Visual Analytics

Giuseppe Santucci

5 – Representation

Thanks to John Stasko, Robert Spence, Ross Ihaka, Marti Hearst, Kent Wittemburg

Outline

- Data types & data complexity
- Encoding of values
  - Univariate data
  - Bivariate data
  - Trivariate data
  - Multidimensional data
- Encoding of relations
- Lines
- Map & Diagrams
- Trees
- Support for design

Outline

- Data types & data complexity
- Encoding of values
  - Univariate data
  - Bivariate data
  - Trivariate data
  - Multidimensional data
- Encoding of relations
- Lines
- Map & Diagrams
- Trees
- Support for design

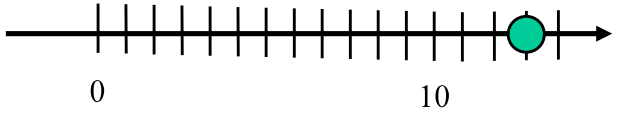
Data types and complexity

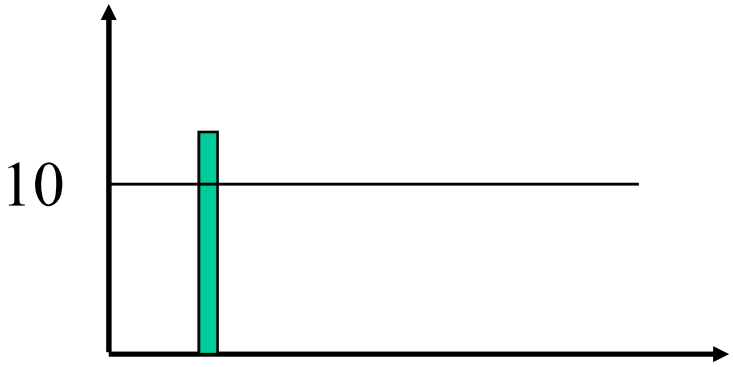
- Attributes are just single values (Numerical /Categorical)
- A single data item (e.g., a single car) has several attributes and the item represent a (mathematical) relation among them
- Attribute pairs may present some patterns (e.g., correlations)
- Data items can have same patterns in their R^n space (e.g., cluster, outliers, etc.)
- Or they can have relationships (E-R model)

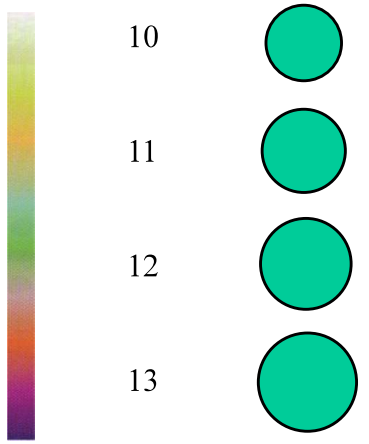
- Visual representation of single values
- Visual representation of relations / relationships
- Visual representation of patterns / functions

Visual representation of a single value

- Representing the price of a car using different encodings





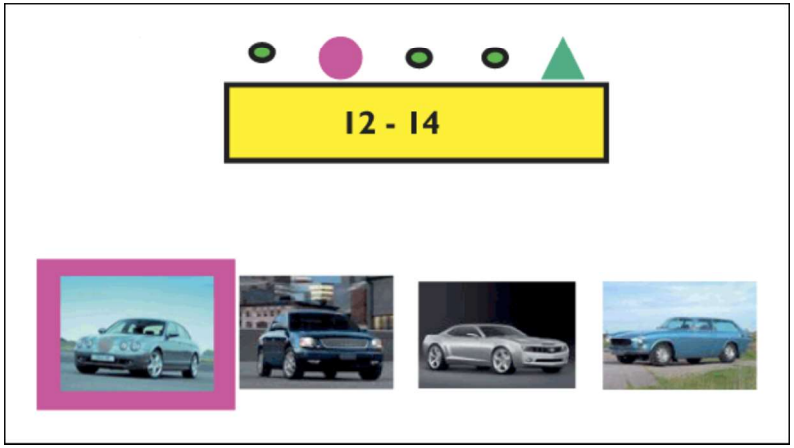


hmmm...

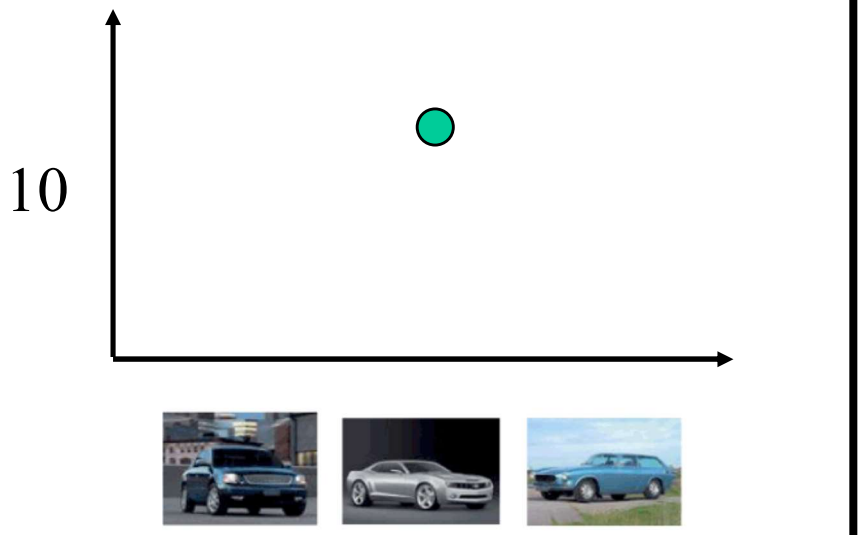
Relations among values

Make	Price (£)	MPG	Rating	Age (yrs)
Ford	15,450	31	*****	3
Chevy	12,450	27	***	4

- A table representing a relation



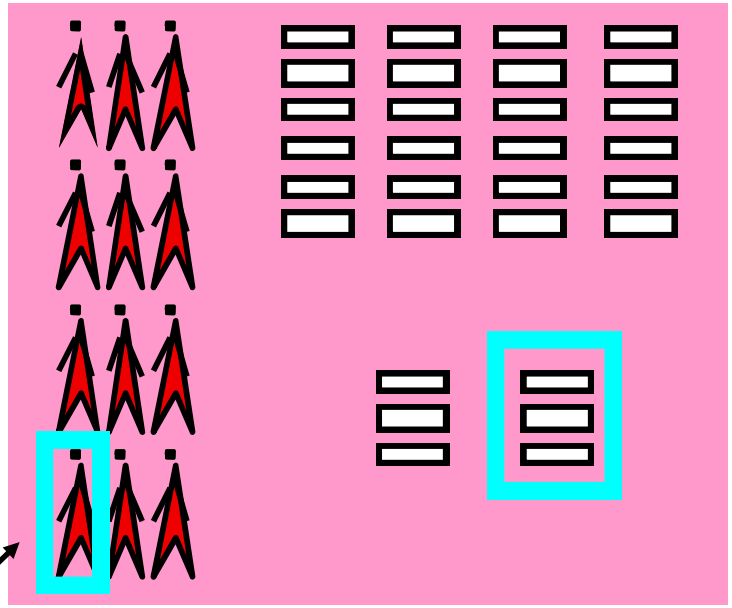
Color coding representing a relation



A scatter plot representing a relation

Brushing

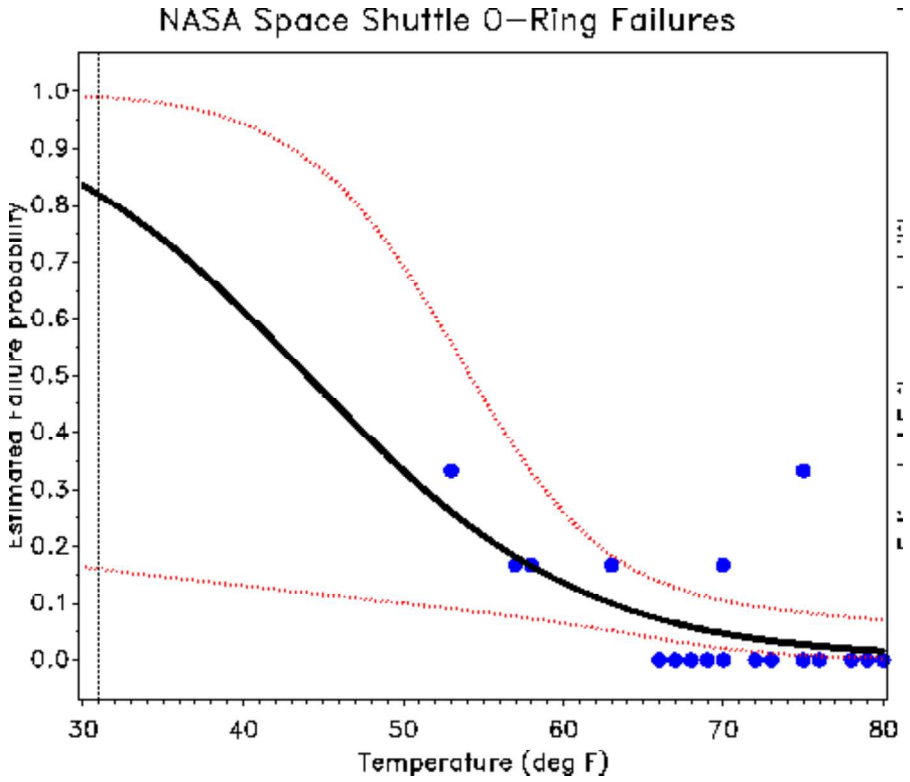
- Brushing, a very useful interaction technique, can be used to visualize a relationship among data items

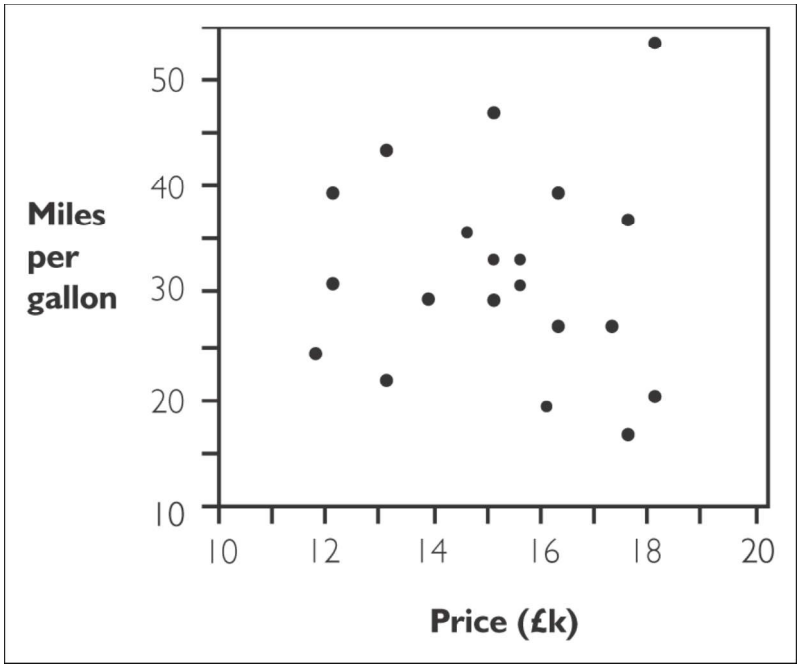


Mouse over

Interaction to identify a doctor highlights the hospital beds under his care, and *vice versa*

Function between two attributes





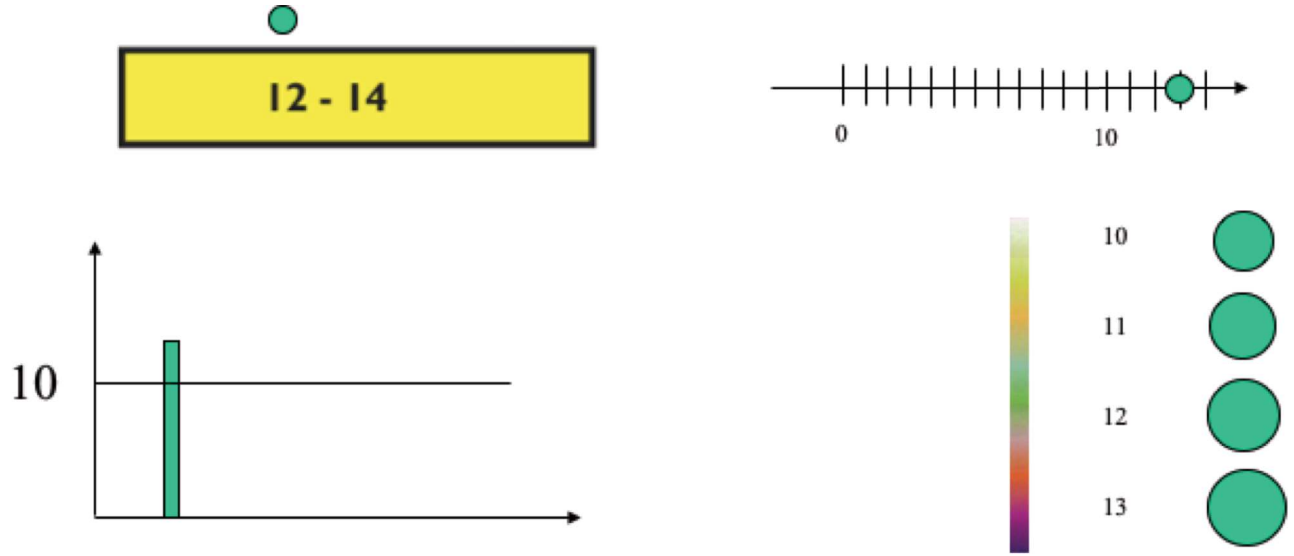
Scatterplot, correlation, regression, etc.

Outline

- Data types & data complexity
- Encoding of values
  - Univariate data
  - Bivariate data
  - Trivariate data
  - Multidimensional data
- Encoding of relations

Univariate data

- Representing single values seem quite easy



- Some human factors make, sometimes, this problem quite hard

An aircraft example

- A basic measure: aircraft height (basic value but very important…)

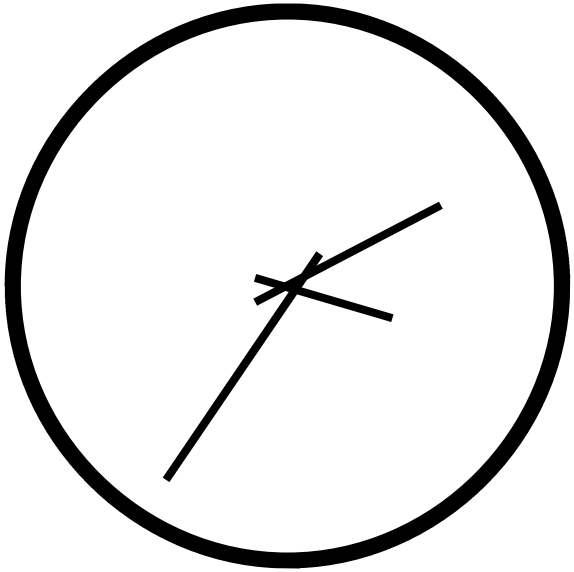
Typical altimeter
13.460 feet

Three problems with it:

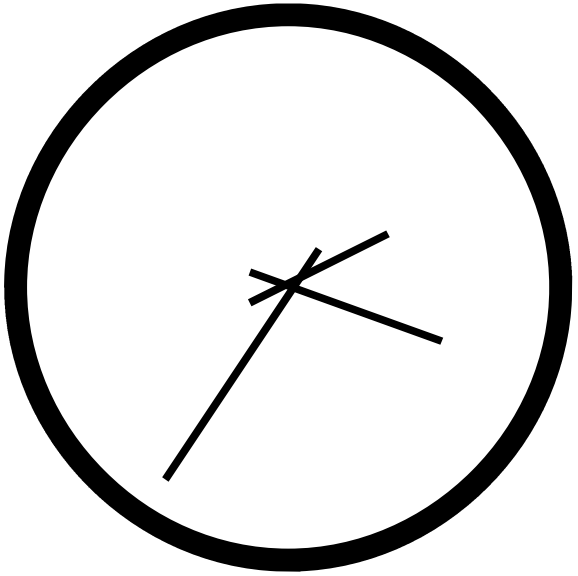
1. The unit : feet (too little) pilots think in terms of hundreds of feet; however 100 feet are still not enough for real usage so altimeters use 3 hands
   - 10.000 feet
   - 1.000 feet
   - 100 feet

Change blindness

2. Pilots gaze moves away from altimeter quite often (piloting, looking at other instruments, etc.) and some, similar representations are sometimes confused (with very bad consequences…)



32.600 feet



24.600 feet (about 3km lower!)


100
FEET
0
1
2
3
4
5
6
7
8
9
ALT
1015

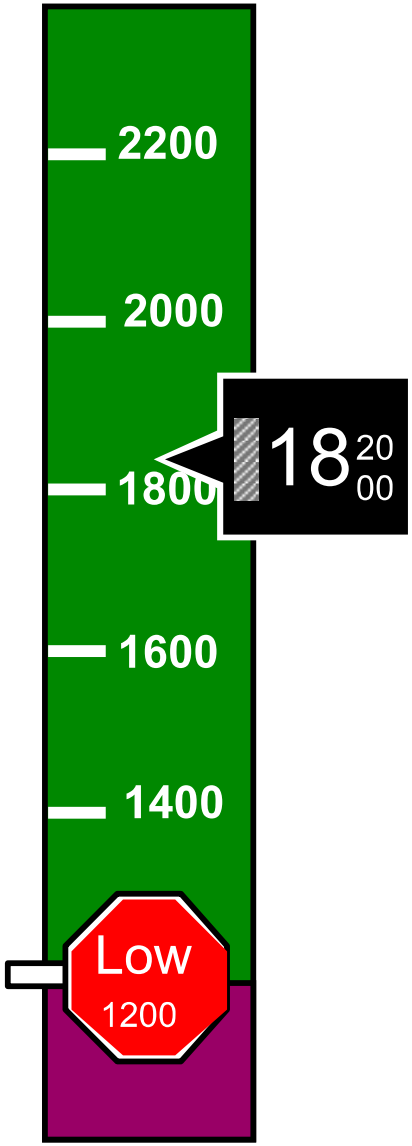
Animation can help...

