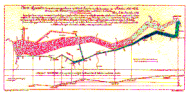
Challenge 2012

Marcel Worring

http://www.vacommunity.org/VAST+Challenge+2012



Challenge 2012

- The setting
  - BankWorld, a planet much like Earth, but with a very different geography.
    - In fact, for this challenge, we are dealing with one large land mass containing several different nation-states

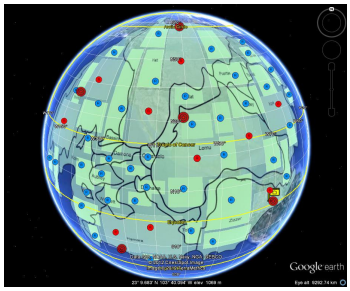
Bankworld

Bankworld and Earth

Challenge 2012

- The setting
  - The most important organization on BankWorld is the Bank of Money (BOM). BOM has many offices of various sizes across BankWorld.
  - Each of these offices has many computers active throughout the day. In fact, we are dealing with about 1,000,000 machines.

Datacentres



Mini Challenge 1

- Cyber Situation Awareness
  - How do you achieve cyber situation awareness across the entire enterprise with such a large number of systems?

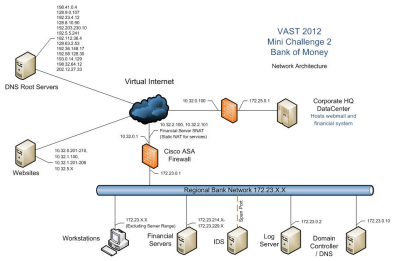
Mini Challenge 1

- Mini challenge 1.1
  - Create a visualization of the health and policy status of the entire Bank of Money enterprise as of 2 pm BMT (BankWorld Mean Time) on February 2. What areas of concern do you observe? (Short Answer)
- Mini challenge 1.2
  - Use your visualization tools to look at how the network's status changes over time. Highlight up to five potential anomalies in the network and provide a visualization of each. When did each anomaly begin and end? What might be an explanation of each anomaly? (Detailed Answer)

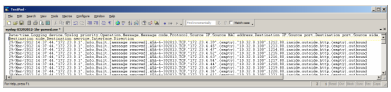
Mini Challenge 1

- Data
  - From 1.000.000 (MC1.1) to 150.000.000 entries (MC1.2) with very basic information
- Components
  - Geovisualization
  - Networks/graphs
  - Multivariate visualizations
- Data analysis
  - Generic mining and statistics

Mini Challenge 2



The data



Firewall logs

Intrusion detection logs

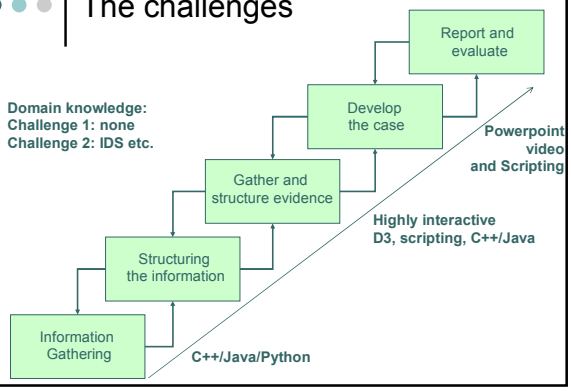
Mini challenge 2

- Bank of Money regional office network operations forensics
  - When something does go awry, can you identify it and the steps needed to resolve the problem?

Mini challenge 2

- Mini challenge 2.1
  - Using your visual analytics tools, can you identify what noteworthy events took place for the time period covered in the firewall and IDS logs? Provide screen shots of your visual analytics tools that highlight the five most noteworthy events of security concern, along with explanations of each event.
- Mini challenge 2.2
  - What do you suspect is (are) the root cause(s) of the events identified in MC 2.1? Understanding that you cannot shut down the corporate network or disconnect it from the internet, what actions should the network administrators take to mitigate the root cause problem(s)?

