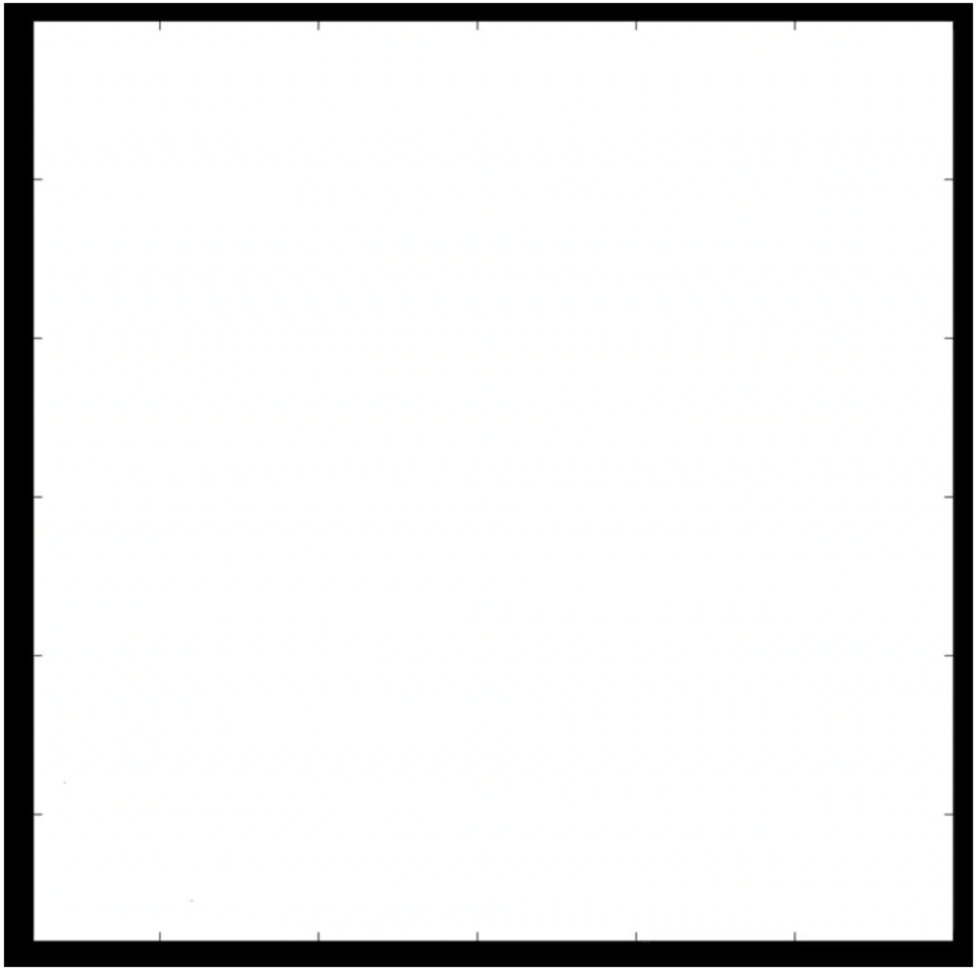


DTW



Compute using dynamic programming

DTW Video



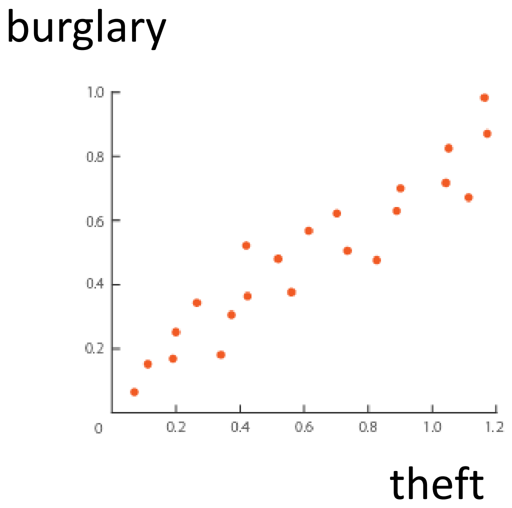
YouTube video

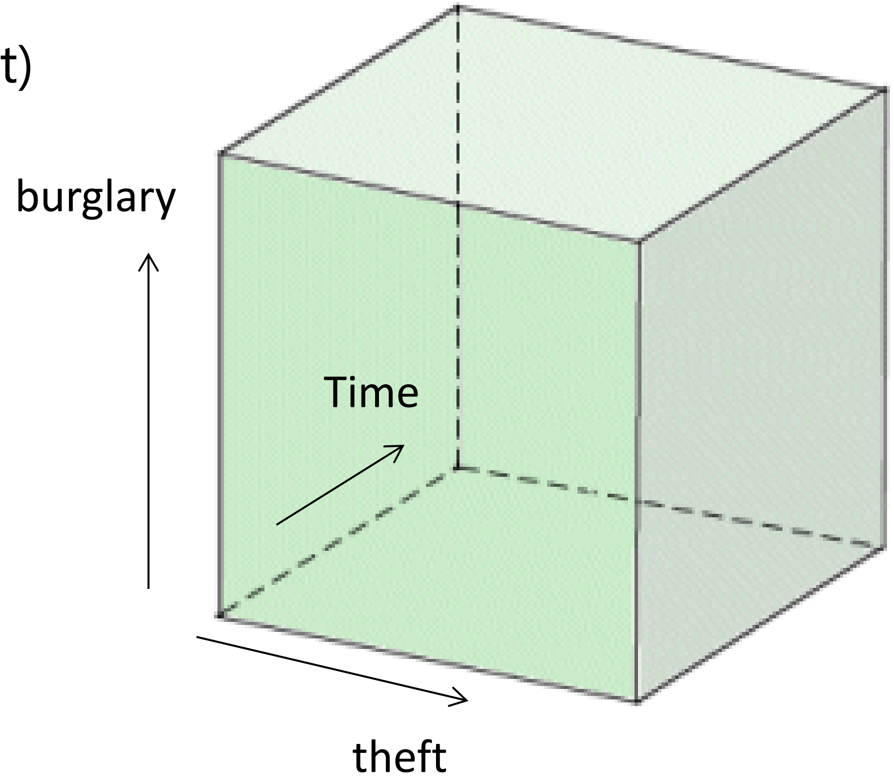
Time Cube

Assume for now we have

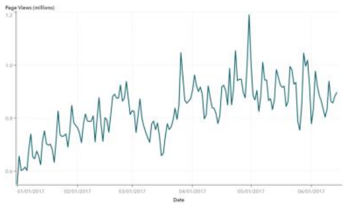
- two attributes (burglary, theft)
- both observed over time

Can visualize





- but each point is a time series!