Stress and fatigue

3. During an emergency, a pilot has to control several instruments and reading the altimeter under stress can make change blindness more easy

European glider pilots take less risks (only two hands and meters)

Modern solution



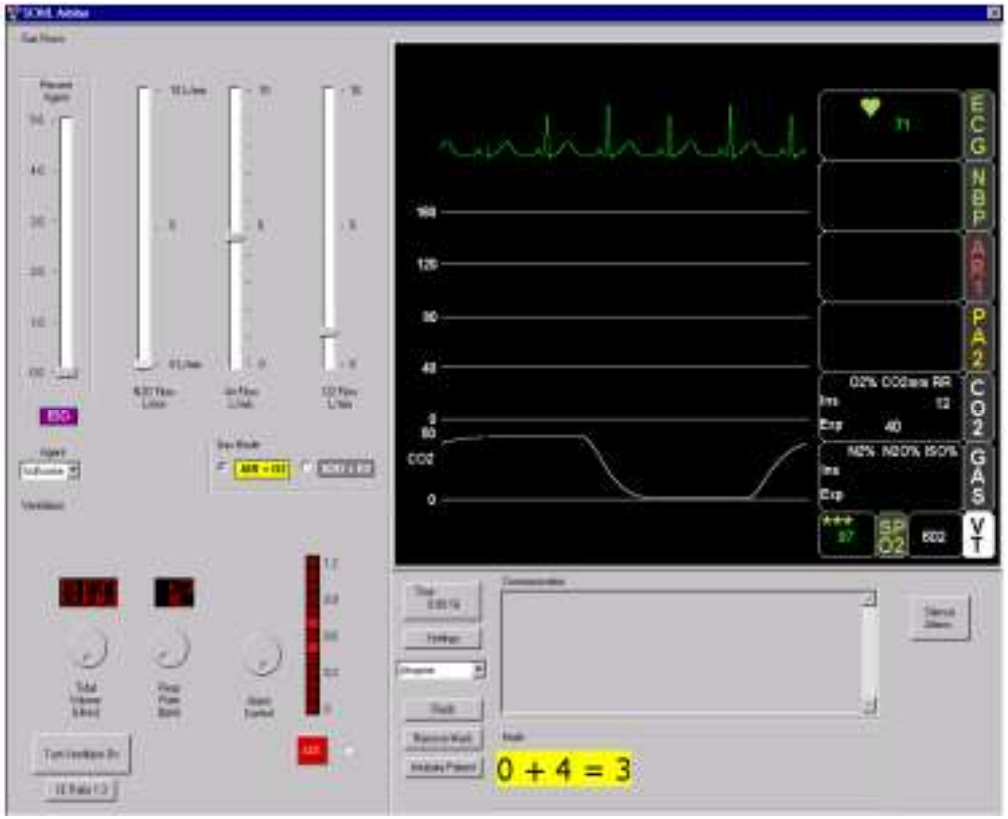
Overview

Details

Bigger digits reflect the pilot mind:

hundreds of feet

Stress and fatigue



Representations of the vital signs of a patient during an operation. The difficulty of paying constant attention to such a display throughout a long operation has led to the encoding of vital signs in the pitch of a frequently repeated ‘beep’. A change in pitch is immediately noticed wherever the gaze of the anesthetists is directed

Note: sometimes aircraft pilots in emergency situations turn off distracting alarm noises…

Back to change blindness

What is the difference between these two picture (if any)?

I prefer the first one …

If it is not a glider … I like engines…

Animation can greatly help!!

Animation and brushing are very powerful tools

Another kind of blindness

In the next movie, the girl with the white t-shirt is going to receive the ball several times

Count how many times she receives the ball (disregarding knocking up)

Ready?

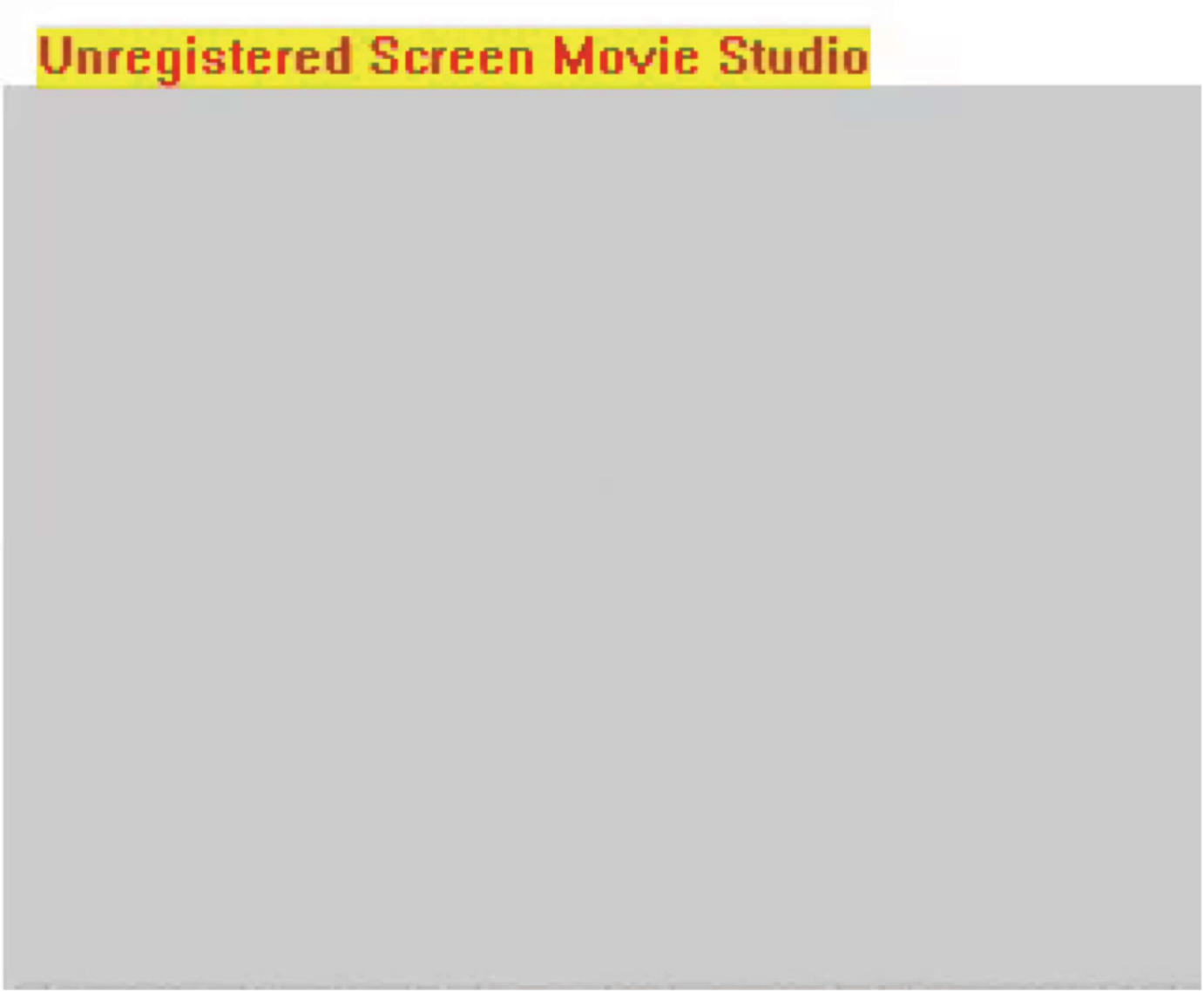
So...

- 6 times ?
- 7 times ?
- 8 times ?
- 9 times ?
- 10 times ?

Fine… and now another question…

- How many gorillas were in the video ?

Inattentional blindness

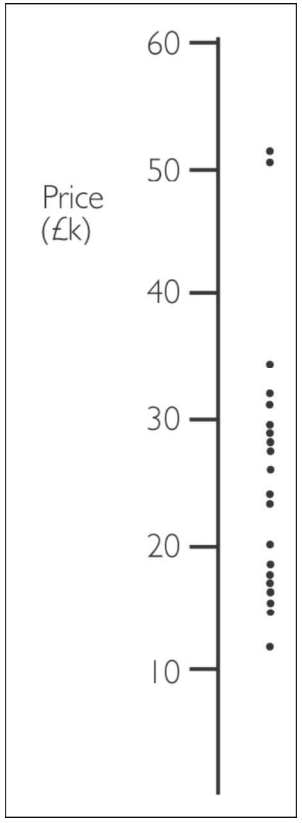


Inattentional blindness

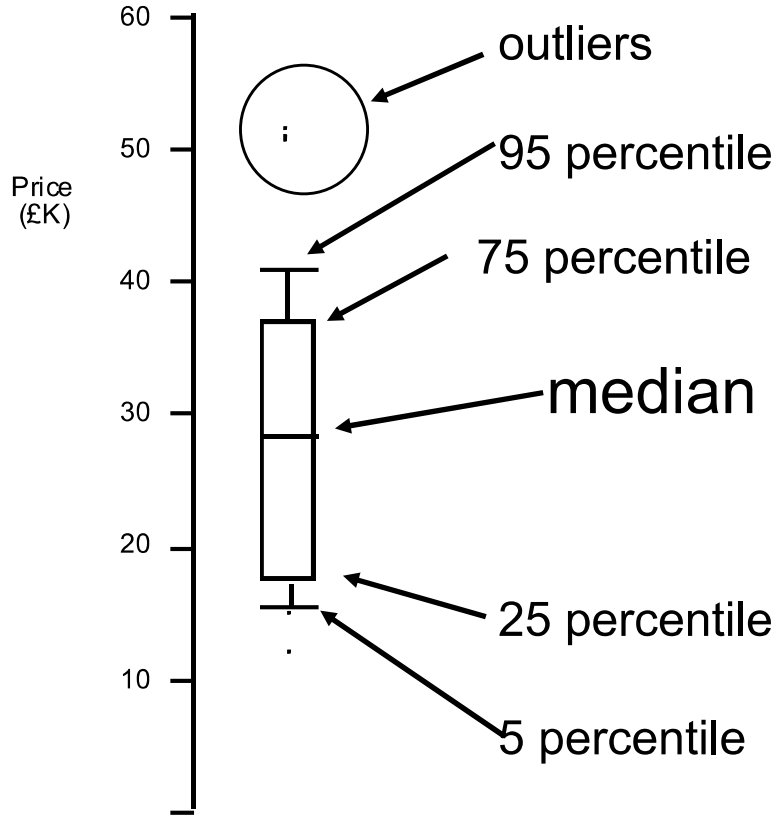
- Just one gorilla…
- Back to univariate data…

Collections of number

- Very often the interest is about a **collection** of numbers



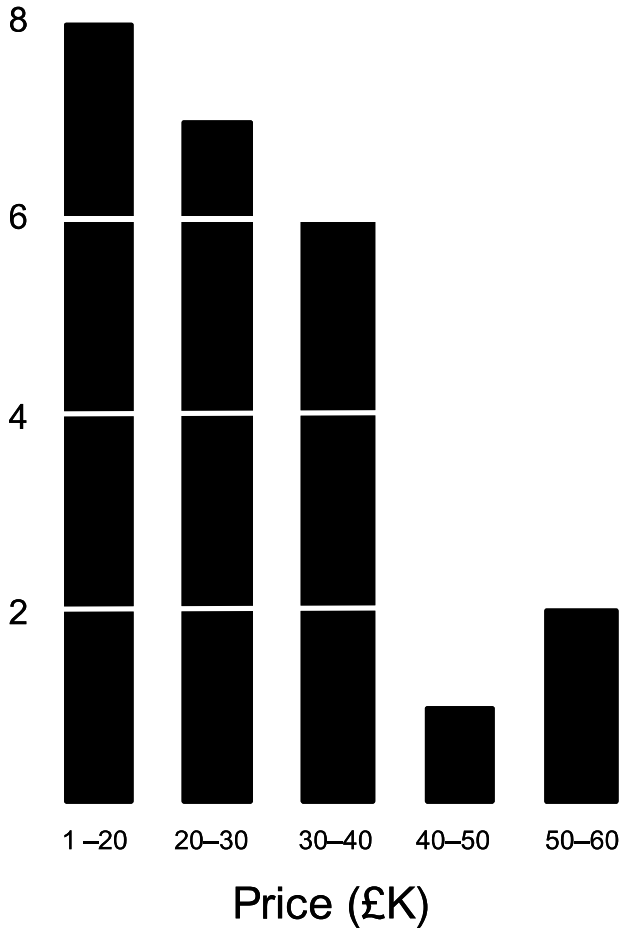
Each dot represents the price of a car

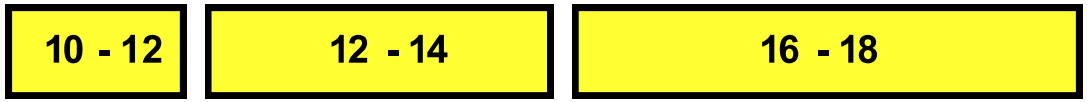


A Box Plot of the same data
Derived value!!!

Collections of number

- Histograms & bargrams (aggregate data)





A bargram representation of univariate data, obtained by 'tipping over' the columns (bars) of a histogram and joining them end-to-end, ignoring any null bins

A quick note:
1-20 20-30 is
[1-20) [20-30) ?

Collections of number

Nissan	Ford	Ferrari	MG	Cadillac

A bargram representation of univariate categorical data

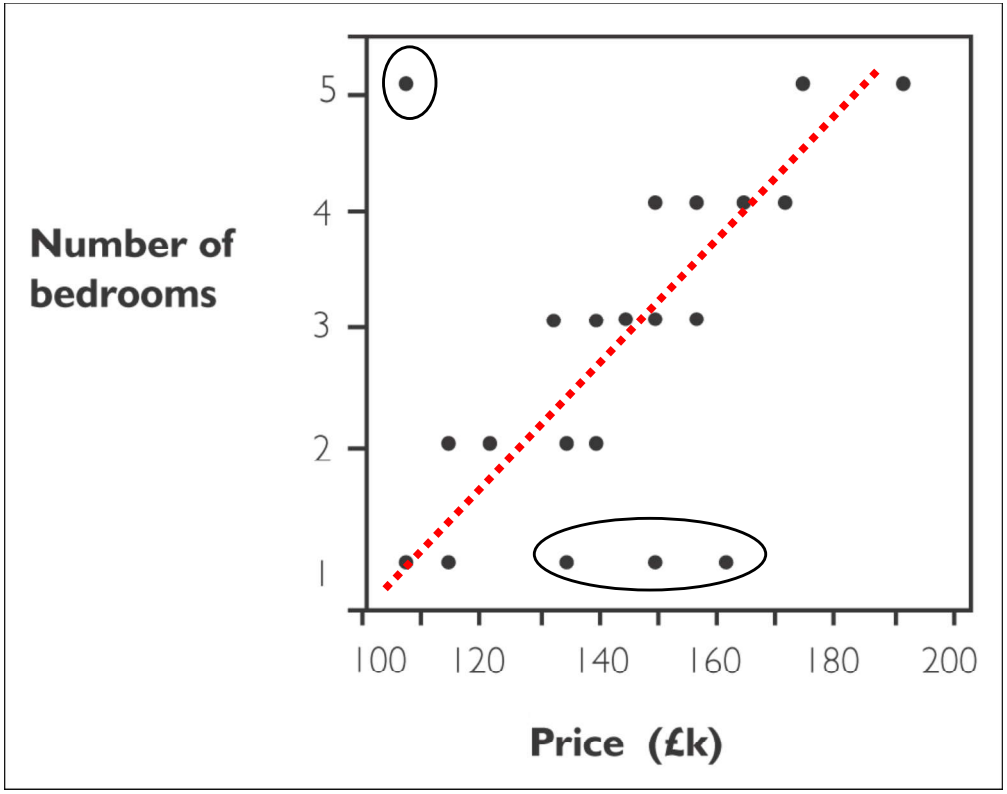
- A frequency distribution

Outline

- Data types & data complexity
- Encoding of values
  - Univariate data
  - Bivariate data
  - Trivariate data
  - Multidimensional data
- Encoding of relations
- Lines
- Map & Diagrams
- Trees
- Support for design

Bivariate data

- The conventional approach to represent data with two attributes is the scatterplot

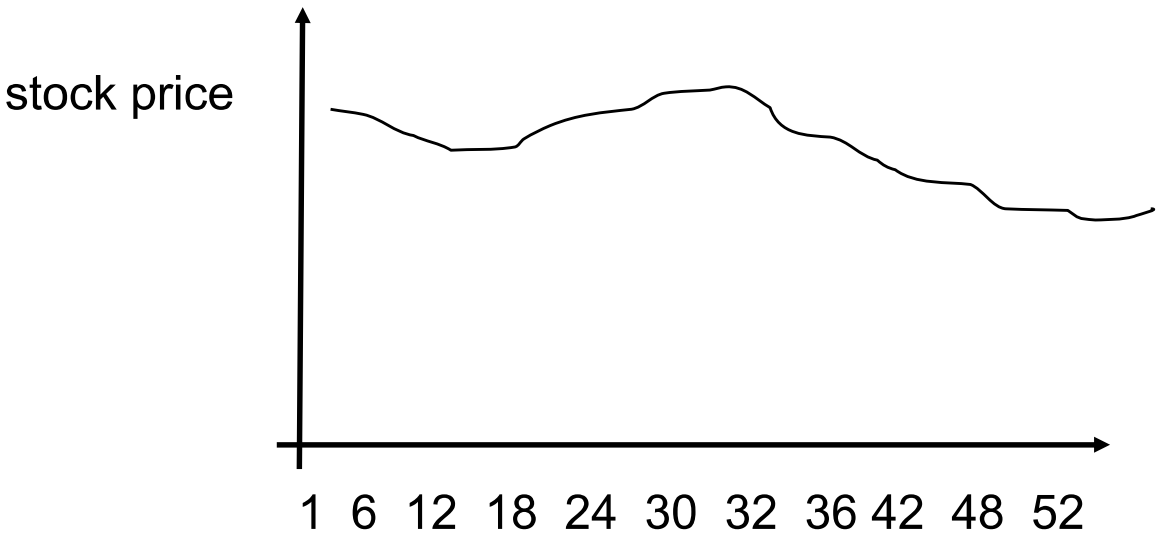


Price	n. of bedrooms
110	1
160	3
180	5
140	2
140	3
160	1
170	5
110	5
…	…

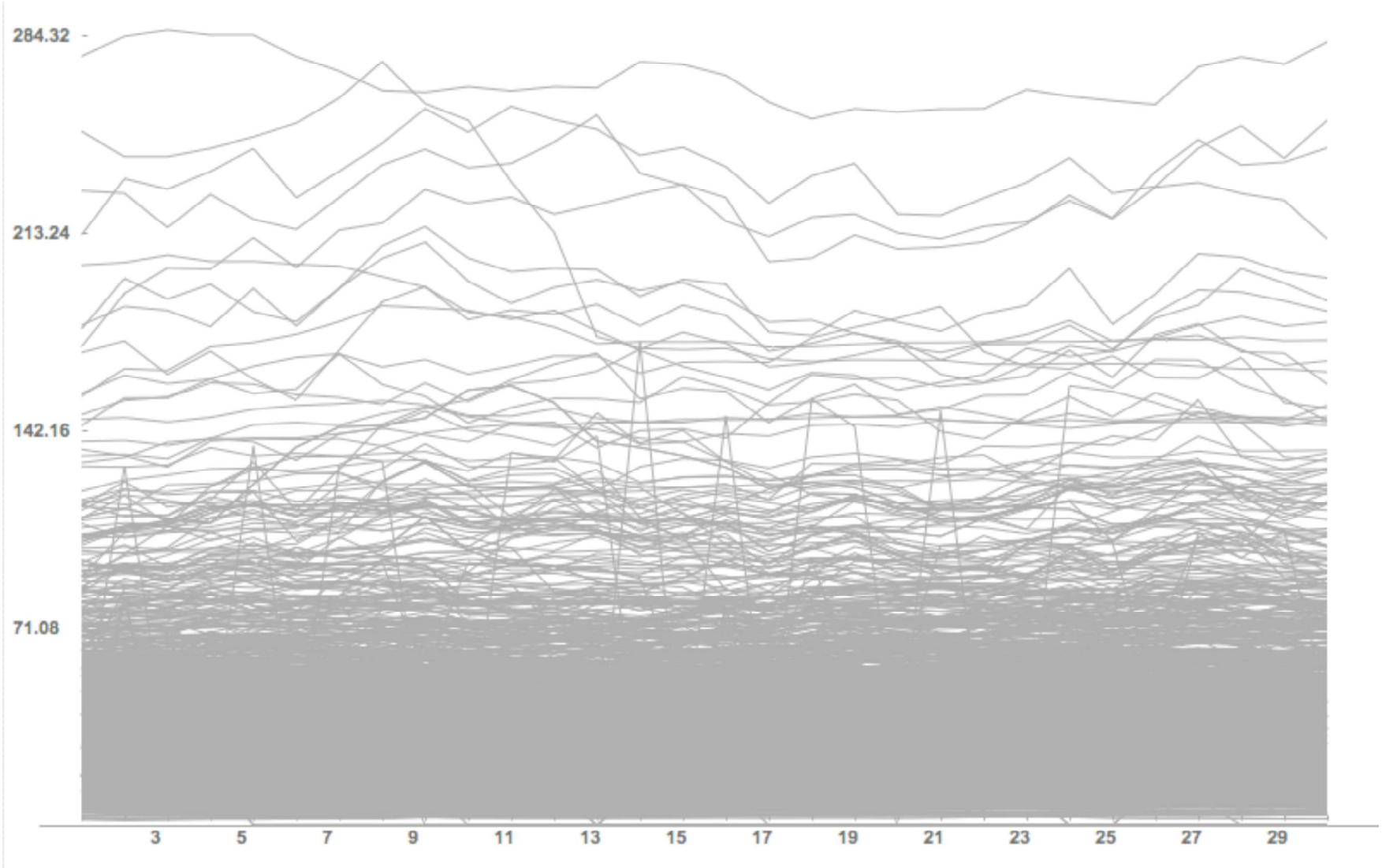
A scatterplot of bivariate data. Each point indicates the Price and Number of bedrooms associated with a house. ⬭ denote outliers