The challenges



Teams

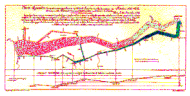
- We will have (one ore more) teams for the following
  - Challenge 1
    - Data mining and large-scale visualization
    - Interaction and presentation
  - Challenge 2
    - Security analysis and large scale visualization
    - Interaction and presentation

Multivariate visualization

Lecture 3:
Marcel Worring

Various sources, among others:

Wong and Bergeron: 30 Years of Multidimensional Multivariate Visualization
Elmqvist et.al. Rolling the Dice: Multidimensional Visual Exploration Using Scatterplot Matrix Navigation



Schneiderman: the eyes have it

The 7 basic datatypes

- According to Schneiderman
  - *1,2,3-dimensional*
  - *multi-dimensional*
  - Temporal
  - Tree
  - Network
- And in a modern age also
  - Multimedia data
    - Text, audio, **image**, video

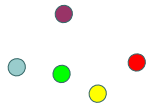
Structures

A tree

An undirected graph

A directed graph

Qualitative: Nominal variables



In this classification, names are assigned to objects as labels. These names come from a given small set, and are meant to identify categories used for classifying the data.

Qualitative: Ordinal variables

In this classification, the numbers assigned to objects represent the rank order (1st, 2nd, 3rd etc.) of the entities measured.

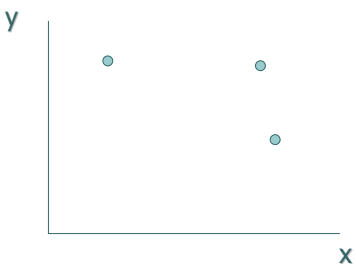
Quantitative: Interval variables

The numbers assigned to objects have all the features of ordinal measurements, and in addition equal differences between measurements represent equivalent intervals. That is, differences between arbitrary pairs of measurements can be meaningfully compared.

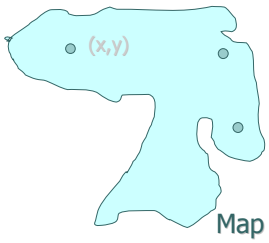
Quantitative: Ratio variables

The numbers assigned to objects have all the features of interval measurement and also have meaningful ratios between arbitrary pairs of numbers.

Quantitative: spatial



Quantitative: geophysical



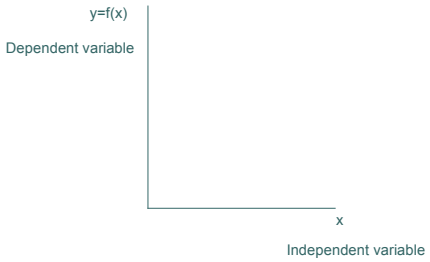
Time

Quantitative

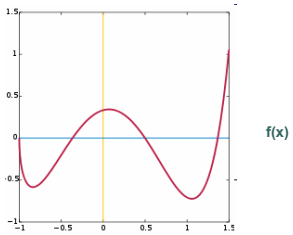
Ordinal

Univariate data

General model: y=f(x)



Example



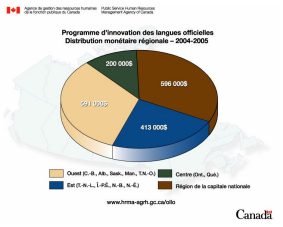
But this is a mathematical function we are dealing with measurements, observations and such.

Example

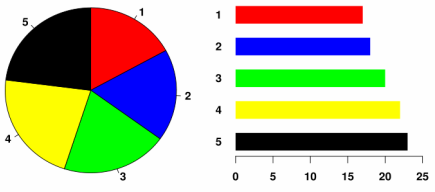
- f(west) = 591000
- f(east) = 413000
- f(centre) = 200000
- f(other) = 596000

How can we visualize this?

