Visual Data Mining:
Background, Techniques, and Drug Discovery Applications

Mihael Ankerst
The Boeing Company

Georges Grinstein
UMass Lowell and AnVil Inc.

Daniel Keim
AT&T Research and University of Konstanz

A color version of the tutorial notes can be found via
http://www.fmi.uni-konstanz.de/~keim

KDD'2002 Conference

Emails and URLs

Mihael Ankerst
- Mihael.Ankerst@boeing.com
- http://www.visualclassification.com/ankerst

Daniel A. Keim
- Keim@research.att.com
- keim@informatik.uni-konstanz.de
- http://www.fmi.uni-konstanz.de/~keim

George Grinstein
- Grinstein@cs.uml.edu
- http://genome.uml.edu
- http://www.anvilinfo.com

Overview

Goals of Visualization Techniques

- Presentation
  - starting point: facts to be presented are fixed a priori
  - process: choice of appropriate presentation techniques
  - result: high-quality visualization of the data to present facts
- Confirmatory Analysis
  - starting point: hypotheses about the data
  - process: goal-oriented examination of the hypotheses
  - result: visualization of data to confirm or reject the hypotheses
- Exploratory Analysis
  - starting point: no hypotheses about the data
  - process: interactive, usually undirected search for structures, trends
  - result: visualization of data to lead to hypotheses about the data

Data Exploration

- Definition

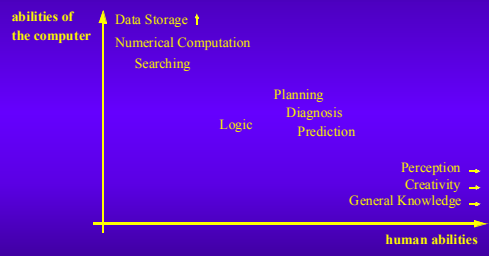
  Data Exploration is the process of searching and analyzing databases to find implicit but potentially useful information

- more formally

  Data Exploration is the process of finding a
  - subset D' of the database D and
  - hypotheses $H_U(D',C)$

  that a *user U* considers useful in an *application context C*

Abilities of Humans and Computers



Brief Historical Overview of Exploratory Data Visualization Techniques (cf. [WB 95])

- **pioneering work of Tufte [Tuf 83, Tuf 90] and Bertin [Ber 81] focuses on**
  - visualization of data with inherent 2D-/3D-semantics
  - general rules for layout, color composition, attribute mapping, etc.
- **development of visualization techniques for different types of data with an underlying physical model**
  - geographic data, CAD data, flow data, image data, voxel data, etc.
- **development of visualization techniques for arbitrary multidimensional data (without an underlying physical model)**
  - applicable to databases and other information resources

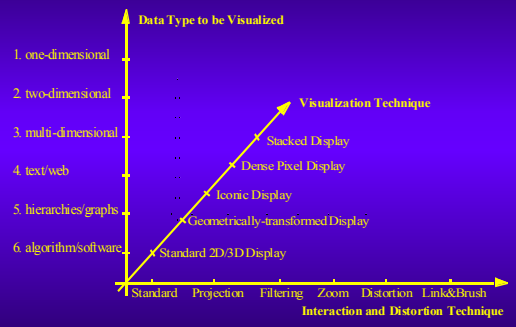
Data Preprocessing Techniques

- **Techniques for Dimension Reduction**
  (Set of d-dim Data Items -> Set of k-dim. Data Items; k<<d)
- **Principal Component Analysis [DE 82]**
  Determines a minimal set of principal components (linear combinations of the original dimensions) which explain the main variations of the data.
- **Factor Analysis [Har 67]**
  Determines a set of unobservable common factors which explain the main variations of the data. The original dimensions are linear combinations of the common factors.
- **Multidimensional Scaling [SRN 72]**
  Uses the similarity (or dissimilarity) matrix of the data as defining coordinate axes in multidimensional space. The Euclidean distance in that space is a measure of the data items.
- **Fastmap [FL 95]**
  Fastmap also operates on a given similarity matrix and iteratively reduces the number of dimensions while preserving the distances as much as possible.

Data Preprocessing Techniques

- **Subsetting Techniques**
  (Set of Data Items -> Subset of Data Items)
  - Sampling (determines a representative subset of a database)
  - Querying (determines a certain, usually a-priori fixed subset of the database
- **Segmentation Techniques**
  (Set of Data-Items -> Set of (Set of Data Items))
  - Segmentation based upon attribute values or attribute ranges
- **Aggregation Techniques**
  (Set of Data-Items -> Set of Aggregate Values)
  - Aggregation (sum, count, min, max,...) based upon
    - attribute values
    - topological properties, etc.
  - Visualization of Aggregations:
    - Histograms
    - Pie Charts, Bar Charts, Line Graphs, etc.

Classification

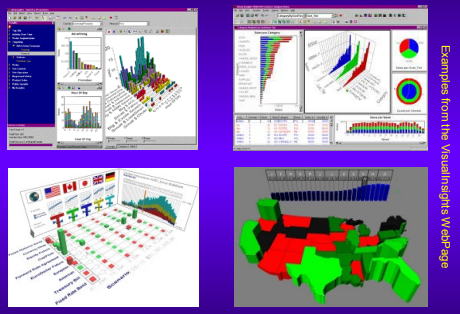


Overview

Visual Data Exploration Techniques

- Standard 2D/3D Displays
- Geometric Transformations
- Iconic Displays
- Dense Pixel Displays
- Stacked Displays

Standard 2D/3D Displays



Examples from the VisualInsights WebPage

Geometric Transformations

Basic Idea:

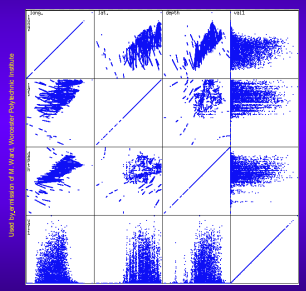
Visualization of geometric transformations and projections of the data

- **Scatterplot-Matrices [And 72, Cle 93]**
- **Landscapes [Wis 95]**
- **Projection Pursuit Techniques [Hub 85]**

  (⇨ techniques for finding meaningful projections of multidimensional data)
- **Prosection Views [FB 94, STDS 95**
- **Hyperslice [WL 93]**
- **Parallel Coordinates [Ins 85, ID 90]**

Geometric Transformations

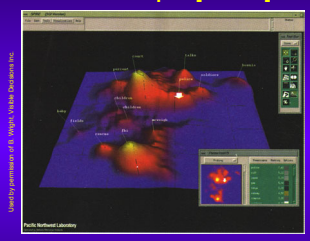
Scatterplot-Matrices [Cle 93]



matrix of scatterplots (x-y-diagrams) of the k-dim. data [total of (k2/2-k) scatterplots]

Geometric Transformations

Landscapes [Wis 95]

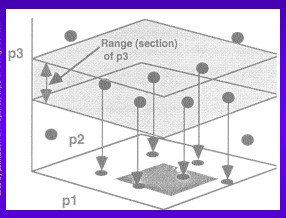


news articles visualized as a landscape

- visualization of the data as perspective landscape
- the data needs to be transformed into a (possibly artificial) 2D spatial representation which preserves the characteristics of the data

Geometric Transformations

Prosection Views [FB 94, STDS 95]

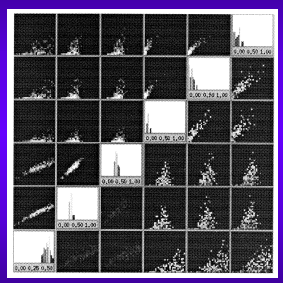


schematic representation example

⇨ matrix of all orthogonal projections where the result of the selected multidimensional range is colored differently (combination of selections and projections)

Geometric Transformations

Hyperslice [93]



matrix of k^2 slices through the k-dim. Data (the slices are determined interactively)

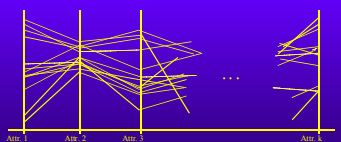
Geometric Transformations

Parallel Coordinates [Ins 85, ID 90]

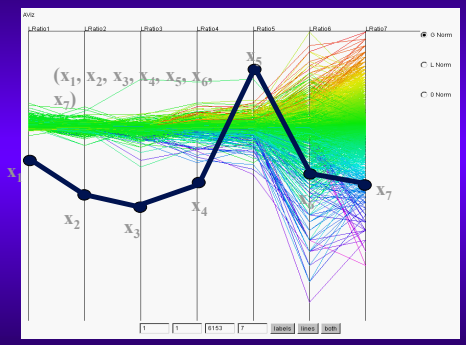
⇨ n equidistant axes which are parallel to one of the screen axes and correspond to the attributes

⇨ the axes are scaled to the [minimum, maximum] - range of the corresponding attribute

⇨ every data item corresponds to a polygonal line which intersects each of the axes at the point which corresponds to the value for the attribute



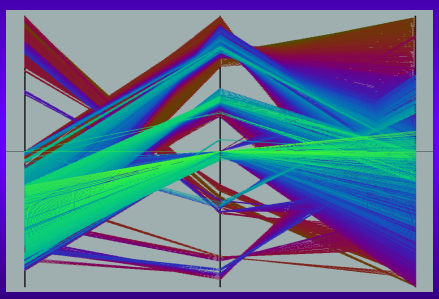
Geometric Transformations



Mihael Ankerst, The Boeing Company -- Daniel A. Keim, AT&T and Univ. of Konstanz
Georges Grinstein, UMass Lowell and AnVil Inc.

Geometric Transformations

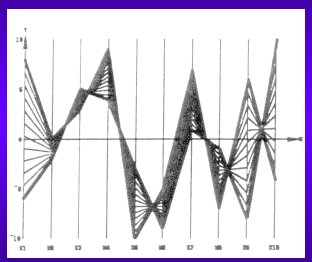
Parallel Coordinates (cont'd)



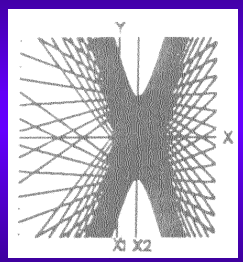
Mihael Ankerst, The Boeing Company -- Daniel A. Keim, AT&T and Univ. of Konstanz
Georges Grinstein, UMass Lowell and AnVil Inc.

Geometric Transformations

Parallel Coordinates (cont'd)



points on a line in 10-dim. space

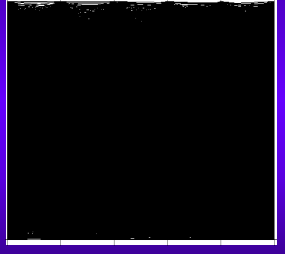


points on a circle in 2-dim. space

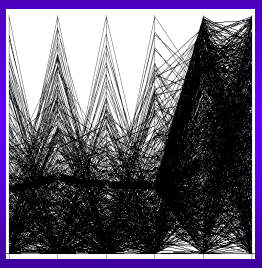
Mihael Ankerst, The Boeing Company -- Daniel A. Keim, AT&T and Univ. of Konstanz
Georges Grinstein, UMass Lowell and AnVil Inc.

Geometric Transformations

Parallel Coordinates (cont'd)



15.000 data items with noise

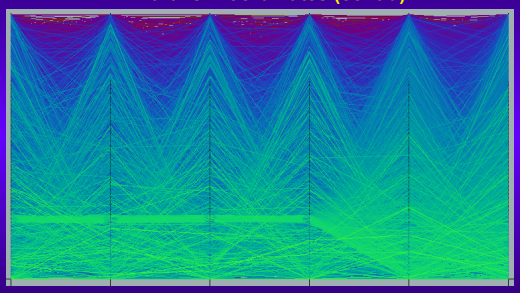


5 % of the data (750 data items)

Mihael Ankerst, The Boeing Company -- Daniel A. Keim, AT&T and Univ. of Konstanz
Georges Grinstein, UMass Lowell and AnVil Inc.

Geometric Transformations

Parallel Coordinates (cont'd)

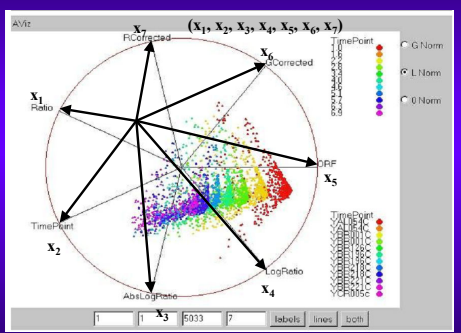


15.000 data items with a query-dependet coloring

Mihael Ankerst, The Boeing Company -- Daniel A. Keim, AT&T and Univ. of Konstanz
Georges Grinstein, UMass Lowell and AnVil Inc.

Geometric Transformations

RadViz™



Mihael Ankerst, The Boeing Company -- Daniel A. Keim, AT&T and Univ. of Konstanz
Georges Grinstein, UMass Lowell and AnVil Inc.

Iconic Displays

Basic Idea: Visualization of the data values as features of icons.

Overview

- **Chernoff-Faces [Che 73, Tuf 83]**
- **Stick Figures [Pic 70, PG 88]**
- **Shape Coding [Bed 90]**
- **Color Icons [Lev 91, KK 94]**
- **TileBars [Hea 95]**
  (a use of small icons representing the relevance feature vectors in document retrieval)

Iconic Displays

Chernoff-Faces [Che 73, Tuf 83]

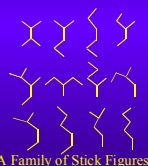
⇨ visualization of the multidim. data using the properties of a face icon (shape of nose, mouth, eyes, and the shape of the face itself)

Iconic Displays

Stick Figures [pic 70, PG 88]

- ⇨ visualization of the multidim. data using the properties of a face icon
- ⇨ two attributes of the data are mapped to the display axes and the remaining attributes are mapped to the angle and/or length of the limbs
- ⇨ texture patterns in the visualization show certain data characteristics

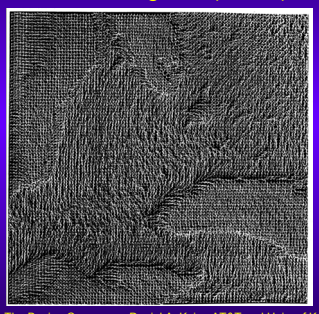
Stick Figure Icon



A Family of Stick Figures

Iconic Displays

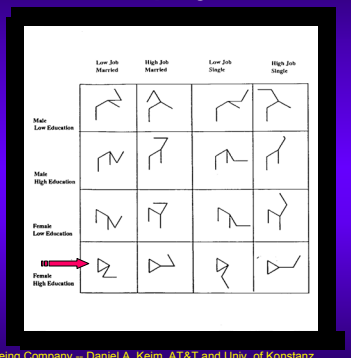
Stick Figures (cont'd)



5-dim. image data from the great lake region

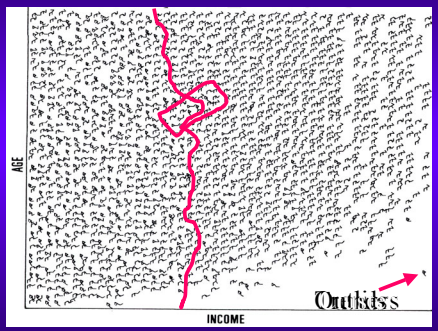
Iconic Displays

ExVis Census Icons



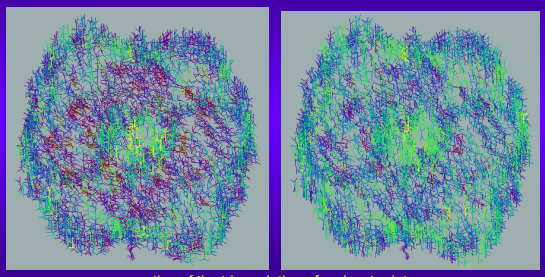
Iconic Displays

Census data image



Iconic Displays

Stick Figures (cont'd)

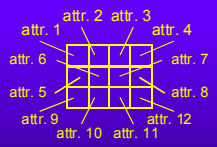


properties of the triangulation of molecute data

Iconic Displays

Shape Coding [Bed 90]

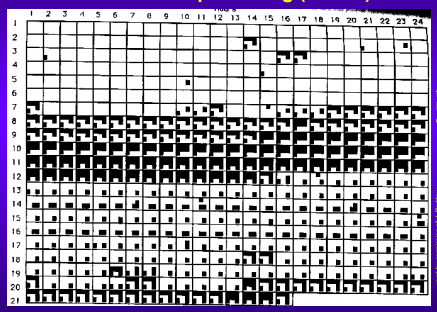
- ⇨ the data are visualized using small arrays of fields
- ⇨ each field represents one attribute value
- ⇨ arrangement of attribute fields (e.g., 12-dimensional data):



- ⇨ arrays are arranged line-by-line according to a given sorting (e.g., the time attribute for time-series data)

Iconic Displays

Shape Coding (cont'd)



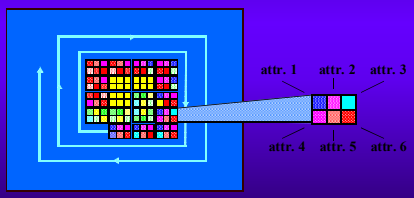
time series of NASA earth observation data

Iconic Displays

Color Icons [Lev 91, KK 94]

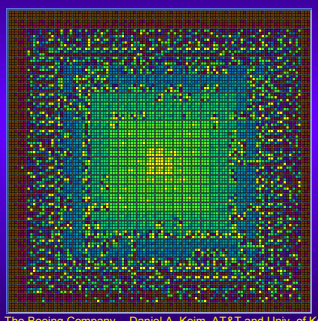
- ⇨ visualization of the data using color icons
- ⇨ color icons are arrays of color fields representing the attribute values
- ⇨ arrangement is query-dependent (e.g., spiral)

schematic representation of 6-dim. data



Iconic Displays

Color Icons (cont'd)

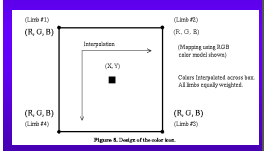


random data containing several clusters

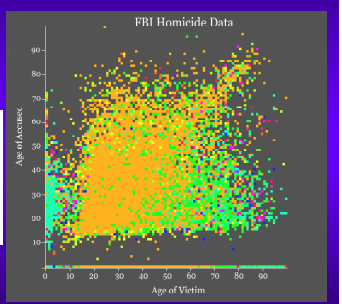
Iconic Displays

Color Icon

FBI Homicide Data

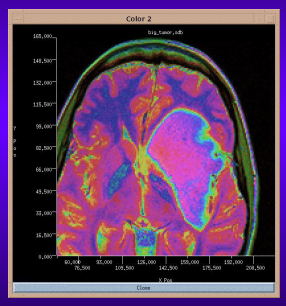


14 Dimensions



Iconic Displays

Color Icon

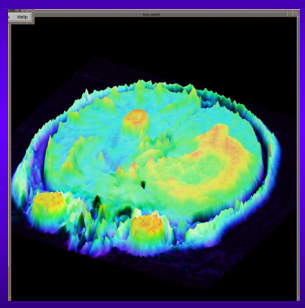


Fused MRI Image data

used by permission of G. Grinstein, University of Massachusetts at Lowell

Iconic Displays

3D Color Icon

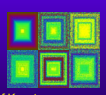


Fused MRI Image data

used by permission of G. Grinstein, University of Massachusetts at Lowell

Dense Pixel Displays

- Task: data exploration and analysis
  ⇨ very large amounts of multidimensional data
- Principle: use the perceptual abilities of humans
  ⇨ adequate presentation of as much information as possible
- Goal: use each pixel of the display to visualize one data value
  ⇨ about 1.3 million data values may be displayed at one point of time
- Idea: map each data value to a colored pixel and arrange them adequately
  ⇨ pixel-oriented visualization techniques

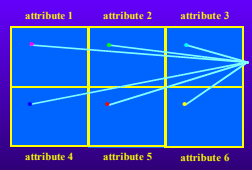


Dense Pixel Displays

Basic Idea

- each attribute value is represented by one colored pixel (the value ranges of the attributes are mapped to a fixed colormap)
- the attribute values for each attribute are presented in separate subwindows
- example:

visualization of six-dim. data



Four Questions

- How should the *pixels be arranged* within the subwindows?

- Are alternative *shapes* of the subwindows possible?
- How can an appropriate ordering of the dimensions be achieved?
- What can be done with *geometry-related data*?

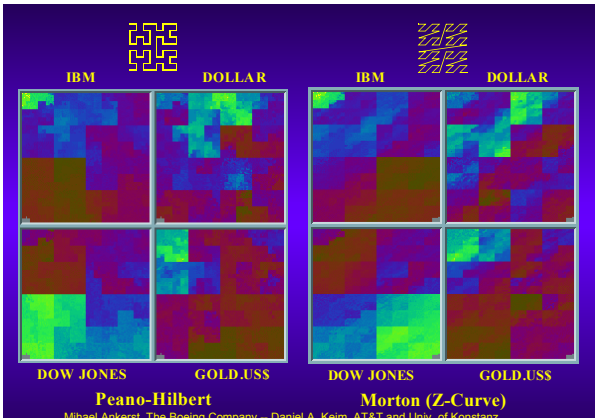
Arrangement of Pixels

Given: Ordered Set of *n* data items $\{a_1, ..., a_n\}$ consisting of *k* data values each $\left(a_i^1, \cdots, a_i^k\right)$

Goal: Two-dim. arrangement of the data values, i.e. bijective mapping $f: \{1...n\} \rightarrow \{1...b\}^*\{1...h\}$ $(n \leq b * h)$, such that the function

$$\sum_{i=1}^{n} \sum_{j=1}^{n} \left| d\left(f(i), f(j)\right) - d\left((0,0), \left(b \cdot \sqrt{\frac{i-1}{n}}, h \cdot \sqrt{\frac{i-1}{n}}\right)\right)\right|$$

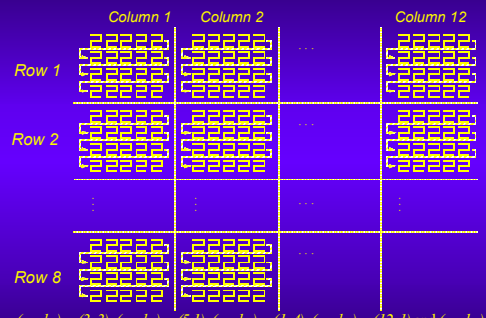
is minimal, where $d(f(i), f(j))$ is the L^p-distance (p=1,2) of the pixels belonging to a_i and a_j



Peano-Hilbert Morton (Z-Curve)

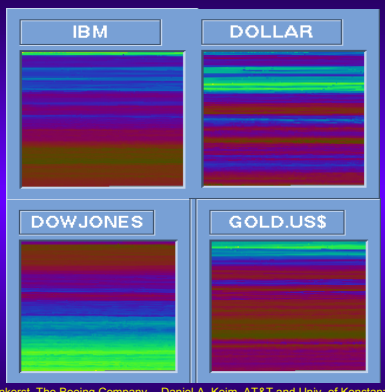
Mihael Ankerst, The Boeing Company -- Daniel A. Keim, AT&T and Univ. of Konstanz
Georges Grinstein, UMass Lowell and AnVil Inc.

Example of a Structured Arrangement

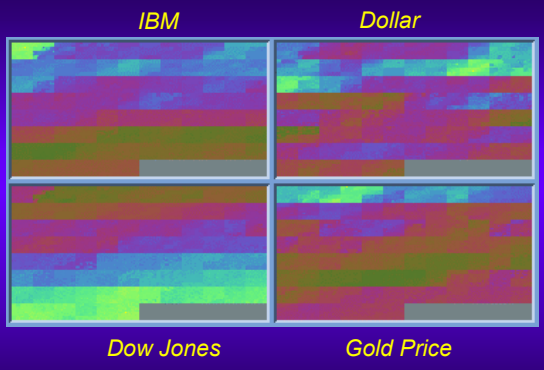


$(w_1, h_1) = (3, 3), (w_2, h_2) = (5,1), (w_3, h_3) = (1, 4), (w_4, h_4) = (12, 1)$ and $(w_5, h_5) = (1, 8)$

Mihael Ankerst, The Boeing Company -- Daniel A. Keim, AT&T and Univ. of Konstanz
Georges Grinstein, UMass Lowell and AnVil Inc.