Time-series

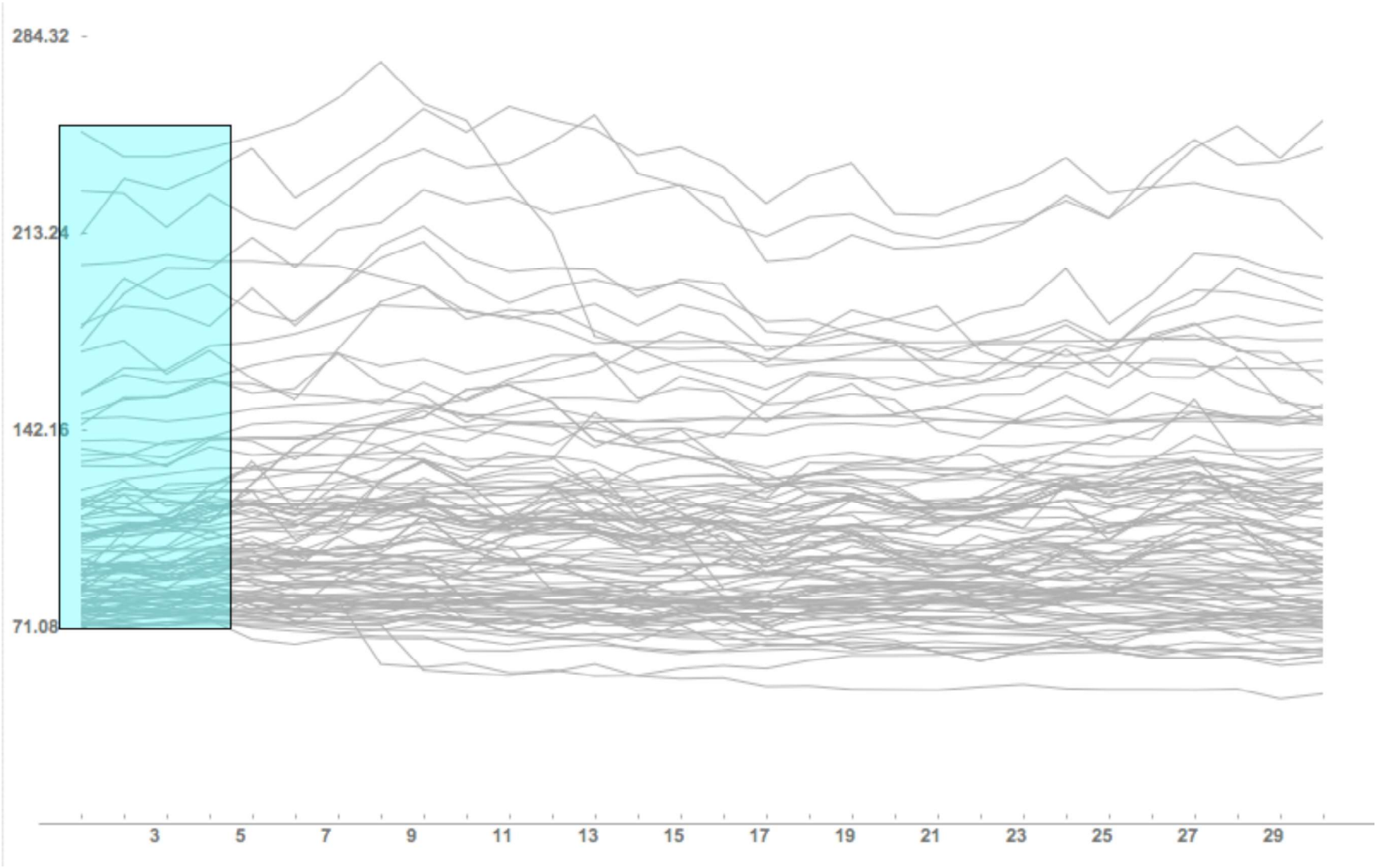
- Special solutions are available when one of the attribute is the **time**
- Typical application are medical, climate studies, and market analysis



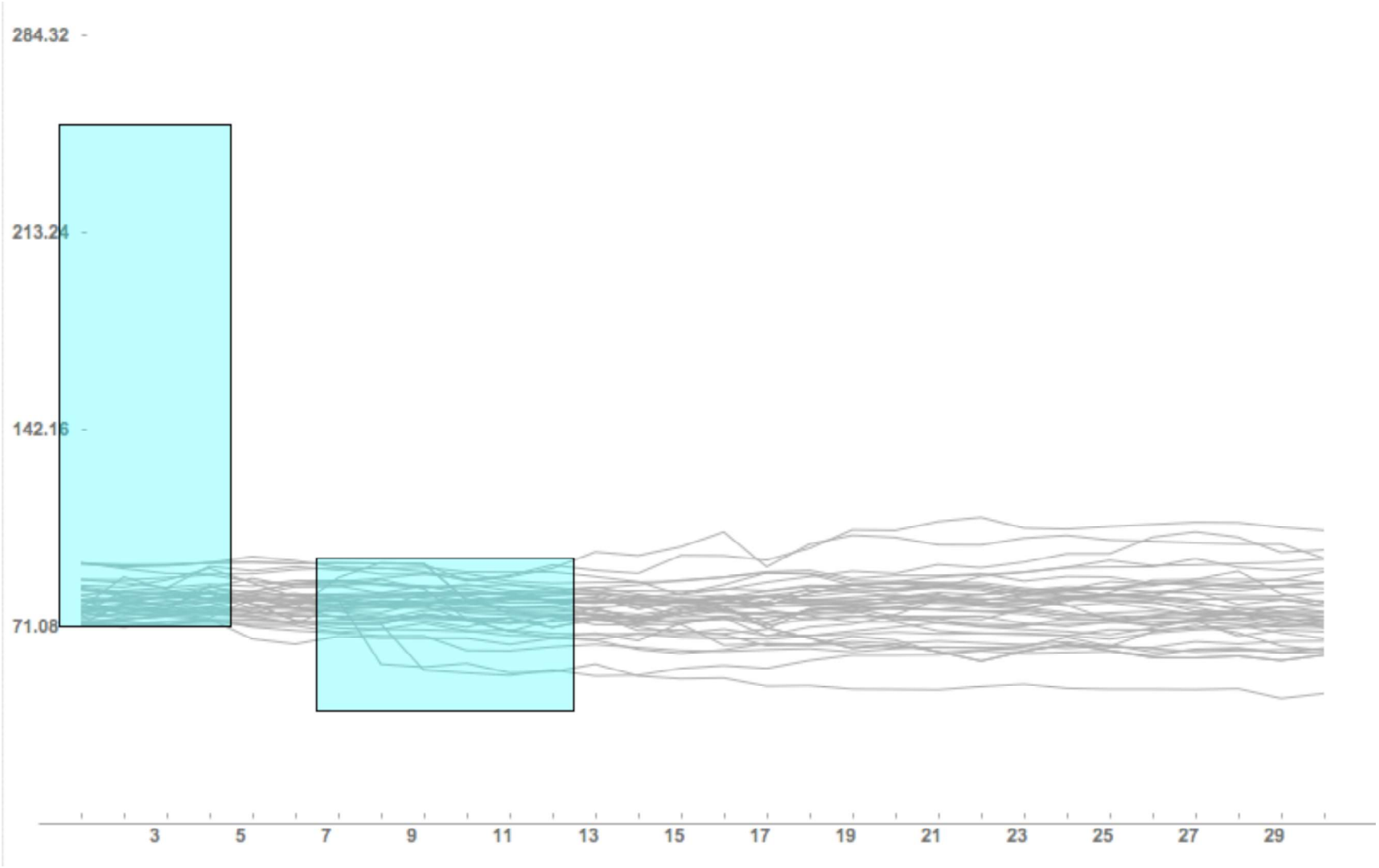
Example: weekly stock prices for 1430 stocks in a year



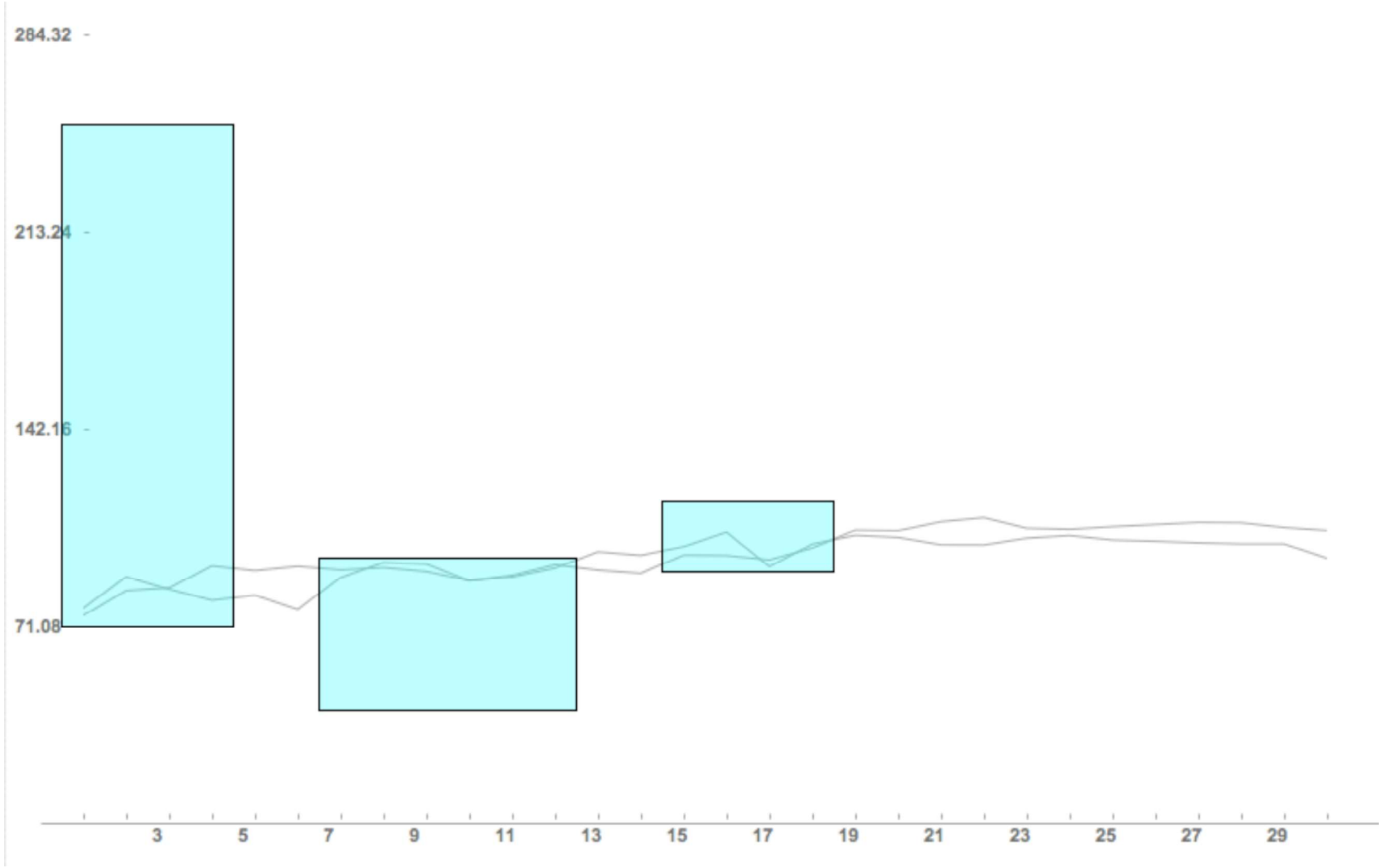
An overview of the entire data set



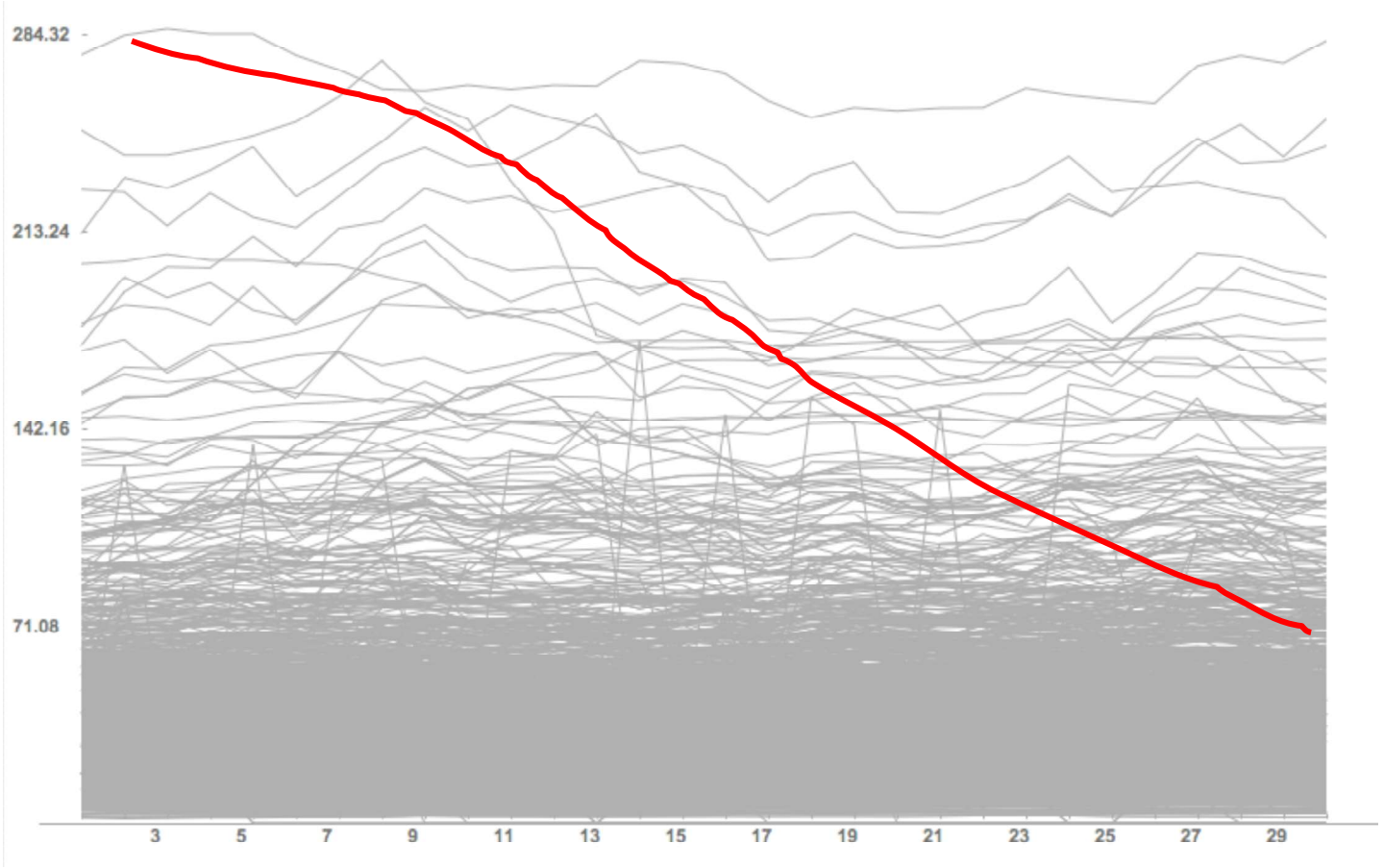
A single time-box limits the display to items with prices between $70 an $250 in the first weeks of the year



A additional constraint selects items with prices between $70 and $95 during weeks 7 to 12

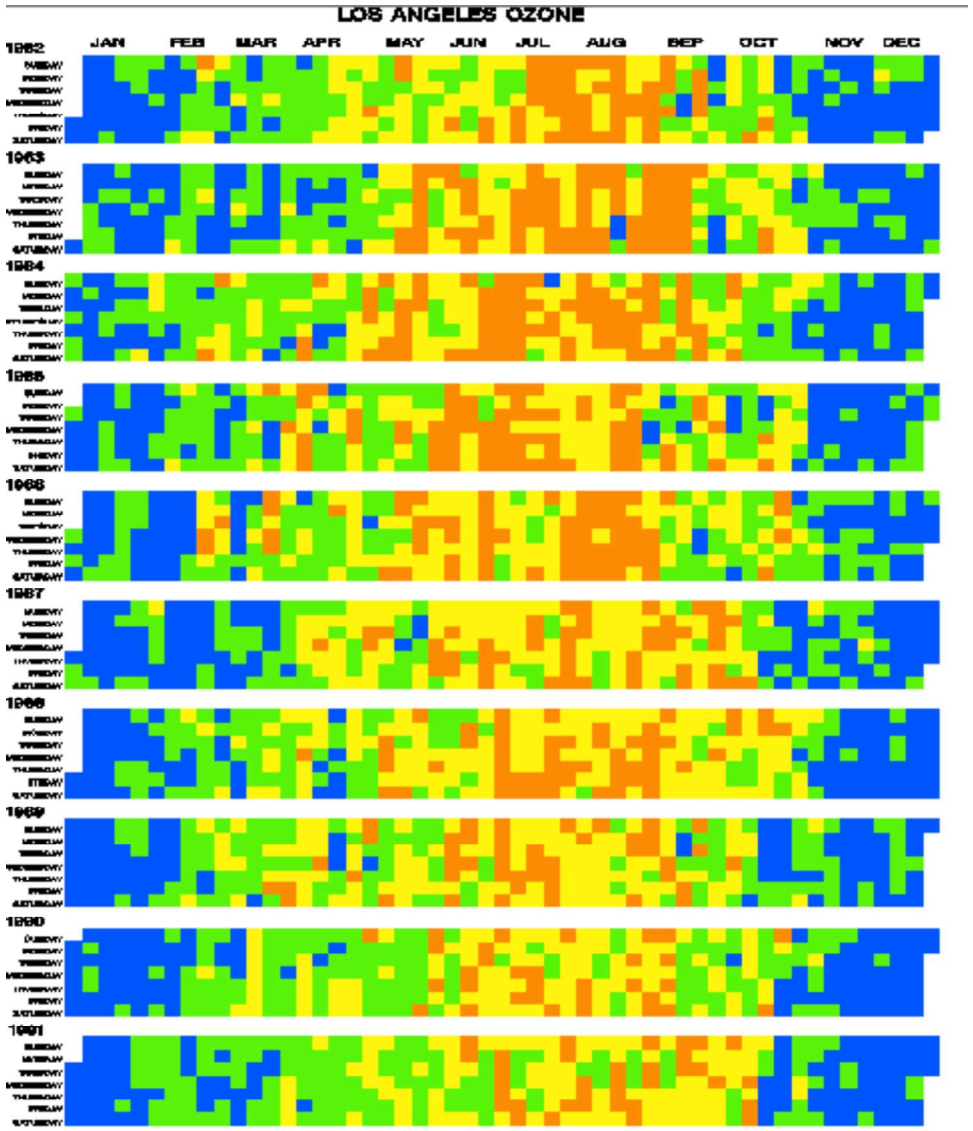


A constraint concerns prices between $90 and $115 for weeks 15 to 18



Or: give me stocks whose trend is “similar to that line”