Numeric data: Pie chart



Two alternative presentations

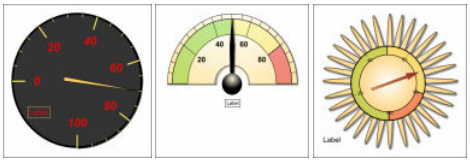


Corda.com: Traffic Lights

Used for visualizing binary indicator

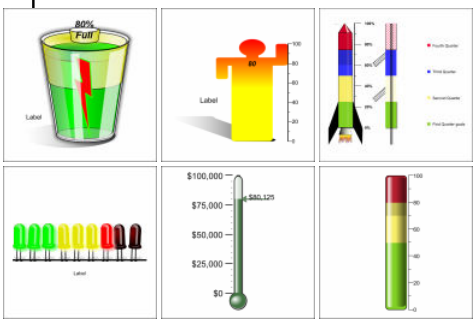
f(x) = 0 or f(x) = 1

Corda.com: Radials



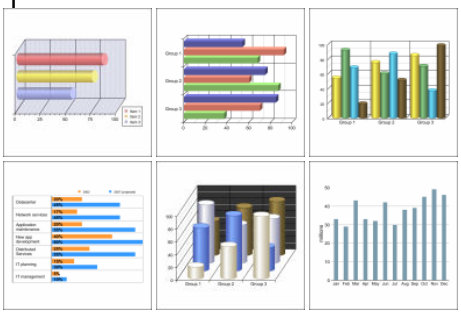
Employing the dashboard metaphor

Corda.com: Filled displays

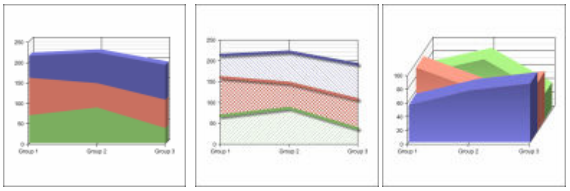


Adding a direct semantic interpretation

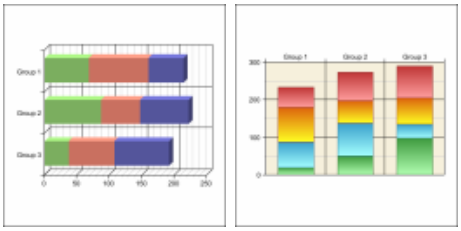
Corda.com: bar charts



Corda.com: Area Graphs



Corda.com: stacked graphs



How about the following

- Measurements
  - 1,5,6,4,5,7,9,5,6,7,4,6,7,4,6,4,7,4,7,4,5,8,,7,9,8,7,6,8,7,6,5,4,5,4,3,4,5,6,7,4,6,3,7,6,7,8,5

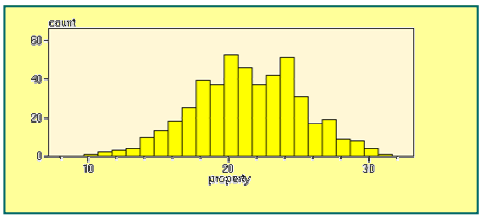
Count the number of occurrences of each measurement

And this one

- Measurement
  - 1.3, 6.5, 8.6, 2.1, 4.4, 12.32, 1.99, 24.67, 32.0, 1.456, 3.776, 9.88, 3.2, 2.45, 11.2, 13.2, 27.56

Define bins (e.g. 0-2, 3-5, 5-7, …..) and count the elements that fall in each of those

Histogram

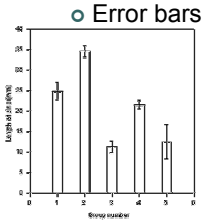


A special case of a bar chart, where the values denote frequencies.

Teach/Me Data analysis

Repeated measurements Bar graphs

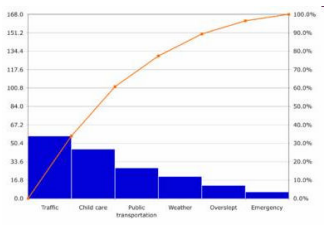
- Error bars



When your measurements have an inherent uncertainty (and they often have) make sure to indicate that properly.

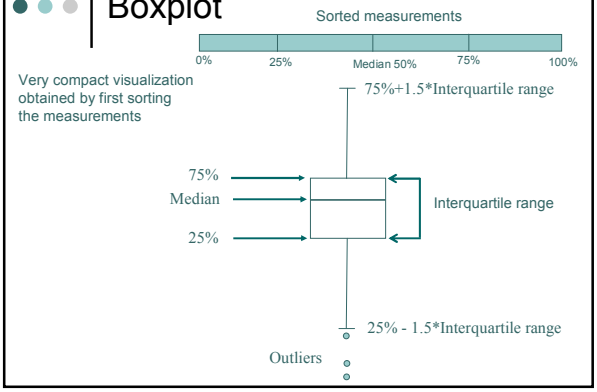
Usually symmetric with one standard deviation to both sides.

