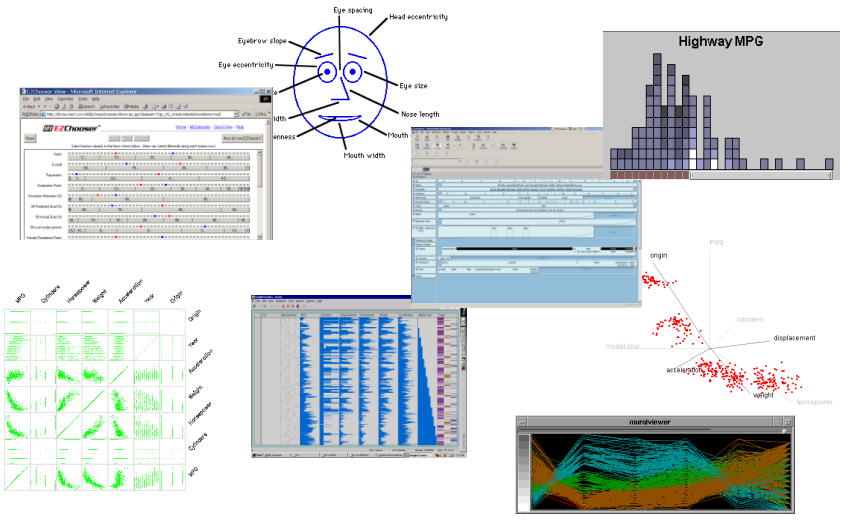
Multivariate Visual Representations 2

CS 7450 - Information Visualization
Sep. 2, 2015
John Stasko

Recap

- We examined a number of techniques for projecting >2 variables (modest number of dimensions) down onto the 2D plane
  - Scatterplot matrix
  - Table lens
  - Parallel coordinates
  - etc.

Varieties of Techniques



Can We Make a Taxonomy?

- D. Keim proposes a taxonomy of techniques
  - Standard 2D/3D display
    - Bar charts, scatterplots
  - Geometrically transformed display
    - Parallel coordinates
  - Iconic display
    - Needle icons, Chernoff faces
  - Dense pixel display
    - What we're about to see…
  - Stacked display
    - Treemaps, dimensional stacking

TVCG'02

Minimum Possible?

- We have data cases with variables
- What's the smallest representation we can use?
  - How?

Dense Pixel Display

- Represent data case or a variable as a pixel
- Million or more per display
- Seems to rely on use of color
- Can pack lots in

- Challenge: What's the layout?

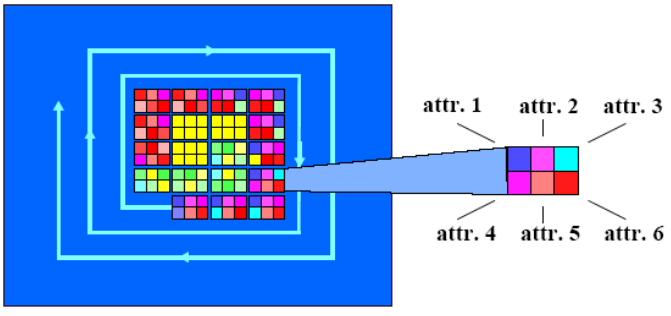
One Representation

- Grouping arrangement
- One pixel per variable
- Each data case has its own small rectangular icon
- Plot out variables for data point in that icon using a grid or spiral layout

Uses color scale

Illustration

schematic representation of 6-dim. data



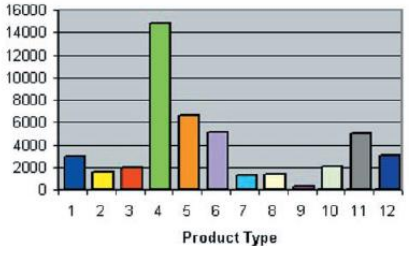
Levkowitz
Vis '91

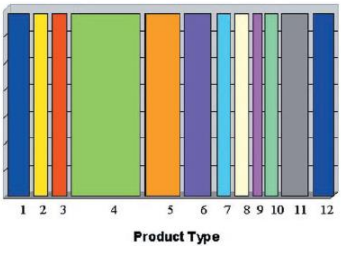
Related Idea

- Pixel Bar Chart
- Overload typical bar chart with more information about individual elements

