

An Introduction to Information Visualization Techniques for Exploring Large Database

Jing Yang
Fall 2005

1

WWW and Internet Visualization

Class 13, Part A
Reference: John Stasko's Infovis class;
<http://www.cs.brown.edu/memex/ACMCSHT/51/51.html>

2

Motivation

- Aid authors and webmasters with production and organization of content
- Assist Web surfers making sense of the information
- Help researchers understand the Web

3

Main Topics

- Presentations of the Internet and WWW
 - Focus on topology and navigation, similar to the graph visualization work
 - Visualizing the evolution of the Web
- Visual aids for browsing and using the WWW and the Internet
 - Assistive visualizations not focusing on presenting net structure and connectivity
 - Visualizing clickstreams
 - Visualizing users
 - Visualizing searches

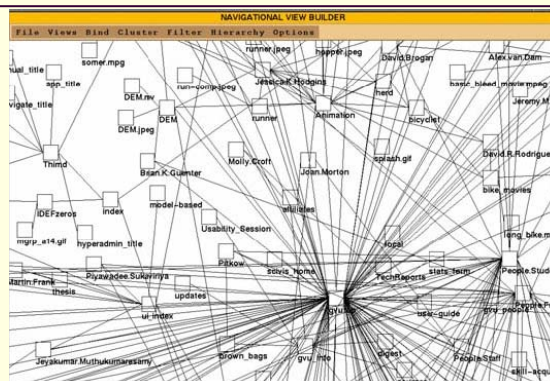
4

Major Challenges

- Websites simply are too big
- Huge graphs
- Layout is challenging

5

The Problem



Mukherjea & Foley
WWW '95

6

CAIDA Macroscopic Topology Measurements project

- Goal: measures connectivity and latency data for a wide cross-section of the commodity Internet
- Method: track global IP level connectivity by sending probe packets from a set of source monitors to hundreds of thousands of destinations stratifying the current IPv4 address space as well as the Earth
- Visualizations are used to analyze the collected data, especially network collectivity

<http://www.caida.org/analysis/topology/macroscopic/>

<http://www.caida.org/tools/measurement/skitter/visualizations.xml> ⁷

AS Internet graph

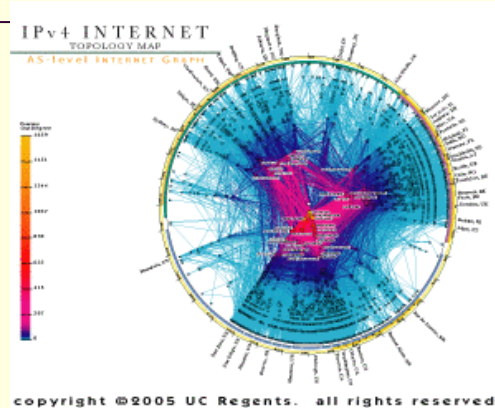
- Input: IP addresses and IP links
 - topology of Autonomous Systems (ASes). Each IP address are mapped to the AS responsible for routing it
- Layout: the position of each AS node is plotted in polar coordinates

$$\text{radius} = 1 - \log\left(\frac{\text{outdegree}(\text{AS}) + 1}{\text{maximum.outdegree} + 1}\right)$$
$$\theta = \left(\text{longitude of the AS headquarters in whois records}\right)$$