Pareto curves

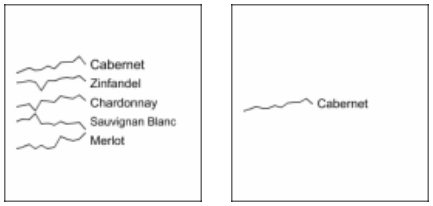


Add a cumulative graph

Boxplot

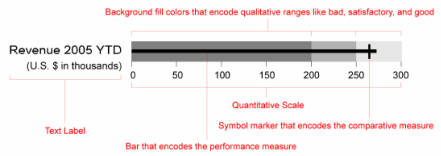


Corda.com: spark lines



Small lines to indicate the general trend in the data

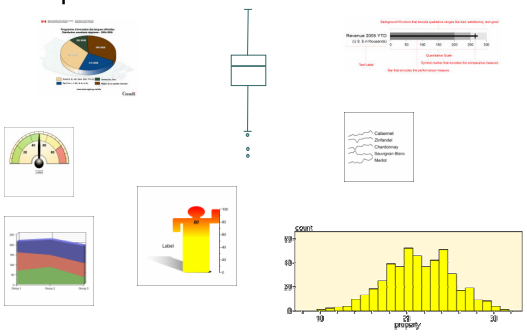
Bullet Graph



More complex

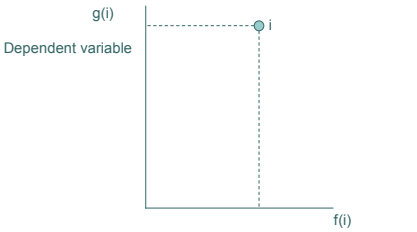
- Multidimensional data
  - How to go beyond 3 dimensions?
  - More than you can easily visualize

Univariate Visualizations



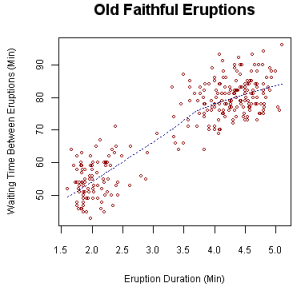
Bivariate data

General model v(i) = (f(i),g(i))



Bivariate data

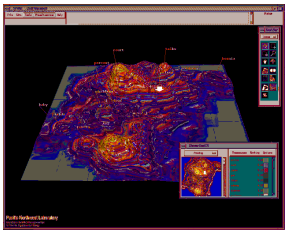
Scatterplot



Good way to see correlations between two variables.

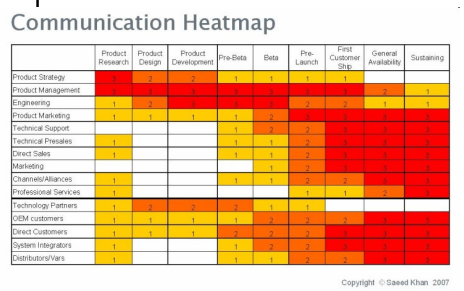
To indicate the difference

General model v(x,y) = f(x,y)