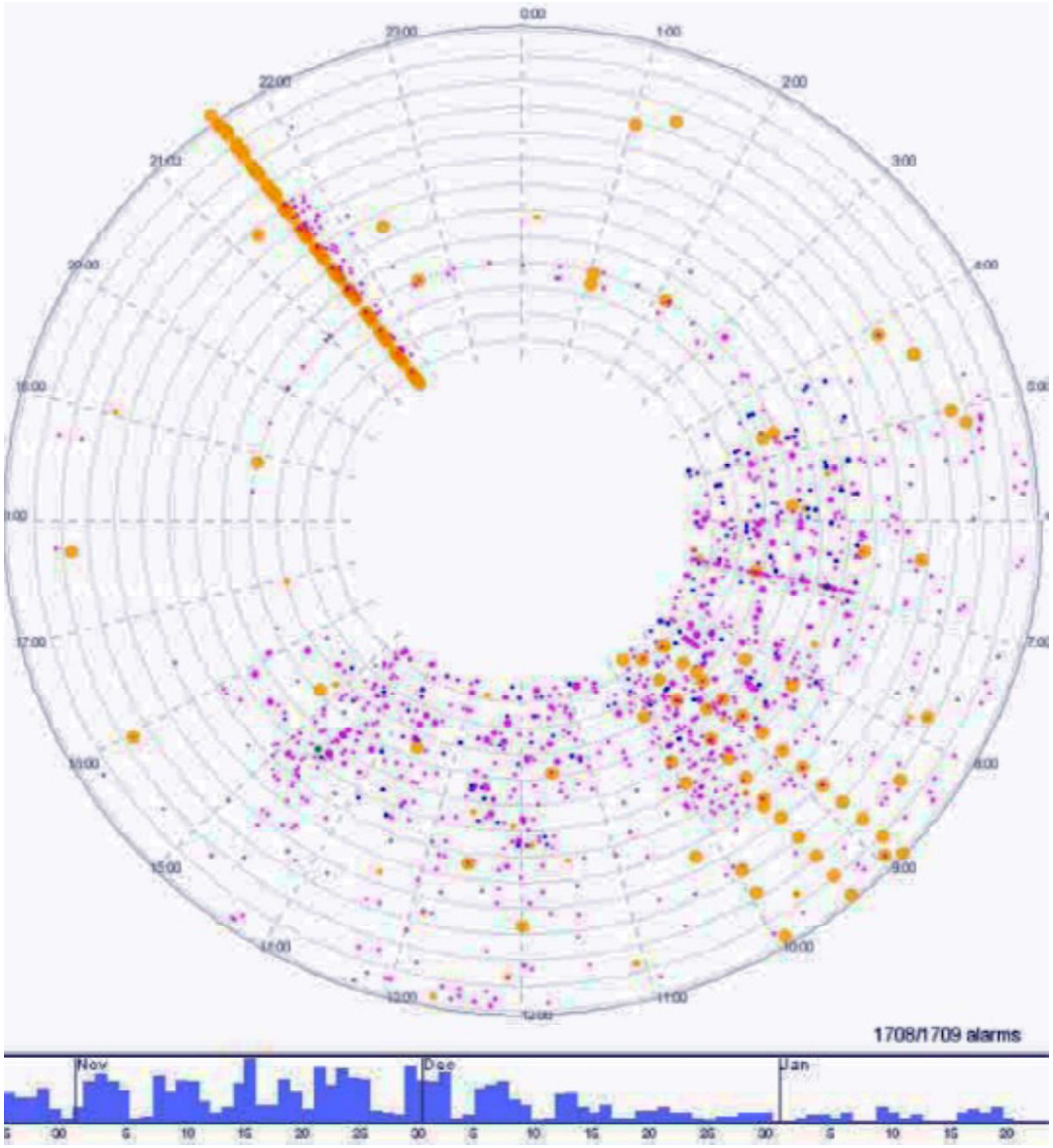
Time-series



- When cyclic aspect of the time are relevant different visualizations may make evident repetitive patterns

Example:
Ozone level in Los Angles over 10 years

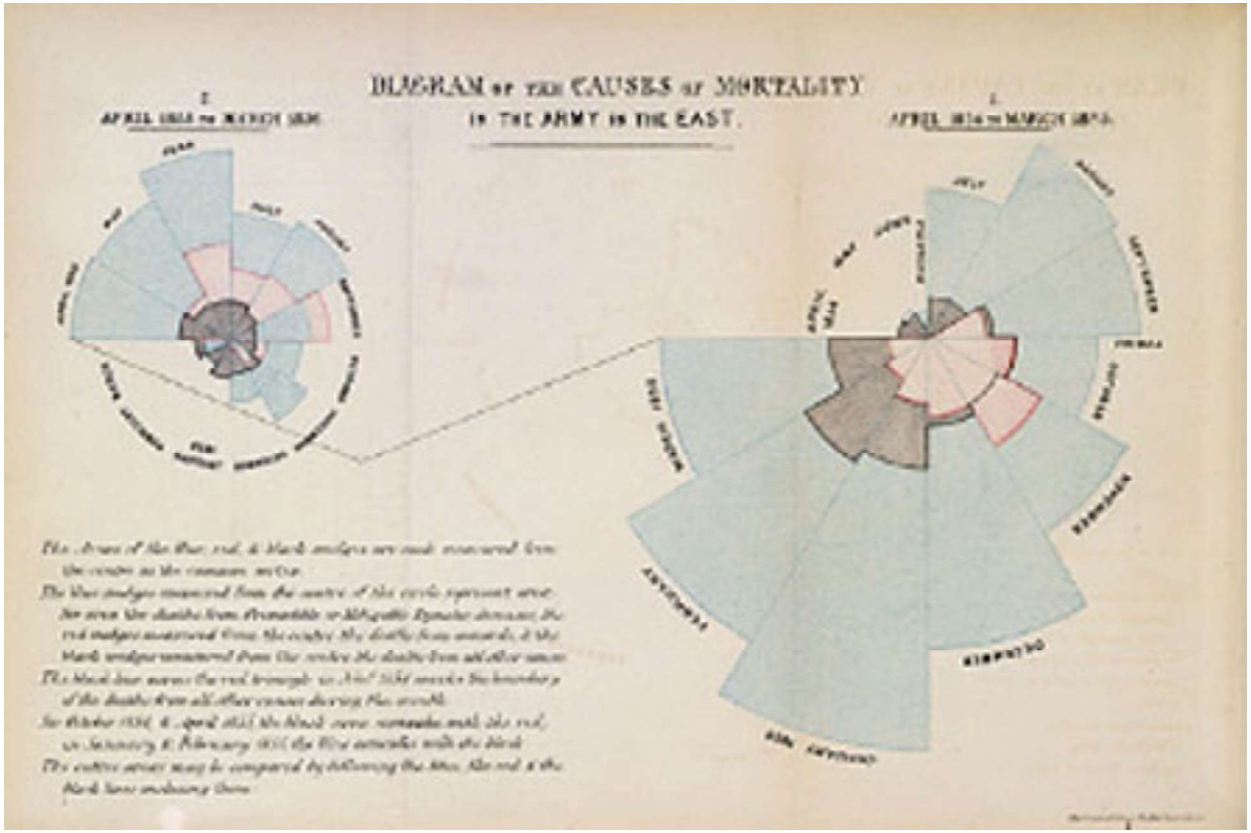
Time-series

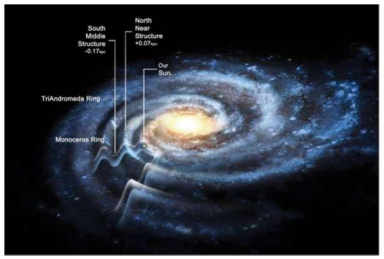


- Or in a spiral /circle fashion

Example: Alarms in a network

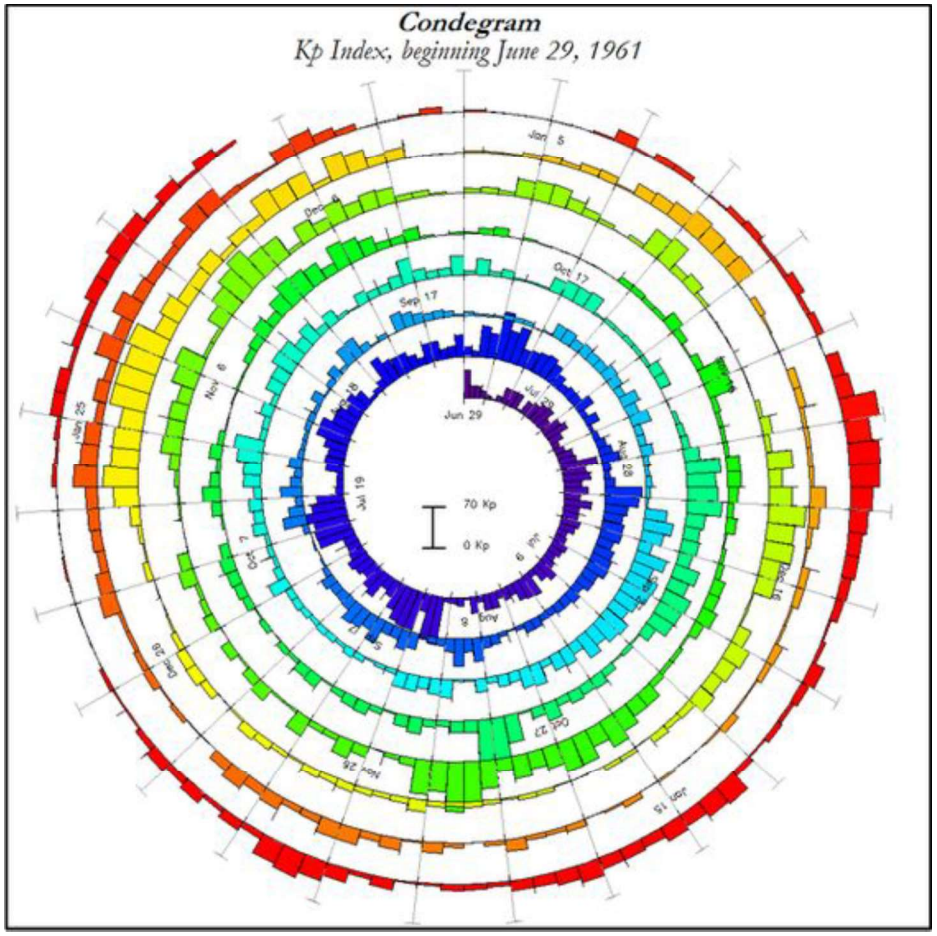
Do you remember Florence Nightingale?





Circular/radial encoding

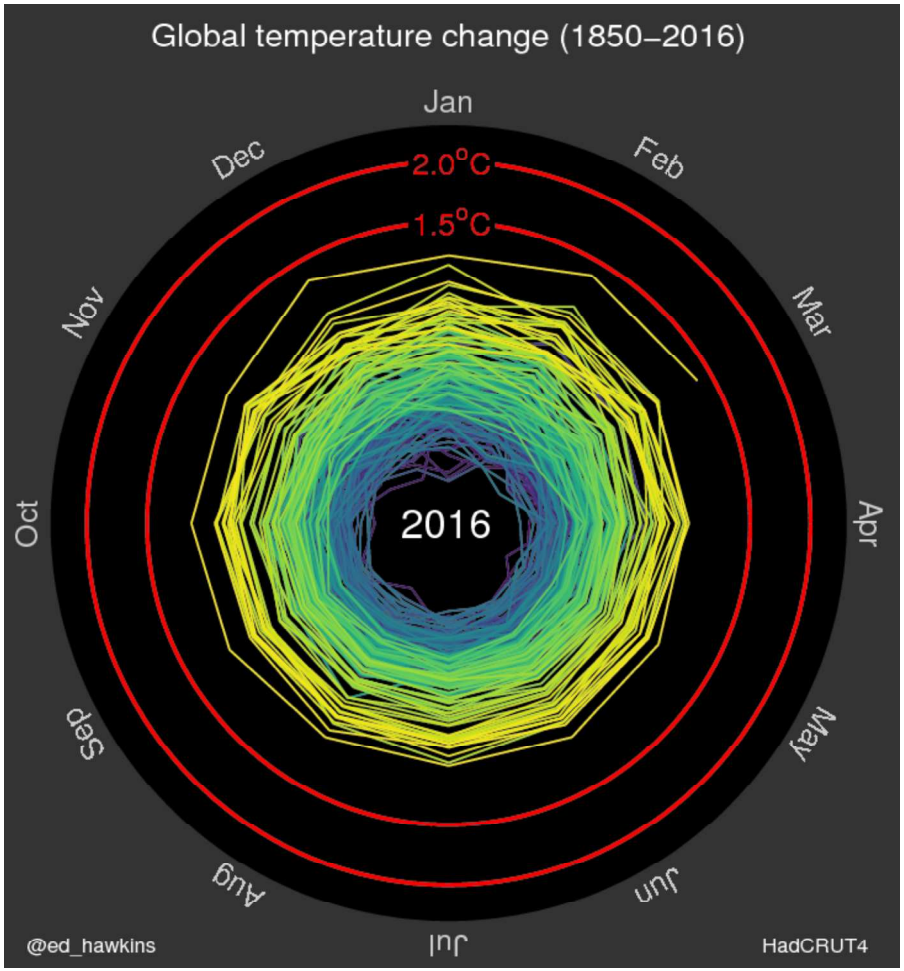
(see, e.g., http://circos.ca/intro/published_images/)



Not focused on repetitive patterns

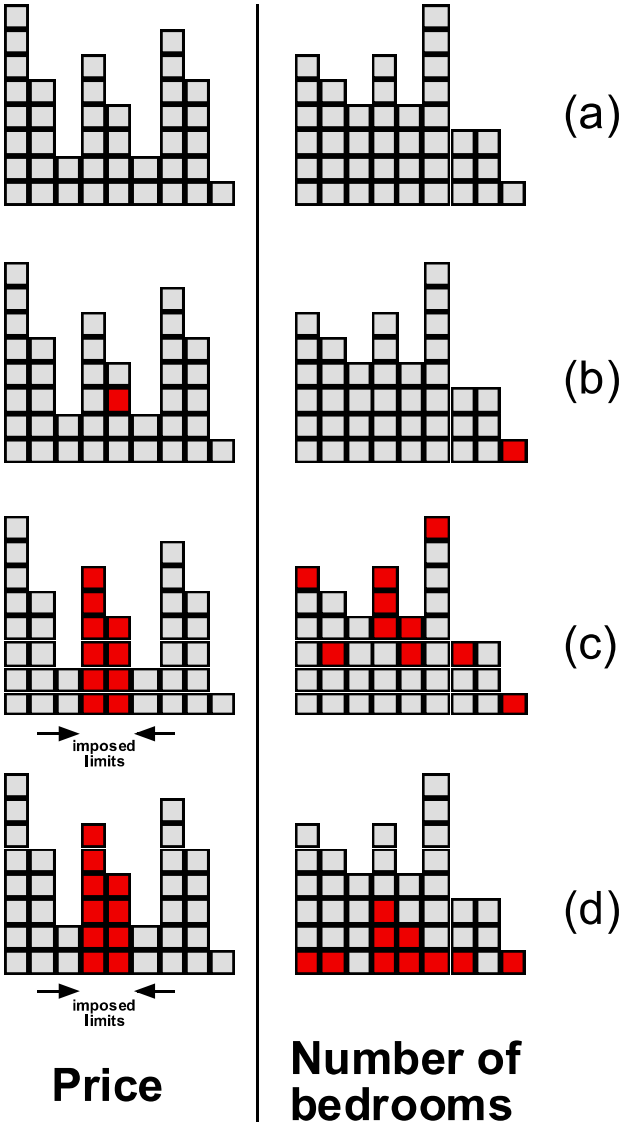
More room for the data

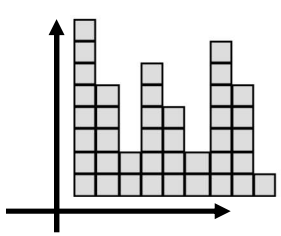
Repetitive temporal patterns

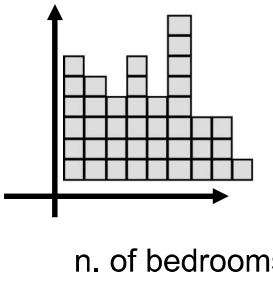


What is wrong with it?

Linked histograms & **brushing**







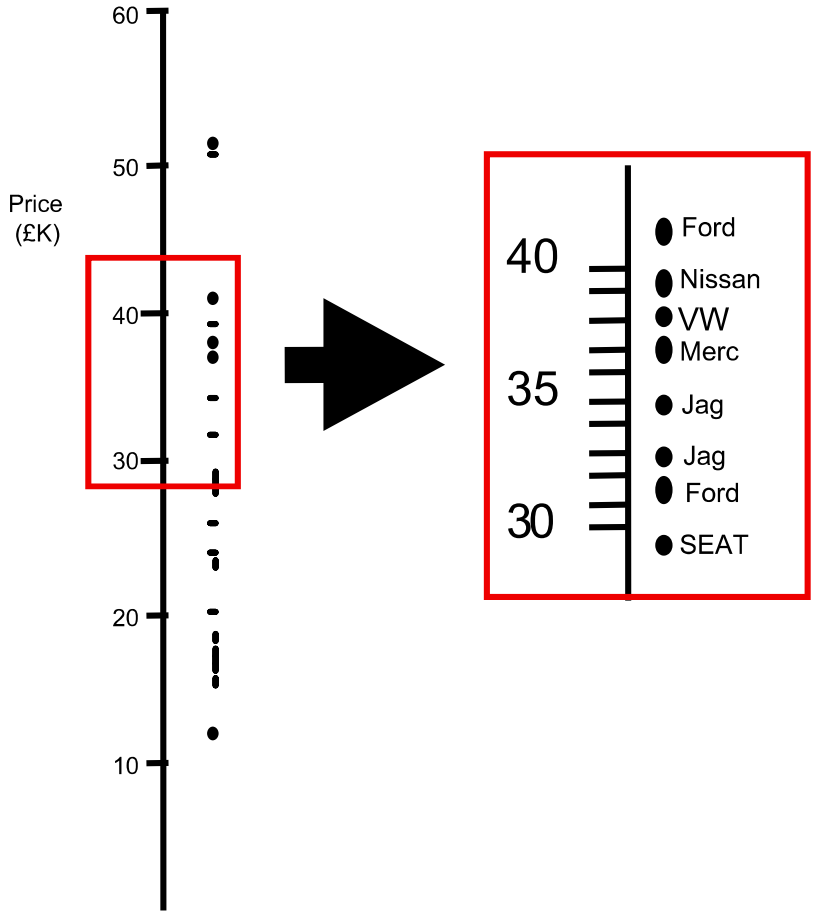
(a) the price and number of bedrooms associated with a collection of houses are represented by separate histograms;

(b) a single house is represented once on each histogram;

(c) upper and lower limits placed on price define a subset of houses which are coded red on *both* histograms;