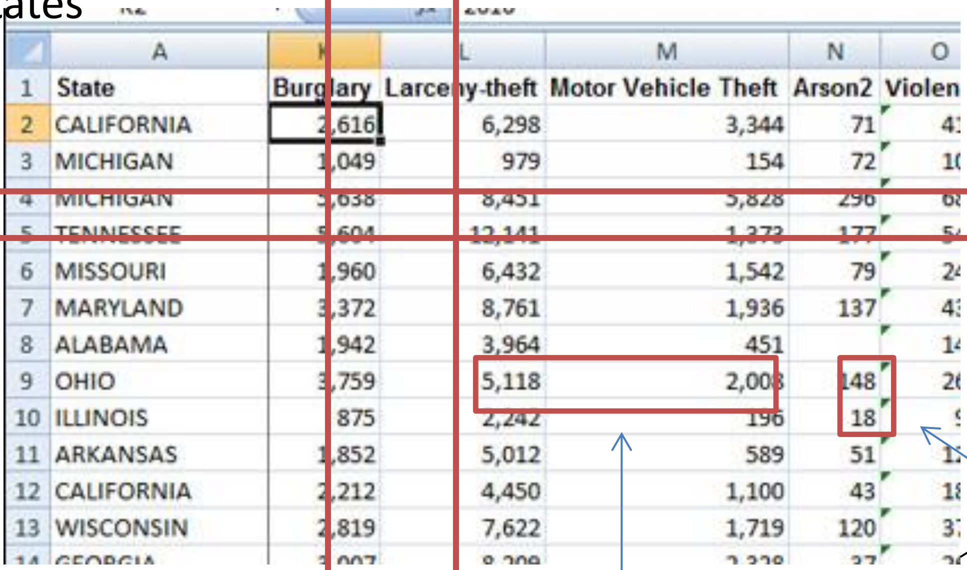
Similarity Measures

Needed it for clustering

- recall Euclidean, correlation, cosine distances

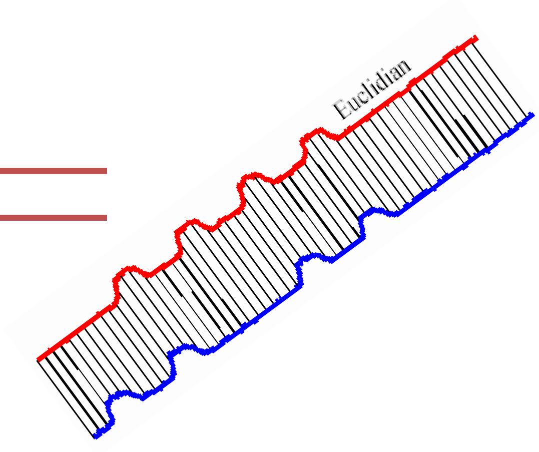
similarity of two states

similarity of two crimes



	A	K	L	M	N	O
1	State	Burglary	Larceny-theft	Motor Vehicle Theft	Arson2	Violen
2	CALIFORNIA	2,616	6,298	3,344	71	41
3	MICHIGAN	1,049	979	154	72	10
4	MICHIGAN	3,638	8,451	5,828	296	68
5	TENNESSEE	5,604	12,141	1,373	177	54
6	MISSOURI	1,960	6,432	1,542	79	24
7	MARYLAND	3,372	8,761	1,936	137	43
8	ALABAMA	1,942	3,964	451		14
9	OHIO	3,759	5,118	2,008	148	26
10	ILLINOIS	875	2,242	196	18	9
11	ARKANSAS	1,852	5,012	589	51	12
12	CALIFORNIA	2,212	4,450	1,100	43	18
13	WISCONSIN	2,819	7,622	1,719	120	37
14	GEORGIA	3,007	8,209	2,328	37	2

two time series

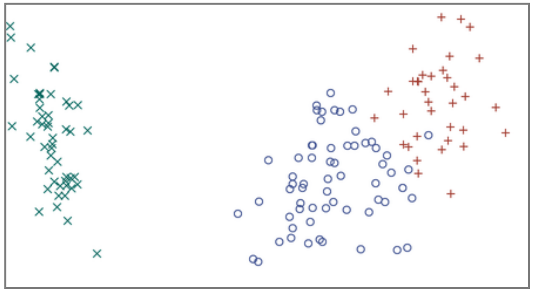


Time

similarity of two crimes in a given state over time

similarity of two states for a given crime over time

Clustering



What can be clustered with these measures?

- crimes (averaged over time)
- states (averaged over time)
- crimes in a given state (taking time series into account)
- states for a given crime (taking time series into account)

You may want to just keep time instances as separate entities

- that will work too
- then you might discover clusters that are sensitive to time
- or you can see how the years relate to another along a trajectory
- as a general rule, when you visualize multivariate data first decide what you will put into the rectangular data matrix (samples, attributes)