Keim et al
Information Visualization `02

Idea 1





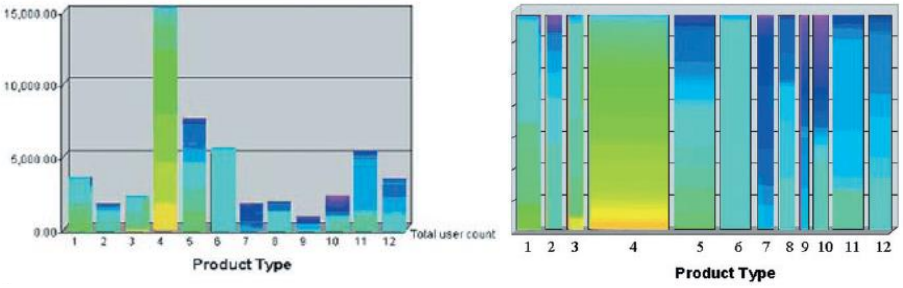
Height encodes quantity

Width encodes quantity

Idea 2

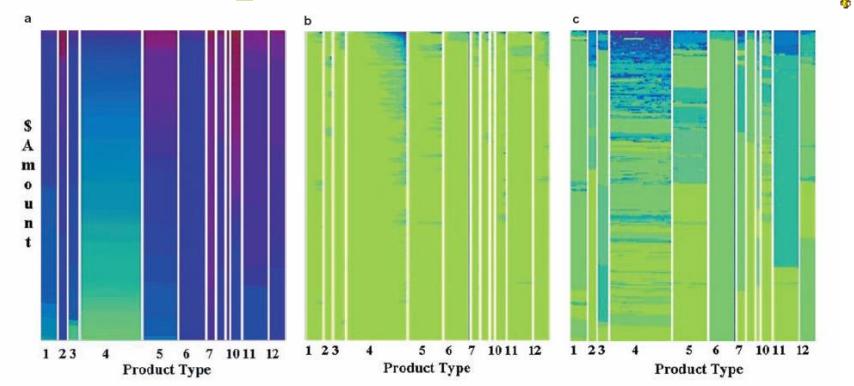
- Make each pixel within a bar correspond to a data point in that group represented by the bar
  - Can do millions that way
- Color the pixel to represent the value of one of the data point's variables

Idea 3



Each pixel is a customer
Color encodes amount spent by that person
High-bright, Low-dark
Ordered by that color attribute too
Right one shows more customers

Idea 4



Product type is x-axis divider
Customers ordered by
- y-axis: dollar amount
- x-axis: number of visits

Color is (a) dollar amount spent, (b) number of visits, (c) sales quantity

Example Application

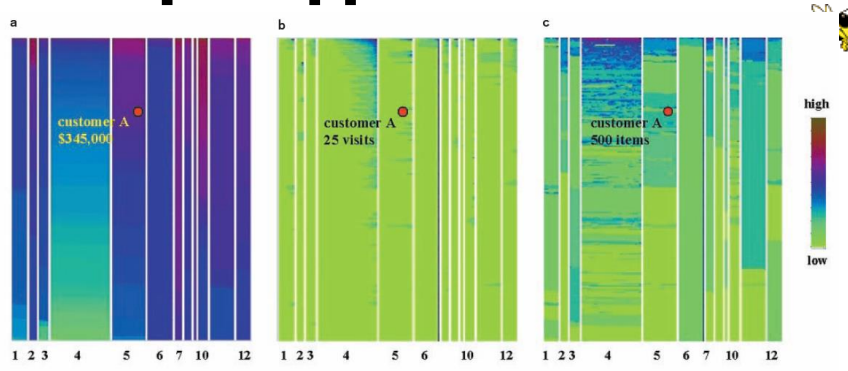


Figure 13 Multi-pixel bar chart for mining 405,000 sales transaction records. (D_x = *Product Type*, $D_y = \perp$, O_x = *no. of visits*, O_y = *dollar amount*, C). (a) Color: dollar amount. (b) Color: no. of visits. (c) Color: quantity.