- Outdegree: number of "next hop" ASes observed accepting traffic from this AS

⁸

AS Internet graph



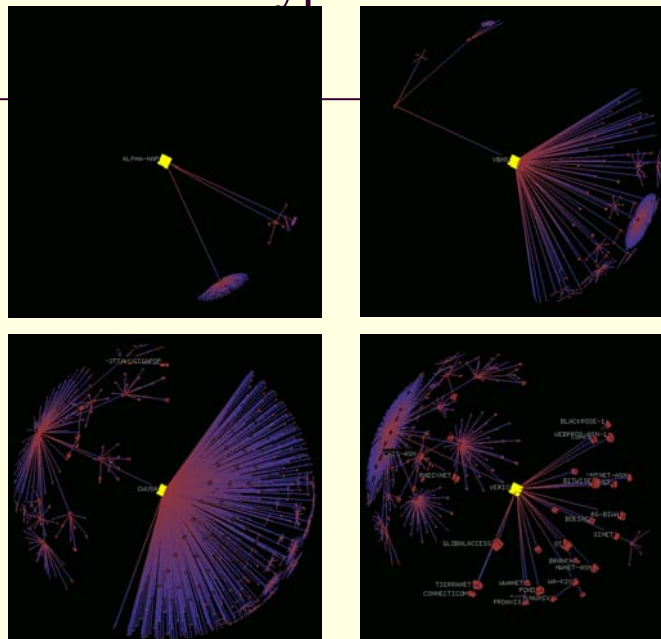
The link color reflects outdegree, from lowest (blue) to highest (yellow).

Insight:

- ISPs in Europe and Asia have many peering relationships with ISPs in the U.S. there are fewer links directly between ISPs in Asia and Europe.

9

AS Paths in Hypviewer



10

Visualizing the Global Topology of the MBone

- <http://graphics.stanford.edu/papers/mbone/>
- Tamara Munzner and Eric Hoffman and K. Claffy and Bill Fenner
- MBone: the Internet's multicast backbone
- Multicast: distributing data from one sender to multiple receivers with minimal packet duplication
- MBone has been extremely popular for efficient transmission across the Internet of real-time video and audio streams such as conferences, meetings, congressional sessions, and NASA shuttle launches
- MBone grew exponentially with no central authority
- video

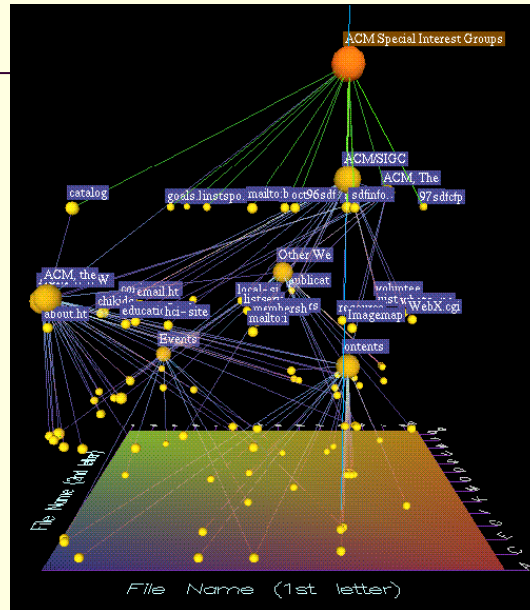
11

Natto [Shiozawa and Matsushita HCI International '97]

- Target: web pages with links
- Initial layout: a flat horizontal plane.
- Node placement: map attributes of the web page (e.g. its size, title, number of images) to the two-axis of the plane.
- Interaction: users can select nodes and raise them vertically to de-occlude the structure. Adjacent (linked) nodes maintain a close proximity to the raised nodes so that the structure is gradually "disentangled" from the plane
- Limitations: the number of nodes that may comfortably occupy the flat plane before selection becomes difficult.

12

Natto [Shiozawa and Matsushita HCI International '97]



Narcissus [Hendley et al InfoVis 95]

- Layout

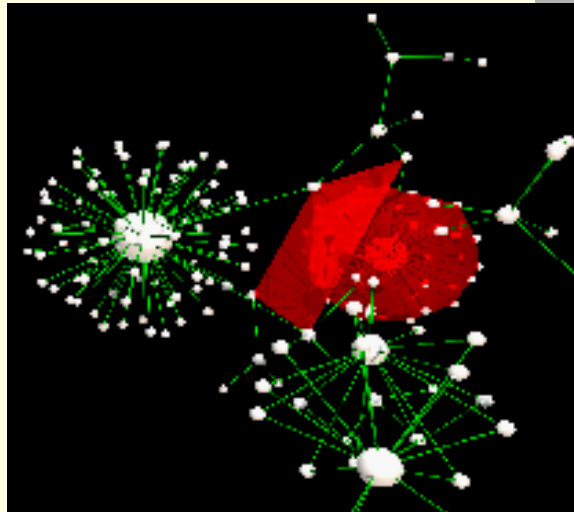
- A graph-like representation of webpages
- Web pages (nodes) exert repulsive forces on each other
- Links between them lead to attractive forces.
- Simulating these forces results in tightly inter-linked sets of pages being grouped into spatial clusters within the visualization