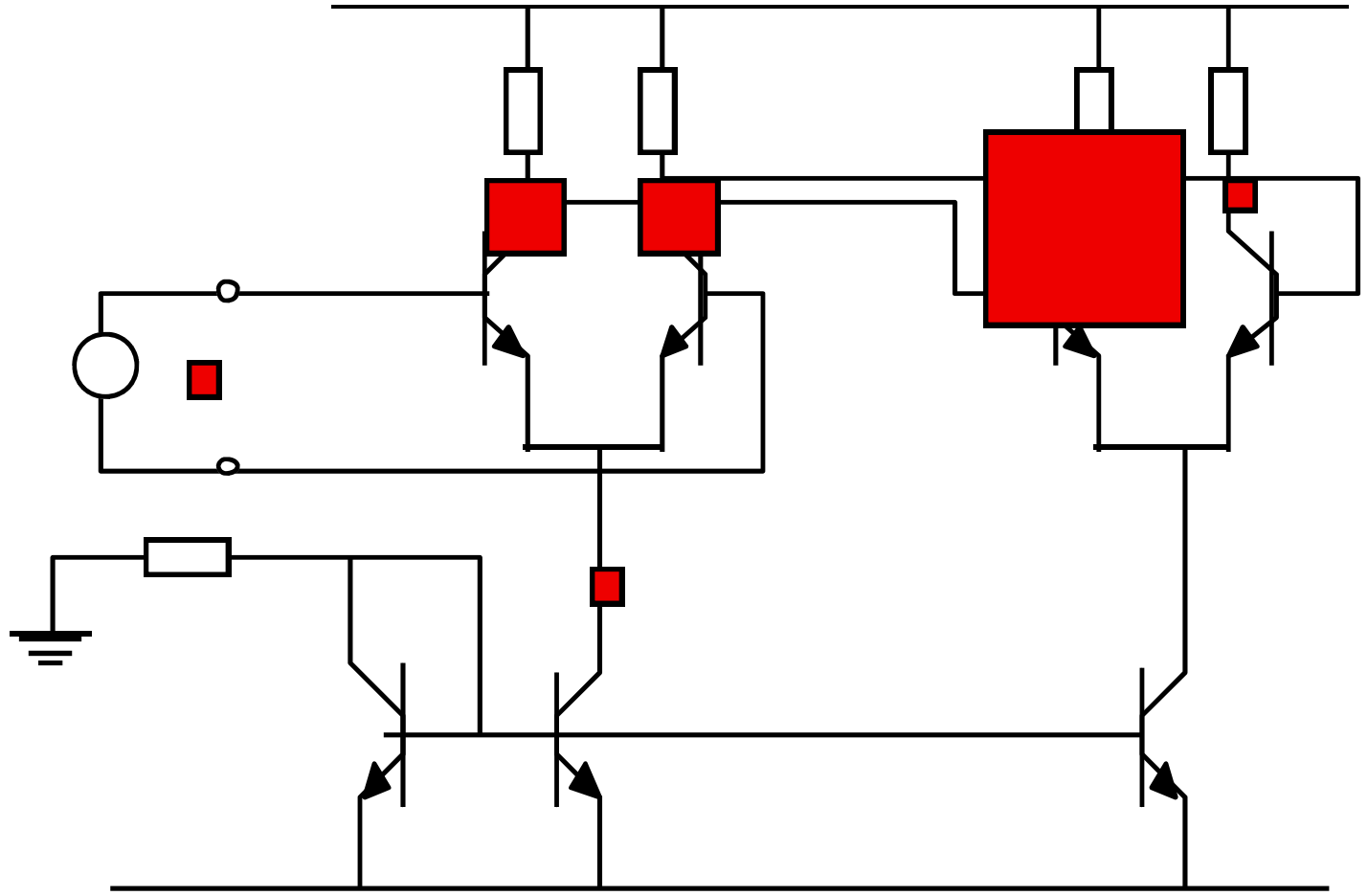
(d) Interpretation is enhanced by 'ranging down' the colour-coded houses, especially if exploration involves the dynamic alteration of limits

Semantic zoom on a collection

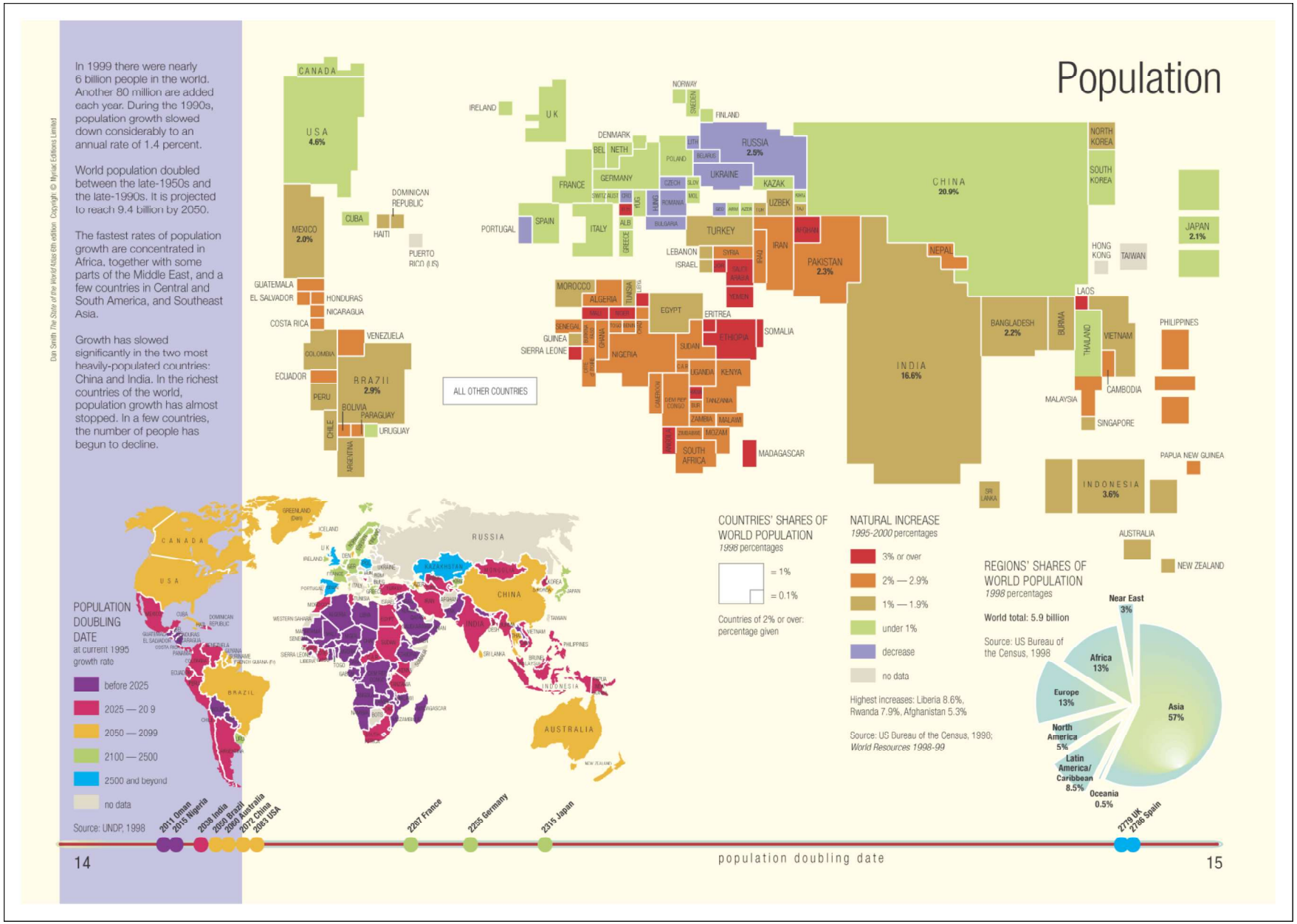


Semantic zoom reveals data about a second attribute

Qualitative understanding



The area of each red square encodes the value of the voltage occurring at the point in the circuit at which the square is located



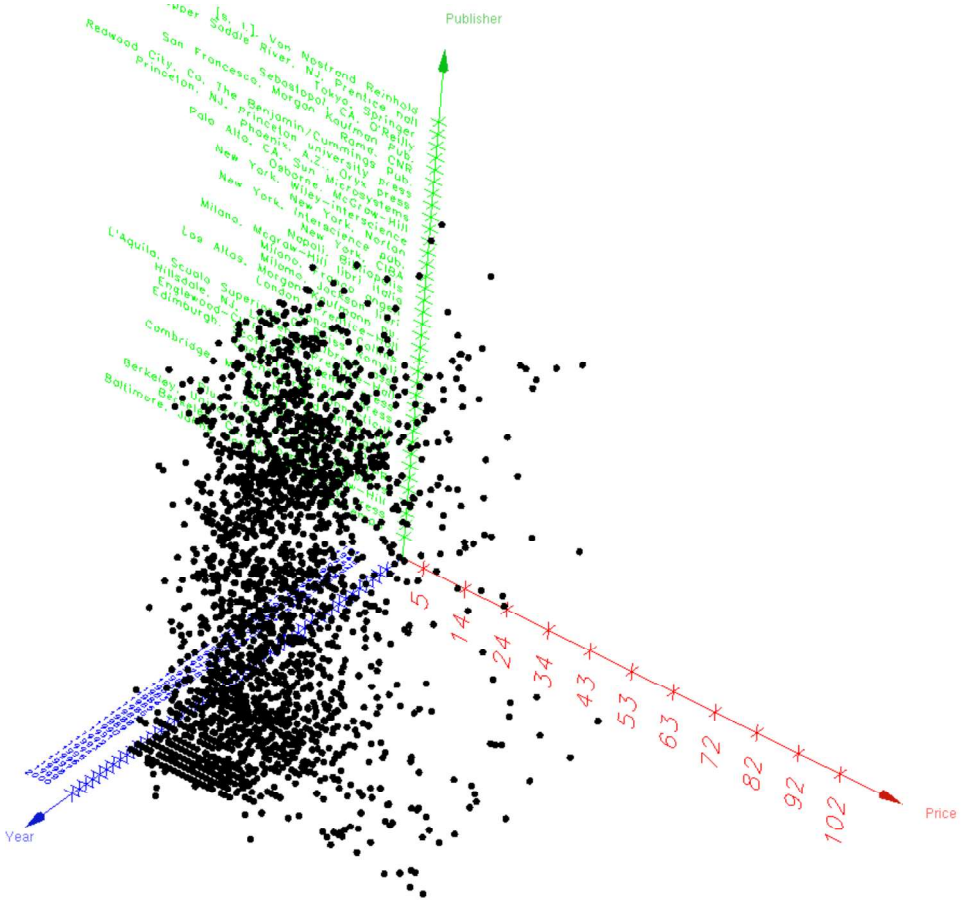
In the *State of the World Atlas*, magnification encoding is used to give a first impression of population densities. Note the reduced ‘size’ of Canada and Australia when compared with a conventional map

Outline

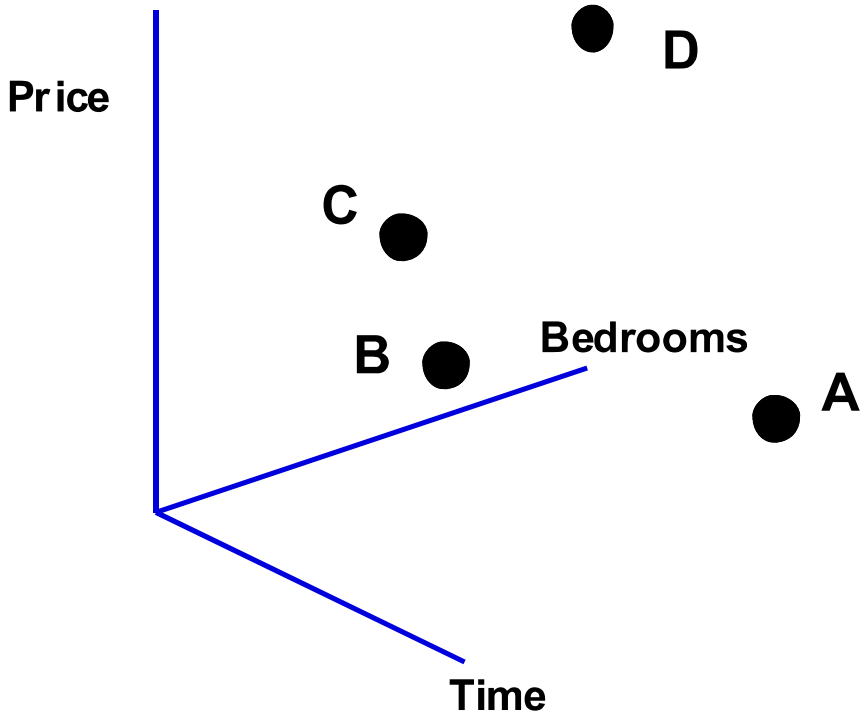
- Data types & data complexity
- Encoding of values
  - Univariate data
  - Bivariate data
  - Trivariate data
  - Multidimensional data
- Encoding of relations
- Lines
- Map & Diagrams
- Trees
- Support for design

Trivariate data

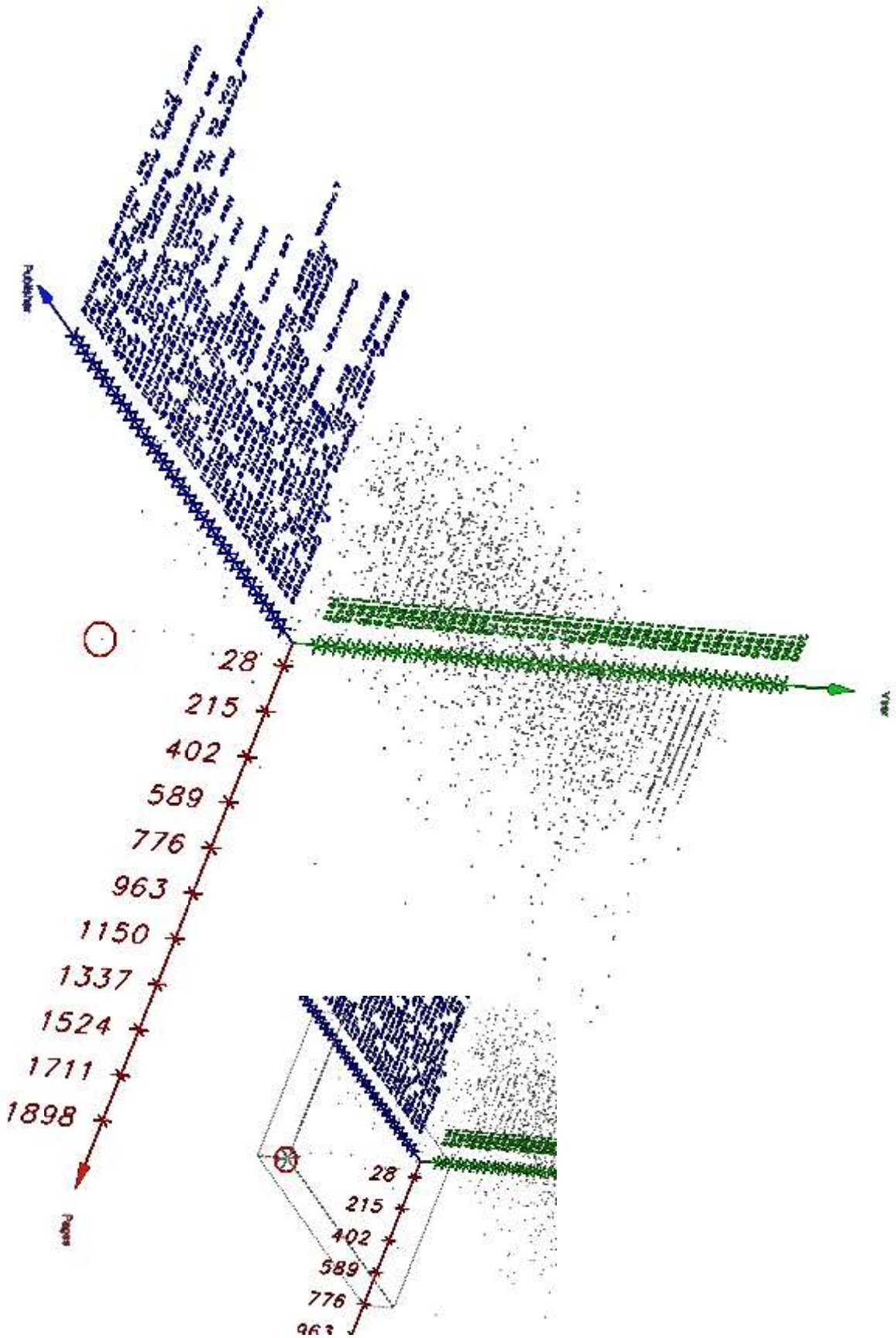
- Extension of 2-dimensional scatterplot to 3-dimensional is straightforward but not very effective



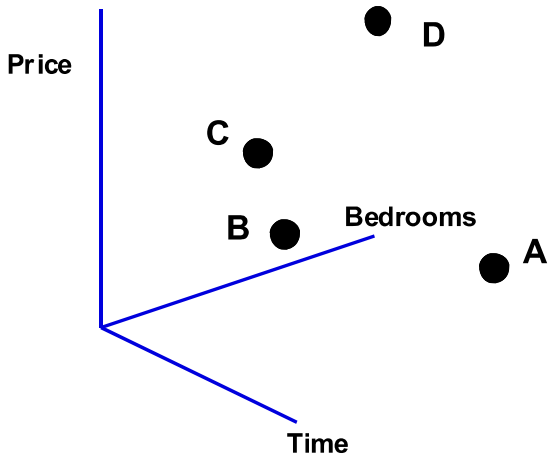
Occlusion

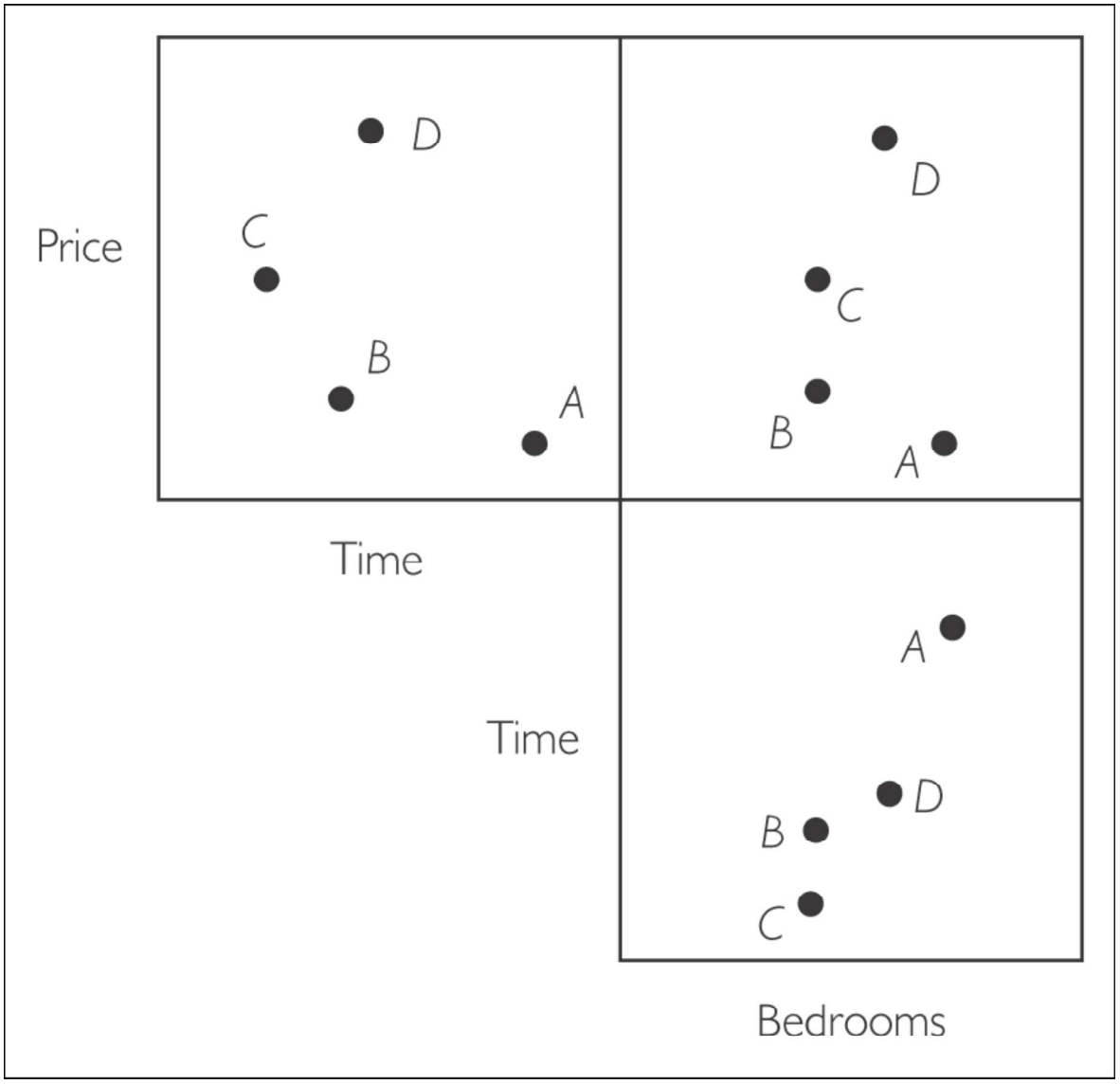


Does A cost less or more than B?