Mihael Ankerst, The Boeing Company -- Daniel A. Keim, AT&T and Univ. of Konstanz
Georges Grinstein, UMass Lowell and AnVil Inc.



Mihael Ankerst, The Boeing Company -- Daniel A. Keim, AT&T and Univ. of Konstanz
Georges Grinstein, UMass Lowell and AnVil Inc.

Recursive Pattern Technik

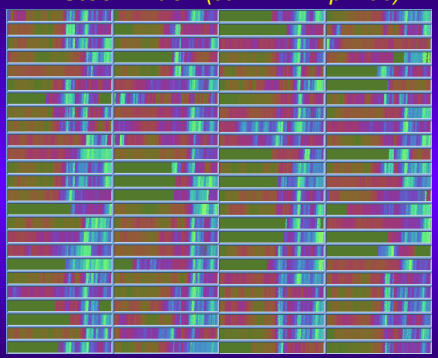
Idea:
- recursive generalization of line- and column-oriented arrangements
- semantic arrangement by allowing user interaction to determine the height h_i and width w_i for each recusion level

Algorithm for recursion level *i*:

Draw w_i pattern of recursion level *(i-1)* in *left-right* direction and repeat this h_i times in *top-down* direction

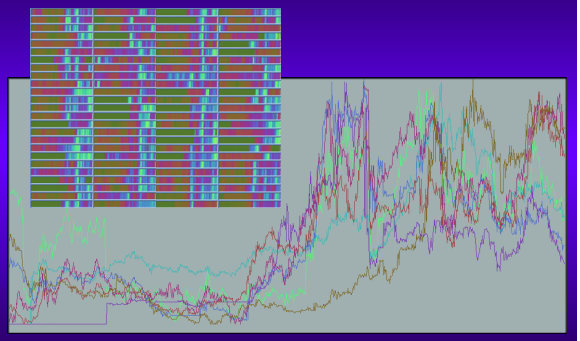
Mihael Ankerst, The Boeing Company -- Daniel A. Keim, AT&T and Univ. of Konstanz
Georges Grinstein, UMass Lowell and AnVil Inc.

FAZ Stock Index (Jan. '74 - Apr. '95)



Mihael Ankerst, The Boeing Company -- Daniel A. Keim, AT&T and Univ. of Konstanz
Georges Grinstein, UMass Lowell and AnVil Inc.

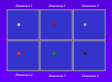
FAZ Aktien Index (Jan. '74 - Apr. '95)



Shape of Subwindows

Idea: Pixel, which belong to the *k* data values of the same data item, should have a small distance.

Goal: Shape of the Subwindows, such that



$$\frac{1}{n}\sum_{k=1}^{n}\left(\frac{1}{k}\sum_{i=1}^{k}\sum_{j=1}^{k} d\left(f(a_k^i), f(a_k^j)\right)\right)$$ is minimal

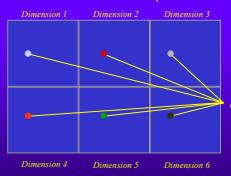
where $d\left(f(a_k^i), f(a_k^j)\right)$ is the L^p-distance of two pixels a_k^i and a_k^j belonging to two different dimensions

Shape of Subwindows

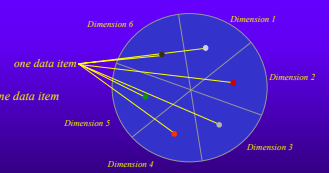
Two Possibilities for the Shape of the Subwindows:

1. two-dim. array of rectangles
2. segmented circle

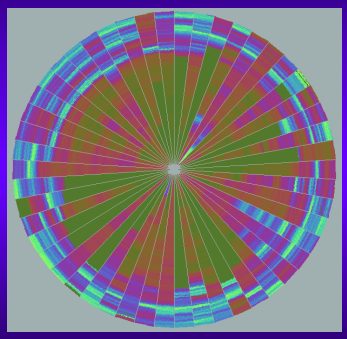
Recursive Pattern & Spiral Technique



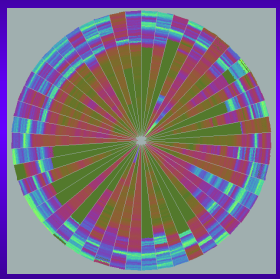
Circle Segments Technique



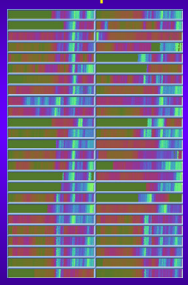
50 Stocks of the FAZ Stock Index



Comparison of Circle Segments and Recursive Pattern Technique



Circle Segments



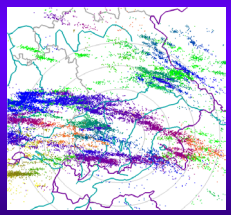
Recursive Pattern

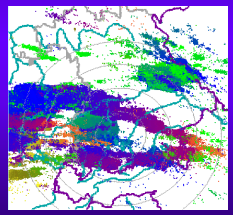
Geometry-related Visualizations

Task: Visualizing Spatial Data

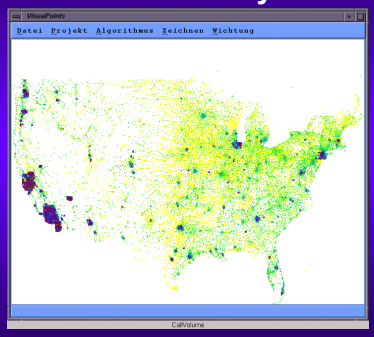
→ Problem: Overlap in the Visualization

Example: Lightning Strikes in Southern Germany





The VisualPoints System



Application of **Gridfit** Alg. to Telcom Data

Stacked Displays

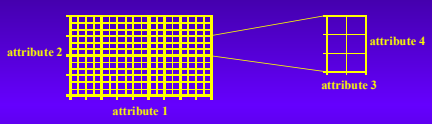
Basic Idea: Visualization of the data using a hierarchical partitioning into subspaces.

Overview

- Dimensional Stacking [LWW 90]
- Worlds-within-Worlds [FB 90a/b]
- Treemap [Shn 92, Joh 93]
- Cone Trees [RMC 91]
- InfoCube [RG 93]

Stacked Displays

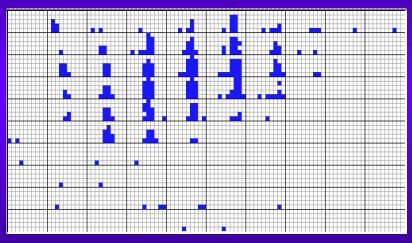
Dimensional Stacking [LWW 90]



- ⇨ partitioning of the n-dimensional attribute space in 2-dimensional subspaces which are 'stacked' into each other
- ⇨ partitioning of the attribute value ranges into classes
- ⇨ the important attributes should be used on the outer levels
- ⇨ adequate especially for data with ordinal attributes of low cardinality

Stacked Displays

Dimensional Stacking [cont'd]

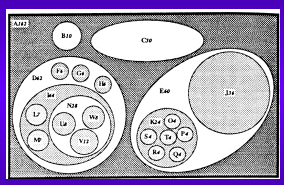


Visualization of oil mining data with longitude and latitude mapped to the outer x-, y-axes and ore grade and depth mapped to the inner x-, y-axes

Used by permission of M. Ward, Worcester Polytechnic Institute

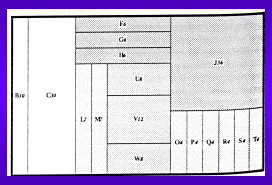
Stacked Displays

Treemap [cont'd]



Used by permission of B. Shneiderman, University of Maryland

Venn Diagram

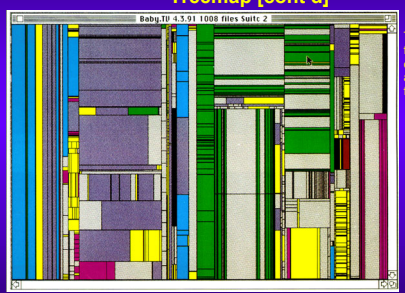


Used by permission of B. Shneiderman, University of Maryland

Tree-Map

Stacked Displays

Treemap [cont'd]



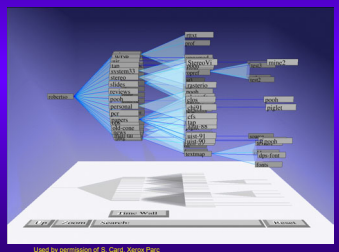
Treemap of a file system containing about 1000 files

Used by permission of B. Shneiderman, University of Maryland

Stacked Displays

Cone Trees [RMC 91, CK 95]

⇨ **animated 3 D Visualizations of hierarchical data**



Used by permission of S. Card, Xerox Parc

File system structure visualized as a cone tree

Hybrid Techniques

Basic Idea: Integrated use of multiple techniques in one or multiple windows to enhance the expressiveness of the visualizations.

⇨ linking diverse visualization techniques may provide additional information

⇨ virtually all visualization techniques are combined with dynamics & interactivity

Examples:

IVEE [AW 95a/b] uses *Starfield Displays* [AS 94] which are scatterplots of icons with dynamic zooming and mapping (combination of geometric, icon-based, and dynamic techniques)

XmDv [War 94] allows to dynamically link and brush scatterplot matrices, star icons, parallel coordinates, and dimensional stacking combination of geometric, icon-based, hierarchical and dynamic techniques)

Comparison of the Technique

Criteria for Comparison [KK 96]

comparison of the described information visualization techniques based on their suitability for certain

⇨ **data characteristics**
(e.g., no. of variates, no. of data items, categorical data, ...)

⇨ **task characteristics**
(e.g., clustering, multi variate hot spots, ...)

⇨ **visualization characteristics**
(e.g., visual overlap, learning curve, ...)

Disclaimer: The following comparison table expresses my personal opinion obtained from reading the literature and experimenting with several of the described techniques. Many of the ratings are arguable and largely depend on the considered data, the exploration task, experience of the user, etc. In addition, implementations of the techniques in real systems usually avoid the drawbacks of a single technique by combining it with other techniques, which is also not reflected in the ratings.

Comparison of the Techniques

Comparsion: An Attempt

		clustering	multi-variate hot spot	no. of variates	no. of data items	categorical data	visual overlap	learning curve
Geometric Transformations	Scatterplot Matrices	++	++	+	+	-	o	++
	Landscapes	+	+	-	o	o	+	+
	Prosection Views	++	++	+	+	-	o	+
	Hyperslices	+	+	+	+	-	o	o
	Parallel Coordinates	o	++	++	-	o	--	o
Iconic Displays	Stick Figures	o	o	+	-	-	-	o
	Shape Coding	o	-	++	+	-	+	-
	Color Icon	o	-	++	+	-	+	-
Pixel Displays	Query-Independent	+	+	++	++	-	++	+
	Query-Dependent	+	+	++	++	-	++	-
Stacked Displays	Dimensional Stacking	+	+	o	o	++	o	o
	Worlds-within-Worlds	o	o	o	+	o	o	o
	Treemaps	+	o	+	o	++	+	o
	Cone Trees	+	+	o	+	o	+	+
	InfoCube	o	o	-	-	o	o	+

Overview

Distortion and Interaction Techniques

- Projection
- Filtering
- Zooming
- Linking and Brushing
- Distortion

Interactive Projections

- dynamic or interactive variation of the projections
- visualization of the remaining parameters in 2D or 3D
- automatic variation results in an animation of the data
- examples:
  - GrandTour [Asi 85]
  - S Plus [BCW 88]
  - XGobi [SCB 92, BCS 96]
  - Influence & Attribute Explorer [STDS 95, SDTS 95]
  - ...

Interactive Filtering

- dynamic or interactive determination of subsets of the database
- distinction between
  - ***selection:*** direct selection of the desired subset
  - ***querying:*** specification of properties of the desired subset
- specific problem: specification of complex boolean conditions
- examples:
  - Magic Lenses [Bie 93] / Moveable Filter [FS 95]
  - Filter-Flow Model [YS 93]
  - InfoCrystal [Spo 93]
  - DEVise [Liv 97]
  - Dynamic Queries [AS 94, Eic 94, GR 94]
  - ...

Interactive Filtering

Magic Lenses / Moveable Filter [Bie 93, SFB 94, FS 95]

- interactive selection using lens-like tools which selectively filter the data in the considered areas
- multiple lenses / moveable filters can be used for a multi-level filtering (allowing complex conditions)

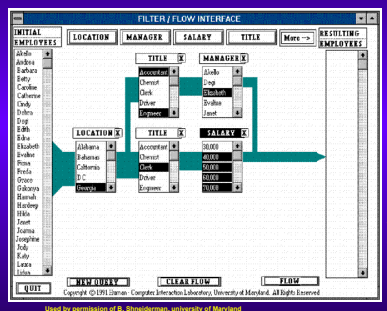
Interactive Filtering

Filter-Flow Model [YS 93]

- selection based on a dataflow-oriented model: the data flows through filter-units which reduce the flow
- especially useful for an intuitive specification of complex boolean queries:
  - AND-connected query portions may be specified using multiple filter units in a pipeline fashion
  - OR-connected query portions may be specified using multiple independent flows which reunite into a single bigger flow

Interactive Filtering

Filter-Flow Model [YS 93]



Used by permission of B. Shneiderman, university of Maryland

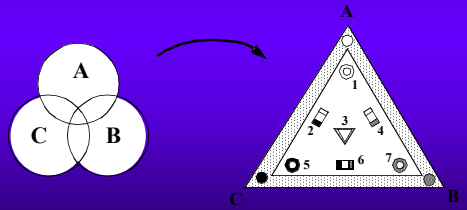
Complex boolean query:

Find the accountants or engineers from Georgia who are managed by Elizabeth or the clerks from Georgia who make more than 30.000!

Interactive Filtering

InfoCrystal [Spo 93]

- specification of complex boolean queries using an intuitive model for specifying complex subsets
- basic idea:

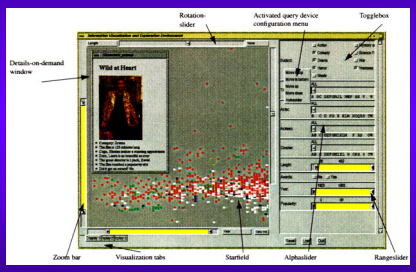


Interactive Zooming

- visualization of large amounts of data in reduced form to provide an overview of the data
- variable zooming of the data with automatic changes of the visualization modes to present more details
- examples:
  - PAD++ [PF 93, Bed 94, BH 94]
  - IVEE [AW 95a/b]
  - DataSpace [ADLP 95]
  - ...
- a comparison of fisheye and zooming techniques can be found in [Sch 93]

Interactive Zooming

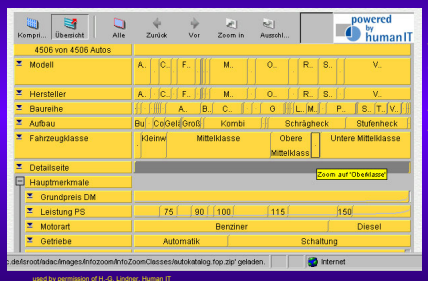
IVEE / Spotfire [AW 95a/b]



used by permission of C. Ahlberg, IVEE Development

Interactive Zooming

InfoZoom [Hum 01]



used by permission of H.-G. Lindner, Human IT

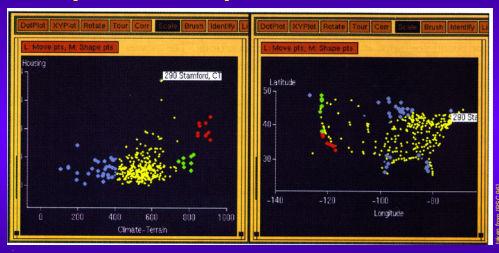
web-pages of the German Automobile Association (ADAC)

Interactive Linking and Brushing

- prerequisite: multiple visualizations of the same data (e.g., visualizations of different projections)
- interactive changes made in one visualization are automatically reflected in the other visualizations
- examples:
  - Xmdv-Tool [War 94]
  - S Plus [BCW 88]
  - XGobi [SCB 92, BCS 96]
  - DataDesk [Vel 92, WUT 95]
  - ...

Interactive Linking and Brushing

XGobi [SCB 92, BSC 96]



climate and housing data of the US

Distortion Techniques

Basic Idea: Distortion of the image to allow a visualization of larger amounts of data

Overview [LA 94]

- Perspective Wall MRC 91]
- Bifocal Displays [SA 82]
- TableLens [RC 94]
- Graph. Fisheye Views [Fur 86, SB 94]
- Hyperbolic Repr. [LR 94, LRP 95]
- Hyperbolic Repr. [LR 94, LRP 95]
- 3D-Hyperbolic Repr. [MB 95]
- Hyperbox [AC 91]

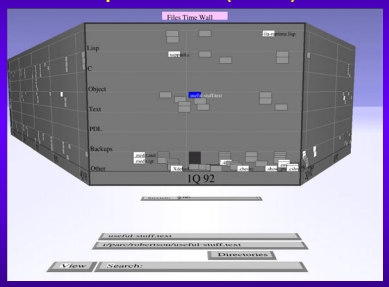
Distortion Techniques

Perspective Wall [MRC 91]

- ⇨ presentation of the data on a perspective wall
- ⇨ the data outside the focal area are perspectively reduced in size
- ⇨ the perspective wall is a variant of the bifocal lens display [SA 82] which horizontally compresses the sides of the workspace by direct scaling

Distortion Techniques

Perspective Wall (cont'd)

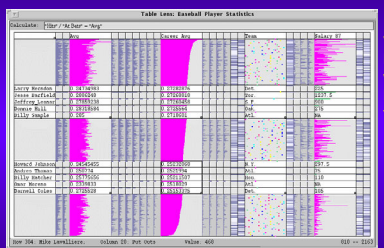


used by permission of S. Card, Xerox Parc

documents arranged on a perspective wall

Distortion Techniques

Table Lens [RC 94]



used by permission of R. Rao, Xerox PARC

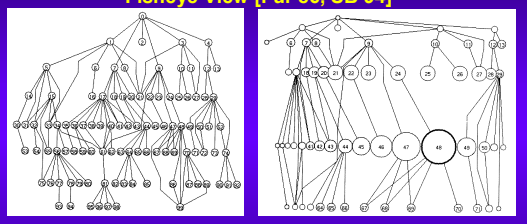
Visualization of a baseball database with a few rows being selected in full detail

- ⇨ compact visualization of a table (spreadsheet/ database) with the possibility of viewing portions of the table in more details

Distortion Techniques

Fisheye View [Fur 86, SB 94]

used by permission of G. Furnas, University of Michigan

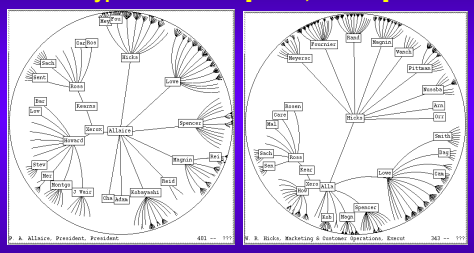


original graph — fisheye view of the graph

- ⇨ shows an area of interest quite large and with detail and the other areas successively smaller and in less detail
- ⇨ graph visualization using a fisheye perspective

Distortion Techniques

Hyperbolic Trees [LR 94, LRP 95]



used by permission of R. Rao, Xerox PARC

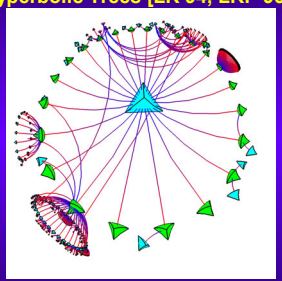
visualization of a large organizational hierarchy

- ⇨ visualization of tree structure in hyperbolic space with different foci

Distortion Techniques

Hyperbolic Trees [LR 94, LRP 95]

used by permission of T. Munzner, Stanford University

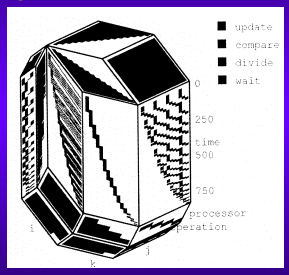


visualization of a large number of connected web-pages

visualization of a graph in 3D hyperbolic conetree-like representation

Distortion Techniques

Hyperbox [AC 91]



Parallel processing performance data visualized as a hyperbox

mapping of scatterplots onto a hyperbox

Overview

Visual Data Mining Systems

Overview

- **Statistics-oriented Systems**
- **Visualization-oriented Systems**
- **Database-oriented Systems**
- **Special Purpose Visualization Systems**

Visual Data Mining Systems

- **Statistics-oriented Systems**
  - ⇨ S Plus [BCW 88] / Trellis [BCS 96]
    (→ generic system for statistical analysis and visualization)
  - ⇨ XGobi [XGobi, SCB 92, BSC 96]
    (→ extensible lisp-based system for statistical analysis and visualization)
  - ⇨ Data Desk [Vel 92, WUT 95]
    (→ commercial system for statistical analysis and visualization;
    features: dyn. linking & brushing of scatterplots and histograms)
  - ⇨ Diamond (SPSS)
    (→ commercial system for statistical analysis and visualization;
    features: dyn. linking & brushing of scatterplots, parallel coordinates, etc.)
  - ⇨ DataSpace [ADLP 95]
    (→ 3D-arrangement of a large number of arbitrary visualizations)

Visual Data Mining Systems

- **Visualization-oriented Systems**
  - ⇨ ExVis [GPW 89]
    (→ features: stick figure and other icon-based techniques)
  - ⇨ Parallel Visual Explorer (IBM)
    (→ features: parallel coordinate technique with query-based coloring, etc.)
  - ⇨ XmDv [War 94, MW 95]
    (→ features: scatterplot matrices, star icons, parallel coordinates, dimensional stacking, dynamic linking and brushing)
  - ⇨ Influence & Attribute Explorer [STDS 95, SDTS 95]
    (→ features: scatterplot and prosection matrices, histograms, dynamic linking and brushing)
  - ⇨ Information Visualizer (Xerox) [HC 86, CRY 96]
    (→ features: diverse information visualization techniques including perspective wall, table lens, cone trees)

Visual Data Mining Systems

- **Database-oriented Systems**
  - ⇨ Hy$^+$ [CM 93]
    (→ features: query and visualizations of hygraphs)
  - ⇨ TreeViz [Joh 93]
    (→ features: treemap technique)
  - ⇨ VisDB [KK 94, KK 95]
    (→ system for interactive slider-based exploration of very large databases
    features: stick figure, parallel coordinate, and pixel-oriented techniques)
  - ⇨ IVEE [AW 95a/b] / Spotfire
    (→ commercial system for database exploration;
    features: generic interactive slider-based visualization environment)
  - ⇨ DEVise [Liv 97]
    (→ system for the generation of interactive special purpose database visualizations)

Visual Data Mining Systems

- **Special Purpose Visualization Systems**
  - ⇨ **Software & Algorithm Visualization**
    (e.g., SeeSoft [ESS 92, BE 96] - a listing of Software & Algorithm Interfaces can be found under "http://wwwbroy.informatik.tu-muenchen.de/~trilk/sv.html" for an overview paper see [SP 92])
  - ⇨ **Web Visualization**
    (e.g., Narcissus [HDWB 95], WebBook and WebForager [CRY 96] - a listing of Web Visualization Interfaces can be found under "http://www.geog.ucl.ac.uk/casa/martin/geography_of_cyberspace.html")
  - ⇨ **Visual Information Retrieval**
    (e.g., Vibe [Ols 93] - a bibliography of Information Retrieval Interfaces can be found under "http://www.pitt.edu/~korfhage/viri_bib.htm"; for an overview paper see [Car 96])

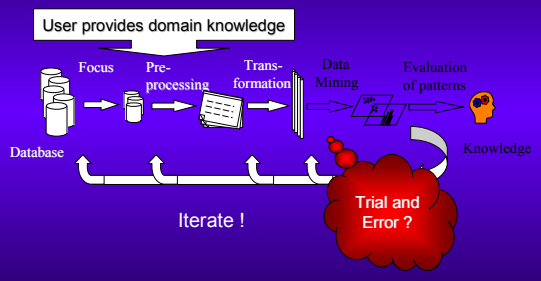
Overview

Specific Visual Data Mining Techniques

- Association Rules
- Classification
- Clustering
- Text Mining
- Tightly Integrated Visualization

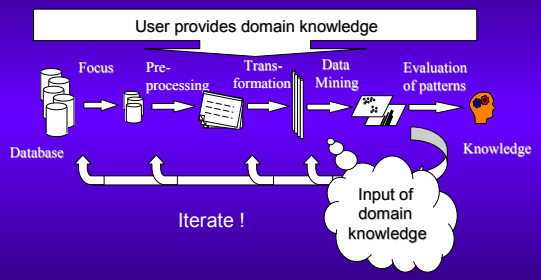
The Human's Role (I)

- Typical KDD Process



The Human's Role (II)

- Human-centered KDD Process



Data Mining <-> Visualization

	Data Mining Algorithms	Visualization
Actionable	+	-
Evaluation	+	-
Flexibility	-	+
User Interaction	-	+

➢ Visual Data Mining

Human Involvement

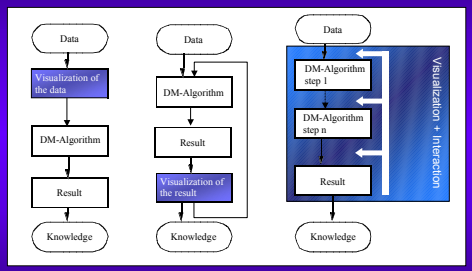
- When ?