- Interaction

- agglomerating clusters into a single, identifiable object using a translucent surface

14

Narcissus [Hendley et al InfoVis 95]



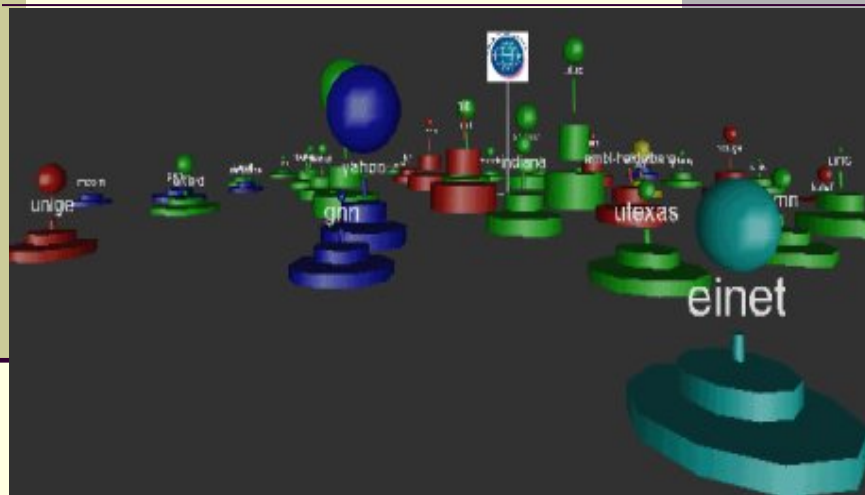
15

Open Text Web Index [Tim Bray 1996]

- Visualize websites as composite objects placed in 3D space
- Represent ariables associated with web sites, such as number of pages, number of links to and from these pages, and domain identifiers
- Use the distance between two objects to represent the degree of connectivity between the two sites
- Work as a map in the site level

16

Open Text Web Index [Tim Bray 1996]



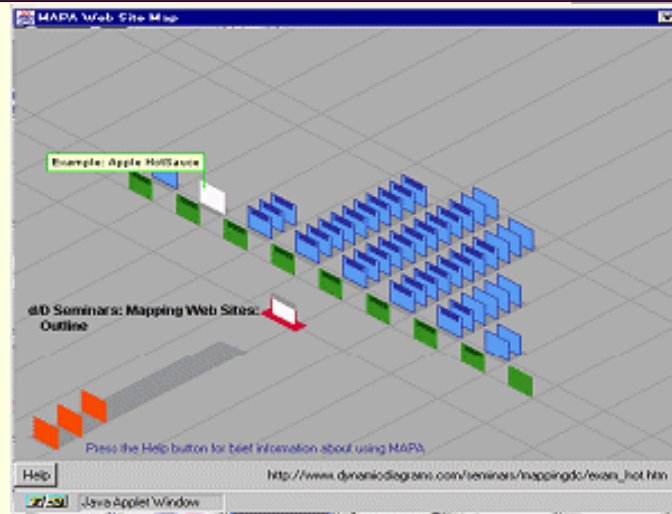
17

MAPA [Durand and Kahn Proc. ACM Hypertext '98]

- Aims to improve navigation in large web sites of between 500 and 50,000 pages
- Presents pages as square icons that stand in rows and columns on a flat plane
- A focus page is placed at the front edge of the plane and its child pages form a row behind the focus page. Each child page then has its children behind it so as to form a single column behind the page

18

MAPA [Durand and Kahn Proc. ACM Hypertext '98]



19

Time Tube [Chi ACM SIGCHI '98]