This general model yields a 2D surface

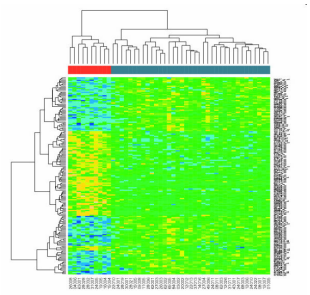
A discrete example



A very simple example: a color scale is used to denote values

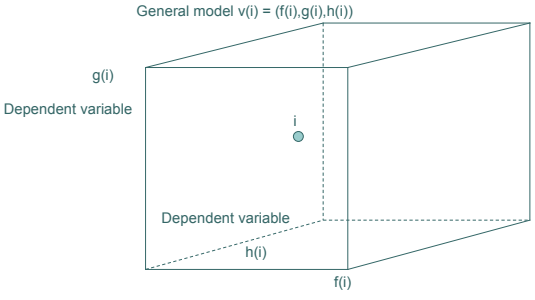
Ref: University of -Warwick

Heatmap

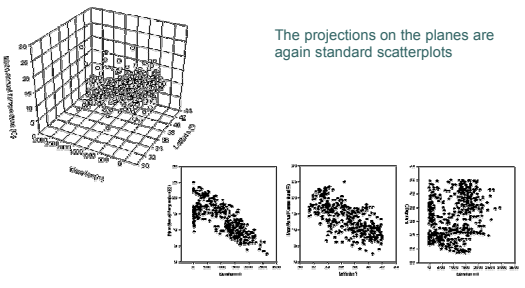


A far more elaborate example where structure is also visualized

Trivariate data

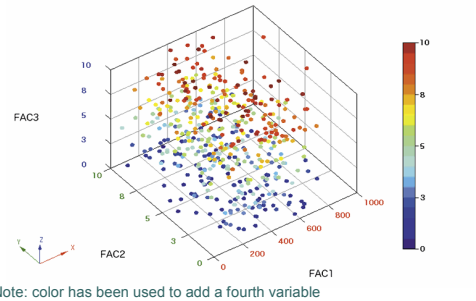


Three dimensional graphs



The projections on the planes are again standard scatterplots

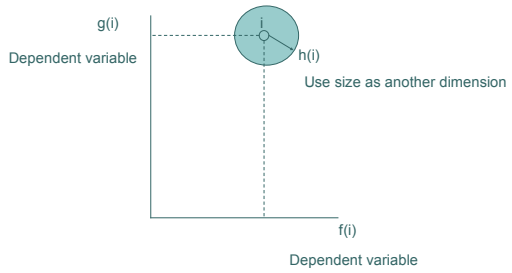
Example 3D scatterplot



Note: color has been used to add a fourth variable
(in general this would be a fourth dimension, but here it is just reinforcing the z-dimension.)

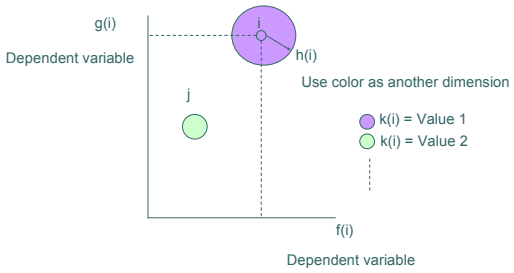
An alternative visualization

General model v(i) = (f(i),g(i),h(i))

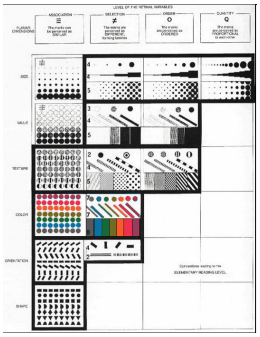


And four dimensional

General model v(i) = (f(i),g(i),h(i),k(i))



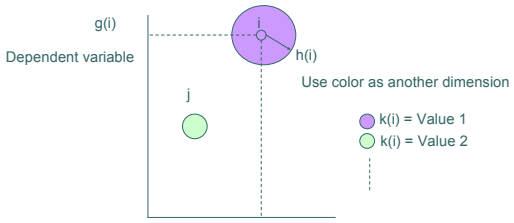
Or any of the other channels



Jacques Bertin

And 5- dimensional

General model v(i) = (f(i),g(i),h(i),k(i),t(i))



Add t(i) as a time dimension using animation

A famous example

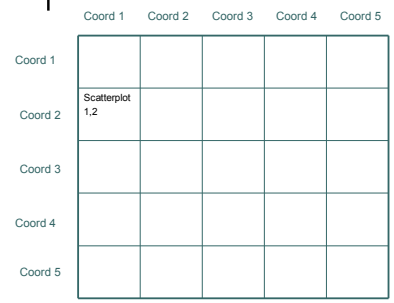
- Gapminder

http://www.gapminder.org/world

And beyond five?