☆ Right before the data mining step
- ⇨ Display initial data
- ⇨ Focus on/ narrow relevant search space

◴ During the data mining step
- ⇨ Display intermediate results
- ⇨ Direct the search

◴ After the data mining step
- ⇨ Display the result

Overview



Preceding Visualization (PV) | Subsequent Visualization (SV) | Tightly integrated Visualization (TIV)

Overview

Association Rules

Definitions:

$I = \{i_1,\ldots,i_m\}$ I *Items*,
$t \subseteq I$ t *Transactions*,
$D = \{t_1,\ldots,t_N\}$, $t_i \subseteq I$ D Database,
$X, Y \subset I$

Support of X, s(X): $\frac{|\{t \in D : X \subseteq t\}|}{|D|}$

Confidence of X and Y, c(X,Y): $\frac{s(X \cup Y)}{s(X)}$

Association Rules

Problem description:

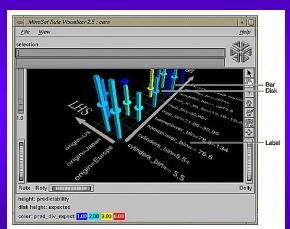
Find all association rules X -> Y with

$s(X \cup Y) \geq s_{min}$

and $c(X, Y) \geq c_{min}$

Subsequent Visualization: Association Rules

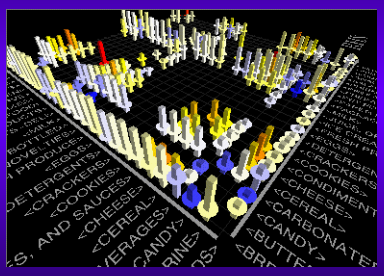
- Rule Visualizer (MineSet) [Min 01]



- LHS and RHS items are mapped to x-, y-axis
- Confidence, support correspond to height of the bar or disc, respectively
- Interestingness is mapped to Color

Subsequent Visualization: Association Rules

- Rule Visualizer (cont'd)



Subsequent Visualization: Association Rules

- *Rule-to-item* visualization [WWT 99]

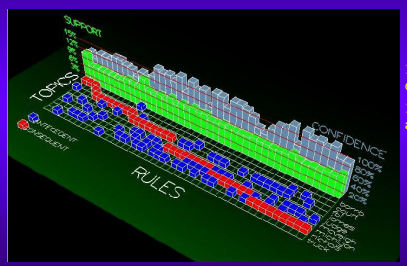
Limitations of item-to-item visualizations:

➢ A->C, B->C or A+B -> C ?

➢ Object occlusion <-> metadata

Subsequent Visualization: Association Rules

- *Rule-to-item* visualization [WWT 99]



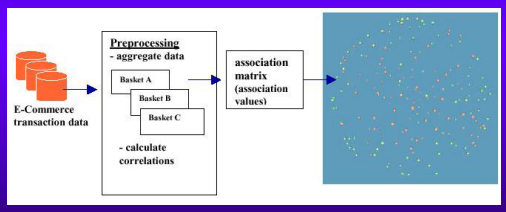
➢ Items are mapped to one axis

➢ Rules can be sorted according to a criterion

Subsequent Visualization: Association Rules

- Market basket Analysis Visualizer [HHD+ 01]

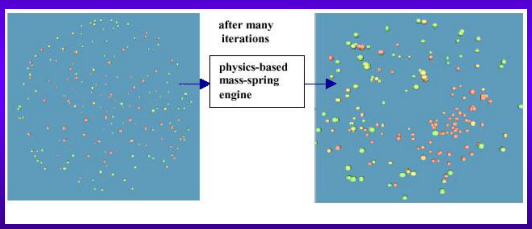
➢ Initialization



Subsequent Visualization: Association Rules

- MAV cont'd

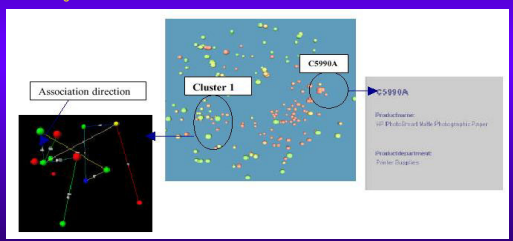
➢ Relaxation



Subsequent Visualization: Association Rules

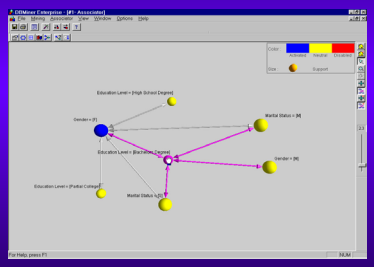
- MAV cont'd

➢ Navigation



Subsequent Visualization: Association Rules

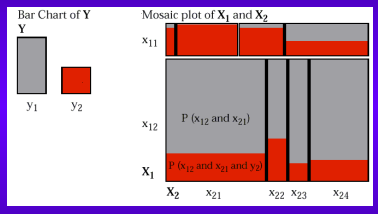
- Association Ball Graph (DBMiner) [DBM 01]



- Items are visualized as balls
- Arrows indicate rule implication
- Size represents support

Subsequent Visualization: Association Rules

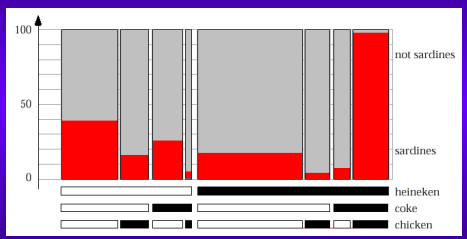
- Interactive Mosaic Plots [HSW 00]



- Visualization of contingency table of attributes within a rule
- Recursive height/width splitting

Subsequent Visualization: Association Rules

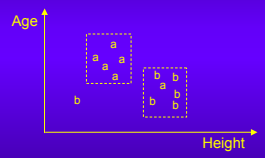
- Double Decker Plots [HSW 00]



- Recursive width splitting

Overview

Classification



Problem description:

Given a set of objects with known class labels.

- Description

  Build model describing the data with respect to the class

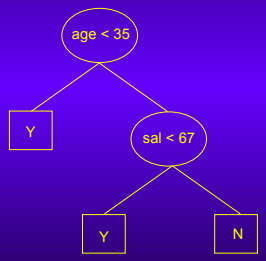
- Prediction

  Use model to predict the class label of objects

Classification

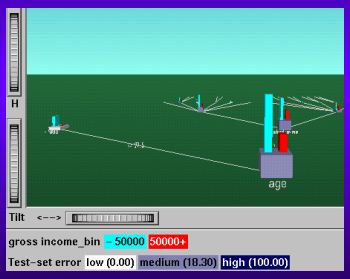
Decision Tree

Age	Salary	Sex	Class
25	15	M	Y
42	40	M	N
29	63	F	Y
81	45	F	N
57	89	M	Y



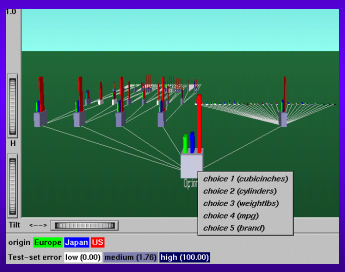
Subsequent Visualization: Classification

- Decision Tree Visualizer (MineSet) [Min 01]



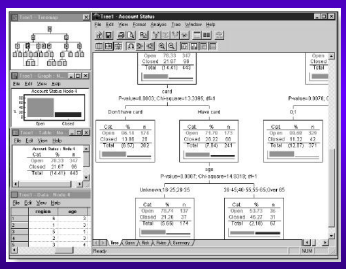
Subsequent Visualization: Classification

- Option Tree Visualizer (MineSet) [Min 01]



Subsequent Visualization: Classification

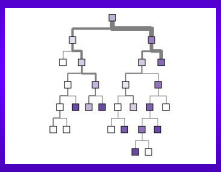
- SPSS AnswerTree [SPS+ 01]



Subsequent Visualization: Classification

- SAS EM [SAS 01]

Tree Viewer

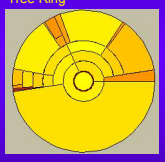


- Color corresponds to relative frequency of a class in a node
- Branch line thickness is proportional to the square root of the objects

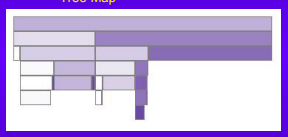
Subsequent Visualization: Classification

- SAS EM [SAS 01]

Tree Ring



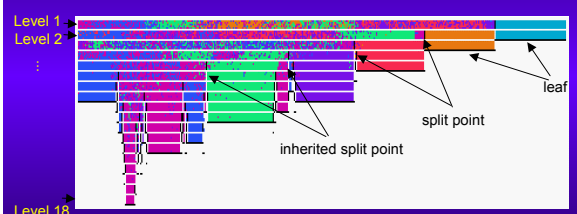
Tree Map



- Color corresponds to relative frequency of a class in a node
- Number of objects in a node are reflected proportionally

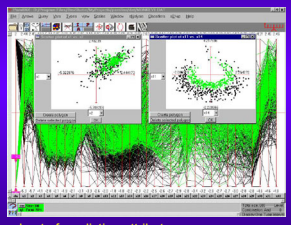
Subsequent Visualization: Classification

- Visual Classification [AEK 00]



Subsequent Visualization: Classification

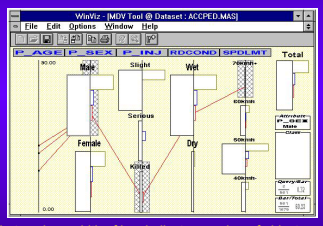
- ParallAX [IA 00]



- Select and order subset of predicting attributes
- Visualize the result based on the parallel coordinates technique

Subsequent Visualization: Classification

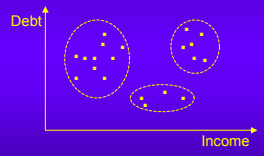
- WinViz [LO 96]



- Left of attribute value: width of box indicates number of objects
- Right of attribute value: class histograms

Overview

Cluster Analysis



Problem description:

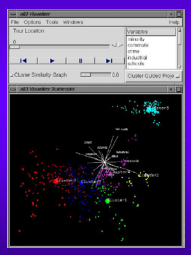
Given a set of objects.

- Group data into *clusters* so that objects within a cluster are very similar

  objects not in the same cluster are dissimilar

Subsequent Visualization: Cluster Analysis

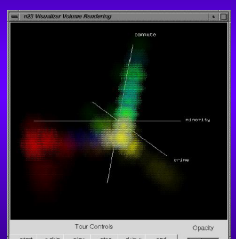
- 3D Dynamic Projections [Yan 00]



- 3D subspace is determined by centroids of 4 clusters 0, 1, 3, 5
- Projection preserves inter-cluster distances
- Projection-determining cluster centroids are visualized as big spheres
- Other cluster centroids are represented as small cubes

Subsequent Visualization: Cluster Analysis

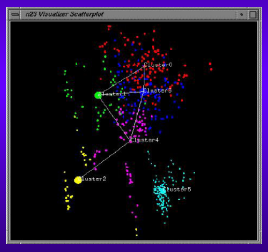
- 3D Dynamic Projections [Yan 00]



- Volume rendering (by splatting) of multi-dimensional volume data to overcome clutter

Subsequent Visualization: Cluster Analysis

- 3D Dynamic Projections [Yan 00]

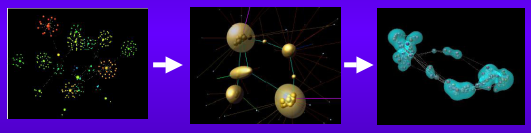


- Cluster similarity graph can be overlaid on to data projections
- User-defined threshold for distance between two cluster centroids

Subsequent Visualization: Cluster Analysis

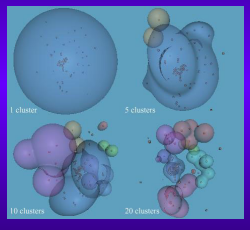
- H-BLOB (Hierarchical BLOB) [SBG 00]

Motivation



Subsequent Visualization: Cluster Analysis

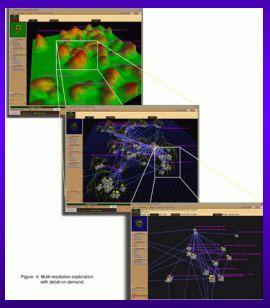
- H-BLOB (Hierarchical BLOB) [SBG 00]



- Cluster hierarchies are shown for 1, 5, 10 and 20 clusters

Subsequent Visualization: Cluster Analysis

- VxInsight [VXI 02]



- Clusters are visualized as hills
- SQL query to database
- Multi-resolution exploration

Subsequent Visualization: Cluster Analysis

- OPTICS [ABKS 99]

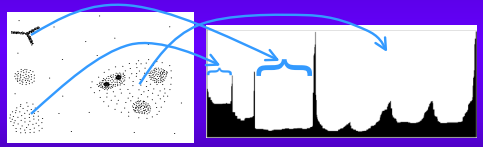
OPTICS = Ordering Points To Identify the Clustering Structure

- Insensitive to Parameters
- Augmented Cluster Ordering
- Reachability-distance: Basis for Interactive Cluster Analysis

Subsequent Visualization: Cluster Analysis

- OPTICS [ABKS 99]

The Reachability-Plot



- Represents the density-based clustering structure
- Easy to analyze
- Independent of the dimension of the data

Overview

Part I: Visualization Techniques
1. Introduction
2. Visual Data Exploration Techniques
3. Distortion and Interaction Techniques
4. Visual Data Mining Systems

Part II: Specific Visual Data Mining Techniques
1. Association Rules
2. Classification
3. Clustering
4. Text Mining
5. Tightly Integrated Visualization

Part III: Drug Discovery Applications
1. Biology and Chemistry
2. Bioinformatics and Cheminformatics
3. Examples
4. Bioinformatics Packages
5. Cheminformatics Packages

Text Mining

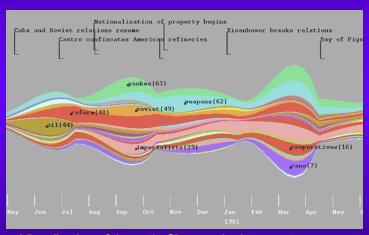
Problem description:

Given unstructured documents.

- Cluster documents
- Summarize documents
- Classify documents

Subsequent Visualization: Text Mining

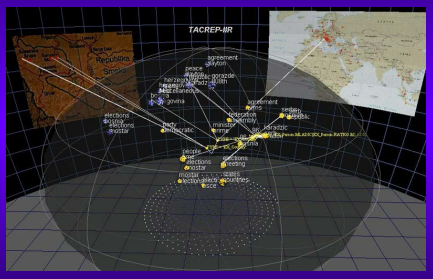
- ThemeRiver [WCF+ 00]



- Visualization of thematic Changes in documents
- Vertical distance indicates collective strength of the themes

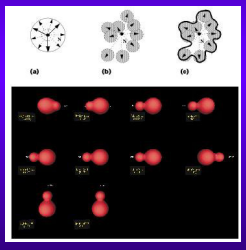
Subsequent Visualization: Text Mining

- Starlight [STA 01]



Subsequent Visualization: Text Mining

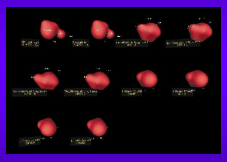
- Shape-based Visual Interface for Text Retrieval [RSE 99]



- Document terms are features
- Term vector proportional to term weight
- Term vectors are spread evenly about the sphere (for all documents)
- Spherical density source field is used to form a surface for each document

Subsequent Visualization: Text Mining

- Shape-based Visual Interface for Text Retrieval



- Shape-based visualization of query result (*Macbeth* query by example)

- Close-up of King Lear

Subsequent Visualization

... and a lot more ... !

e.g. Homepage of visualization group of PNNL:
http://www.pnl.gov/InfoViz/technologies.html

or SOM-based Visualization:
http://www.cis.hut.fi/~juuso/

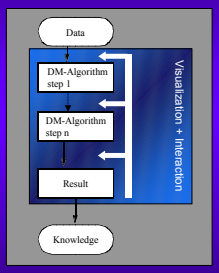
Overview

Part I: Visualization Techniques
1. Introduction
2. Visual Data Exploration Techniques
3. Distortion and Interaction Techniques
4. Visual Data Mining Systems

Part II: Specific Visual Data Mining Techniques
1. Association Rules
2. Classification
3. Clustering
4. Text Mining
5. Tightly Integrated Visualization

Part III: Drug Discovery Applications
1. Biology and Chemistry
2. Bioinformatics and Cheminformatics
3. Examples
4. Bioinformatics Packages
5. Cheminformatics Packages

Tightly Integrated Visualization (TIV)



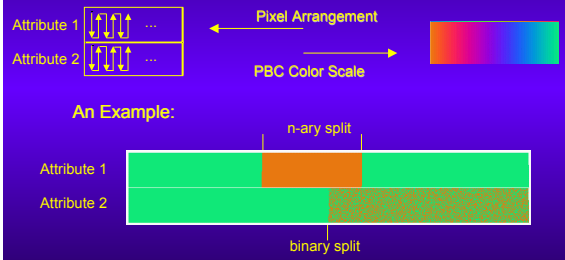
Visualization of algorithmic decisions

- Data and patterns are better understood
- User can make decisions based on perception
- User can make decisions based on domain knowledge
- Visualization of result enables user specified feedback for next algorithmic run

TIV: Visual Classification [AEK 00]

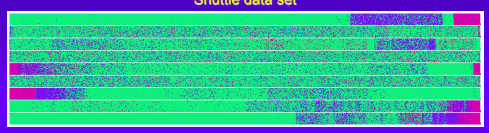
- Each attribute is sorted and visualized separately
- Each attribute value is mapped onto a unique pixel
- The color of a pixel is determined by the class label of the object
- The order is reflected by the arrangement of the pixels

TIV: Visual Classification [AEK 00]

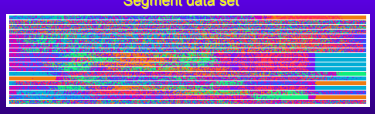


TIV: Visual Classification [AEK 00]

Shuttle data set

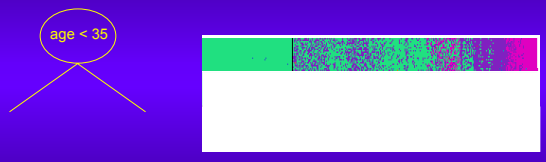


Segment data set



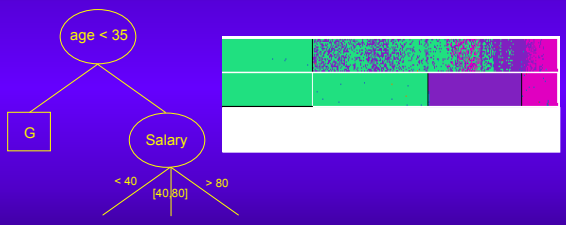
Visual Classification

❑ A New Visualization of a Decision Tree



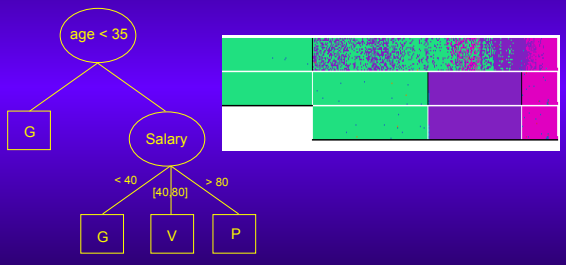
Visual Classification

❑ A New Visualization of a Decision Tree



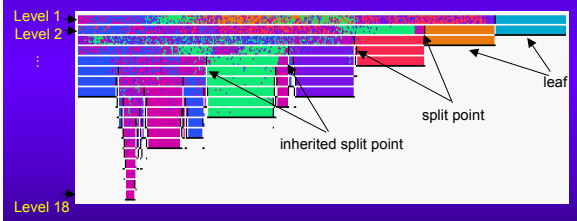
Visual Classification

❑ A New Visualization of a Decision Tree



Visual Classification

- Decision Tree Visualization for the Segment Dataset



TIV: HD-Eye [HKW 99]

- Combining a high-dimensional clustering algorithm with visualization techniques

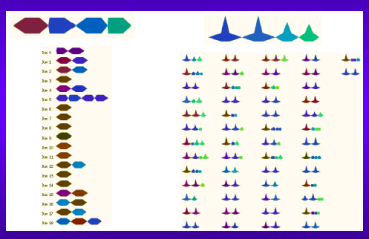
The OptiGrid Algorithm

1. Determine a set of contracting projections $\{P_0,\ldots,P_k\}$
2. Determine the best q separators $\{H_0,\ldots H_q\}$ in the projections
3. If there are no good separators exit;
   Otherwise:
4. Determine a multi-dimensional grid based on $\{H_0,\ldots, H_q\}$
5. Find Clusters C_i in the grid by determining highly-populated grid cells
6. For each C_i: OptiGrid(C_i)

- -> map problem space of finding projections and separators onto visualization space

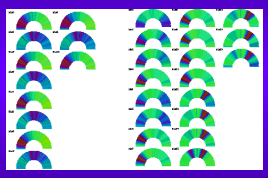
TIV: HD-Eye [HKW 99]

- Visualization techniques for finding contracting projections

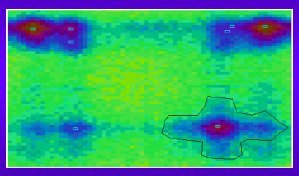


TIV: HD-Eye [HKW 99]

- Visualization techniques for finding separators



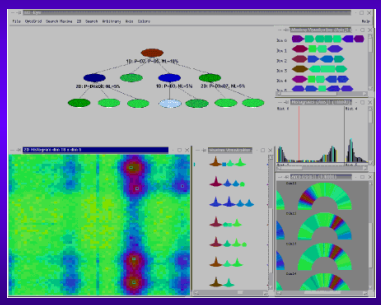
in 1D projections



in 2D projections

TIV: HD-Eye [HKW 99]