1. Product type 7 and product type 10 have the top dollar amount customers (dark colors of bar 7 and 10 in Figure 13a)
2. The dollar amount spent and the number of visits are clearly correlated, especially for product type 4 (linear increase of dark colors at the top of bar 4 in Figure 13b)
3. Product types 4 and 11 have the highest quantities sold (dark colors of bar 4 and 11 in Figure 13c)
4. Clicking on pixel A shows details for that customer

Thoughts?

- Do you think that would be a helpful exploratory tool?

High Dimensions

- Those techniques could show lots of data, but not so many dimensions at once
  - Have to pick and choose

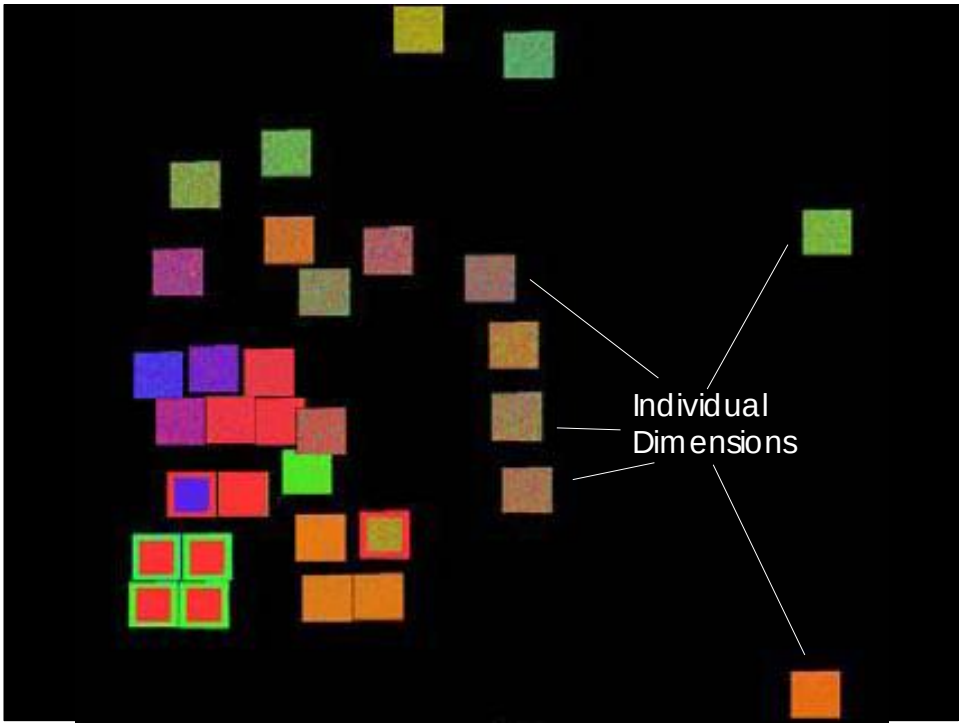
Another Idea

- Use the dense pixel display for showing data and dimensions, but then project into 2D plane to encode more information
- VaR – Value and relation display

Yang et al
InfoVis '04

Algorithm

- Find a correlation function for comparing dimensions
- Calculate distances between dimensions (similarities)
- Make each dimension into a dense pixel glyph
- Assign position for each glyph in 2D plane using multi-dimensional scaling



Questions

- What order are the data cases in each dimension-glyph?
  - Maybe there is a predefined order
  - Choose one dimension as "important" then order data cases by their values in that dimension
    - "Important" one may be the one in which many cases are similar