- Visualises web site evolution
- Disk tree represents the tree structure of a website
 - a disk shape
 - center is the root page
 - hyperlink trees branch out around the root
- A number of disk trees in parallel along a time axis to form a *time tube*, thus demonstrating the evolution of the web site over time
- Interactions:
 - extract slices from the tube and rotate the tube.
 - zoom into each slice to reveal greater detail
 - A conventional Web browser can also be invoked to view specific pages of interest

20

Visual aids for browsing and using the WWW and the Internet

- Potential web-related tasks
 - How and when has info been accessed?
 - Where do people enter and spend time?
 - How do they move about?
 - What paths aren't traversed?
 - Where are they coming from?
 - What has been added, changed, deleted?
 - Do changes affect navigation patterns?
 - Do we need to do a redesign?

21

Data Set

- Each server request is a data case
- Example variables
 - IP Address/Client host
 - Timestamp
 - URL requested
 - HTTP status (success, not found, ...)
 - Bytes delivered
 - Referencing URL (HTTP-Referrer)
 - User agent (browser and OS info)

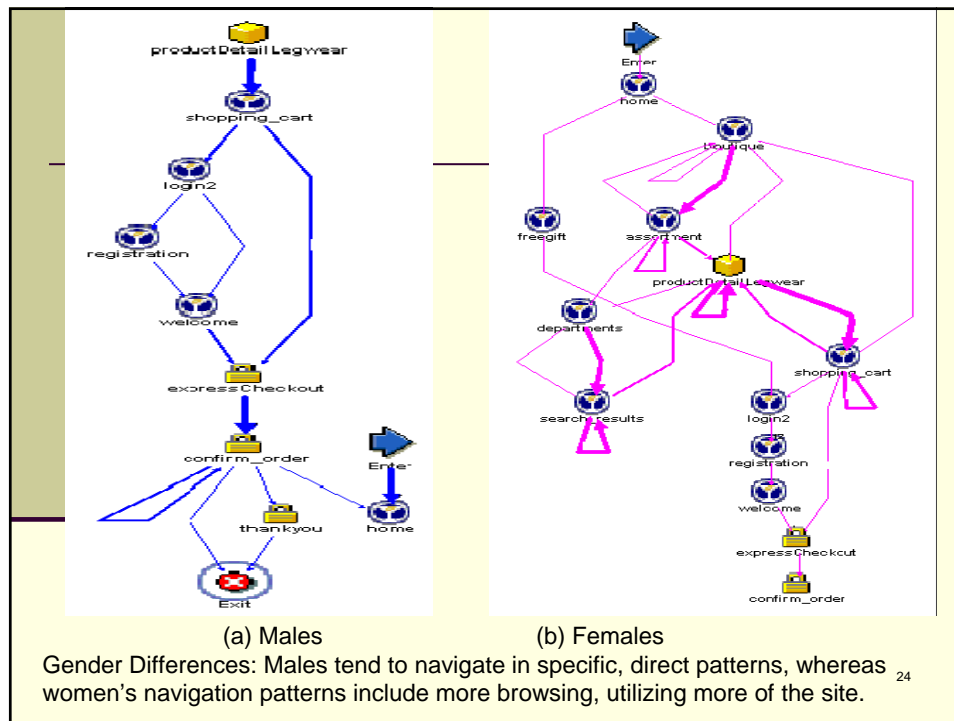
22

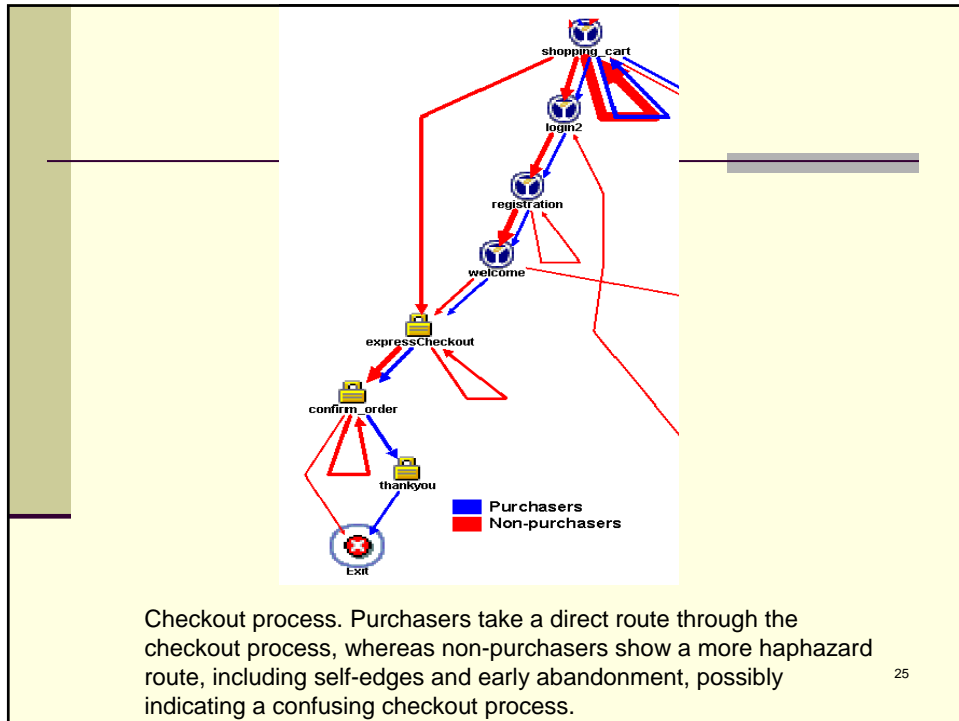
E-Commerce Clickstream

Visualizaiton [Brainerd and Becker Infovis 2001]