- The HD-Eye system



Summary

Visual data mining architectures

Benefits of Visualization/ Interaction	Preceding Visualization	Subsequent Visualization	Tightly int. Visualization
Present/ display patterns		●	●
Search problem space with perception	●		●
Incorporate domain knowledge	●		●
Provide trust and understandability of patterns		●	●

Overview

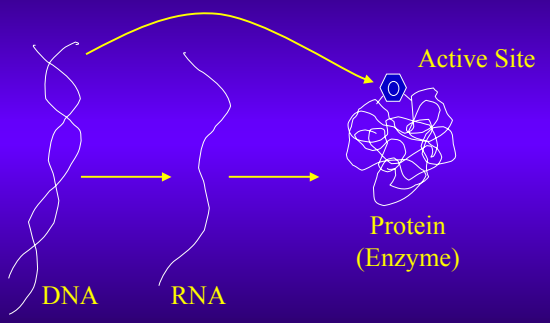
Biology 101 (Slide 1)

- Genomics is the study of the entire genome sequence of an organism
  - Genes are turned on and off at different times in response to different environmental and biological factors
  - Gene expression determines for example
    - The differentiation of organs (ontogeny)
    - The susceptibility and onset to specific diseases
- Proteomics is the study of the entire protein output from all genes and related activities

Biology 101 (Slide 2)

- Cells
  - fundamental working units of every living system
- DNA
  - Arrangement of bases along a strand
- Genome
  - Organism's complete DNA
- Gene
  - Basic physical and functional unit of heredity
- Chromosome
  - Physically separated linear molecules consisting of genes

Biology 101 (Slide 3)



Chemistry 101 (Slide 1)

- Proteins
  - Perform most life functions
  - Make up majority of cellular structures
  - Made up of linear array of amino acids (from 20 different types) which folds into specific 3D structures
- Proteome
  - Collection of all proteins in a cell

Chemistry 101 (Slide 2)

- Proteins include
  - Structural materials (muscle), enzymes (carry out chemistry), receptors (recognize and release messengers), antibodies (recognize foreign objects), peptides (small proteins that act as hormones and messengers)
- Proteins carry out most of the biochemical processes of the living organism
  - Some are bound to DNA, RNA, and other molecules forming very large macromolecules
- They must be present at a the right time in the right amount

Terminology

- Information technology based
  - Bioinformatics
  - Cheminformatics (Chemoinformatics)
- Computational technology based
  - Computational biology
  - Computational chemistry
- Lots of others
- Lots of overlap

Overview

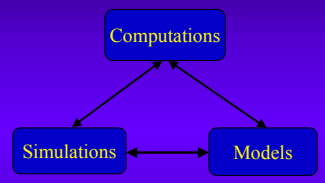
State of the Landscape

- Tremendous IT demands due to
  - Clean room robotics
  - Nanotechnology
  - High performance computing
- Few tools integrate analysis and visualization
- Very few tools integrate biology and chemistry!
- Both present a great opportunity

Classic Informatics as IT Applied to Biology and Chemistry

- Information management, access, manipulation, and retrieval
  - Biology and Chemistry guides and textbooks
  - Journal and database access
- Searches
  - Author, Corporate Name, Subject, Patent, Gene Name (various), SNP, Chemical Name and Formula, Structure, Analytical Chemistry, Physical Property, Chemical Synthesis or Reactions, Chemical Safety or Toxicology Information, …
  - Exact and similarity matches

Modern Computational Informatics



- And then
  - Search results
  - Mine results

Current Computational Chemistry and Cheminformatics

- Lots of 3D modeling and visualizations
- Queries, analyses, predictions
  - Structures from NMR and X-ray
  - Physical properties
  - Chemical synthesis or reactions (chemical and enzymatic)
  - Chemical safety or toxicology
  - …
- Lots of algorithms
- Information mining

Current Bioinformatics and Computational Biology

- Information mining
- More mature than similarity querying
  - Sequence similarity
  - Expression array analyses
  - Lots of analysis tools
    - Statistical
    - Machine learning
  - Some visualizations
- Few integrated analysis and visualizations

Cheminformatics Goals

- Predict a compound's benefit and liabilities (ADMET) as early as possible
- Select (predict)
  - Most promising lead candidates
  - Based on genomic, biological, and chemical databases
- Run simulations and experiments
  - To validate drug classes and activities

Cheminformatics Goals (2)

- Target compounds to a specific protein
  - Induce binding
  - Modulate amount
  - Inhibit binding or produce competition
- Identify proteins responsive to a compound (or classes of proteins)
- Identify compounds responsive to a specific protein (or classes of proteins)

Bioinformatics Goals (1)

- Generate high-quality reference sequences for the entire human genome
- Identify
  - All human genes
  - Their variants
  - Gene control elements and networks
  - Haplotypes
- Sequence and identify the genomes of all living organisms
- Study human variations

Bioinformatics Goals (2)

- Molecular medicine
  - Gene tests
  - Gene therapy
  - Create drugs based on molecular information
- Risk assessment
- Bioarcheology
- DNA identification
- Agriculture and livestock breeding
- Bioprocessing

Others: Proteomics

- Isolate, identify, determine proteins
- Determine function, interactions and pathways
- Key issues
  - Large number of proteins (many more than genes)
  - One gene can form possibly 8 alternative spicing variants
  - The proteome varies continuously in each of the 200 or so types of cells over time
  - Some enzymes cut and rearrange proteins!
  - Many proteins interact indirectly via messengers
  - Post translational modifications (5 per protein)

Others: Macromolecular Structure

- Determine the 3D structure of a molecule
- Four approaches
  - X-ray crystallography
  - Nuclear magnetic resonance
  - Mass spectrometry
  - Computational chemistry
    - Heuristics
    - Molecular mechanics
    - Quantum mechanics
- Lots of new tools (docking programs, pattern discovery, database searching, QSAR, …)

Integrated Analysis and Visualization Systems

- KXEN
- NeuroGenetic
- Optimizer
- AnswerTree
- Clementine
- Alice
- CART, MARS
- Cubist
- See5 (C5.0)
- R, S-Plus
- SAS
- Matlab
- CrossGraphs
- Intelligent Miner
- DecisionSite
- Partek
- SOMine
- OmniVis

Bio/Cheminformatics Tools

- Examples
  - Sequence matching
  - Gene expression analysis
  - QSAR and 3D-QSAR modeling
  - ADME and toxicology prediction
  - Pattern recognition
  - Molecular similarity analysis
  - Diversity analysis
    - Population
    - Structure
  - …

Overview

Part I: Visualization Techniques
1. Introduction
2. Visual Data Exploration Techniques
3. Distortion and Interaction Techniques
4. Visual Data Mining Systems

Part II: Specific Visual Data Mining Techniques
1. Association Rules
2. Classification
3. Clustering
4. Text Mining
5. Tightly Integrated Visualization

Part III: Drug Discovery Applications
1. Biology and Chemistry
2. Bioinformatics and Cheminformatics
3. Examples
4. Bioinformatics Packages
5. Cheminformatics Packages

Bioinformatics Specific Example:

Gene Expression Array

Array Instrumentation

- Array chips
  - Silicon-based, Oligos, 10,000+ genes
  - Custom or potted cDNA arrays, glass, 1,000-5,000+ genes
  - PE glass beads
  - Fiber optics
- Lots of other arrays

Simple Array Data

- Each assay provides expression of genes
- Each measures different conditions
- The array is a matrix of images which are converted to a matrix of numbers
- Informatics activities
  - Manipulate, convert, store, validate, compute, identify, search, mine, present, …

Computations

- Statistics that are meaningful are hard to get – need replicates
- Goal is to come up with hypotheses
  - Classes, clusters, relationships, outliers
- In many cases only a small number of genes change expression levels significantly
- Lots of noise
- Some missing values

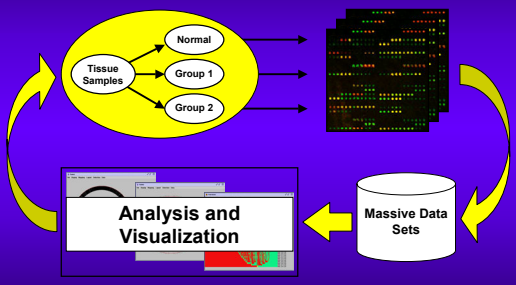
Define Expression Levels

- Threshold differences
- N-fold ratio change
- Statistics (f, T, non-parametric)
  - T assumes normal distributions
  - Large data set requires Bonferroni correction
  - Too small (p = .05, 10,000 genes, yields p/tests= .05/10,000 = .000005)
- Use novel techniques
- Use known-data

Compute

- Analyze and mine
  - Define similarity and other metrics
  - Use Euclidean or other cost functions
  - Reduce dimensionality (sigh)
- Use external information (databases, structures, domain experts)
- Predict or classify or …
  - Cluster
  - Identify associations
  - Delineate outliers

In Summary



Golub & Slonim *et al.*

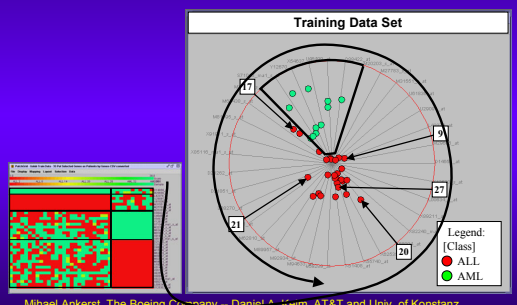
Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring – Golub & Slonim, *et al.*, (1999) SCIENCE 286 531-537.

Whitehead Institute/MIT Center for Genomic Research

ALL - Acute Lymphoblastic Leukemia
AML - Acute Myeloid Leukemia

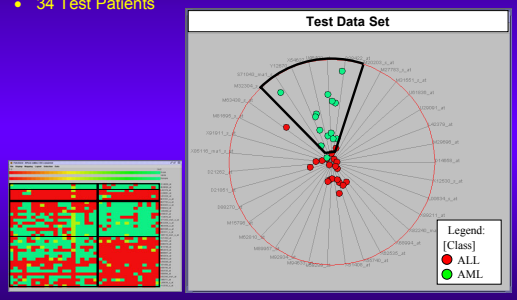
Cancer Classifier

- 38 Training Patients



Cancer Classifier

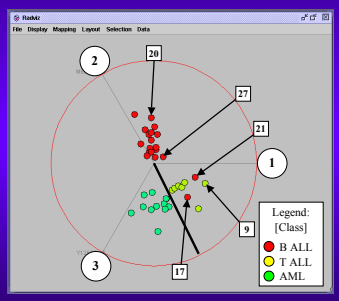
- 34 Test Patients



Class Discovery & Identifying Gene Targets

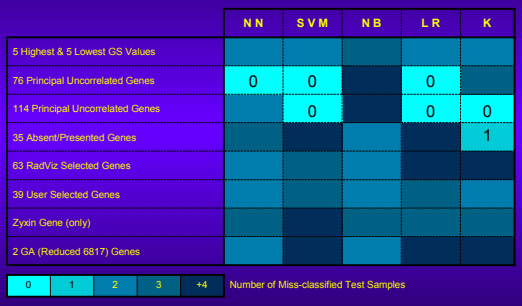
- GA reduction to 3 genes

1. KIAA0102 gene
2. IGB Immunoglobulin-associated beta (B29)
3. LEPR Leptin receptor



Training Set

Picking Best Answers

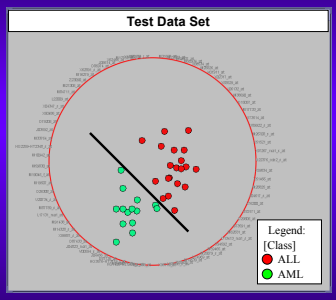


Cheminformatics Specific Example:

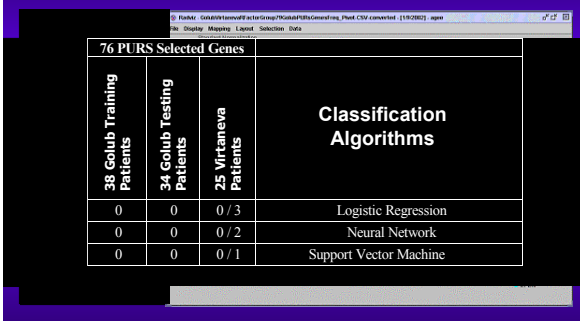
Structural Predictive Toxicology

PURS™ Cancer Classifier

$p < 0.10$	7 genes
$p < 0.05$	6 genes
$p < 0.01$	4 genes
$p < 0.001$	2 genes



Testing the Classifier



76 PURS Selected Genes			
38 Golub Training Patients	34 Golub Testing Patients	25 Virtaneva Patients	Classification Algorithms
0	0	0 / 3	Logistic Regression
0	0	0 / 2	Neural Network
0	0	0 / 1	Support Vector Machine

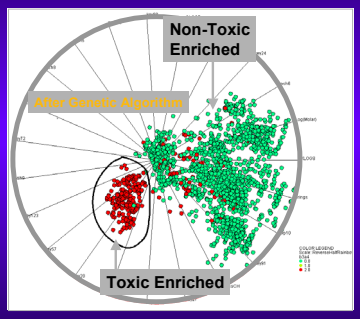
Gene Expression and Drug Toxicity

- Technical Objective
  - Find marker genes informing on drug hepatotoxicity by animal dose and gender.
- Client's Data
  - Affymetrix DNA microarray data
  - Dose and gender correlations to PK/TK and histopathology
  - Pilot study with experimental design problems and noisy data

Example – Predicative Toxicology

- 100,000 chemicals (records)
- 280 data fields (variables)
  - 1 biological assay
  - 4 liver isozyme assays
  - 275 chemical descriptors
    - 166 Substructure Search Keys – ISIS/Host
    - 109 Electro-topological State Indicators - MolConnZ

Toxicity Distinction



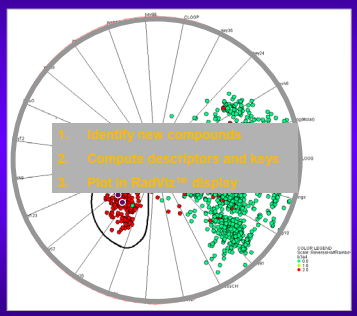
test

Example – Web Site

Centre for Molecular and Biomolecular Informatics (CMBI)

University of Nijmegen, NL
Http://www.Kun.nl/

Toxicity Prediction

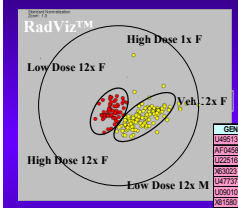


test

CMBI Bioinformatics Services

- Provide access to very fast database searches
  - Smith & Waterman, FrameSearch and Profile searches
  - on the Bioccelerators, dedicated computers for high-speed, sensitive searches
- Sequence Data Distribution Service
  - For commercial users
  - Provides up-to-date sequence databases for local (in house) access
  - may use this service for data transfer

Efficacy & Toxicity



Drug Dos (Hi/Lo) | Gender (F/M)

Marker Genes for Dose & Gender

GENE_ACC	FC(Hi/Lo)	FC(F/M)	Description
U48513	2.6	0.9	Small inducible cytokine A9
AF045887	1.6	1.3	Angiotensinogen
U22516	1.6	1.5	Angiogenin
X63023	1.5	1.0	Cytochrome P450, steroid inducible 3a13
U47737	1.5	1.3	thymic shared antigen-1 (TSA-1)
U09010	1.4	0.6	Mannose binding lectin, liver (A)
X81580	1.3	1.2	Insulin-like growth factor binding protein 2
X04480	1.0	-1.2	Insulin-like growth factor 1
D89613	-1.1	1.0	Cytokine inducible SH2-containing protein
X74938	-1.8	2.3	Hepatocyte nuclear factor 3 gamma (winged helix transcription factor)
X62940	3.2	-2.6	Transforming growth factor beta 1 induced transcript 4
M10062	2.3	-1.8	Mouse IgE-binding factor mRNA
U04428	1.5	-2.0	mu-class glutathione s-transferase (mGSTM5)
AF109906	1.1	-1.4	MHC class III region RD
Y08135	1.1	-1.4	ASM-like phosphodiesterase 3a
Y14004	1.1	-2.2	acyl-CoA thioesterase
X15903	-1.8	-1.3	Cytochrome C oxidase, subunit Va
X70764	-2.0	2.0	ELKL motif kinase
U31993	-2.3	3.2	Interleukin 17 receptor
X63451	-2.8	1.4	Glutathione S-transferase, pi 2
AF109905	-4.5	8.0	major histocompatibility locus class III regions Hsc70t
J02980	-4.8	1.6	Alkaline phosphatase 2, liver

Gene Clusters
Dose
Gender

Programs

- General analysis packages
- Phylogeny inference
- Database search
- Multiple alignment
- Linkage analysis
- Modeling and 3D structure retrieval
- Online documentation

Databases

- Major sequence and related databases
- Genome databases

General Analyses Packages

- The GCG package (a.k.a. The WISCONSIN package), version 10.1 (8.1 still accessible)
- EMBOSS version 1.0
- EGCG version 8.1, extensions to the GCG package (rice et al.)
- The STADEN package (Staden), version 2000.0

Phylogeny Inference Packages

- The PHYLIP package for inferring phylogenies (Felsenstein)
- The PAUP* package (Swofford)
- The MOLPHY program suite (Hasegawa)
- PUZZLE (Strimmer & von Haeseler)

Database Search Packages

- SRS (Sequence Retrieval System, Etzold)
  - The CMBI SRS Server features one of the largest collections of interlinked databases available for this system
- Bioccelerator software & hardware (Compugen)
- NCBI software, like BLAST and ENTREZ
- The FASTA suite of programs (Pearson)
- The ACeDB database system (Durbin & Thierry-Mieg)

Alignment Algorithms

- Oxford Molecular (OML) Software
  - CAMELEON and IDITIS
- CLUSTALW (Higgins et al.)
- T-Coffee (Notredame et al.)
- Principal Coordinate Analysis programs (Higgins)
- TkDCSE
  - a multiple alignment editor (De Rijk)
- Many local extensions and interfaces

Secondary Structure Prediction

- PredictProtein (Rost & Sander)
- GOR IV (Garnier et al.)
- DSSP (Kabsch & Sander)
- Stride, Predator (Argos et al.)
- DSC (King)
- ASSP, Alscript (Barton)
- SecCons (Leunissen et al.)

Linkage Analysis

- CRI-MAP
  - constructs multilocus linkage maps (Green)
- The MapMaker programs EXP and QTL (Lincoln, Daly, and Lander)
- Fastlink
  - faster versions of the standard Linkage programs
  - parallel version also available (Schaffer)

Modeling and 3D Structure Retrieval

- The WhatIf program (Vriend)
- SYBYL
  - Tripos MM/MD program with Basic, BioPolymer, Advanced Computation, Dynamics, Composer and QSAR/CoMFA Modules
- MolScript
  - a program to generate publication-quality plots of protein structures
- ProCheck
  - to verify protein structures

Modeling and 3D Structure Retrieval

- CSD Cambridge Crystallographic Database with programs Quest, Pluto, Gstat, and Vista
- PDB (Protein, PDBBrowse)
  - X-ray protein structure data
- MMCC Results
  - A browse and search service for the issues Vol. 3 to Vol. 6 (1994 - 1997) of the MMCC Results abstracts

Online Documentation

- GCG Program Manual
- EGCG Program Manual
- PHYLIP Manual
- Clustal W Manual
- And more

Major Sequence and Sequence-Related Databases

- EMBL
- EMBLNEW
- GENBANK
- SWISS-PROT
- SWISSNEW
- PIR
- OWL
- PDB
- GENESEQ
- TREMBL
- SBASE
- PRODOM
- KABAT
- PROSITE
- PRINTS
- REBASE
- HOVERGEN
- BLOCKS

Genome Databases

- The Dutch mirror of the Human Genome Database (GDB) from the HosPital of Sick Children in Toronto
- The Dutch mirror of the Mouse Genome Database (MGD) from The Jackson Laboratory
- A mirror of the GeneCards system from the Weizmann Institute of Science in Israël

Other Specific Databases

- Arabidopsis thaliana (AAtDB)
- Caenorhabditis elegans (ACeDB)
- Saccharomyces cerevisiae (AScDB)
- Mycobacterium (MycDB)

Bioinformatics Services

- Sequence Retrieval Server (SRS 6)
- Human Genome Database (GDB)
- Mouse Genome Database (MGD)
- GeneCards
  - human genes, proteins and diseases
- Bioinformatics Toolkit
  - BLAST, Clustal, PHD, …

Local Tools

- Database searches
  - BLAST (assembled human genome or completed bacterial genomes), FASTA, PSI-BLAST (profile searching)
- Multiple alignment
  - Clustal W, T-Coffee, "Roll your own" PAM matrix
- Protein 2D structure prediction
  - PredictProtein
- Restriction enzyme mapping
  - TACG

Remote Tools

- SeWeR
  - Sequence Analysis using Web Resources (CMBI mirror)
- Sequence elongation, expression and function
  - Compugen's Lab On Web
- Literature search, alert, and document delivery
  - TheScientificWorld
- And more

Cheminformatics Services

- Synthesis Planning
- Reaction and Structure Retrieval
- 3-D Structure Retrieval
- Modeling
- Computational Chemistry
- Visualization Tools and Utilities

Synthesis Planning

- LHASA, an expert system to assist organic chemists in the design of efficient routes to target molecules

Reaction and Structure Retrieval

- Beilstein CrossFire
  - chemical database system containing a very large database of compounds and their properties, reactions, and citations (including abstracts)
- ISIS
  - chemical information management system containing a compound database and several reaction databases
- SYNLIB
  - organic reaction retrieval system containing a current-literature database

3-D Structure Retrieval

- CSD
  - Cambridge Crystallographic Database with programs ConQuest, Quest, Pluto, and Vista.
  - X-ray and neutron diffraction studies on organic compounds, organometallic compounds and metal complexes
- PDB
  - PDBBrowse, X-ray protein structure data
- ICSD
  - X-ray analyses of inorganic compounds
  - Searchable using CRYSTIN software from Hundt and Sievers, UofBonn, or Hewat's web interface

Modeling

- SYBYL
  - Tripos' MM/MD program with Basic, BioPolymer, Advanced Computation, Dynamics, Composer and QSAR/CoMFA Modules
- MacroModel
  - Prof. Clark Still, Columbia University, NY.
  - MM/MD program including Batchmin and Xcluster
- CSECM
  - Crystal Symmetry Modeler

Computational Chemistry

- MOPAC
  - semi-empirical molecular orbital package
  - solid state and molecular structures and reactions
- DeFT
  - density functional deMon LCGTO-DFT-based program
- GAMESS-UK
  - ab initio molecular electronic structure program
  - SCF- and MCSCF-gradient calculations with post Hartree-Fock calculations
- QCLDB
  - Quantum Chemistry Literature Database (web access)

Visualization Tools and Utilities

- MOLDEN
  - molecular and electronic structures visualizations
- Draw_Plot
  - PLUTON-based
  - visualization of small molecules retrieved from 3D databases or model-based
- Molscript
  - publication-quality plots of protein structures
- RasMol
  - proteins and nucleic acids visualizations

Limited Access Tools

- Quanta/CHARMm
  - from Molecular Simulations Inc.
- Insight II/Discover
  - from Molecular Simulations Inc. (formerly Biosym Technologies Inc.)