Alternative

- Instead of each glyph being a dimension, it can be a data case

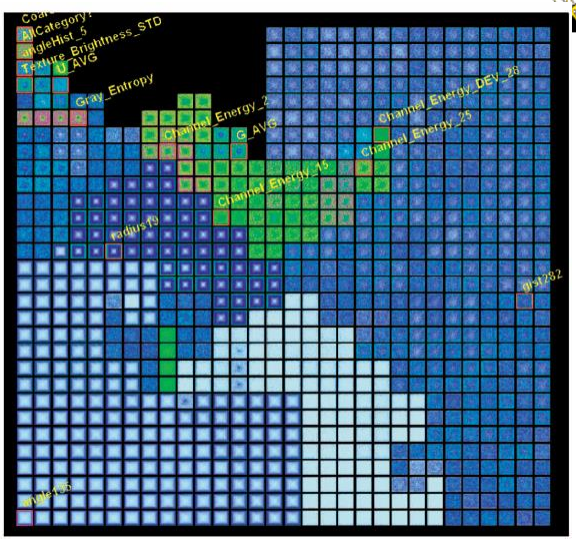
Follow-on Work

- Use alternate positioning strategies other than MDS
- Use Jigsaw map idea (Wattenberg, InfoVis '05) to lay out the dimensions into a grid
  - Removes overlap
  - Limits number that can be plotted

Yang et al
TVCG '07

New Layout

Plot the glyphs into the grid positions



Very Different Metaphor

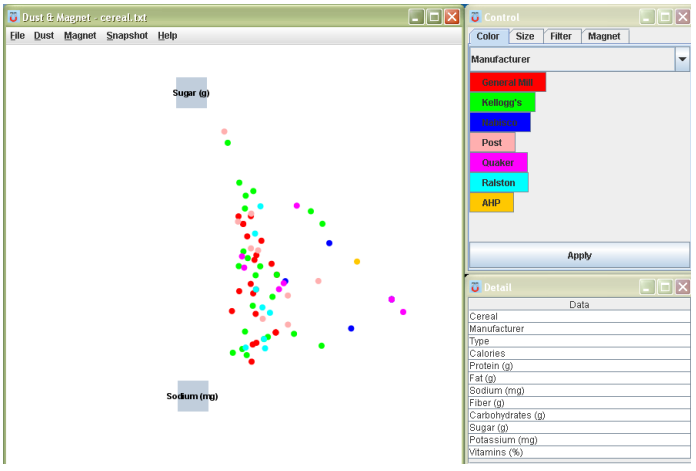
- Represent each data case as a small glyph
- Make interaction be a crucial part of the visualization

Dust & Magnet

- Altogether different metaphor
- Data cases represented as small bits of iron dust
- Different attributes given physical manifestation as magnets
- Interact with objects to explore data

Yi, Melton, Stasko & Jacko
Information Visualization '05

Interface



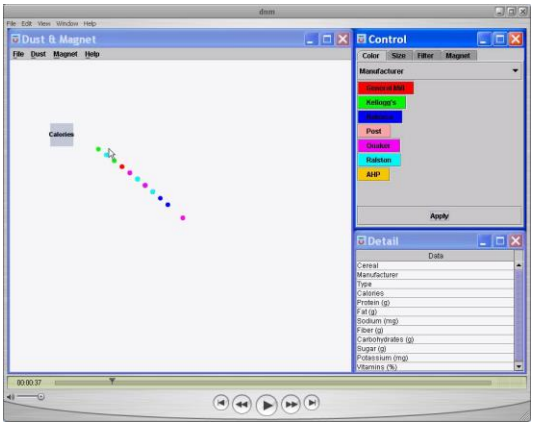
Interaction

- Iron bits (data) are drawn toward magnets (attributes) proportional to that data element's value in that attribute
  - Higher values attracted more strongly
- All magnets present on display affect position of all dust
- Individual power of magnets can be changed
- Dust's color and size can connected to attributes as well

Interaction

- Moving a magnet makes all the dust move
  - Also command for shaking dust
- Different strategies for how to position magnets in order to explore the data