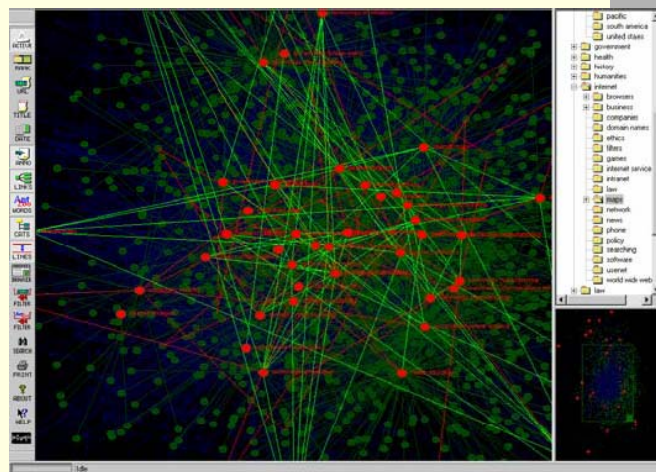
- Brainerd and Becker Infovis 2001
- Goal: analyzing user behavior of a web site
 - Understand the interactions between users and web site
- Visualization:
 - shows site topology and traffic flow
 - Presents a more complete picture of web site usage by segmenting site traffic data based on user attributes, including demographic data and purchase history

23





Web Site Visitations



Download the software from www.inventix.com

Other Applications

Class 13, Part B

27

Microarray Visualization

- Biologists use high-throughput experiments to answer complex biological research questions
- Experiments, such as gene-expression microarrays, result in datasets that are very large
- Reference: An evaluation comparing microarray tools [Saraiya et al InfoVis 2004]

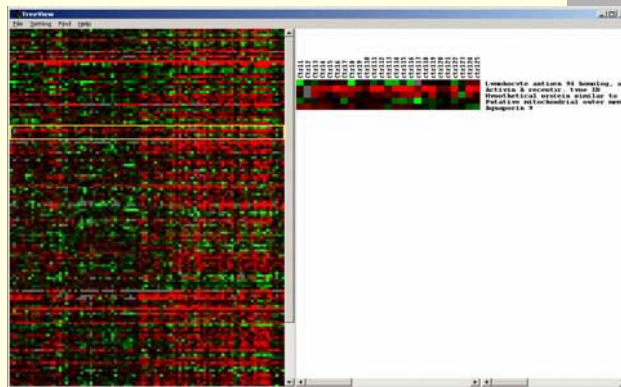
28

Microarray Dataset Examples

- Time Series - Data for 1060 genes over 5 time points of a viral infection cycle in human embryonic kidney cells (1060 rows, 5 columns)
- Viral Conditions - Data for 861 genes for 3 related viral infections at 8 hrs post infection of human lung epithelial cells (861 rows, 3 columns)
- Lupus vs. Control - Data for 170 genes from 42 control (healthy) people and 48 people suffering from systemic lupus erythematosus (SLE), an autoimmune disease (170 rows, 90 columns)