Local Tools

- Cambridge Crystallographic Database (CSD)
  - Free, though limited (15 CPU minutes/month) access is available to academics from continental Europe
- Crystal Symmetry Modeler (CSECM)
  - Examine space groups of arbitrary dimension
  - Manipulate and inspect Wyckoff positions, point group symmetry and systematic extinctions
  - Visualize space groups or space group elements (thru VRML - Virtual Reality Modeling Language)
- Quantum Chemistry Literature Data Base
- MOLDEN) Visualization of molecular and electronic structure

Molecular Mining

http://www.molecularmining.com/GeneLinker/Gold/

Overview

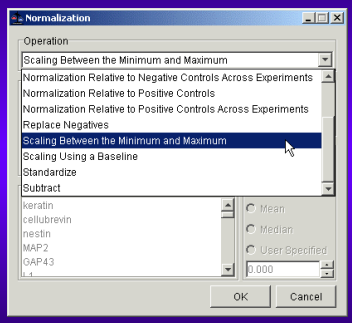
Features

- A set of analysis tools which handle a variety of tasks, including
  - data filtering and transformation
  - sorting and clustering
  - classification and prioritization of genes
- A suite of visualization tools which provides easily understood presentation of data mining results

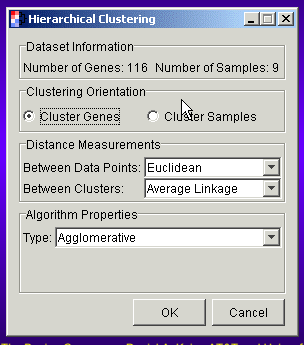
Packages Described

- **Molecular Mining**
  - GeneLinker
- **Silicon Genetics**
  - GeneSpring
- **BioDiscovery**
  - Genesight
- **Iobion**
  - GeneTraffic
- **Gene-IT**
  - LASSAP
- **Rosetta Biosciences (Merck)**
  - Resolver

Normalization

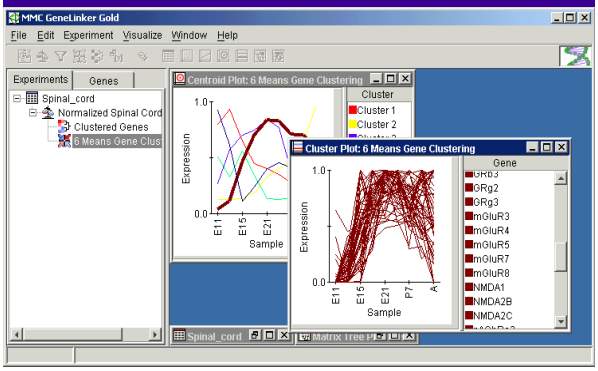


Clustering

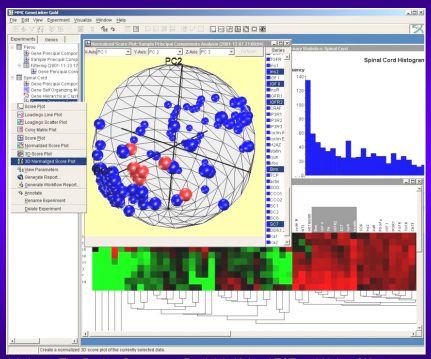


Mihael Ankerst, The Boeing Company -- Daniel A. Keim, AT&T and Univ. of Konstanz
Georges Grinstein, UMass Lowell and AnVil Inc.

Centroid and Clustering Plots



Gene-Linker

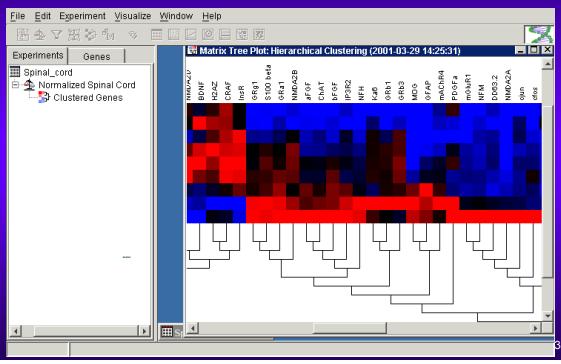


Mihael Ankerst, The Boeing Company -- Daniel A. Keim, AT&T and Univ. of Konstanz
Georges Grinstein, UMass Lowell and AnVil Inc.

GeneSpring

http://www.sigenetics.com/

Matrix Tree Plot

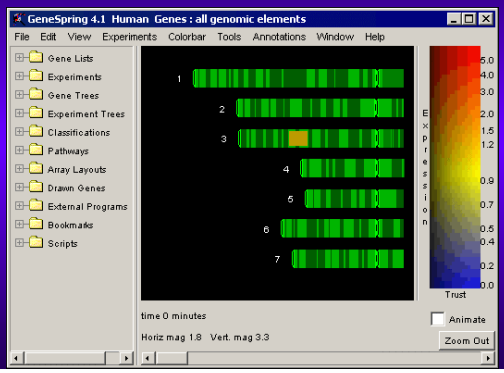


Features

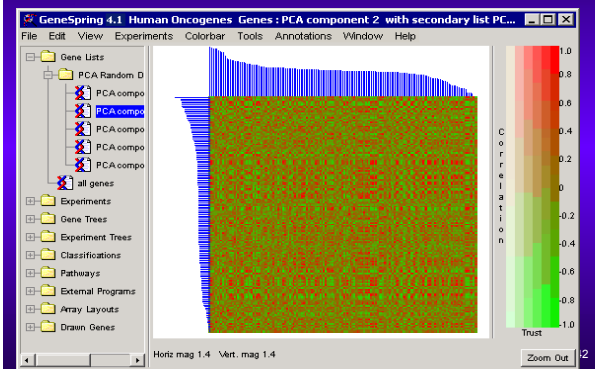
- Easy data loading
- Straightforward normalization
- Powerful analysis tools
- Numerous visualization tools
- Automated gene ontology construction and gene annotation tools
- Convenient data management

Mihael Ankerst, The Boeing Company -- Daniel A. Keim, AT&T and Univ. of Konstanz
Georges Grinstein, UMass Lowell and AnVil Inc.

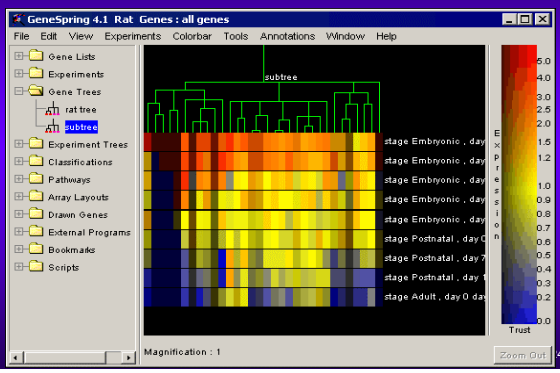
Physical Position



Gene Comparison View



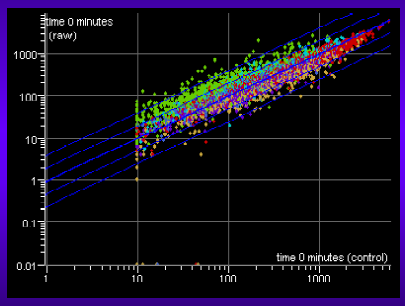
Hierarchical Clustering



GeneSight

http://www.biodiscovery.com/

Scatterplot



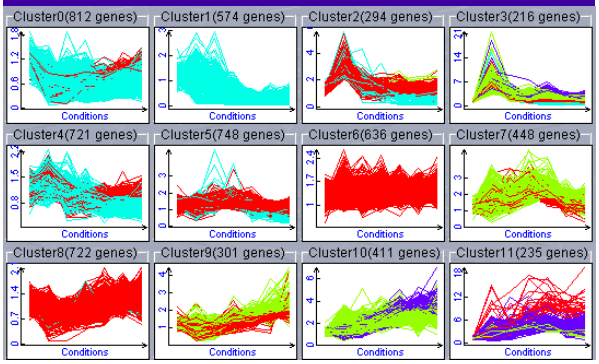
Mihael Ankerst, The Boeing Company -- Daniel A. Keim, AT&T and Univ. of Konstanz
Georges Grinstein, UMass Lowell and AnVil Inc.

Features

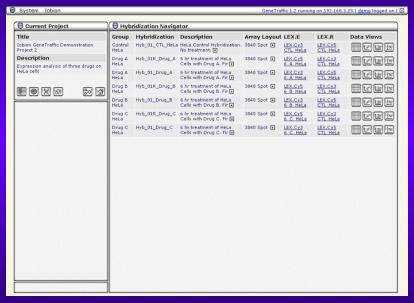
- Dataset builder
- Pre-processing of data
- Unregulated gene finding
- Clustering of genes
- Statistical significance analysis
- Time series analysis

Mihael Ankerst, The Boeing Company -- Daniel A. Keim, AT&T and Univ. of Konstanz
Georges Grinstein, UMass Lowell and AnVil Inc.

2D Self Organizing Maps



Hybridization Navigator

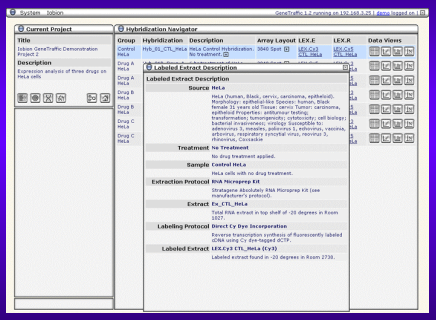


Mihael Ankerst, The Boeing Company -- Daniel A. Keim, AT&T and Univ. of Konstanz
Georges Grinstein, UMass Lowell and AnVil Inc.

GeneTraffic

http://www.iobion.com/

Labeled Extract Description



Mihael Ankerst, The Boeing Company -- Daniel A. Keim, AT&T and Univ. of Konstanz
Georges Grinstein, UMass Lowell and AnVil Inc.

GeneTraffic

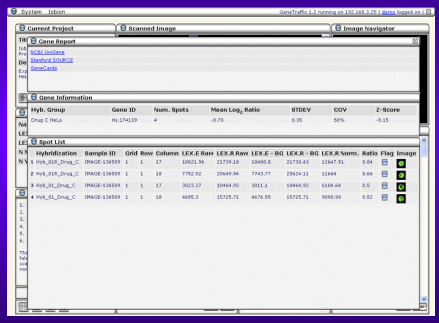
- A microarray gene expression data mining system which
  - Stores
  - Analyzes
  - Visualizes
- Compliant with MIAME & MAML standard
- Features normalization, clustering, PCA, multidimensional scaling, …

Mihael Ankerst, The Boeing Company -- Daniel A. Keim, AT&T and Univ. of Konstanz
Georges Grinstein, UMass Lowell and AnVil Inc.

Scanned Image View

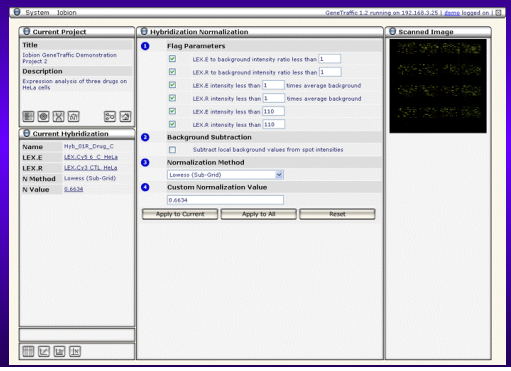
Mihael Ankerst, The Boeing Company -- Daniel A. Keim, AT&T and Univ. of Konstanz
Georges Grinstein, UMass Lowell and AnVil Inc.

Gene Report

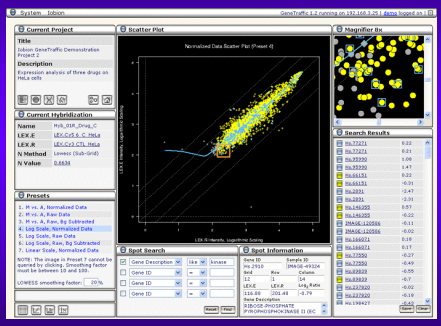


Mihael Ankerst, The Boeing Company -- Daniel A. Keim, AT&T and Univ. of Konstanz
Georges Grinstein, UMass Lowell and AnVil Inc.

Hybridization Normalization

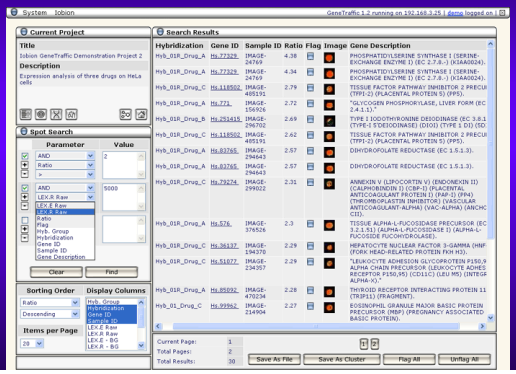


Scatter Plot

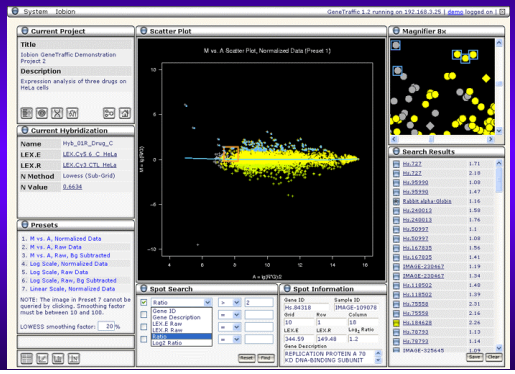


Mihael Ankerst, The Boeing Company -- Daniel A. Keim, AT&T and Univ. of Konstanz
Georges Grinstein, UMass Lowell and AnVil Inc.

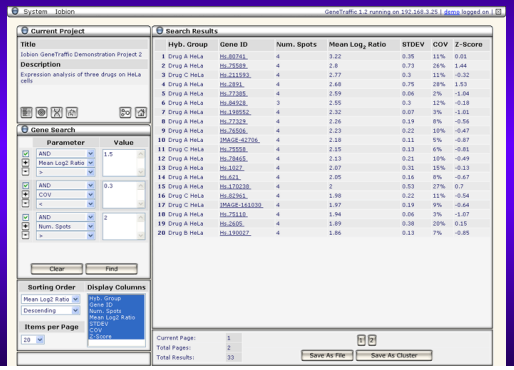
Spot View



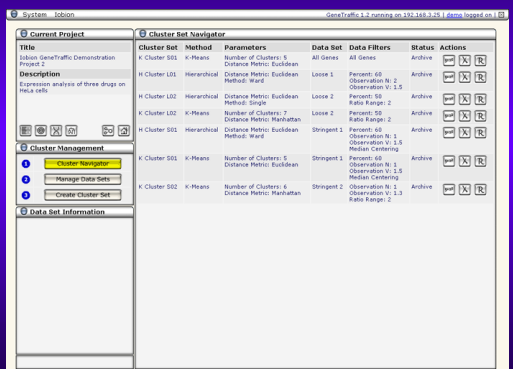
M vs A Scatter Plot



Gene View

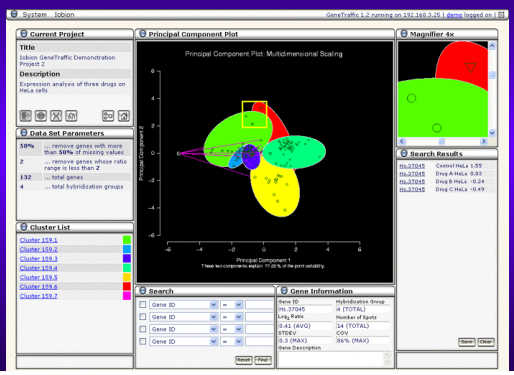


Cluster View



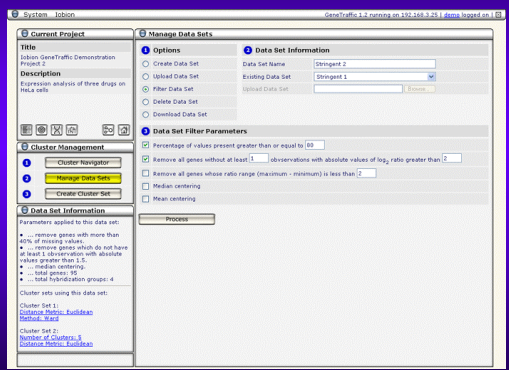
257

Multidimensional Scaling View



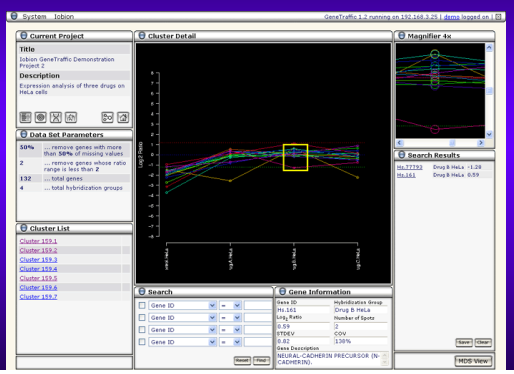
260

Manage Data Set



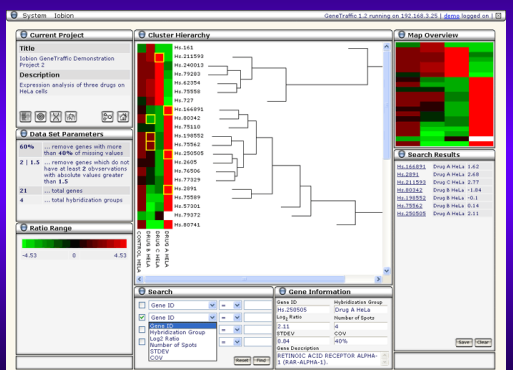
258

K-means Clustering View



261

Hierarchical Clustering



259

LASSAP

Http://www.geneit.Com/lassap.html

Functionality

- Comparison of entire sequence databases
- Simultaneous handling of sequence and annotation
- Sequence comparison algorithms
- Parallel execution of all algorithms on multiprocessor platforms
- Efficient handling of large amounts of data without size limitation
- Results can be used as input to further analysis
- Extensive result analysis capabilities

Features

- Sequence database management
- High-throughput sequence comparison engine
- Result analysis

Rosetta BioSciences

http://www.rosettabio.com/products/resolver/

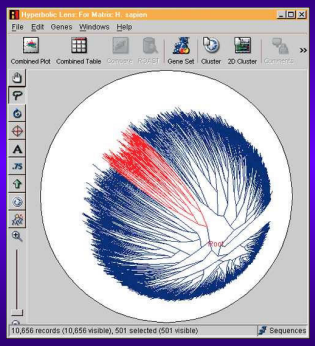
Features

- Multi-platform capability
- Error-models and quality statistics
- Multi-dimensional analysis
- Analysis and data visualization

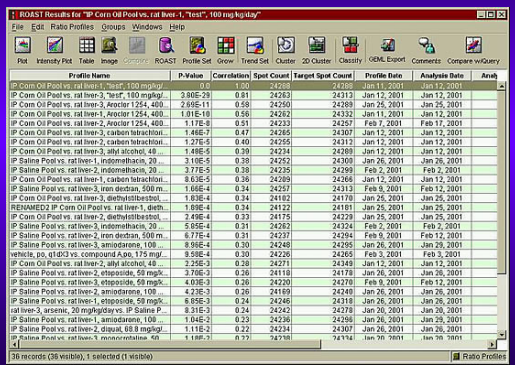
Clustering Analysis Tools

- Analysis Algorithms
  - Agglomerative, Divisive, K-means, K-medians, Self-Organizing Maps
  - Custom algorithms integrated using the Rosetta Resolver SDK
- Analysis Visualization
  - Flat Tree Viewer, 2D Matrix or Heat Map, Hyperbolic Lens Viewer, Table Viewer

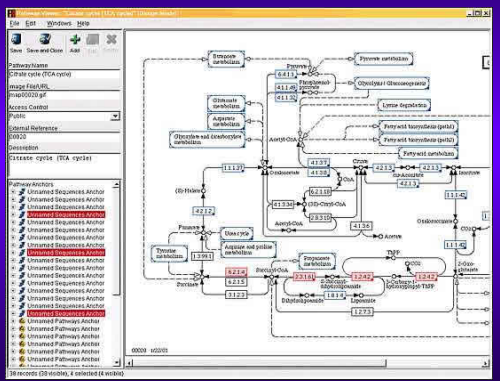
Hyperbolic Lens Viewer



ROAST rm



Pathway Viewer

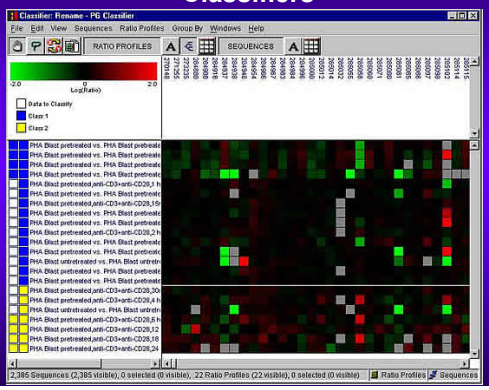


Analysis and Data Visualization

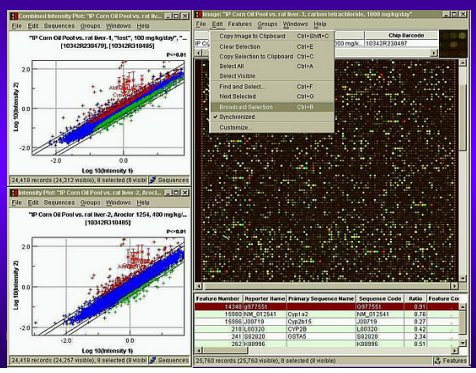
- Image viewer
- Histogram viewer
- Intensity plot viewer
- Synchronization and broadcast
- Trend viewer

Mihael Ankerst, The Boeing Company – Daniel A. Keim, AT&T and Univ. of Konstanz
Georges Grinstein, UMass Lowell and AnVil Inc.