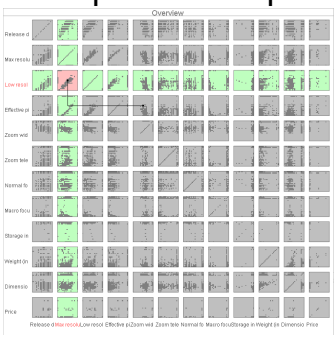
- Just a little bit more
  - Let's say 5-10
- Too many
  - >10

Scatterplot matrices



Stay within the scatterplot paradigm, but make many of them

Example: Scatterplot Matrix



Elmqvist: Rolling the Dice: Multidimensional Visual Exploration using Scatterplot Matrix Navigation, 2008

Demo

SCATTER DICE

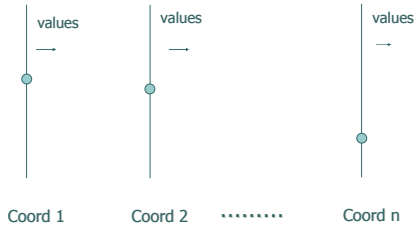
http://www.aviz.fr/~fekete/scatterdice/

Rolling the Dice: Multidimensional Visual Exploration using Scatterplot Matrix Navigation

Table lens

N dimensional vector 5-10 dimensions

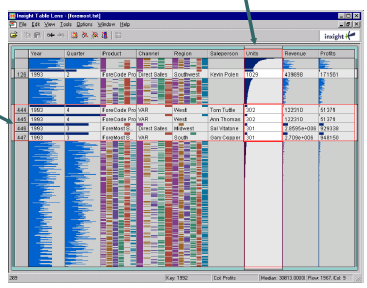
General model v(i) = (f(i),g(i),h(i),k(i), ……)



TableLens

Actions
- Sort by 1 column

Actions
Show instance (without losing context)

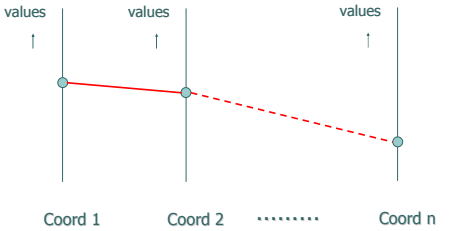


Actions
- Change order columns
- Hide/show columns
- Sort by 1 column

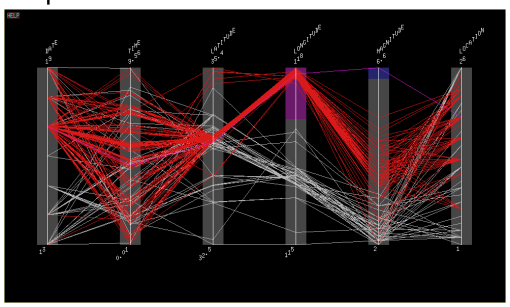
Parallel coordinates

N dimensional vector 5-10 dimensions

General model v(i) = (f(i),g(i),h(i),k(i), ……)

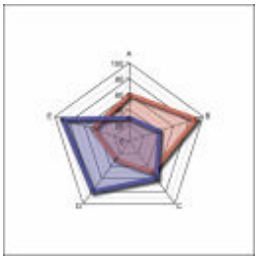


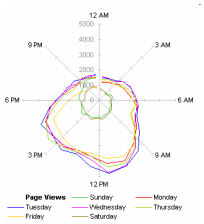
Parallel coordinates



Schall 95

Corda.com: radar curves





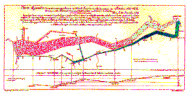
Ref: Sit4eMeter

Dashboard design

Marcel Worring

Various sources, among others:

Few: Information Dashboard Design



Dashboards

- Definition:
  - A dashboard is a visual display of the most important information needed to achieve one or more objectives; consolidated and arranged on a single screen so the information can be monitored at a glance.

Oracle

A case study

- Imagine
  - You are working as a manager at a marketing company launching a new product and you want to monitor how your 10 marketeers are doing in terms of reaching their targets

Dashboard design

Source	Type

Check
1
2
3
4
5
6
7
8
9
10
11
12
13

Various visualizations