28
215
402
589
776
963
1150
1337
1524
1711
1898
28
215
402
589
776

Scatterplot matrix

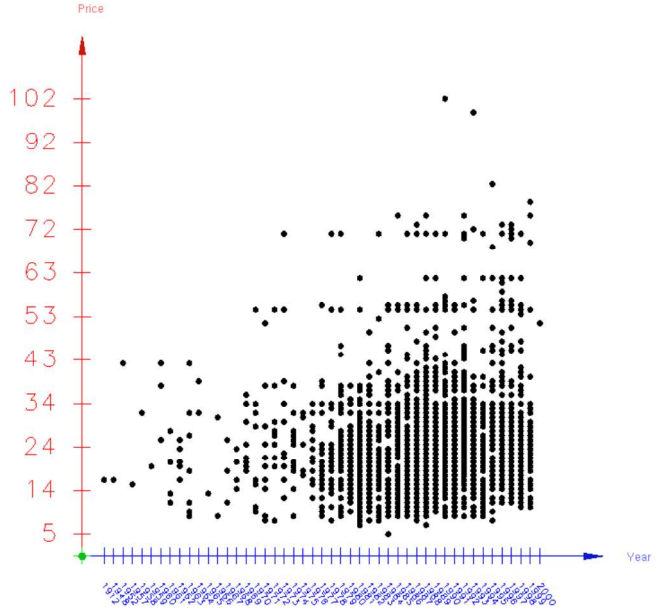


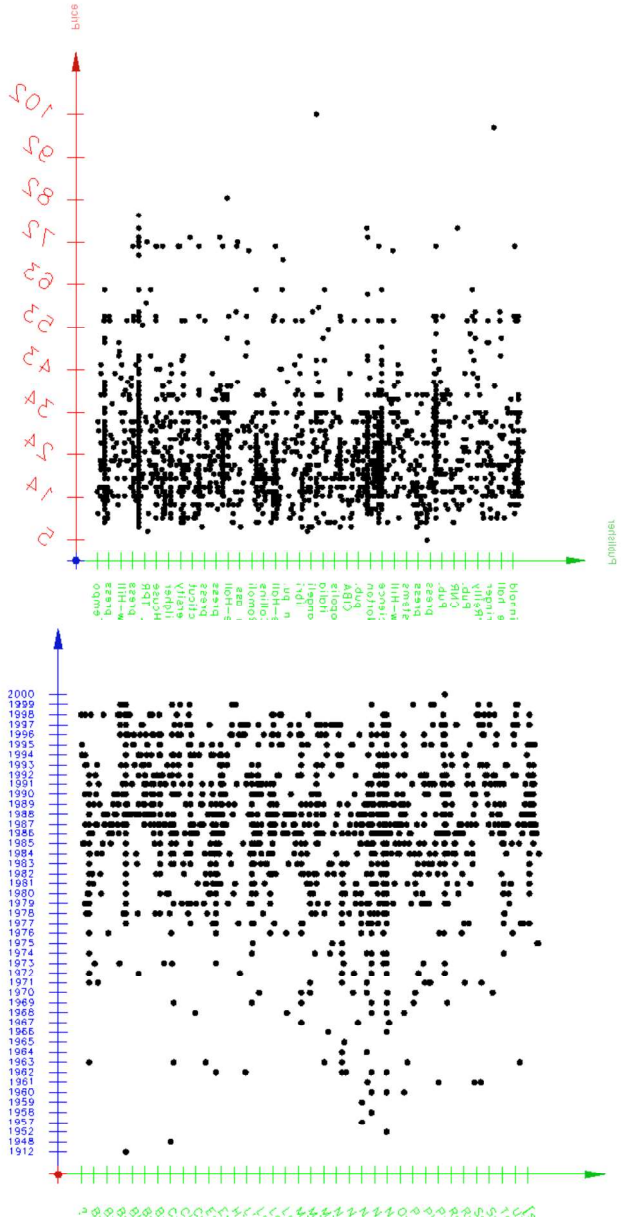


Scatterplot matrix & brushing

The highlighting of houses in one plane is brushed into the remaining planes

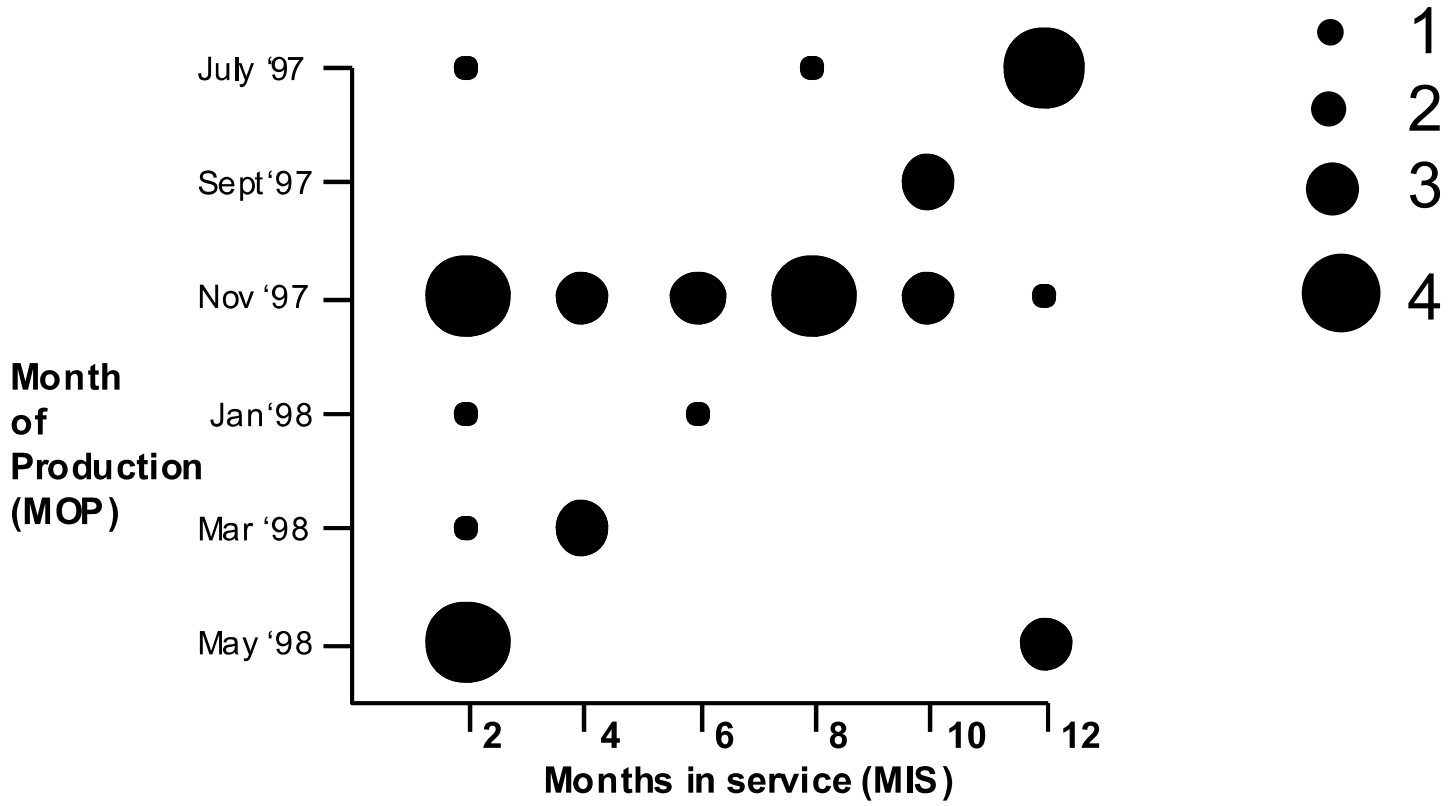
Scatterplot matrix





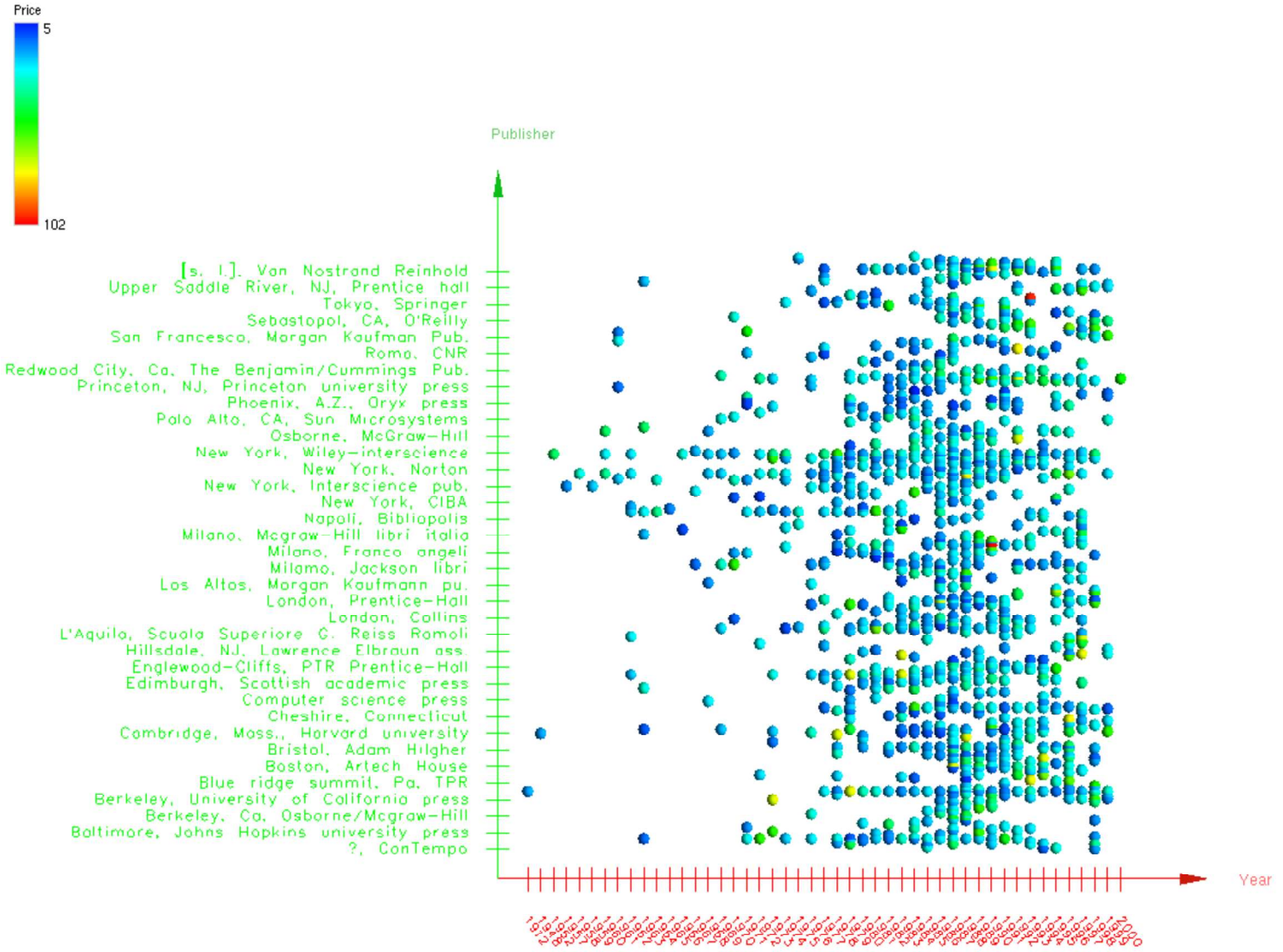
No brushing implemented ☹

Scatterplot + dimension

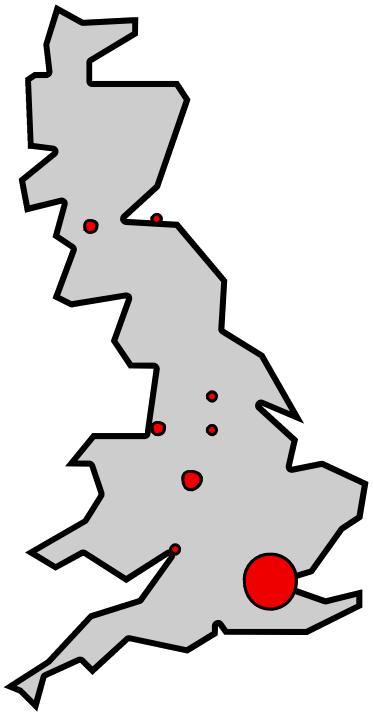


A representation of reported product failure, based on month of production (MOP) of the failed product, and total months in service (MIS) before the fault occurred. The radius of each circle indicates the number of faults reported for a given MOP and MIS (0 failures are not represented...)

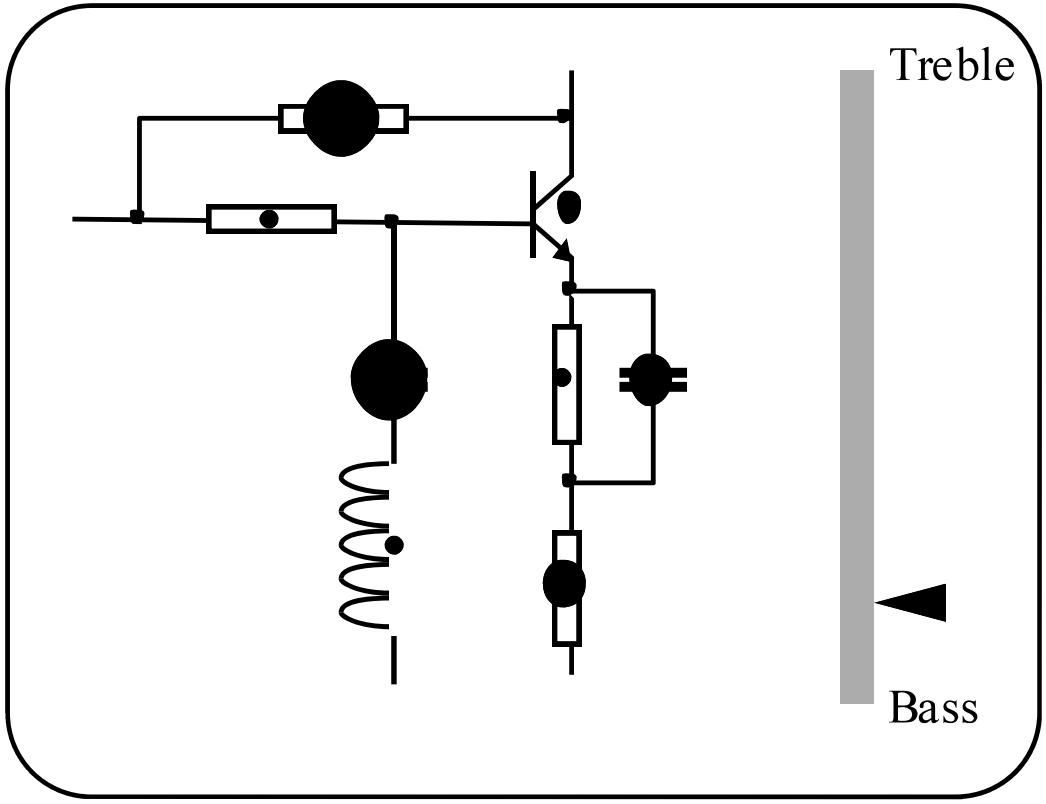
Scatterplot + color



2D drawing + visual attribute



A representation of the population of major cities in England, Wales and Scotland. Circle area is proportional to population



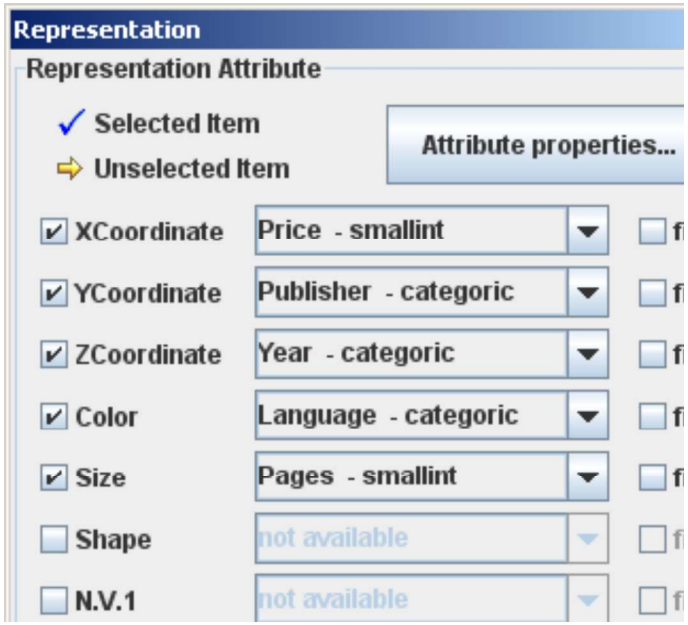
Circles indicate the extent of the effect of a component on some property of the circuit, and change in size as the frequency cycles up and down the range from bass to treble

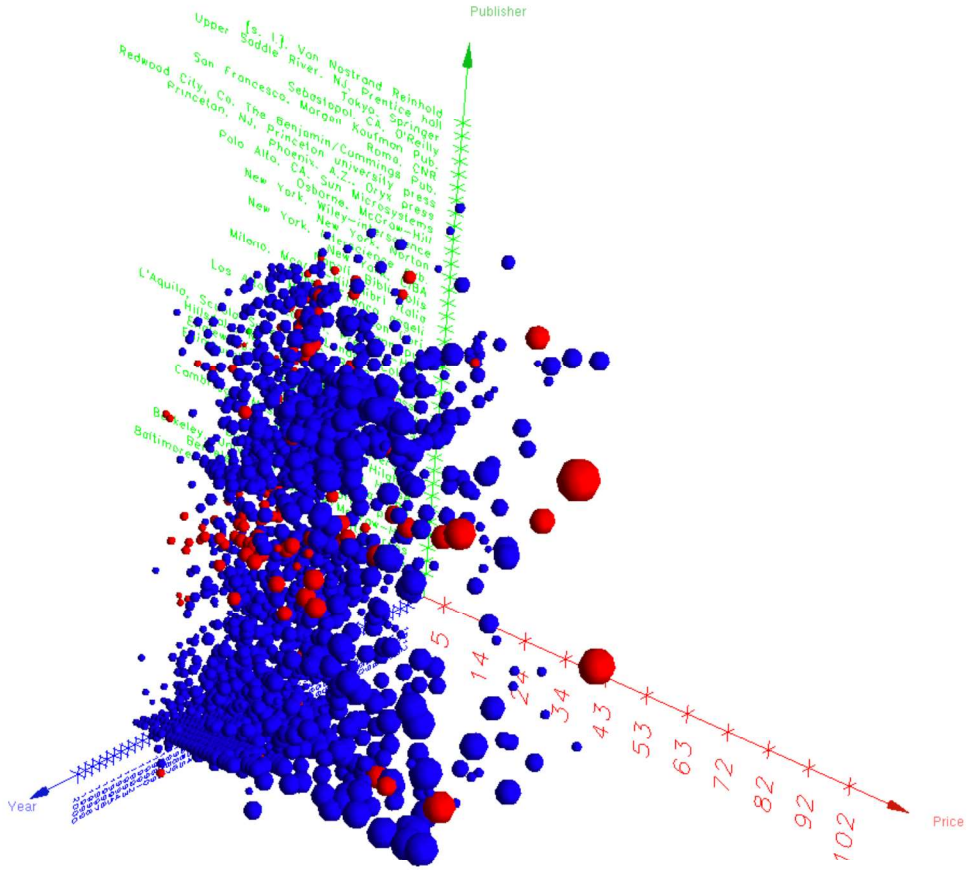
Outline

- Data types & data complexity
- Encoding of values
  - Univariate data
  - Bivariate data
  - Trivariate data
  - Multidimensional data
- Encoding of relations
- Lines
- Map & Diagrams
- Trees
- Support for design