Class Prediction Algorithms Bayesian Classifiers

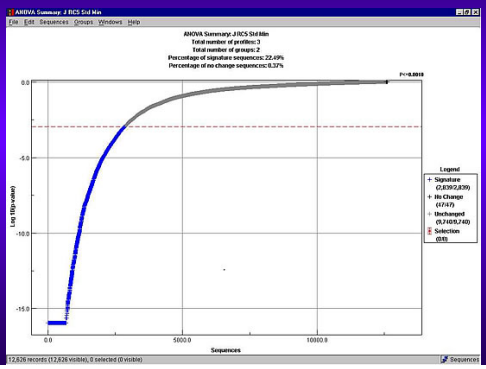


Synchronization and Broadcast



Pattern Growing Algorithm

Analysis of Variance (ANOVA)



HyperChem

At the Thomson Instrument Company

http://www.hplc1.com/hyperchem/hc5_features.html

Overview

Part I: Visualization Techniques
1. Introduction
2. Visual Data Exploration Techniques
3. Distortion and Interaction Techniques
4. Visual Data Mining Systems

Part II: Specific Visual Data Mining Techniques
1. Association Rules
2. Classification
3. Clustering
4. Text Mining
5. Tightly Integrated Visualization

Part III: Drug Discovery Applications
1. Biology and Chemistry
2. Bioinformatics and Cheminformatics
3. Examples
4. Bioinformatics Packages
5. Cheminformatics Packages

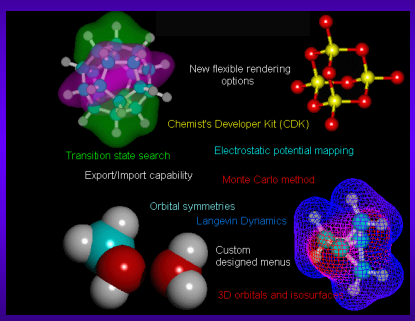
HyperChem 5

- Molecular modeling program
- 3-D visualization and animation
- Quantum chemical calculations
- Customizable with user controls
- Interfaces easily to other programs

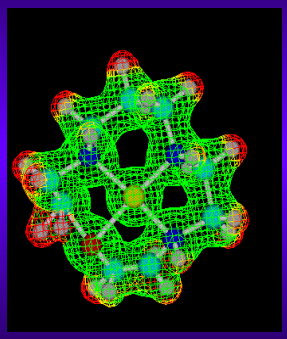
Packages Described

- HyperChem 5
- Accelrys
- CACTVC
- ACD/HNMR
- Tripos
- Meteor
- CCDC
- Elsevier
  - MDL
  - Beilstein
- Scitegic
- BioReason

HyperChem Capabilities



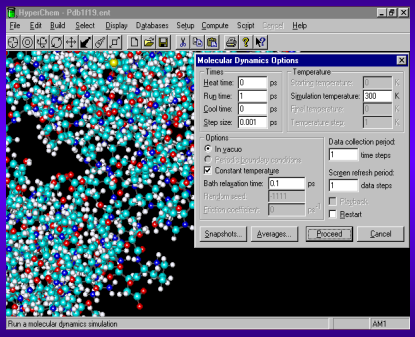
Examples: Electrostatic Potential



Accelrys

http://www.accelrys.com/

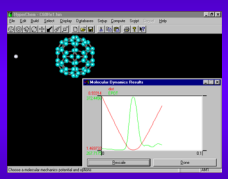
Molecular Dynamics

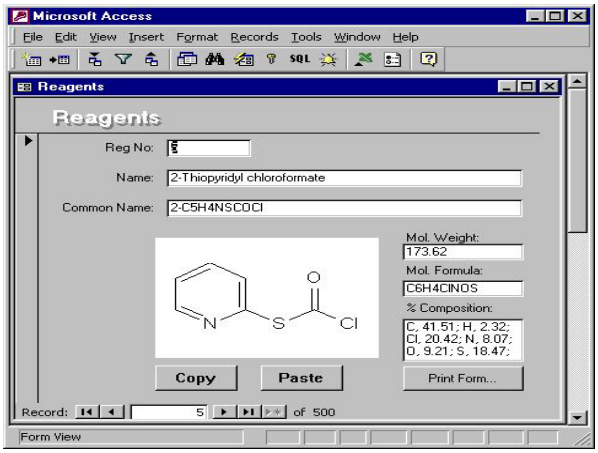


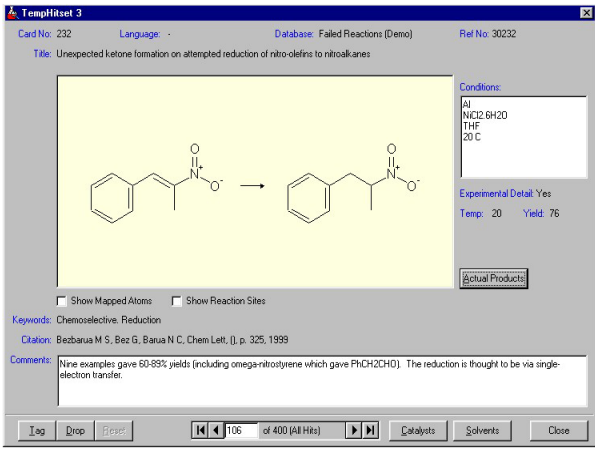
Accelrys

- Desktop productivity tools
- Developer tools
- Core databases and clients
- Chemistry databases
- ADME/Tox tools
- Chemistry workflow products

Examples: Display of results



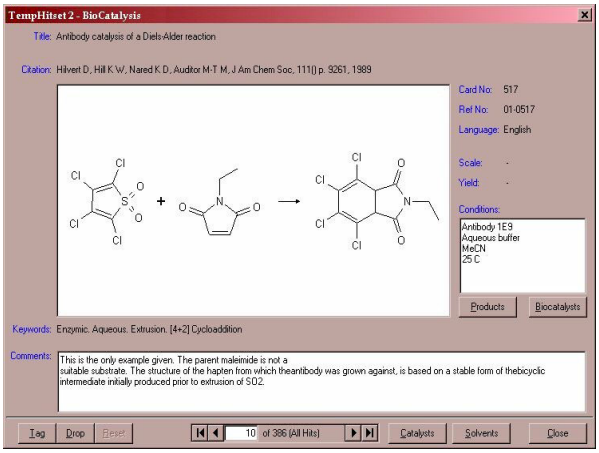


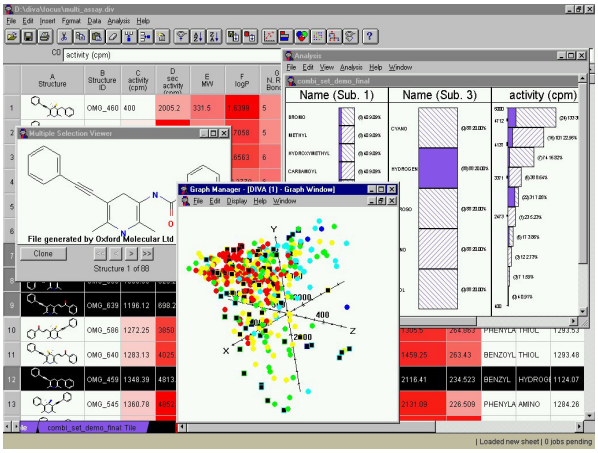


DIVA Capabilities

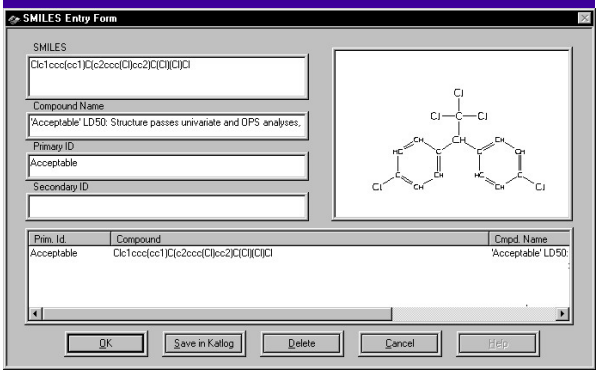
- Assemble data
- Generate Different Views
- Observe Patterns and Trends
- Plot Graphs and Charts
- Statistical Analyses
- Generate custom Reports

Mihael Ankerst, The Boeing Company -- Daniel A. Keim, AT&T and Univ. of Konstanz
Georges Grinstein, UMass Lowell and AnVil Inc.

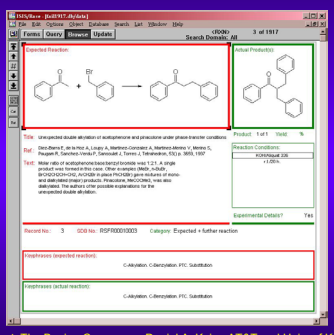




TOPKAT

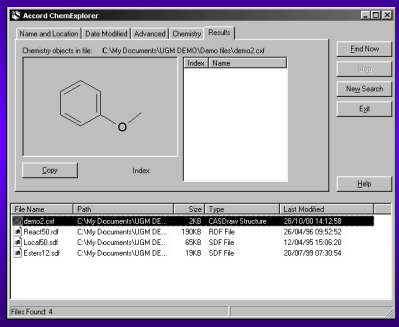


Accord Database Explorer



Mihael Ankerst, The Boeing Company -- Daniel A. Keim, AT&T and Univ. of Konstanz
Georges Grinstein, UMass Lowell and AnVil Inc.

Accord ChemExplorer



CACTVS

http://www2.chemie.uni-erlangen.de/ software/cactvs/

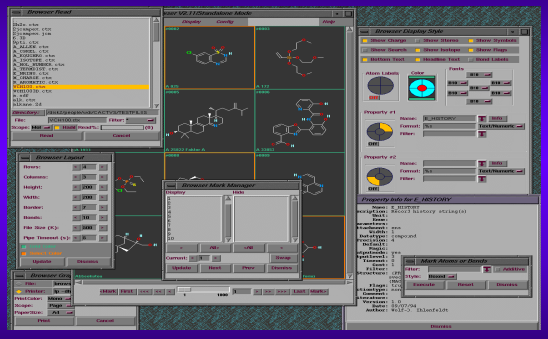
CACTVS

- Distributed client/server system;
- Uses a world-wide network of databases with many attribute
- Visual programming paradigm
- Searches its network of databases to retrieve information

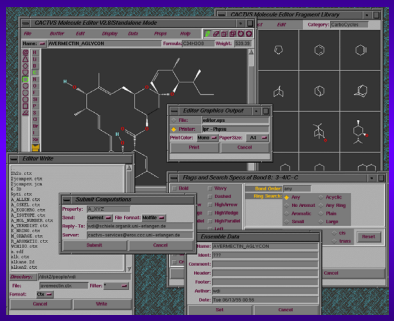
CACTVS Capabilities

- 2D structure editor with advanced features
- 2D structure browser
- Spectra display
  - IR, MS, Cont. MS, NMR, EPR, UV/VIS,GC
  - for JCAMP files
- XY data plotter
- WWW Chemical MIME access tool
- Chemical GIF image viewer and generator
- Networked WWW substructure search tool
- Networked NMR shift archive access tool

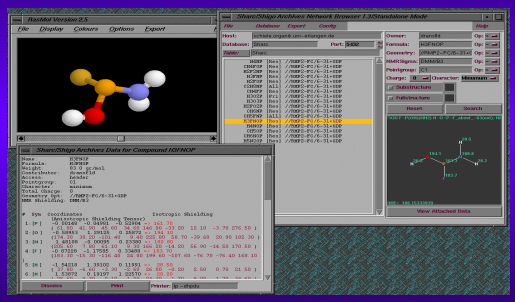
2D Structure Browser



2D Structure Editor



Sharc/Shigo ab initio NMR Shift Archives

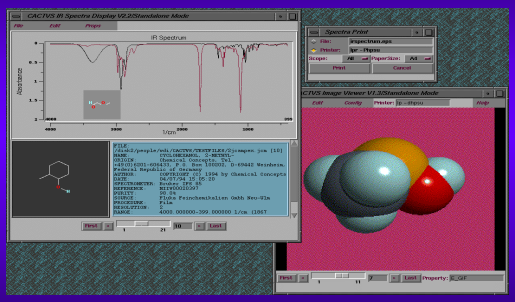


Mihael Ankerst, The Boeing Company -- Daniel A. Keim, AT&T and Univ. of Konstanz
Georges Grinstein, UMass Lowell and AnVil Inc.

ACD/HNMR

http://www.acdlabs.com/products/spec_lab/predict_nmr/hnmr/

Infrared Spectra Display and Image Viewer

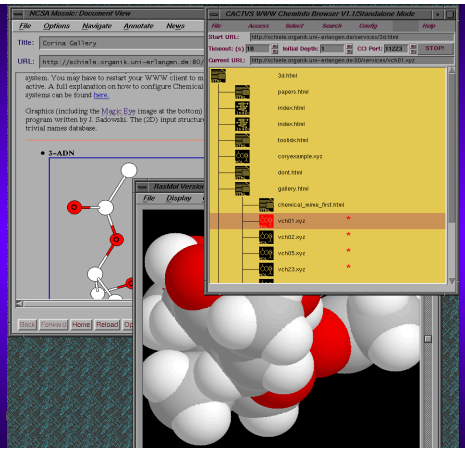


Mihael Ankerst, The Boeing Company -- Daniel A. Keim, AT&T and Univ. of Konstanz
Georges Grinstein, UMass Lowell and AnVil Inc.

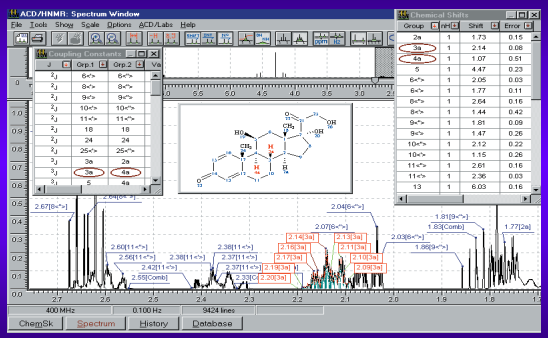
ACD/HNMR

- General Features & Prediction Power
  - Calculates accurate HNMR spectra under any basic frequency
  - Explains the predicted shift
  - Allows user to increase the accuracy of prediction using a Self Training System
  - Uses 3D molecular structure minimization and Karplus relationships to predict proton-proton coupling constants

Mihael Ankerst, The Boeing Company -- Daniel A. Keim, AT&T and Univ. of Konstanz
Georges Grinstein, UMass Lowell and AnVil Inc.

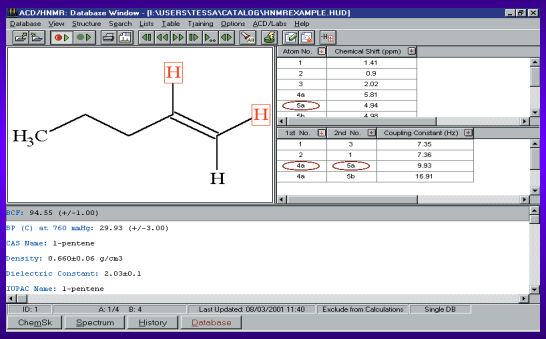


ACD/HNMR - Examples

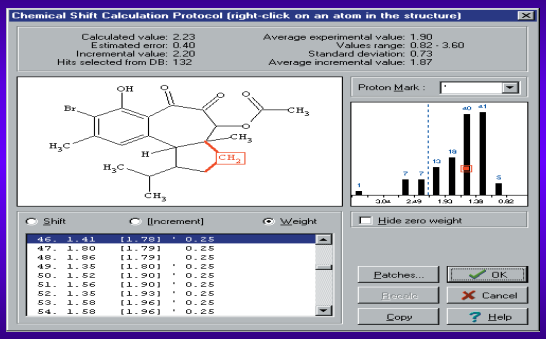


Mihael Ankerst, The Boeing Company -- Daniel A. Keim, AT&T and Univ. of Konstanz
Georges Grinstein, UMass Lowell and AnVil Inc.

ACD/HNMR - Examples



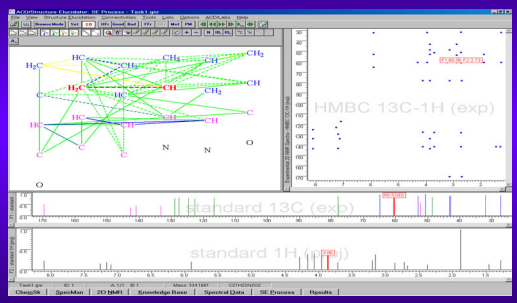
ACD/HNMR - Examples



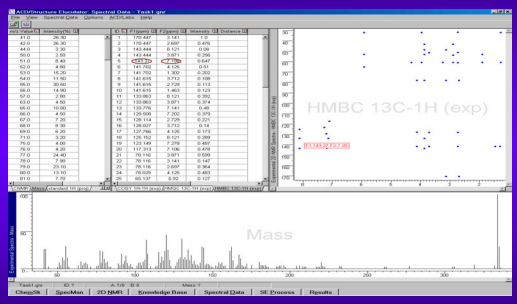
ACD/Structure Elucidator

- Determine the chemical structure from experimental spectral data for an unknown compound
- Uses spectral search, structure generation and filtration by spectral data
- Ultimately propose a small set of structures which fully correspond to the input spectral data

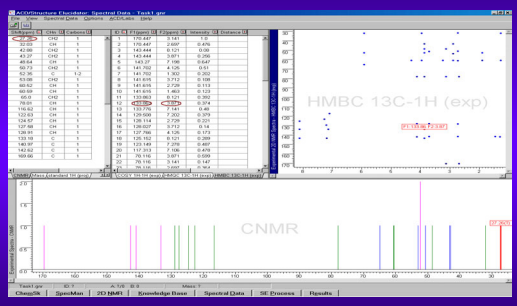
ACD/Structure Elucidator Example



ACD/Structure Elucidator Example



ACD/Structure Elucidator Example

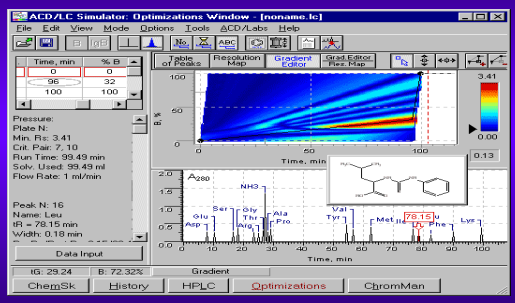


ACD/LC Simulator

- Simulates elution order and retention times for LC separation of organic chemical mixtures
- Uses experimentally measured retention times for a set of compounds and calculates predicting equation
- This equation can be used to extrapolate retention times for new compounds under given experimental conditions

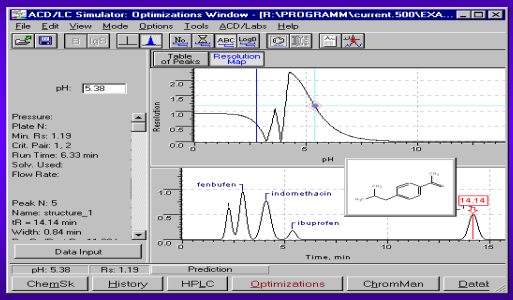
Mihael Ankerst, The Boeing Company -- Daniel A. Keim, AT&T and Univ. of Konstanz
Georges Grinstein, UMass Lowell and AnVil Inc.

Gradient Optimization



Mihael Ankerst, The Boeing Company -- Daniel A. Keim, AT&T and Univ. of Konstanz
Georges Grinstein, UMass Lowell and AnVil Inc.

Structure-based pH Robustness Prediction

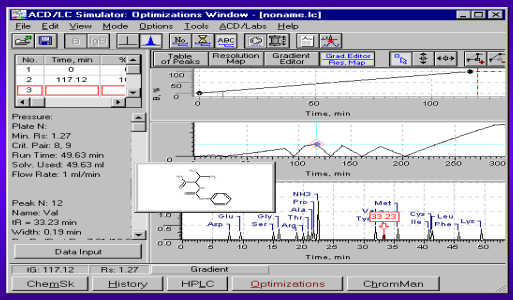


Mihael Ankerst, The Boeing Company -- Daniel A. Keim, AT&T and Univ. of Konstanz
Georges Grinstein, UMass Lowell and AnVil Inc.

Tripos

http://www.tripos.com/index.html

Temperature Optimization



Mihael Ankerst, The Boeing Company -- Daniel A. Keim, AT&T and Univ. of Konstanz
Georges Grinstein, UMass Lowell and AnVil Inc.

MM3 (Tripos)

- Molecular dynamics
- Normal Mode Vibrational Analysis
- Stochastic conformational searches
- Thermodynamic Property Calculations
- Molecular mechanics calculations
- IR Frequencies
- Integrated with SYBYL Interface

Mihael Ankerst, The Boeing Company -- Daniel A. Keim, AT&T and Univ. of Konstanz
Georges Grinstein, UMass Lowell and AnVil Inc.

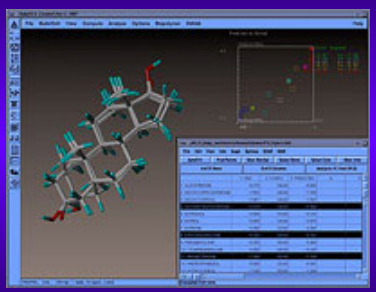
Sybyl

- Molecular design and analysis
- Comprehensive computational toolkit
- Molecular spreadsheet with tools
  - Conformational searching
  - Biopolymer modeling
  - Combinatorial Chemistry and Library design
  - QSAR and advanced CoMFA
  - Protein homology modeling and analysis

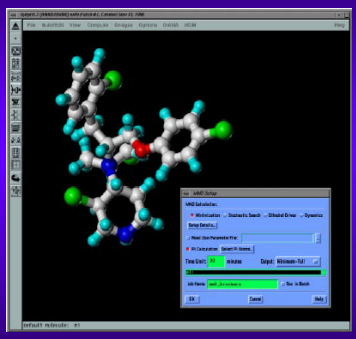
Conformational Analysis

- Conformational analysis
  - intramolecular contacts and strain energies
  - distance constraints or constrain conformers based on results from previous searches on similar molecules
  - Ring searching
- Random searching for systems too large to handle systematically
- Torsion driver for energy profiling of specific rotatable bonds
- Molecular Spreadsheet for analysis

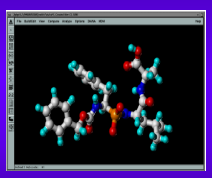
Molecular Spreadsheet



Integrated with SYBYL



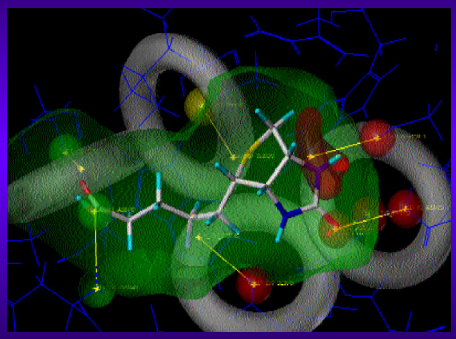
Phosphoramide Peptide Analog



UNITY

- 2D and 3D database retrieval
- Stores compound structure and associated data
- Searches and retrieves Oracle™ data
- Enterprise-wide access to structures and data
- Management, post-processing and analysis of search results

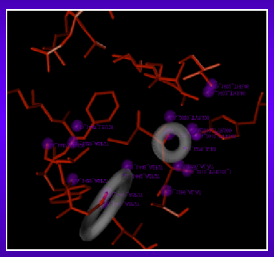
UNITY-3D Search

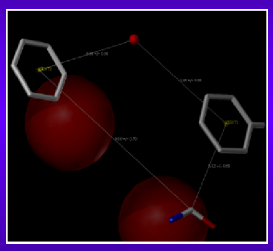


Meteor

http://www.chem.leeds.ac.uk/LUK/meteor/

UNITY - Examples

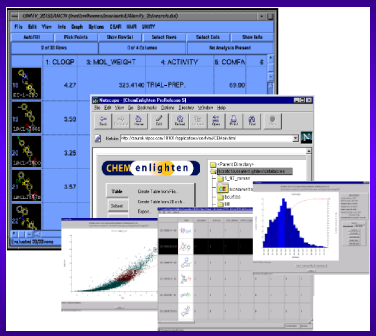




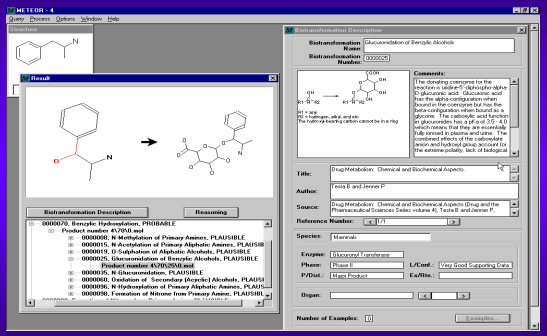
LHASA Meteor

- Contains a wide range of both Phase I and Phase II biotransformations
- Justifies predictions with relevant literature references
- Visualization of metabolic sequences
- Knowledge base editor for the development of proprietary knowledge
- Uses Log P to estimate the lipophilicity of metabolites
- External links to Derek for toxicity prediction