29

Clusterview



Heat-map

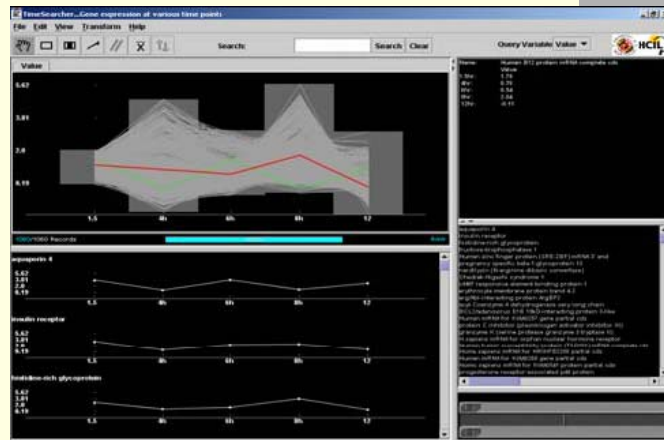
Increased gene-expression values: red brightness scale

Decreased gene-expression values: green brightness scale

Nochange: black.

30

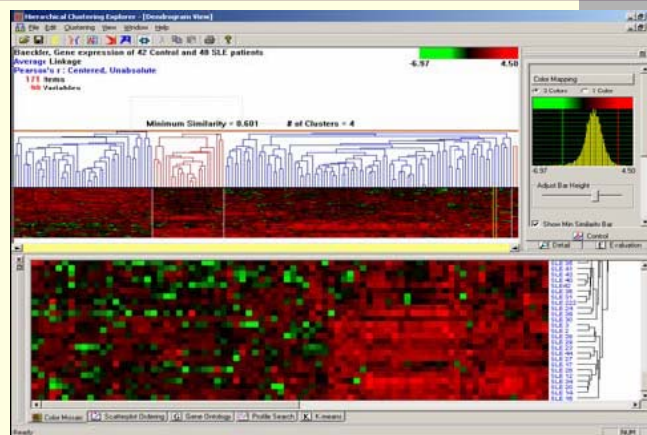
TimeSearcher



Parallel Coordinates for both overview and individual views
Dynamic query for filtering