Multidimensional data

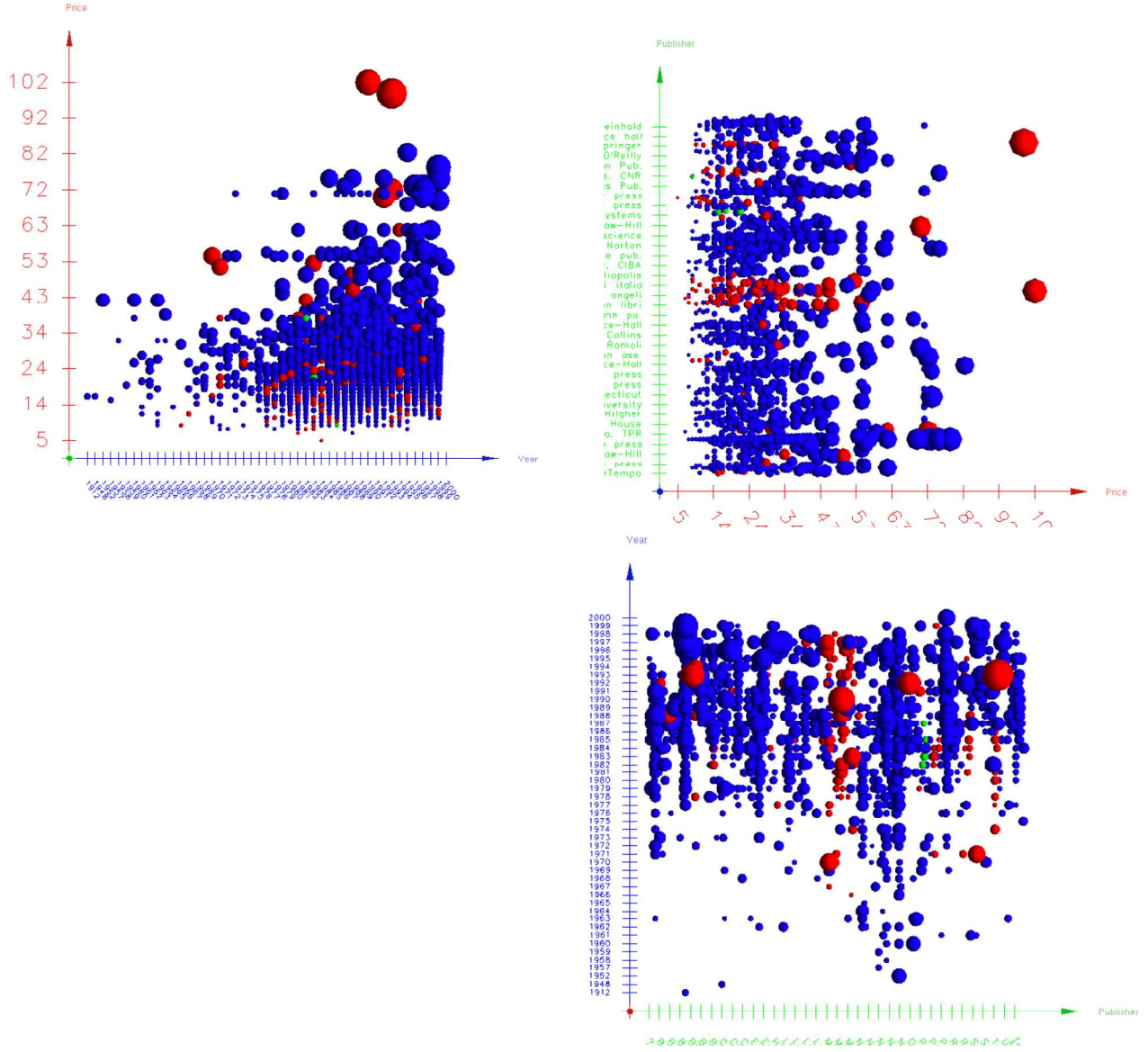
- Data attributes > 3
- We need to represent data attributes with other attributes than X,Y, Z



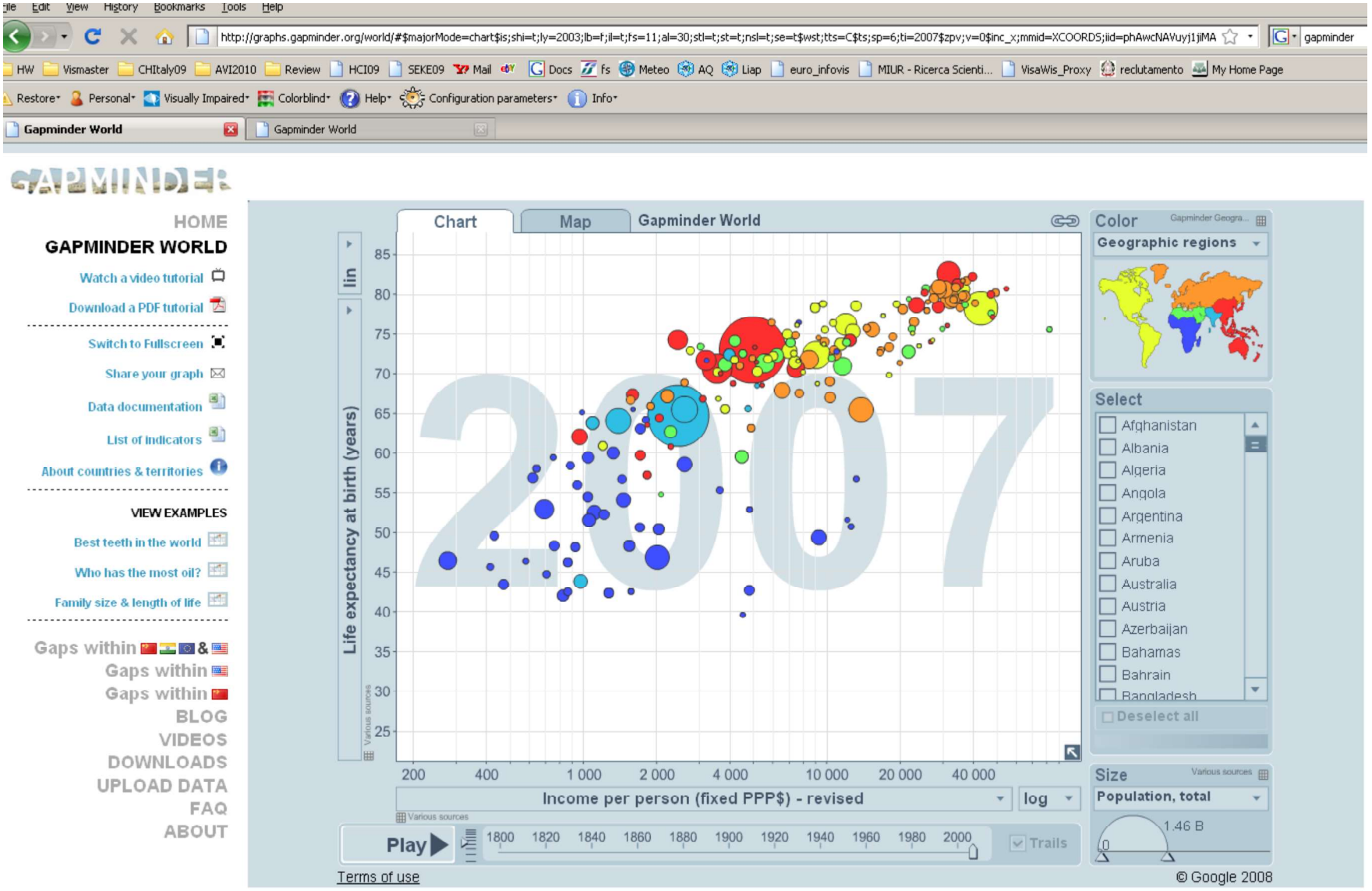


3D scatterplot+size+color

Scatterplot matrix + size+color



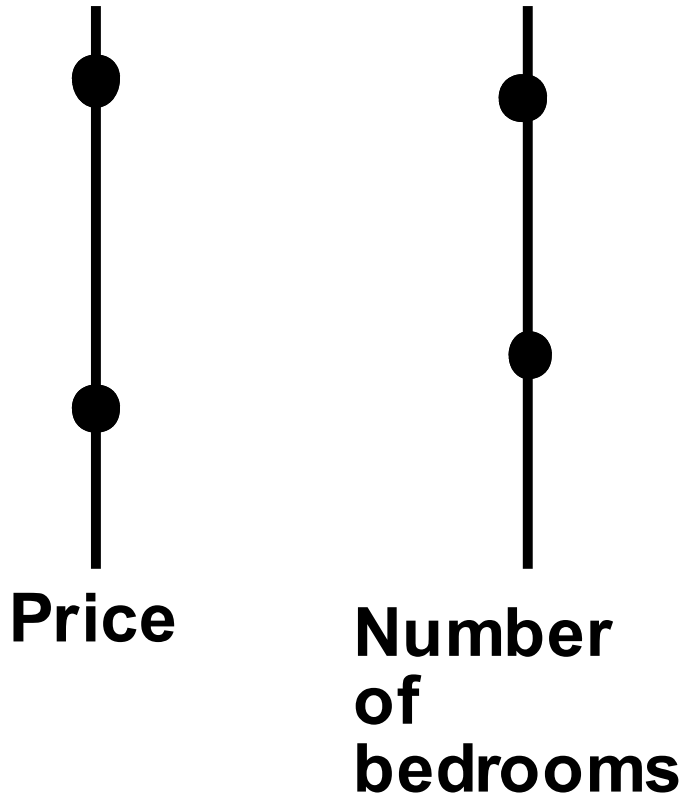
Google & Gapminder



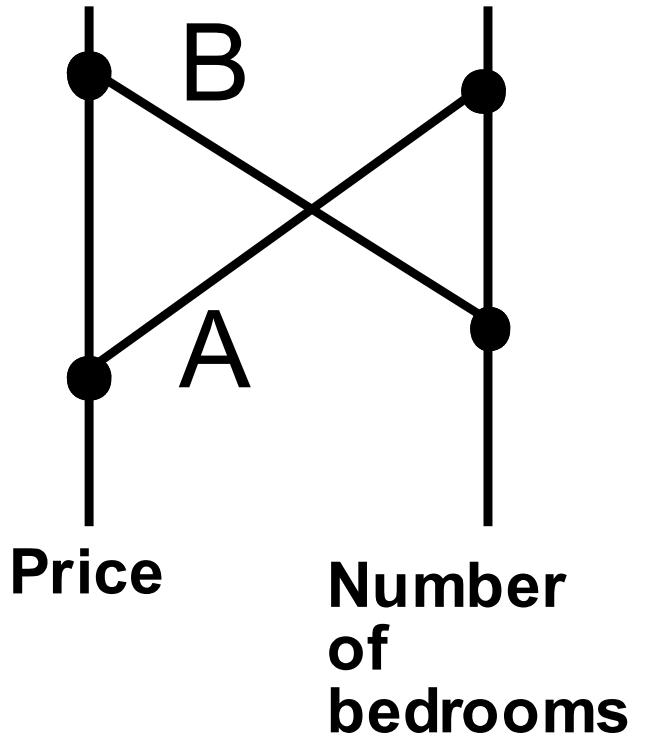
How far can we go?

- X,Y, Z+
- Color+
- Size +
- Shape +
- Pattern +
- Orientation +
- …
- It is clear that we cannot manage in efficient way more than 7, 8 attributes
- We need different approaches

Parallel coordinates

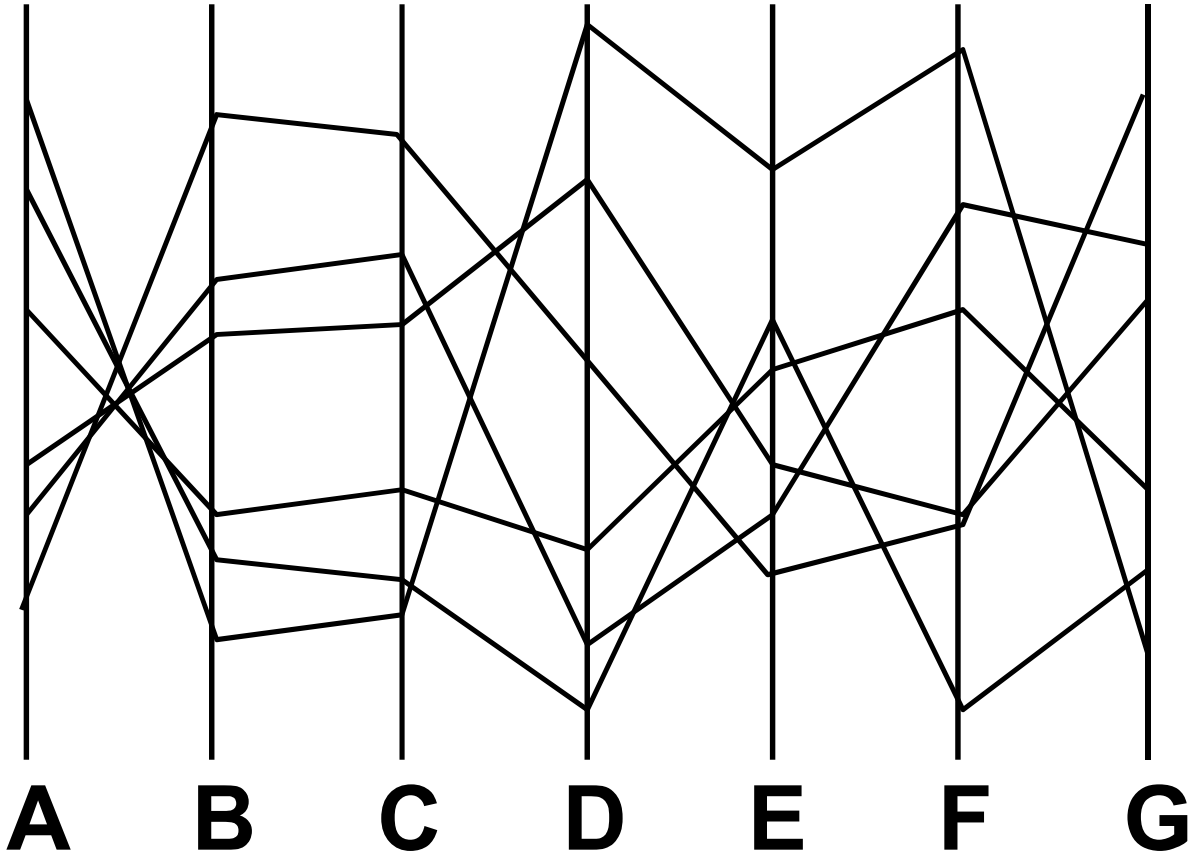


An alternative representation to the scatterplot in which the two attribute scales are presented in parallel, thereby requiring two points to represent each house



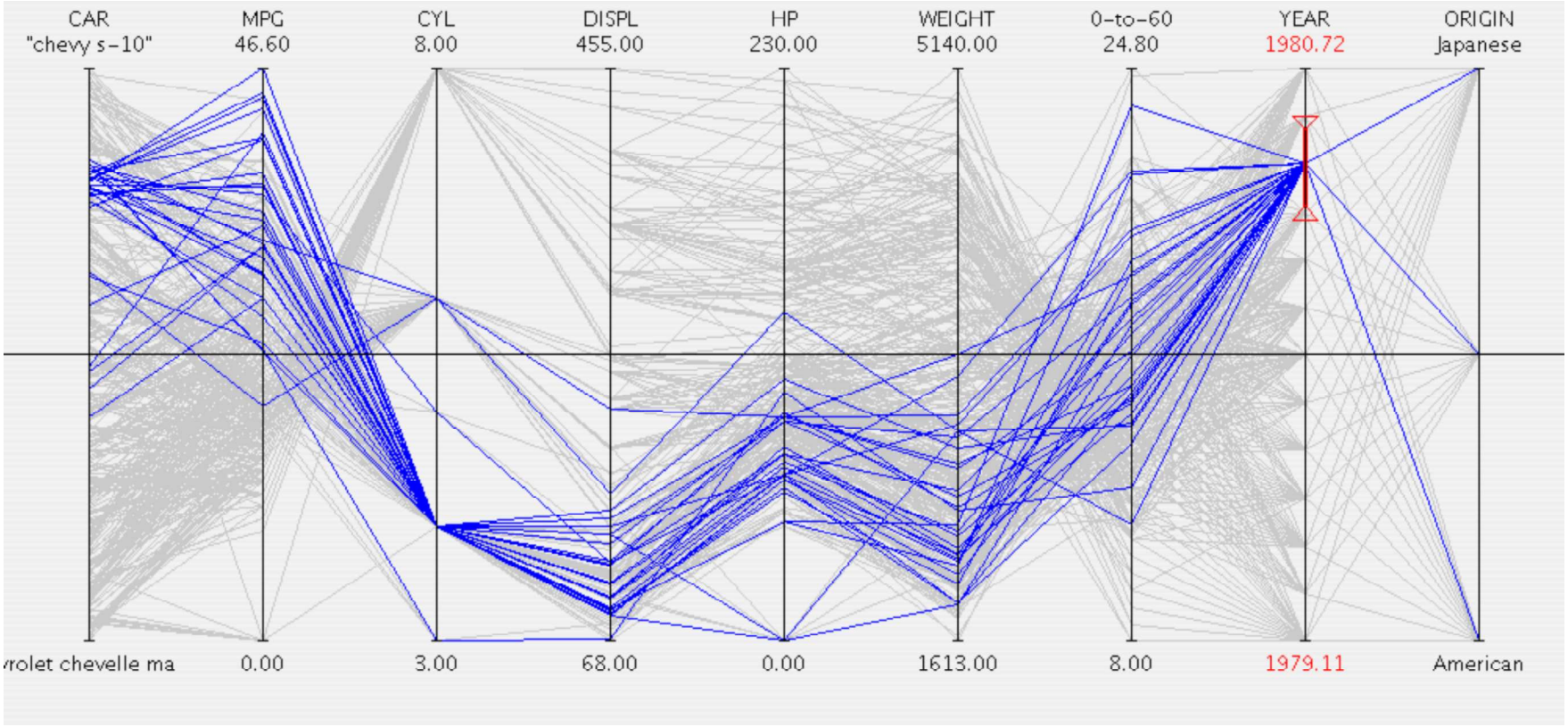
To avoid ambiguity the pair of points representing a house are joined and labelled