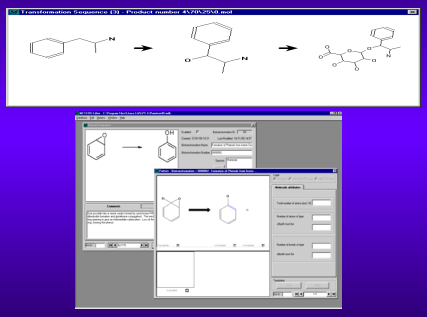
UNITY - Example



METEOR Analysis



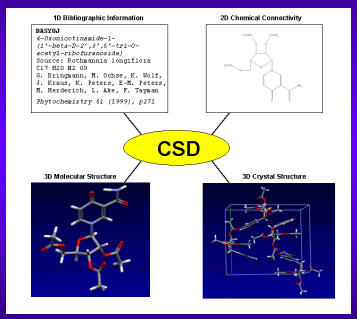
METEOR Example



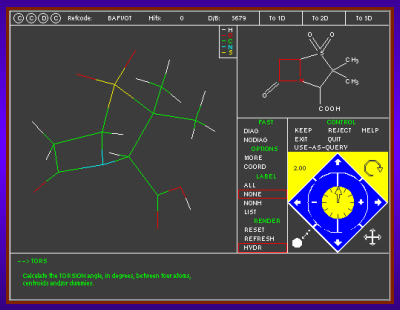
Cambridge Crystallographic Data Centre

http://www.ccdc.cam.ac.uk/

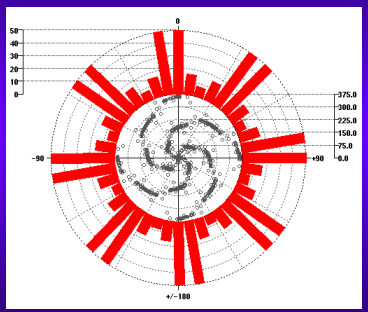
The Cambridge Structural Database



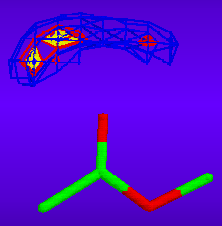
QUEST



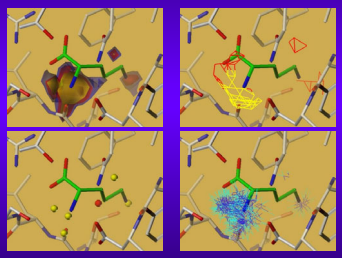
VISTA



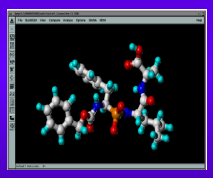
ISOSTAR



SuperStar



Examples: Expanded Applications



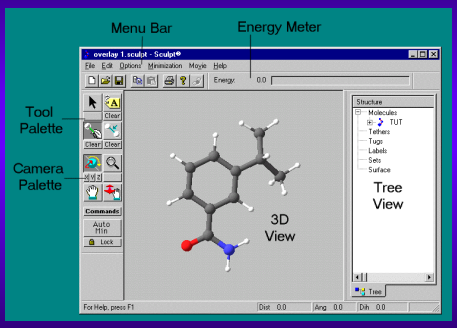
Elsevier
Molecular Design, Ltd.

http://www.mdli.com/

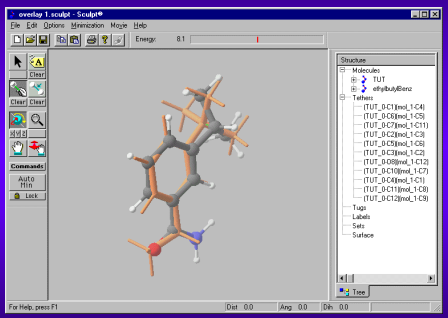
Features

- Obtain 3D structural information
- Explore and sample conformations
- Align compounds
- Visualize 3D relationships
- Facilitate communication

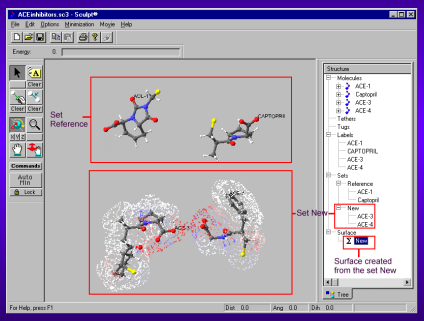
SCULPT



Paste Align

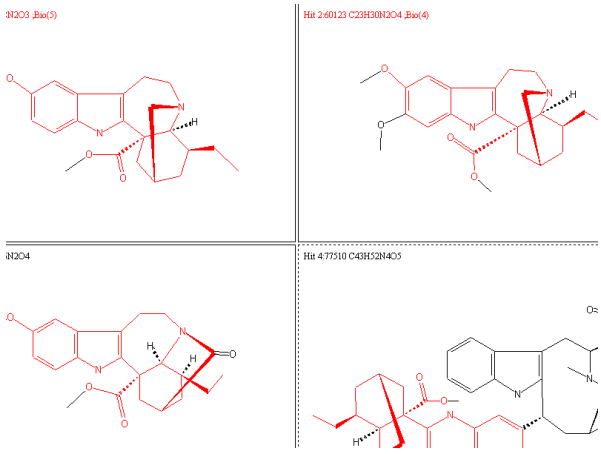


Set



Mihael Ankerst, The Boeing Company -- Daniel A. Keim, AT&T and Univ. of Konstanz
Georges Grinstein, UMass Lowell and AnVil Inc.

341

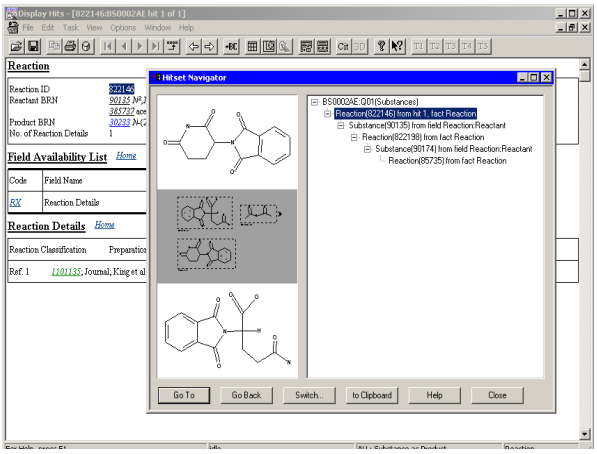


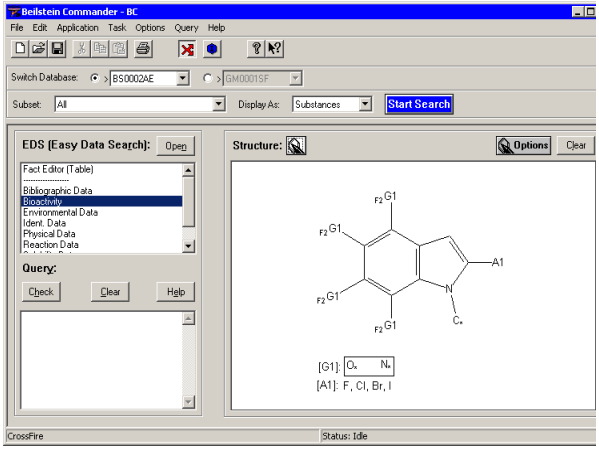
Elsevier
Beilstein CrossFire

http://www.beilstein.com/beilst_1.shtml

or

http://www.mimas.ac.uk/crossfire/overview.html





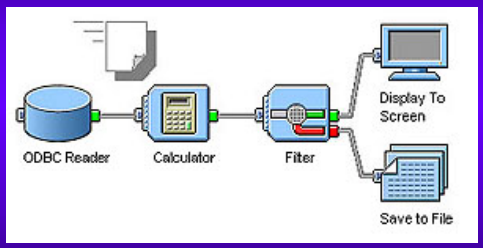
Scitegic

http://www.scitegic.com/

Pipeline Pilot

- Streamlines
  - The analysis and
  - The management of data

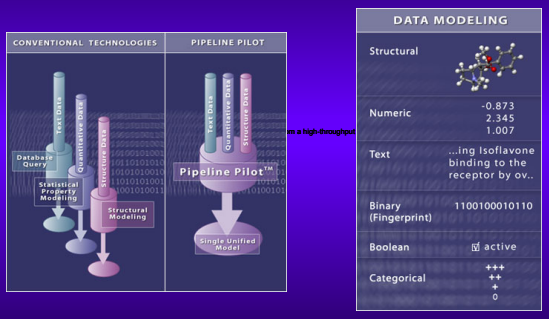
Data Pipelining



Capabilities

- Avoids data integration
- Learns from data
- Easy customization and extensibility

Data Modeling



LeadPharmer and ClassPharmer

http://www.bioreason.com

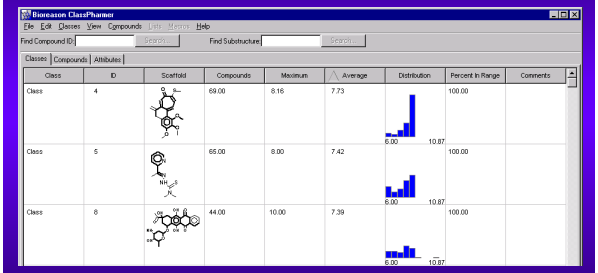
Packages

- ClassPharmer
  - Chemical family analyzer
- LeadPharmer
  - Classifies compound data according to learned common substructures
- DataPharmer
  - Identifies and isolates duplicate compound ID's and structures
  - Identifies data exceptions (metals, salts, solvents)
- DrugPharmer
  - view analysis results

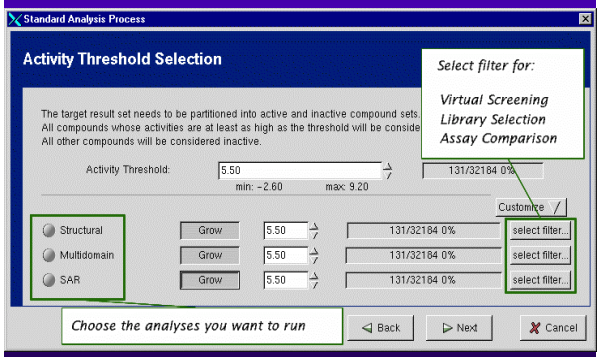
LeadPharmer

- Automated high throughput screening
- Identifies drug leads and how they leads work.
- Three-level analysis

Classification Results Using ClassPharmer™ Suite



Three-Level Analysis



Summary and Conclusions

Underlying Technology

- Self-organizing maps
- Genetic algorithms
- Automated production of GASP (genetic algorithm similarity program) models
- Decision trees
  - Rules
    - To guide tree growth
    - To choose input sets
  - Knowledge-based systems
    - To analyze tree results
- Spreadsheet visualizations

Summary (1)

- Lots of commercial and academic products
- Key differentiating factors include
  - Usability
  - Number of tools
  - Flexibility
  - Robustness
  - Sensitivity
  - Interoperability
- Some have newer and more novel visualization techniques which are applicable to database exploration

Summary (2)

- Some have newer and more novel visualization techniques which are applicable to database exploration
- Different techniques apply to different types of data (relational tables, hierarchies, graphs, etc.)
  - no guarantee of success
  - many of the techniques are applicable to traditional relational information sources
- Customization of tools is still necessary

Research Issues

- Develop integrated information visualization and exploration systems
  - with techniques from statistics, machine learning, databases, ...
- Perform in-depth evaluations and comparisons of visualization techniques for database exploration
  - there are possibilities for improvement
- Use more dynamics & interaction to steer the mining process
- Perform more case studies in a variety of application areas

Bibliography

[ABKS 99] Ankerst M., Breunig M., Kriegel H.P., Sander J.:"*OPTICS: Ordering Points To Identify the Clustering Structure*", Proc. ACM SIGMOD'99, Int. Conf on Management of Data, Philadelphia, PA, 1999, pp. 49-60.

[AC 91] Alpern B., Carter L.: *'Hyperbox'*, Proc. Visualization '91, San Diego, CA, 1991, pp. 133-139.

[ADLP 95] Anupam V., Dar S., Leibfried T., Petajan E.: *'DataSpace: 3-D Visualization of Large Databases'*, Proc. Int. Symp. on Information Visualization, Atlanta, GA, 1995, pp. 82-88.

[AEK 00] Ankerst M., Ester M., Kriegel H.P.:"Towards an Effective Cooperation of the Computer and the User for Classification", Proc. ACM SIGKDD Int. Conf. On Knowledge Discovery & Data Mining (KDD 2000), Boston, MA, 2000, pp. 179-188.

[AKK 96] Ankerst M., Keim D.A., Kriegel H.P.: *'Circle Segments: A Technique for Visually Exploring Large Multidimensional Data Sets'*, Visualization '96, Hot Topic Session, San Francisco, CA, 1996.

[And 72] Andrews D. F.: *'Plots of High-Dimensional Data'*, in: Biometrics, Vol. 29, 1972, pp. 125-136.

[And 95] Andrews K.: *'Visualizing Cyberspace: Information Visualization in the Harmony Internet Browser'*, Proc. Int. Symp. on Information Visualization, Atlanta, GA, 1995, pp. 97-104.

[AS 94] Ahlberg C., Shneiderman B.: *'Visual Information Seeking: Tight Coupling of Dynamic Query Filters with Starfield Displays'*, Proc. Human Factors in Computing Systems CHI '94 Conf., Boston, MA, 1994, pp. 313-317.

[AS 95] Ayers E. Z., Stasko J. T.: *'Using Graphic History in Browsing the World Wide Web'*, Proc. 4th Int. World Wide Web Conf., Boston, MA, 1995.

[Asi 85] Asimov D.: *'The Grand Tour: A Tool For Viewing Multidimensional Data'*, SIAM Journal of Science & Stat. Comp., Vol. 6, 1985, pp. 128-143.

[AW 95a] Ahlberg C., Wistrand E.: *'IVEE: An Environment for Automatic Creation of Dynamic Queries Applications'*, Proc. Human Factors in Computing Systems CHI '95 Conf., Demo Program, Denver, CO, 1995.

[AW 95b] Ahlberg C., Wistrand E.: *'IVEE: An Information Visualization and Exploration Environment'*, Proc. Int. Symp. on Information Visualization, Atlanta, GA, 1995, pp. 66-73.

Bibliography

[BCS 96] Becker R. A., Cleveland W. S., Shyu M.-J.: *'The Visual Design and Control of Trellis Display'*, Journal of Computational and Graphical Statistics, Vol. 5, No. 2, 1996, pp. 123-155.

[BSC 96] Buja A., Swayne D. F., Cook D.: *'Interactive High-Dimensional Data Visualization'*, Journal of Computational and Graphical Statistics, Vol. 5, No. 1, 1996, pp. 78-99.

[BCW 88] Becker R., Chambers J. M., Wilks A. R.: *'The New S Language'*, Wadsworth & Brooks/Cole Advanced Books and Software, Pacific Grove, CA, 1988.

[BE 96] Ball T., Eick S.: *'Software Visualization in the Large'*, Computer, Vol. 29, No. 4, 1996, pp. 33-43.

[Bed 90] Beddow J.: *'Shape Coding of Multidimensional Data on a Mircocomputer Display'*, Visualization '90, San Francisco, CA, 1990, pp. 238-246.

[Bed 94] Bederson B.: *'Pad++: Advances in Multiscale Interfaces'*, Proc. Human Factors in Computing Systems CHI '94 Conf., Boston, MA, 1994, p. 315.

[BEMW 90] Becker R. A., Eick S. G., Miller E. O., Wilks A. R.: *'Dynamic Graphics for Network Visualization'*, Proc. Visualization '90, San Francisco, CA, 1990, pp. 93-95.

[Ben 95] Benford S., Snowdon D., Greenhaigh C., Ingram R., Knox I., Brown C.: *'Vr-vibe: A Virtual Environment for Co-operative Information Retrieval'*, Proc. Eurographics '95, Maastricht, The Netherlands, 1995, pp. 349-360.

[Ber 81] Bertin J.: *'Graphics and Graphic Information Processing'*, Berlin, 1981.

[BETT 94] Battista G. D., Eades P., Tamassia R., Tollis I.: *'Annotated Bibliography on Graph Drawing Algorithms'*, Computational Geometry: Theory and Applications, Vol. 4, 1994, pp. 235-282.

[BEW 95] Becker R. A., Eick S. G., Wilks A. R.: *'Visualizing Network Data'*, Transactions on Visualization and Computer Graphics, Vol. 1, No. 1, 1995, pp. 16-28.

[BF 93] Beshers C., Feiner S.: *'AutoVisual: Rule-Based Design of Interactive Multivariate Visualizations'*, IEEE Computer Graphics and Applications, Vol. 13, No. 4, 1993, pp. 41-49.

Bibliography

[BFN 86] Boecker H.-D., Fischer G., Nieper H.: *'The Enhancement of Unterstanding through Visual Representations'*, Proc. Human Factors in Computing Systems CHI '86 Conf., Boston, MA, 1986, pp. 44-58.

[BH 86] Bryce D., Hull R.: *'SNAP: A Graphics Based Schema Manager'*, Proc. 2nd Int. Conf. on Data Engineering, Los Angeles, CA, 1986, pp. 151-164.

[BH 91] Brown M. H., Hershberger J.: *'Color and Sound in Algorithm Animation'*, Proc. IEEE Symp. on Visual Languages, 1991, pp. 10-17.

[BH 94] Bederson B. B., Hollan J. D.: *'Pad++: A Zooming Graphical Interface for Exploring Alternate Interface Physics'*, Proc. UIST, 1994, pp. 17-26.

[Bie 93] Bier E. A., Stone M. C., Pier K., Buxton W., DeRose T.: *'Toolglass and Magic Lenses: The see-through Interface'*, Proc. SIGGRAPH '93, Anaheim, CA, 1993, pp. 73-80.

[BMMS 91] Buja A., McDonald J. A., Michalak J., Stuetzle W.: *'Interactive Data Visualization Using Focusing and Linking'*, Visualization '91, San Diego, CA, 1991, pp. 156-163.

[BN 93] Brown M. H., Najork M. A.: *'Algorithm Animation Using 3D Interactive Graphics'*, Proc. UIST, 1993, pp. 93-100.

[Bro 91] Brown M. H.: *'Zeus: A System for Algorithm Animation and Multi-View Editing'*, Proc. IEEE Symp. on Visual Languages, 1991, pp. 4-9.

[Car 96] Card S. K.: *'Visualizing Retrieved Information: A Survey'*, IEEE CG&A, 1996, pp. 63-67.

[CE 95] Cox K. C., Eick S. G.: *'3D Displays of Internet Traffic'*, Proc. Symp. on Information Visualization, Atlanta, GA, 1995, pp. 129-131.

[CELR 94] Cohen R. F., Eades P., Lin T., Ruskey F.: *'Three-dimensional Graph Drawing'*, Proc. Graph Drawing, 1994, in: Lecture Notes in Computer Science, Vol. 894, Springer, 1994, pp. 1-11.

[Che 73] Chernoff H.: *'The Use of Faces to Represent Points in k-Dimensional Space Graphically'*, Journal Amer. Statistical Association, Vol. 68, pp. 361-368.

Bibliography

[CK 95] Carrière J., Kazman R.: *'Interacting with Huge Hierarchies: Beyond Cone Trees'*, Proc. Symp. on Information Visualization, Atlanta, GA, 1995, pp. 90-96.

[Cle 93] Cleveland W. S.: *'Visualizing Data'*, AT&T Bell Laboratories, Murray Hill, NJ, Hobart Press, Summit NJ, 1993.

[Cle 94] Cleveland W. S.: *'The Elements of Graphing Data'*, Hobart Press, Summit, NJ, revised edition.

[CM 93] Consens M. P., Mendelzon A. O.: *'Hy+: A Hygraph-based Query and Visualization System'*, Proc. ACM SIGMOD Int. Conf. on Management of Data, Washington, DC, 1993, pp. 511-516.

[Con 94] Consens M. P., Eigler F. Ch., Hasan M. Z., Mendelzon A. O., Noik E. G., Ryman A. G., Vista D.: *'Architecture and Applications of the Hy+ Visualization System'*, IBM Systems Journal, Vol. 33, No. 3, 1994, pp. 458-476.

[CPS 95] Chan M. C., Pacifici G., Stadler R.: *'Managing Real-Time Services in Multimedia Networks Using Dynamic Visualization and high-level Controls'*, Proc. Multimedia '95, San Francisco, CA, 1995.

[CR 92] Cox K. C., Roman G.-C.: *'Abstraction in Algorithm Animation'*, Proc. Symp. on Visual Languages, 1992, pp. 18-23.

[CRMK 95] Chuah M. C., Roth S. F., Mattis J., Kolojejchick J.: *'SDM: Malleable Information Graphics'*, Proc. Symp. on Information Visualization '95, Atlanta, GA, 1995, pp. 36-42.

[CRY 96] Card S. K., Robertson G. G., York W.: *'The WebBook and the WebForager: An Information Workspace for the World Wide Web'*, Proc. Human Factors in Computing Systems CHI '96 Conf., 1996.

[CT 94] Cruz I. F., Tamassia R.: *'How to Visualize a Graph: Specification and Algorithms'*, 1994.

[DBM 01] http://www.dbminer.com

[DE 82] Dunn G., Everitt B.: *'An Introduction to Mathematical Taxonomy'*, Cambridge University Press, Cambridge, MA, 1982.

[DH 95] Drew N. S., Hendley R. J.: *'Visualizing Complex Interacting Systems'*, Proc. Human Factors in Computing Systems CHI'95 Conf., Denver, CO, 1995, pp. 204-205.

Bibliography

[Doe 94] Doemel P.: *'Webmap - A Graphical Hypertext Navigation Tool'*, Proc. 2nd Int. World Wide Web Conf., Chicago, IL, 1994.

[Eic 94] Eick S. G.: *'Data Visualization Sliders'*, Proc. ACM UIST, 1994, pp. 119-120.

[Eic 96] Eick S. G.: *'Aspects of Network Visualization'*, Proc. IEEE CG&A, 1996, pp. 69-72.

[ES 92] Eick S. G., Steffen J. L.: *'Visualizing Code Profiling Line Oriented Statistics'*, Proc. Visualization '92, Boston, MA, 1992, pp. 210-217.

[ESS 92] Eick S. G., Steffen J. L., Sumner E. E.: *'SeeSoft - A Tool for Visualizing Software'*, IEEE Transactions on Software Engineering, Vol. 18, No. 11, 1992, pp. 957-968.

[EW 93] Eick S. G., Wills G. J.: *'Navigating Large Networks with Hierarchies'*, Visualization '93, San Jose, CA, 1993, pp. 204-210.

[FB 90a] Feiner S., Beshers C.: *'Visualizing n-Dimensional Virtual Worlds with n-Vision'*, Computer Graphics, Vol. 24, No. 2, 1990, pp. 37-38.

[FB 90b] Feiner S., Beshers C.: *'World within World: Metaphors for Exploring n-dimensional Virtual Worlds'*, Proc. UIST, 1990, pp. 76-83.

[FB 94] Furnas G. W., Buja A.: *'Prosections Views: Dimensional Inference through Sections and Projections'*, Journal of Computational and Graphical Statistics, Vol. 3, No. 4, 1994, pp. 323-353.

[FB 95] Furnas G. W., Bederson B. B.: *'Space-Scale Diagrams: Unterstanding Multiscale Interfaces'*, Proc. Human Factors in Computing Systems CHI '95 Conf., Denver, CO, 1995.

[FL 95] Faloutsos C., Lin K.: 'Fastmap: A fast Algorithm for Indexing, Data-Mining and Visualization of Traditional and Multimedia Datasets', Proc. ACM SIGMOD Int. Conf. on Management of Data, San Jose, CA, 1995, pp. 163-174.

[FPF 88] Fairchild K., Poltrok S., Furnas G.: *'SemNet: Three Dimensional Graphic Representations of Large Knowledge Bases'*, Lawrence Erlbaum (ed.), 1988, pp. 201-233.

Bibliography

[FS 95] Fishkin K., Stone M. C.: *'Enhanced Dynamic Queries via Movable Filters'*, Proc. Human Factors in Computing Systems CHI '95 Conf., Denver, CO, 1995, pp. 415-420.

[Fur 86] Furnas G.: *'Generalized Fisheye Views'*, Proc. Human Factors in Computing Systems CHI'86 Conf., Boston, MA, 1986, pp. 18-23.

[FW 94] Frolich M., Werner M.: *'Demonstration of the Interactive Graph Visualization System da vinci'*, Proc. Graph Drawing, 1994, in: Lecture Notes in Computer Science, Vol. 894, Springer, 1994, pp. 266-269.