31

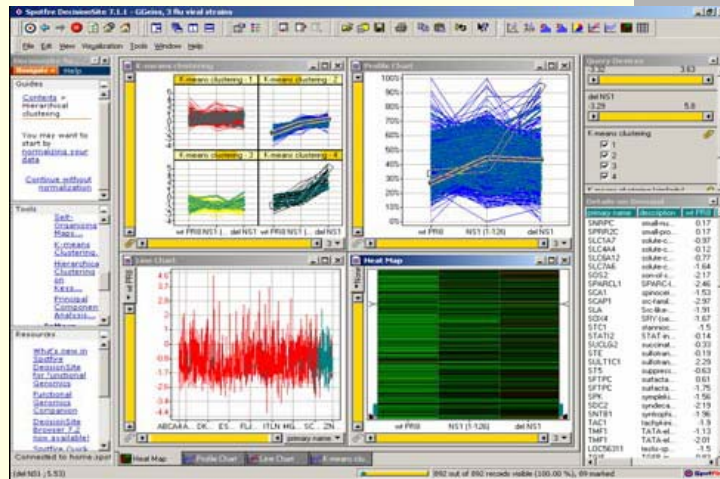
Hierarchical Clustering Explorer



Dendrogram visualization with heat-map

32

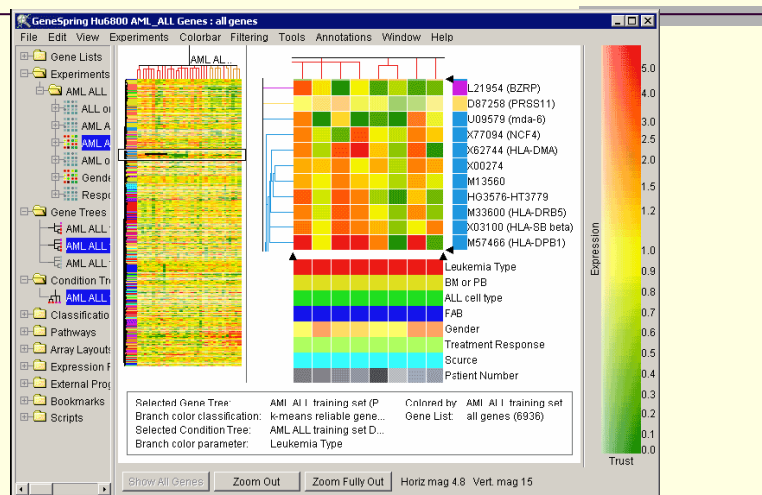
Spotfire



Place each cluster in a separated parallel coordinates window

33

GeneSpring



The largest variety of visualizations for microarray data analysis

34

Visual Analytics

- Visual Analytics is the Science of Analytical Reasoning Facilitated by Interactive Visual Interfaces.

People use visual analytics tools and techniques to

- **Synthesize information and derive insight from massive, dynamic, ambiguous, and often conflicting data.**
 - **Detect the expected and discover the unexpected.**
 - **Provide timely, defensible, and understandable assessments.**
 - **Communicate assessment effectively for action.**
- Reference: *Illuminating the Path: The Research and Development Agenda for Visual Analytics* [Thomas et al 2005]

35

Homeland Security Missions That the Science Will Support

- Preventing Terrorist Attacks
 - Intelligence Analysis
- Reducing Vulnerability to Terrorism
 - Safeguarding Borders
- Minimizing Damage and Recovering from Attacks
 - Emergency Preparation and Response

36

Data - Broad Character of the Problem

- Our ability to Collect data is Increasing at a Faster Rate than our Ability to Analyze it.
- Our Investment in Collection is much Larger than our Investment in Analysis.
- Data is Massive, Multi-Dimensional, Multi-Source, Time Varying, Low Information Density, and "Messy".
- The Analysts, Emergency Response Teams and Border Protection Teams are Nearly Overwhelmed.
- Visual Analytics will Create Methods to Understand the Data Stream to make Decisions in a Time Critical Manner.

37

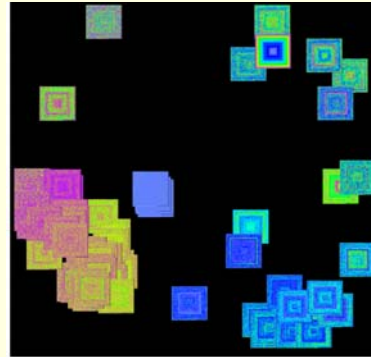
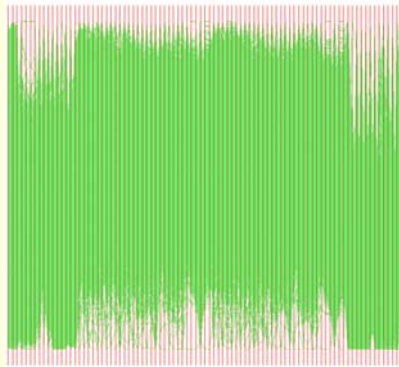
Data - Some Data Types and Volumes

- Textual Data - Reports, Documents, Speeches, E-mails, Web Pages.
Target Rate: One Million New Unstructured "Documents" per Hour
- Databases of Transactional Information - Many corporate and government entities have constructed huge transaction databases with a wealth of information. DHS Alone currently has over 1000 distributed databases.
Target Rate: One Billion New Transactions per Hour
- Geospatial Data - "Imagery" of the Earth, commercial satellites at 1 meter resolution.
- Sensor Data - Miniaturized and Low Cost Computer Systems Enable Vast Distributed Sensor Systems. They Collect Information about their Environment, can Analyze it, and can Communicate. These collectively create huge data rates.
- Video Data - Enhance the Effectiveness of Security in High-Risk Operations.