See It Live

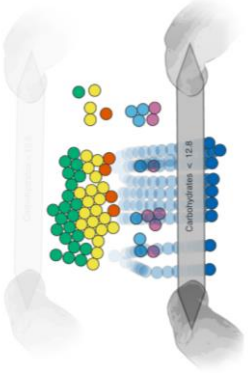


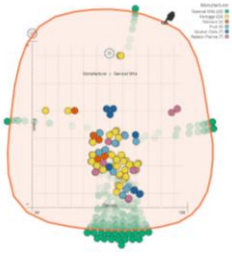
ftp://ftp.cc.gatech.edu/pub/people/stasko/movies/dnm.mov

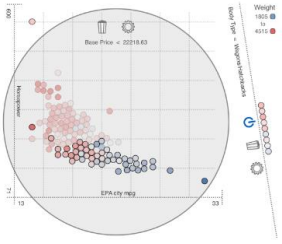
Video & Demo

Kinetica

Video







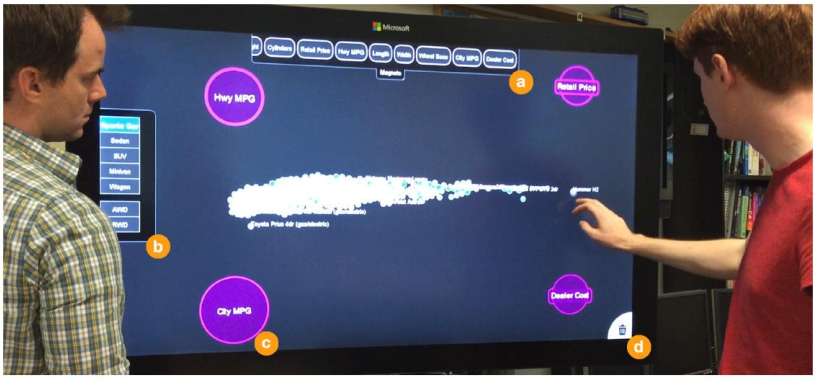
Stress physics metaphor
Touch interaction on tablet

Rzeszotarski & Kittur CHI '14

Go Big

Video

Dust & Magnet on a large multitouch display



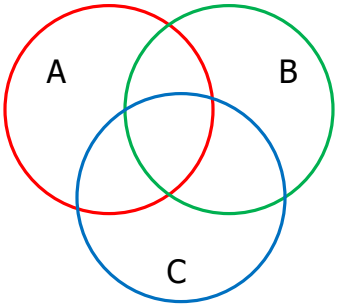
Dai, Sadana, Stolper & Stasko
InfoVis '15 Poster

Set Data & Operations

- Different type of problem
  - Large set of items, each can be in one or more sets
  - How do we visually represent the set membership?

Standard Technique

Venn Diagram

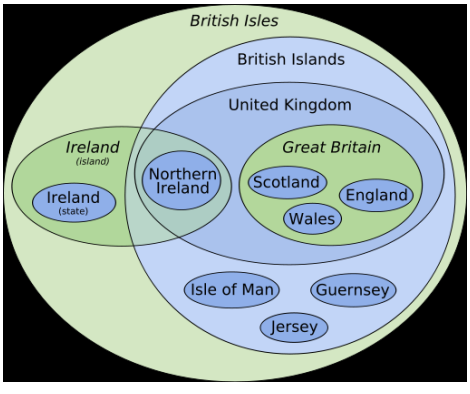


Contains all possible zones of overlap

Alternately

Euler Diagram

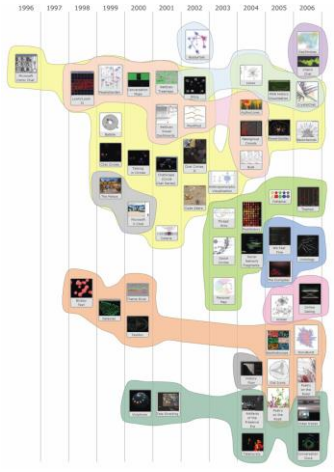
Does not necessarily show all possible overlap zones



http://en.wikipedia.org/wiki/File:British_Isles_Euler_diagram_15.svg

But what's the problem?