[FZ 94] Furnas G. W., Zacks J.: *'Multitrees: Enriching and Reusing Hierarchical Structures'*, Proc. Human Factors in Computing Systems CHI '94 Conf., Boston, MA, 1994, pp. 330-336.

[Ger 95] Gershon N., LeVasseur J., Winstead J., Croall J., Pernick A., Ruh W.: *'Visualizing Internet Resources'*, Proc. Symp on Information Visualization '95, Atlanta, GA, 1995, pp. 122-128.

[Ger 96] Gershon N.: *'Moving Happily Through the World Wide Web'*, Proc. IEEE CG&A, 1996, pp. 72-74.

[GKNV 93] Gansner E. R., Koutsofois E., North S. C., Vo K.-P.: *'A Technique for Drawing Directed Graphs'*, IEEE Transactions on Software Engineering, Vol. 19, No. 3, 1993, pp. 214-229.

[GPW 89] Grinstein G, Pickett R., Williams M. G.: *'EXVIS: An Exploratory Visualization Environment'*, Proc. Graphics Interface '89, London, Ontario, Canada, 1989.

[GR 94] Goldstein J., Roth S. F.: *'Using Aggregation and Dynamic Queries for Exploring Large Data Sets'*, Proc. Human Factors in Computing Systems CHI '94 Conf., Boston, MA, 1994, pp. 23-29.

[GT 96] Garg A., Tamassia R.: *'Giotto3d: A System for Visualizing Hierarchical Structures in 3d'*, Proc. Graph Drawing, '96, in: Lecture Notes in Computer Science, Springer, 1996.

[Har 67] Harman H. H.: *'Modern Factor Analysis'*, University of Chicago Press, 1967.

[HC 86] J.D.A. Henderson and S.K. Card: 'Rooms: The Use of Multiple Virtual Workspaces to Reduce Space Contention in Window-based Graphical User Interfaces', ACM Trans. on Graphics, Vol. 5, No. 3, July 1986, pp. 211-241.

Bibliography

[HDWB 95] Hendley R. J., Drew N. S., Wood A. M., Beale R.: *'Narcissus: Visualizing Information'*, Proc. Int. Symp. on Information Visualization, Atlanta, GA, 1995, pp. 90-94.

[HDD+ 01] Hao M., Hsu M ., Dayal U., Wei S.F., Sprenger T., Holenstein T.:"*Market Basket Analysis Visualization on a Spherical Surface*", http://www.hpl.hp.com/techreports/2001/HPL-2001-3.pdf

[Hea 95] Hearst M.: *'TileBars: Visualization of Term Distribution Information in Full Text Information Access'*, Proc. of ACM Human Factors in Computing Systems Conf. (CHI'95), 1995, pp. 59-66.

[Him 94] Himsolt M.: *'Graphed: A Graphical Platform for the Implementation of Graph Algorithms'*, Proc. Graph Drawing '94, in: Lecture Notes in Computer Science, Vol. 894, Springer, pp. 182-193.

[HKW 99] Hinneburg A., Keim D.A., Wawryniuk M.: „*HD-Eye: Visual Mining of High-Dimensional Data*", IEEE Computer Graphics and Applications, Vol. 19, No. 5, 1999.

[HSW 00] Hofmann H., Siebes A., Wilhelm A.:"*Visualizing Association Rules with Interactive Mosaic Plots*", Proc. ACM SIGKDD Int. Conf. On Knowledge Discovery & Data Mining (KDD 2000), Boston, MA, 2000.

[Hub 85] Huber P. J.: *'Projection Pursuit'*, The Annals of Statistics, Vol. 13, No. 2, 1985, pp. 435-474.

[Hum 01] HumanIT Web Page: „http://www.humanit.de/infozoom", Feb. 2001.

[IA 00] Inselberg A., Avidan T.:"*Classification and Visualization for High-Dimensional Data*", Proc. ACM SIGKDD Int. Conf. On Knowledge Discovery & Data Mining (KDD 2000), Boston, MA, 2000.

[Ins 85] Inselberg A.: *'The Plane with Parallel Coordinates, Special Issue on Computational Geometry'*, The Visual Computer, Vol. 1, 1985, pp. 69-97.

[ID 90] Inselberg A., Dimsdale B.: *'Parallel Coordinates: A Tool for Visualizing Multi-Dimensional Geometry'*, Visualization '90, San Francisco, CA, 1990, pp. 361-370.

[JJ 94] Jeron T., Jard C.: *'3D Layout of Reachability Graphics of Communicating Processes'*, Proc. Graph Drawing '94, in: Lecture Notes in Computer Science, Vol. 894, 1994, pp. 25-32.

[Joh 93] Johnson B.: *'Visualizing Hierarchical and Categorical Data'*, Ph.D. Thesis, Department of Computer Science, University of Maryland, 1993.

Bibliography

[JS 91] Johnson B., Shneiderman B.: *'Treemaps: A Space-filling Approach to the Visualization of Hierarchical Information'*, Proc. Visualization '91 Conf., 1991, pp. 284-291.

[JS 95] Jerding D. F., Stasko J. T.: *'The Information Mural: A Technique for Displaying and Navigating Large Information Spaces'*, Proc. Symp. on Information Visualization '95, Atlanta, GA, 1995, pp. 43-50.

[Kei 94] D. A. Keim: *'Visual Support for Query Specification and Data Mining'*, Shaker-Verlag, 1994, ISBN 3-8265-0594-8.

[Kei 95] D. A. Keim: *'Enhancing the Visual Clustering of Query-dependent Database Visualization Techniques using Screen-Filling Curves'*, Proc. Workshop on Database Issues for Data Visualization, Atlanta, GA, 1995, Springer, in: Lecture Notes in Computer Science, No. 1183, Springer, pp. 101-110.

[Kei 95] D. A. Keim: *'Enhancing the Visual Clustering of Query-dependent Database Visualization Techniques using Screen-Filling Curves'*, Proc. Workshop on Database Issues for Data Visualization, Atlanta, GA, 1995, Springer, in: Lecture Notes in Computer Science, No. 1183, Springer, pp. 101-110.

[Kei 96] Keim D. A.: *'Pixel-oriented Visualization Techniques for Exploring Very Large Databases'*, Journal of Computational and Graphical Statistics, Vol. 5, No. 1, 1996, pp. 58-77.

[KK 94] Keim D. A., Kriegel H.-P.: 'VisDB: Database Exploration using Multidimensional Visualization', Computer Graphics & Applications, Sept. 1994, pp. 40-49.

[KK 95] Keim D. A., Kriegel H.-P.: *'VisDB: A System for Visualizing Large Databases'*, System Demonstration, Proc. ACM SIGMOD Int. Conf. on Management of Data, San Jose, CA, 1995, p. 482.

[KKA 95] Keim D. A., Kriegel H.-P., Ankerst M.: *'Recursive Pattern: A Technique for Visualizing Very Large Amounts of Data'*, Proc. Visualization '95, Atlanta, GA, 1995, pp. 279-286.

[KKS 93] D. A. Keim, H.-P. Kriegel, T. Seidl: *'Visual Feedback in Querying Large Databases'*, Proc. Visualization '93, San Jose, CA, 1993, pp. 158-165.

[KKS 94] Keim D. A., Kriegel H.-P., Seidl T.: *'Supporting Data Mining of Large Databases by Visual Feedback Queries'*, Proc. 10th Int. Conf. on Data Engineering, Houston, TX, 1994, pp. 302-313.

Bibliography

[KK 96] Keim D. A., Kriegel H.-P.: 'Visualization Techniques for Mining Large Databases: A Comparison', Transactions on Knowledge and Data Engineering, Vol. 8, No. 6, Dec. 1996, pp. 923-938.

[KL 93] Keim D. A., Lum V.: *'GRADI: A Graphical Database Interface for a Multimedia DBMS'*, Proc. Int. Workshop on Interfaces to Databases, Glasgow, England, 1992, in: Workshops in Computing, Springer, 1993, pp. 95-112.

[KRB 94] Kaugars K., Reinfelds J., Brazma A.: *'A Simple Algorithm for Drawing Large Graphs on Small Screens'*, Proc. Graph Drawing '94, in: Lecture Notes in Computer Science, Vol. 894, Springer, 1994, pp. 278-281.

[LA 94] Leung Y., Apperley M.: *'A Review and Taxonomy of Distortion-oriented Presentation Techniques'*, Proc. Human Factors in Computing Systems CHI '94 Conf., Boston, MA, 1994, pp. 126-160.

[Lev 91] Levkowitz H.: *'Color icons: Merging color and texture perception for integrated visualization of multiple parameters'*, In Visualization '91, San Diego, CA, October 22-25 1991.

[Liv 97] M. Livny, R. Ramakrishnan, K. Beyer, G. Chen, D. Donjerkovic, S. Lawande, J. Myllymaki, and K. Wenger: 'DEVise: Integrated Querying and Visual Exploration of Large Datasets', Proc. of ACM SIGMOD, May, 1997.

[LO 96] Lee H.-Y., Ong K.-L.:"*Visualization Support for Data Mining*", IEEE Expert, Vol. 11, I. 5, pp.69-75, 1996.

[LR 94] Lamping J., Rao R.: *'Laying out and Visualizing Large Trees Using a Hyperbolic Space'*, Proc. UIST, 1994, pp. 13-14.

[LRP 95] Lamping J., Rao R., Pirolli P.: *'A Focus + Context Technique Based on Hyperbolic Geometry for Visualizing Large Hierarchies'*, Proc. Human Factors in Computing Systems CHI '95 Conf., Denver, CO, 1995, pp. 401-408.

[LSR 96] Lamm S. E., Scullin W. H., Reed D. A.: *'Real-time Geographic Visualization of World Wide Web Traffic'*, Proc. WWW, 1996.

[LWW 90] LeBlanc J., Ward M. O., Wittels N.: *'Exploring N-Dimensional Databases'*, Visualization '90, San Francisco, CA, 1990, pp. 230-239.

Bibliography

[MB 95] Munzner T., Burchard P.: *'Visualizing the Structure of the World Wide Web in 3D Hyperbolic Space'*, Proc. VRML '95 Symp, San Diego, CA, 1995, pp. 33-38.

[MF 95] Mukherjea S., Foley J. D.: *'Showing the Context of Nodes in the World-Wide Web'*, Proc. ACM SIGCHI, Denver, CO, 1995.

[MF 95] Mukherjea S., Foley J. D.: *'Visualizing Complex Hypermedia Networks through Multiple Hierarchical Views'*, Proc. ACM SIGCHI, Denver, CO, 1995.

[MF 95] Mukherjea S., Foley J. D.: *'Visualizing the World-Wide Web with the Navigational View Builder'*, Proc. 3rd Int. World Wide Web Conf., Darmstadt, Germany, 1995.

[MGTS 90] Mihalisin T., Gawlinski E., Timlin J., Schwendler J.: *'Visualizing A Scalar Field on an N-dimensional Lattice'*, Visualization '90, San Francisco, CA, 1990, pp. 255-262.

[MHCF 96] Munzner T., Hoffman E., Claffy K., Fenner B.: *'Visualizing the Global Topology of the mbone'*, Proc. Symp. on Information Visualization, 1996.

[Min 01] http://www.sgi.com/software/mineset/

[MMIK 95] Masui T., Minakuchi M., IV G. R., Kashiwagi K.: *'Multiple-View Approach for Smooth Information Retrieval'*, Proc. UIST, 1995, pp. 199-206.

[MRC 91] Mackinlay J. D., Robertson G. G., Card S. K.: *'The Perspective Wall: Detail and Context Smoothly Integrated'*, Proc. Human Factors in Computing Systems CHI '91 Conf., New Orleans, LA, 1991, pp. 173-179.

[MSG 94] McCreary C., Shieh F.-S., Gill H.: *'Cg: A Graph Drawing System Using Graph-Grammar Parsing'*, Proc. Graph Drawing '94, in: Lecture Notes in Computer Science, Vol. 894, Springer, 1994, pp. 270-273.

[MW 95] Martin A. R., Ward M. O.: *'High Dimensional Brushing for Interactive Exploration of Multivariate Data'*, Visualization '95, Atlanta, GA, 1995, pp. 271-278.

Bibliography

[Mye 94] Myers B. A.: *'Visual Programming, Programming by Example and Program Visualization: A Taxonomy'*, Proc. Human Factors in Computing Systems CHI '94 Conf., Boston, MA, 1994, pp. 59-66.

[NK 94] North S. C., Koutsofois E.: *'Applications of Graph Visualization'*, Proc. Graphics Interface, 1994, pp. 235-245.

[Noi 93] Noik E. G.: *'Layout-independent Fisheye Views of Nested Graphs'*, Proc. Symp. on Visual Languages, 1993, pp. 336-341.

[Noi 94] Noik E. G.: *'Encoding Presentation Emphasis Algorithms for Graphs'*, Proc. Graph Drawing '94, in: Lecture Notes in Computer Science, Vol. 894, Springer, 1994, pp. 428-235.

[Noi 94] Noik E. G.: *'A Space of Presentation Emphasis Techniques for Visualizing Graphs'*, Proc. Graphics Interface, 1994, pp. 225-233.

[Ols 93] Olsen K.A. et al.: *'Visualization of a Document Collection: The VIBE System'*, Information Processing and Management, Vol. 29, No. 1, 1993, pp. 69-81.

[OW 93] Ozoyoglu G., Wang H.: *'Example-Based Graphical Database Query Languages'*, Computer, Vol. 26, No. 5, 1993, pp. 25-38.

[PB 94] Pitkow J., Bharat K.: *'Webviz: A Tool for World-Wide Web Access Log Visualization'*, Proc. 1st Int. World Wide Web Conf., Geneva, Switzerland, 1994.

[PF 93] Perlin K., Fox D.: *'Pad: An Alternative Approach to the Computer Interface'*, Proc. SIGGRAPH, Anaheim, CA, 1993, pp. 57-64.

[PG 88] Pickett R. M., Grinstein G. G.: *'Iconographic Displays for Visualizing Multidimensional Data'*, Proc. IEEE Conf. on Systems, Man and Cybernetics, IEEE Press, Piscataway, NJ, 1988, pp. 514-519.

[Pic 70] Pickett R. M.: *'Visual Analyses of Texture in the Detection and Recognition of Objects'*, in: Picture Processing and Psycho-Pictorics, Lipkin B. S., Rosenfeld A. (eds.), Academic Press, New York, 1970.

Bibliography

[Rao 92] Rao R., Card S. K., Jellinek H. D., Mackinlay J. D., Robertson G. G.: *'The Information Grid: A Framework for Information Retrieval and Retrieval-centered Applications'*, Proc. UIST, 1992, pp. 23-32.

[RC 94] Rao R., Card S. K.: *'The Table Lens: Merging Graphical and Symbolic Representation in an Interactive Focus+Context Visualization for Tabular Information'*, Proc. Human Factors in Computing Systems CHI '94 Conf., Boston, MA, 1994, pp. 318-322.

[RCM 93] Robertson G. G., Card S. K., Mackinlay J. D.: *'Information Visualization Using 3D Interactive Animation'*, Communications of the ACM, Vol. 36, No. 4, 1993, pp. 57-71.

[Reg 93] Regan B.: *'Two Algorithms for Drawing Trees in Three Dimensions'*, Proc. Graph Drawing, 1993.

[Rei 93] Reiss S. P.: *'A Framework for Abstract 3D Visualization'*, Proc. Symp. on Visual Languages, 1993, pp. 108-115.

[Rei 94] Reiss S. P.: *'3-D Visualization of Program Information'*, Proc. Graph Drawing, 1994, in: Lecture Notes in Computer Science, Vol. 894, Springer, 1994, pp. 12-24.

[Ren 94] Rennison E.: *'Galaxy of News: An Approach to Visualizing and Understanding Expansive News Landscapes'*, Proc. UIST, 1994, pp. 3-12.

[RG 93] Rekimoto J., Green M.: *'The Information Cube: Using Transparency in 3d Information Visualization'*, Proc. 3rd Annual Workshop on Information Technologies & Systems (WITS '93), 1993, pp. 125-132.

[RKJ+ 99] Ribarsky W., Katz J., Jiang F., Holland A.: *„Discovery Visualization using Fast Clustering"*, IEEE Computer Graphics and Applications, Vol. 19, No. 5, 1999.

[RM 93] Robertson G. G., Mackinlay J. D.: *'The Document Lens'*, Proc. UIST, 1993, pp. 101-108.

[RMC 91] Robertson G. G., Mackinlay J. D., Card S. K.: *'Cone Trees: Animated 3D Visualizations of Hierarchical Information'*, Proc. Human Factors in Computing Systems CHI '91 Conf., New Orleans, LA, 1991, pp. 189-194.

[RSE 99] Rohrer R.M., Sibert J.L., Ebert D.S.:*"Shape-based Visual Interface for Text Retrieval"*, IEEE Computer Graphics and Applications, Vol. 19, No. 5, 1999.

Bibliography

[SA 82] Spence R., Apperley M.: *'Data Base Navigation: An Office Environment for the Professional'*, Behaviour and Information Technology, Vol. 1, No. 1, pp. 43-54.

[SAS 01] http://www.sas.com

[San 94] Sander G.: *'Graph Layout through the VCG Tool'*, Proc. Graph Drawing, 1994, in: Lecture Notes in Computer Science, Springer, Vol. 894, Springer, 1994, pp. 194-205.

[SB 94] Sarkar M., Brown M.: *'Graphical Fisheye Views'*, Communications of the ACM, Vol. 37, No. 12, 1994, pp. 73-84.

[SCB 92] Swayne D. F., Cook D., Buja A.: *'User's Manual for XGobi: A Dynamic Graphics Program for Data Analysis'*, Bellcore Technical Memorandum, 1992.

[Sch 93] Schaffer, Doug, Zuo, Zhengping, Bartram, Lyn, Dill, John, Dubs, Shelli, Greenberg, Saul, Roseman, Mark: *Comparing Fisheye and Full-Zoom Techniques for Navigation of Hierarchically Clustered Networks'*, Proc. Graphics Interface (GI '93), Toronto, Ontario, 1993, in: Canadian Information Processing Soc., Toronto, Ontario, Graphics Press, Cheshire, CT, 1993, pp. 87-96.

[SDTS 95] Su H., Dawkes H., Tweedie L., Spence R.: *'An Interactive Visualization Tool for Tolerance Design'*, Technical Report, Imperial College, London, 1995.

[SFB 94] Stone M. C., Fishkin K., Bier E. A.: *'The Movable Filter as a User Interface Tool'*, Proc. Human Factors in Computing Systems CHI '94 Conf., Boston, MA, 1994, pp. 306-312.

[SGB 00] Sprenger T.C., Brunella R., Gross M.H.:"*Hierarchical Visual Clustering Method using Implicit Surfaces*", ETH Zurich, Tech report, http://graphics.ethz.ch, 2000

[Shn 92] Shneiderman B.: *'Tree Visualization with Treemaps: A 2D Space-Filling Approach'*, ACM Transactions on Graphics, Vol. 11, No. 1, pp. 92-99, 1992.

[SM 94] Sugiyama K., Misue K.: *'A Simple and Unified Method for Drawing Graphs: Magnetic-spring Algorithm'*, Proc. Graph Drawing, 1994, in: Lecture Notes in Computer Science, Vol. 894, Springer, 1994, pp. 364-375.

Bibliography

[SP 92] Stasko, John T. and Patterson, Charles, "Understanding and Characterizing Software Visualization Systems", Proc. of the 1992 IEEE Workshop on Visual Languages, Seattle, WA, September 1992, pp. 3-10.

[Spo 93] Spoerri A.: *'InfoCrystal: A Visual Tool for Information Retrieval'*, Visualization '93, San Jose, CA, 1993, pp. 150-157.

[SPS+ 01]http://www.spss.com

[STA 01] http://www.pnl.gov/nsd/commercial/starlight, 2001

[STDS 95] Spence R., Tweedie L., Dawkes H., Su H.: *'Visualization for Functional Design'*, Proc. Int. Symp. on Information Visualization (InfoVis '95), Atlanta, GA, 1995, pp.4-10.

[SRN 72] Shepard R. N., Romney A. K., Nerlove S. B.: *'Multidimensional Scaling'*, Seminar Press, New York, 1972.

[SW 93] Stasko J. T., Wehrli J. F.: *'Three-dimensional Computation Visualization'*, Proc. Symp. on Visual Languages, Bergen, Norway, 1993, pp. 100-107.

[TD 95] Tal A., Dobkin D.: *'Visualization of Geometric Algorithms'*, Transactions on Visualization and Computer Graphics, Vol. 1, No. 2, 1995, pp. 194-204.

[Tuf 83] Tufte E. R.: *'The Visual Display of Quantitative Information'*, Graphics Press, Cheshire, CT, 1983.

[Tuf 90] Tufte E. R.: *'Envisioning Information'*, Graphics Press, Cheshire, CT, 1990.

[Vas 94] Vasudevan V.: *'Supporting High Bandwidth Navigation in Object-Bases'*, Proc. 10th Int. Conf. on Data Engineering, Houston, TX, 1994, pp. 294-301.

[Vel 92] Velleman P. F: *'Data Desk 4.2: Data Description'*, Ithaca, NY, 1992.

[VXI 02] http://www.cs.sandia.gov/projects/VxInsight.html, 2001

[War 94] Ward M. O.: *'XmdvTool: Integrating Multiple Methods for Visualizing Multivariate Data'*, Visualization'94, Washington, DC, 1994, pp. 326-336.

Bibliography

[WCF+ 00]Wong P., Cowley W., Foote H., Jurrus E., Thomas J.:"*Visualizing Sequential Patterns for Text Mining*", InfoVis'00.

[Wis 95] Wise J. A., Thomas J. J., Pennock K., Lantrip D., Pottier M., Schur A., Crow V.: *'Visualizing the Non-Visual: Spatial Analysis and Interaction with Information from Text Documents'*, Proc. Symp. on Information Visualization, Atlanta, GA, 1995, pp. 51-58.

[WB 95] Wong P.C., Bergeron R. D.: *'30 Years of Multidimensional Multivariate Visualization'*, Proc. Workshop on Scientific Visualization, IEEE Computer Society Press, 1995.

[WDBH 95] Wood A., Drew N., Beale R., Hendley B.: *'Hyperspace: Web Browsing with Visualisation'*, Proc. 3rd Int. World-Wide Web Conf., Poster, 1995, Darmstadt, Germany.

[WL 93] van Wijk J. J., van Liere R.. D.: *'Hyperslice'*, Proc. Visualization '93, San Jose, CA, 1993, pp. 119-125.

[Wri 95] Wright W.: *'Information Animation Applications in the Capital Markets'*, Proc. Int. Symp. on Information Visualization, Atlanta, GA, 1995, pp. 19-25.

[WUT 95] Wilhelm A., Unwin A.R., Theus M.: *'Software for Interactive Statistical Graphics - A Review'*, Proc. Int. Softstat '95 Conf., Heidelberg, Germany, 1995.

[WWT 99]Wong P., Whitney P., Thomas J.:"*Visualizing Association Rules for Text Mining*", InfoVis'99.

[XGobi] XGobi Web-Page including pointers to publications and the most recent release of the XGobi software: "http://www.research.att.com/~andreas/xgobi/".

[Yan 00] Yang L.:"*Interactive Exploration of Very Large Relational Data Sets Through 3D Dynamic Projections*", Proc. ACM SIGKDD Int. Conf. On Knowledge Discovery & Data Mining (KDD 2000), Boston, MA, 2000.

[YS 93] Young D. and Shneiderman B.: 'A graphical filter/flow model for boolean queries: An implementation and experiment', Journal of the American Society for Information Science Vol. 44,No. 6, 1993, pp. 327-339.

[Zlo 77] Zloof M. M.: *'Query-By-Example: A Data Base Language'*, IBM Systems Journal, Vol. 4, 1977, pp. 324-343.