38

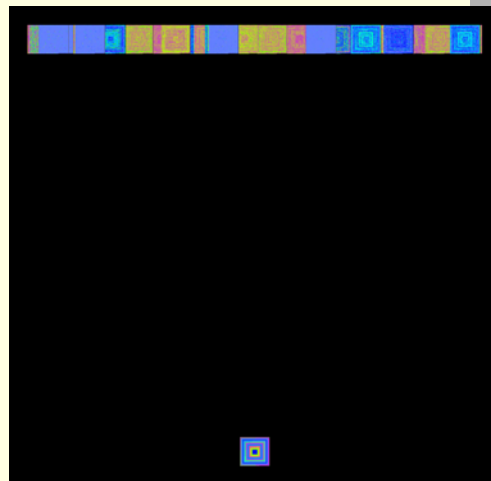
Case Study

- Image classification dataset (89 dimensions, 10,731 data items) [Fan & Luo ACM Multimedia, 2004]



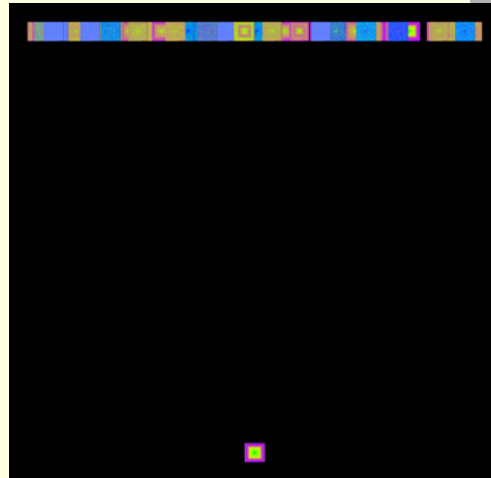
39

Rainfall VaR



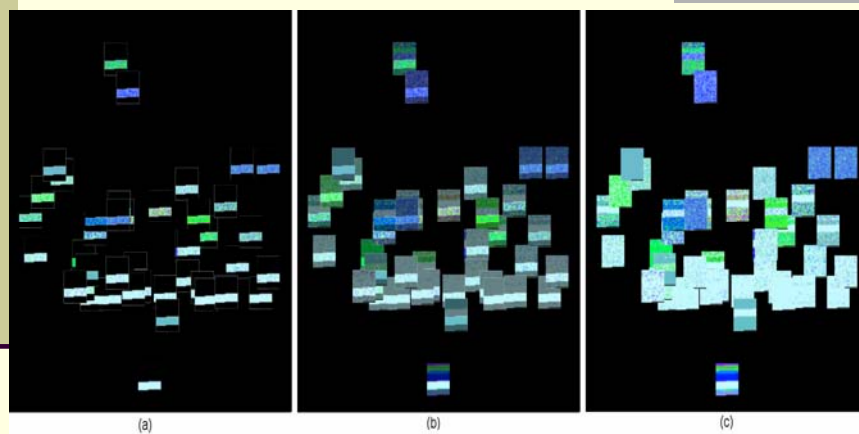
40

Rainfall VaR



41

Masking



42

A Useful Link

- http://www.cc.gatech.edu/classes/AY2005/cs7450_spring/syll.html
- There are some topics not included in this course
- Cognitive tasks
- Color usage
- Automatic design of infovis systems
- Informative art - The class in Dec. 15th!

43

The Last things

- Thank you!
- Good luck in your future study and work!
- Feel free to find me in STECH 435 C, call me at 7046878375, or email me jyang13@uncc.edu, if you have any infovis related problems, or any other things to discuss with me.
- The slides of all classed have been put into WebCT! Feel free to download it!
- Remember, we have another class at Dec. 13th. Tuesday, 7pm – 10 pm
- Final project presentations!

44