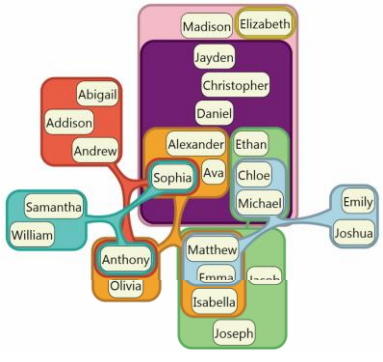
Bubble Sets



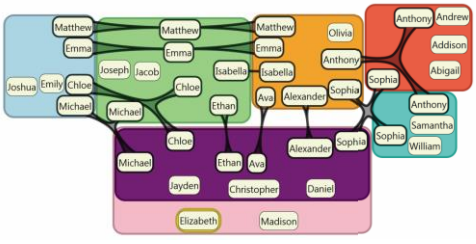
Video

Collins et al
TVCG (InfoVis) '09

ComED & DupED



Item appears once

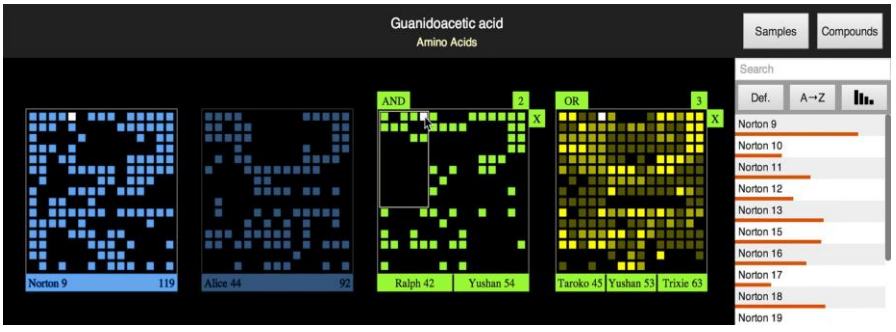


Item can appear more than once

Video

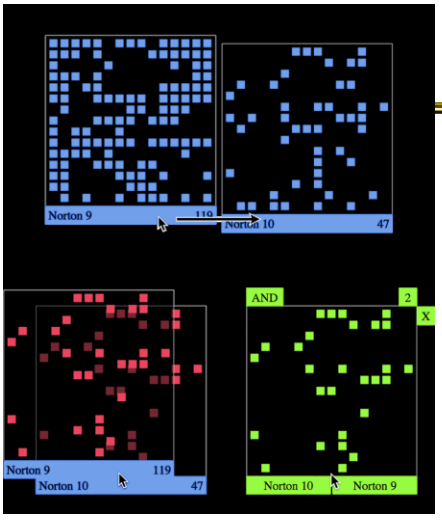
Riche & Dwyer
TVCG (InfoVis) '10

OnSet

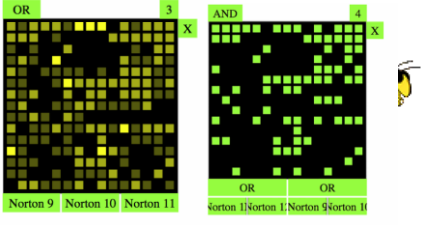


Represent set as a box, elements are spots in that box
Use interaction to do set union, intersection

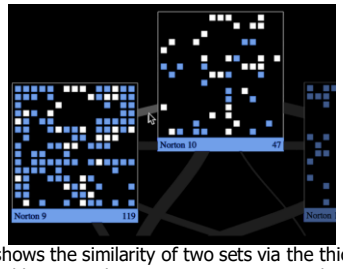
Sadana, Major, Dove & Stasko
TVCG (InfoVis) '14



Dragging and dropping a PixelLayer to create a new AND MultiLayer.



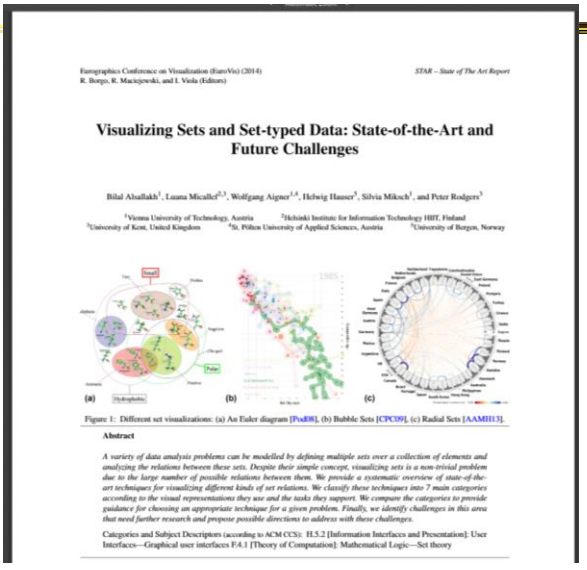
A MultiLayer OR with three sets.
A MultiLayer AND of nested OR layers.



OnSet shows the similarity of two sets via the thickness of a band between them. Hovering over a similarity band highlights the common elements between two sets.

Demo/video

Nice Review



Eurographics Conference on Visualization (EuroVis) (2014)
R. Borgo, R. Maciejewski, and I. Viola (Editors)

STAR – State of The Art Report

Visualizing Sets and Set-typed Data: State-of-the-Art and Future Challenges

Bilal Alsallakh[1], Luana Micallef[2,3], Wolfgang Aigner[1,4], Helwig Hauser[5], Silvia Miksch[1], and Peter Rodgers[3]

[1]Vienna University of Technology, Austria [2]Helsinki Institute for Information Technology HIIT, Finland
[3]University of Kent, United Kingdom [4]St. Pölten University of Applied Sciences, Austria [5]University of Bergen, Norway

Figure 1: Different set visualizations: (a) An Euler diagram [Pod88], (b) Bubble Sets [CPC09], (c) Radial Sets [AAMH13].

Abstract

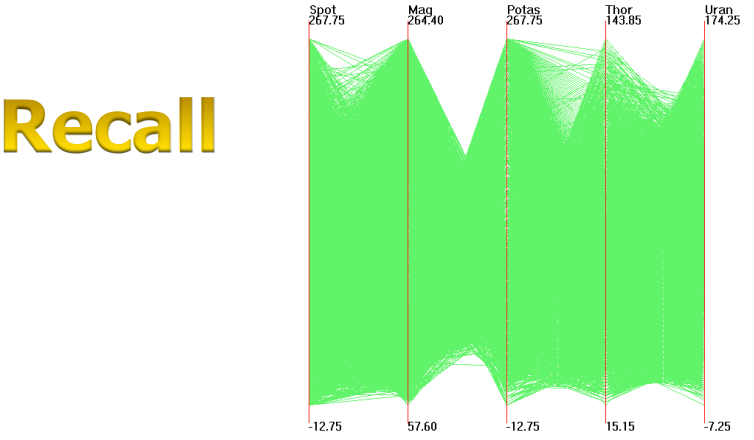
A variety of data analysis problems can be modelled by defining multiple sets over a collection of elements and analyzing the relations between these sets. Despite their simple concept, visualizing sets is a non-trivial problem due to the large number of possible relations between them. We provide a systematic overview of state-of-the-art techniques for visualizing different kinds of set relations. We classify these techniques into 7 main categories according to the visual representations they use and the tasks they support. We compare the categories to provide guidance for choosing an appropriate technique for a given problem. Finally, we identify challenges in this area that need further research and propose possible directions to address with these challenges.

Categories and Subject Descriptors (according to ACM CCS): H.5.2 [Information Interfaces and Presentation]: User Interfaces—Graphical user interfaces F.4.1 [Theory of Computation]: Mathematical Logic—Set theory

Step Back

- Most of the techniques we've examined work for a modest number of data cases or variables
  - What happens when you have lots and lots of data cases and/or variables?

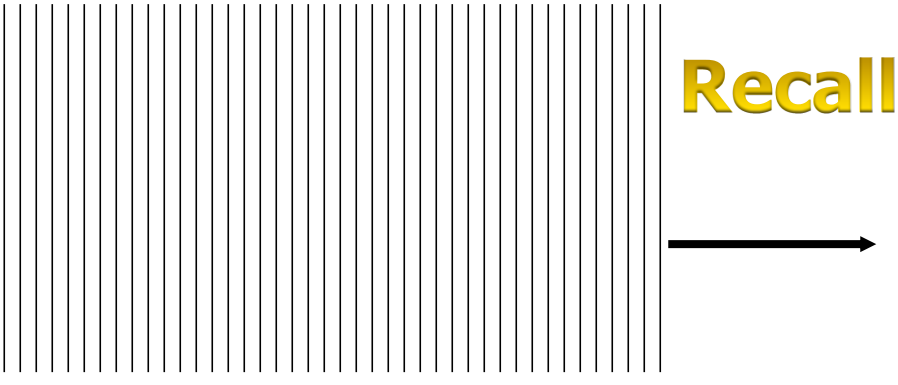
Many Cases



Out5d dataset (5 dimensions, 16384 data items)

(courtesy of J. Yang)

Many Variables



Strategies

- How are we going to deal with such big datasets with so many variables per case?
- Ideas?

General Notion

- Data that is similar in most dimensions ought to be drawn together
  - Cluster at high dimensions
- Need to project the data down into the plane and give it some ultra-simplified representation
- Or perhaps only look at certain aspects of the data at any one time

Mathematical Assistance 1

- There exist many techniques for clustering high-dimensional data with respect to all those dimensions
  - Affinity propagation
  - k-means
  - Expectation maximization
  - Hierarchical clustering

Mathematical Assistance 2

- There exist many techniques for projecting n-dimensions down to 2-D (dimensionality reduction)
  - Multi-dimensional scaling (MDS)
  - Principal component analysis
  - Linear discriminant analysis
  - Factor analysis

Comput Sci & Eng courses
Data & Visual Analytics, Prof. Chau

Data mining
Knowledge discovery

Other Techniques

- Other techniques exist to manage scale
  - Sampling – We only include every so many data cases or variables
  - Aggregation – We combine many data cases or variables
  - Interaction (later)
    - Employ user interaction rather than special renderings to help manage scale

Use?

- What kinds of questions/tasks would you want such techniques to address?
  - Clusters of similar data cases
  - Useless dimensions
  - Dimensions similar to each other
  - Outlier data cases
  - …
- Think about the "cognitive tasks" we want to accomplish

Recap

- We've seen many general techniques for multivariate data these past two days
  - Know strengths and limitations of each
  - Know which ones are good for which circumstances
  - We still haven't explored interaction much

Visualization of the Day

- Everyone posts one
- Use tumblr
  - Overview on class webpages
  - Details on t-square
- Please comment & share thoughts
- Part of participation grade

Project

- Overview
- Topics
  - Last.fm example
- Teams

- Teams & Topics due Monday 14th
  - You must meet me or TA before then
  - Bring 3 copies

HW 1

- Recap

Design Challenge

year	os	units
2007	Symbian	77.7
2007	RIM	11.8
2007	iPhone	3.3
2007	Windows	14.7
2007	Android	0
2007	Other	14.9
2008	Symbian	72.9
2008	RIM	23.1
2008	iPhone	11.4
2008	Windows	16.5
2008	Android	0.6
2008	Other	15.3
2009	Symbian	80.9
2009	RIM	34.3
2009	iPhone	24.9
2009	Windows	15
2009	Android	6.8
2009	Other	10.4
2010	Symbian	107.7
2010	RIM	46.9
2010	iPhone	41.5
2010	Windows	12.7
2010	Android	47.5
2010	Other	12.6
2011	Symbian	141.3
2011	RIM	62.2
2011	iPhone	70.7
2011	Windows	21.3
2011	Android	91.9
2011	Other	26

Smart Phones sold by OS

Challenge: Help someone understand the competitive landscape in this area

Projections

Source: Gartner

Upcoming

- Labor Day holiday
- Visualization Programming Tutorial
  - Reading
    Murray online book
- InfoVis Systems & Toolkits
  - Reading:
    Viegas et al, '07