Parallel coordinates



A parallel coordinate plot for six objects, each characterised by seven attributes. The trade-off between A and B, and the correlation between B and C, are immediately apparent. The trade-off between B and E, and the correlation between C and G, are not

Parallel coordinates

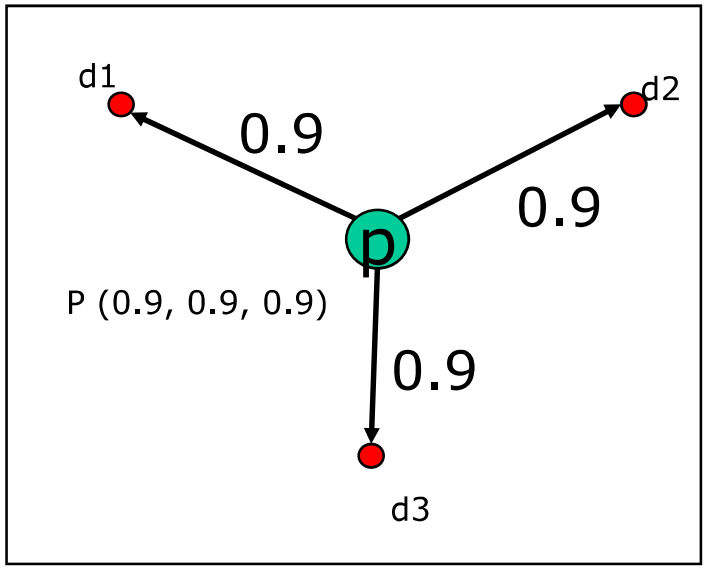


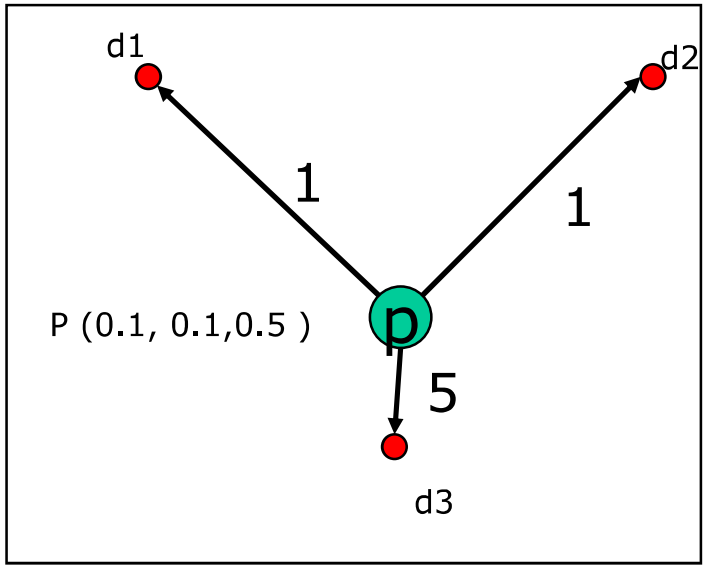
A parallel coordinate plot representation of a collection of cars, in which a range of the attribute *Year* has been selected to cause all those cars manufactured during that period to be highlighted

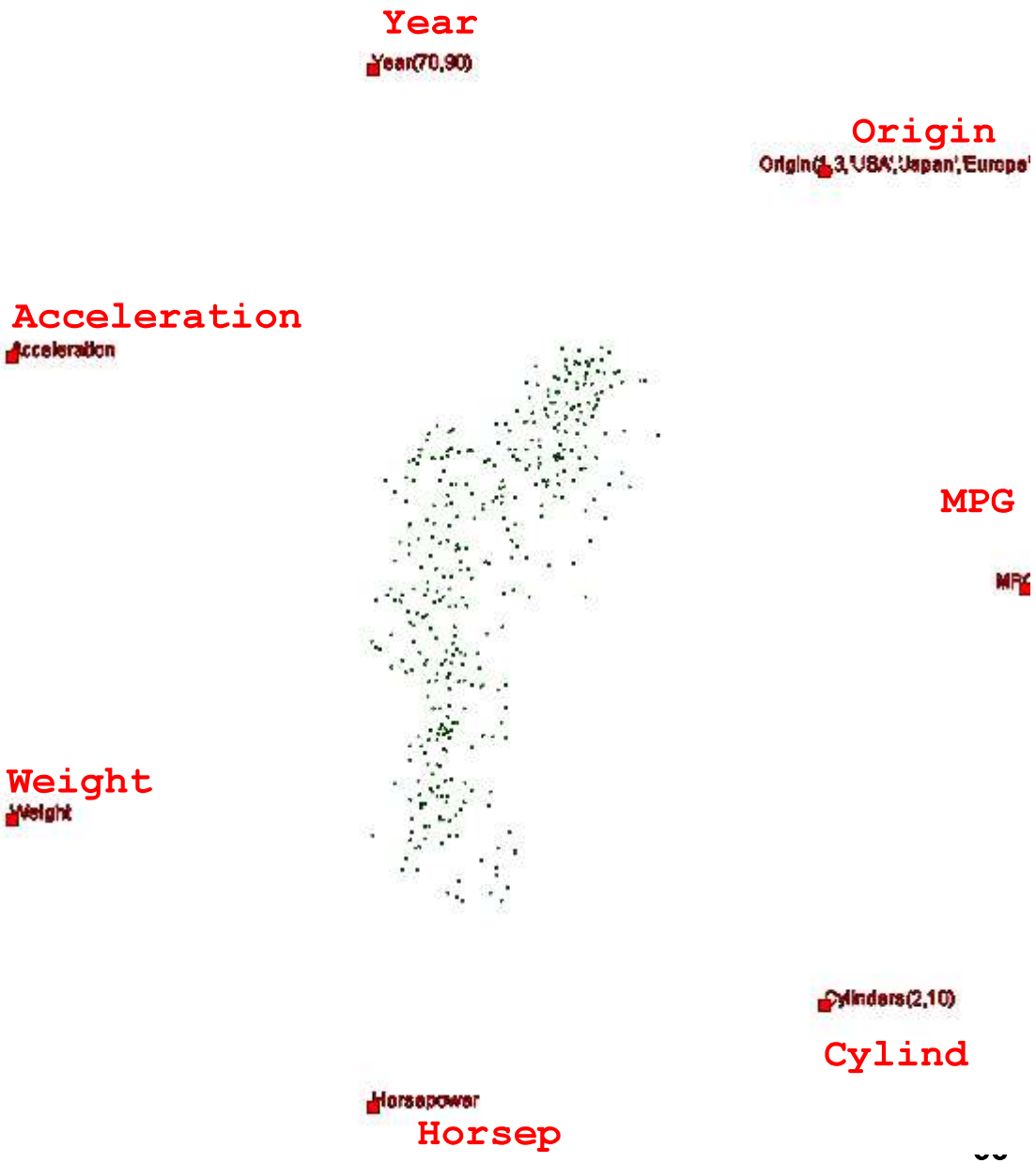
Radviz

- 7 dimensions (391 cars)
  - miles per gallon (M.P.G.)
  - number of cylinders
  - horsepower
  - weight
  - acceleration (time from 0 to 60 mph)
  - year
  - origin (USA, Europe, Japan)

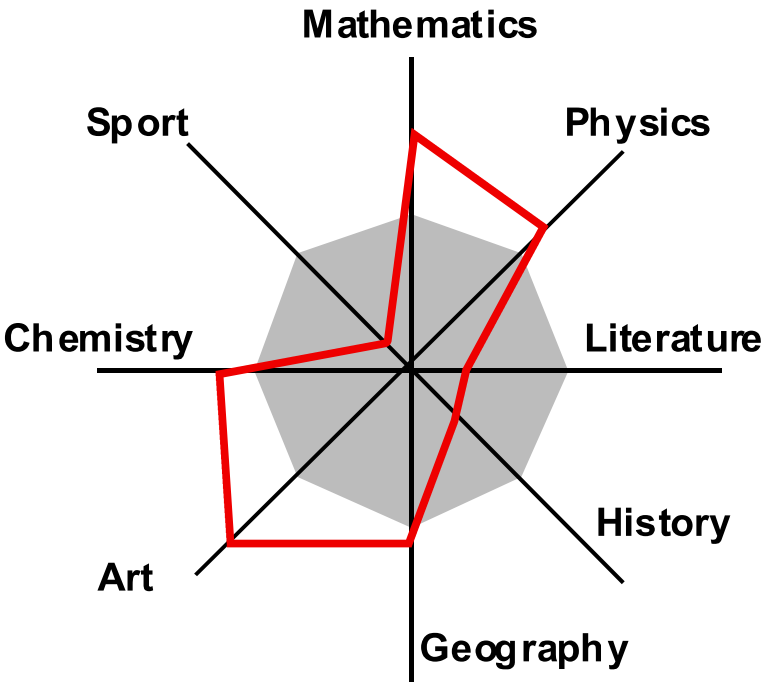
Radviz

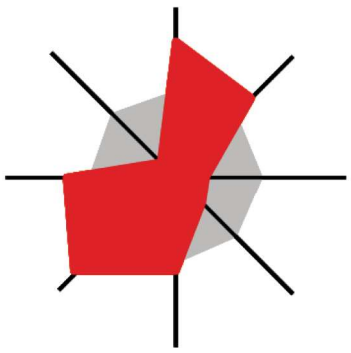

d1
d2
0.9
0.9
p
P (0.9, 0.9, 0.9)
0.9
d3


d1
d2
1
1
p
P (0.1, 0.1,0.5)
5
d3

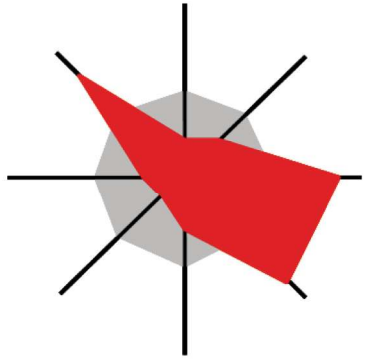

Year
Year(70,90)
Origin
Origin(1,3,'USA','Japan','Europe'
Acceleration
Acceleration
MPG
MPG
Weight
Weight
Cylinders(2,10)
Cylind
Horsepower
Horsep

Star plots (or radar diagrams)





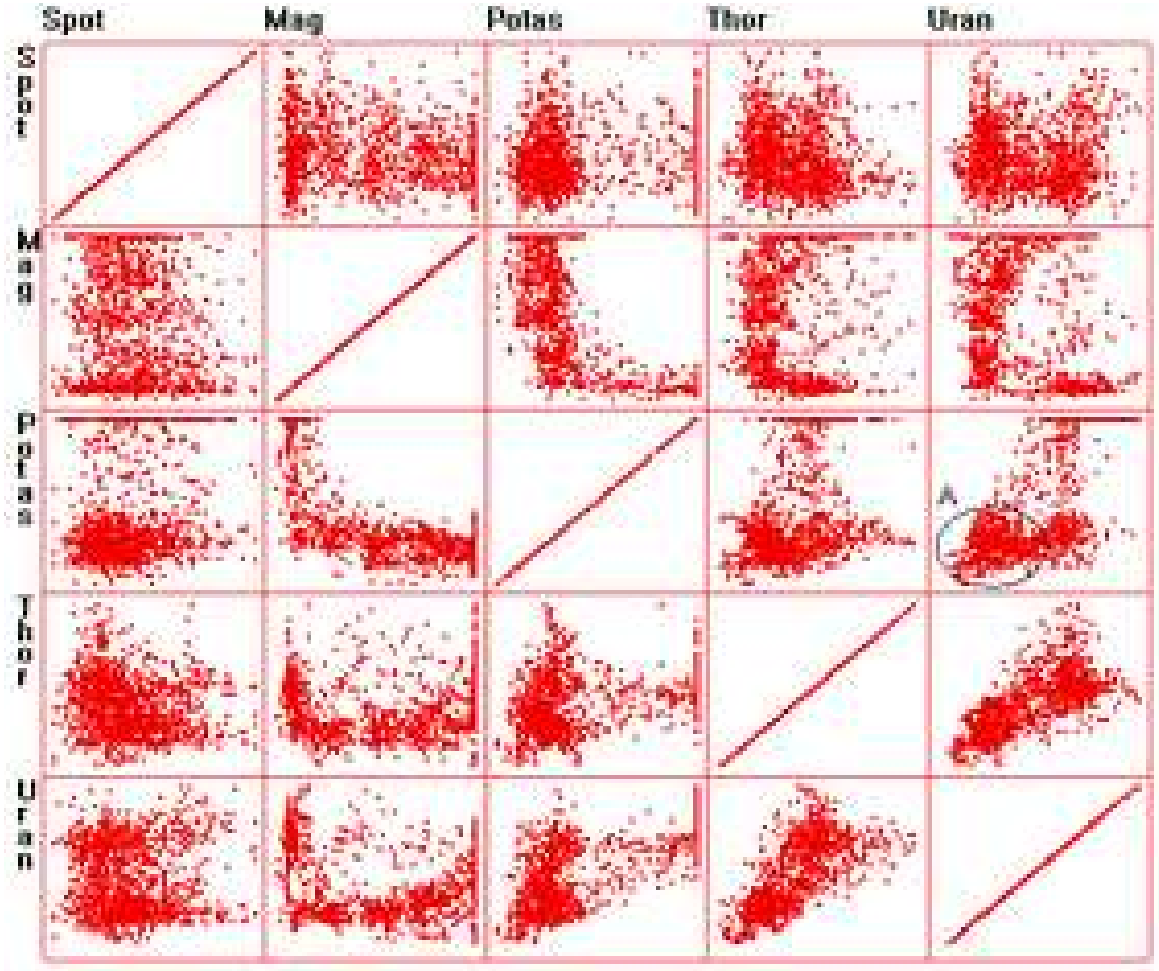
Bob's performance



Tony's performance

In a star plot attribute scales radiate from a common origin. Because shape can often effectively represent the combined attribute values of a single object, the points on each attribute scale can usefully be joined. Other useful information such as average values or thresholds can be encoded on the star plot

Scatterplot matrix (splom)



Scagnostic Wilkinson_2005@Infovis.pdf

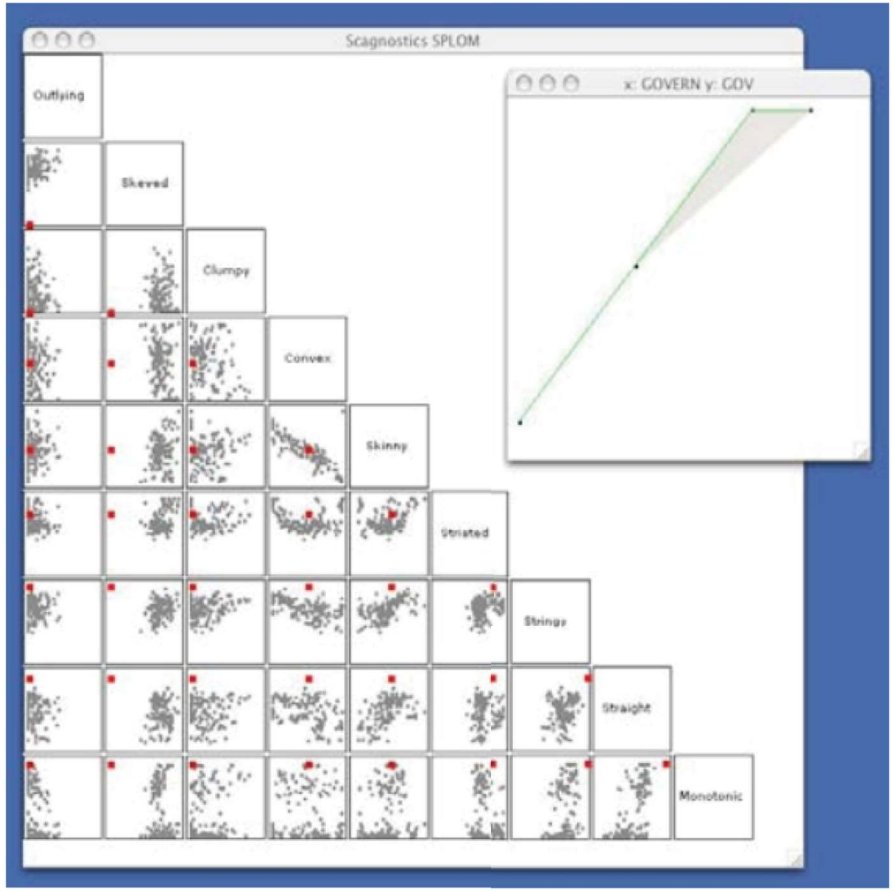


Figure 6: Scagnostics SPLOM of world countries data

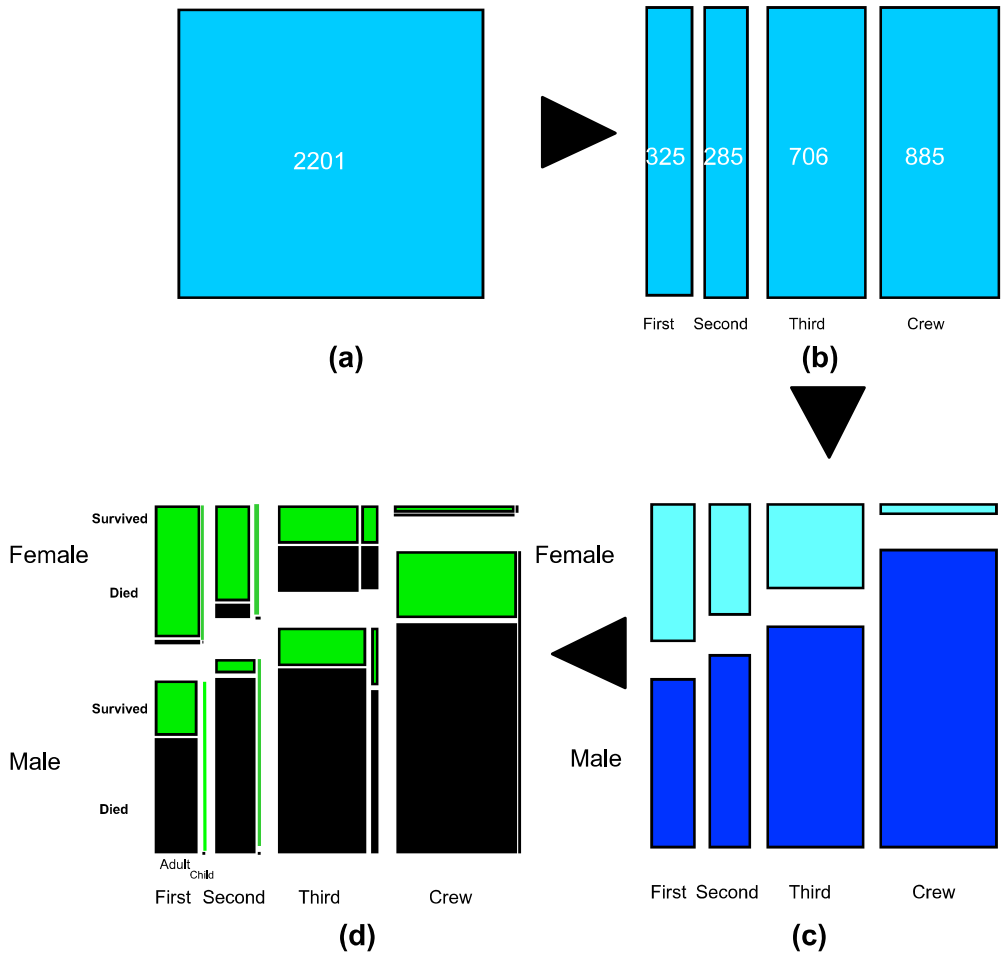
Abstract
We introduce Tukey and Tukey *scagnostics* and develop graph- theoretic methods for implementing their procedure on large datasets.

Mosaic plots

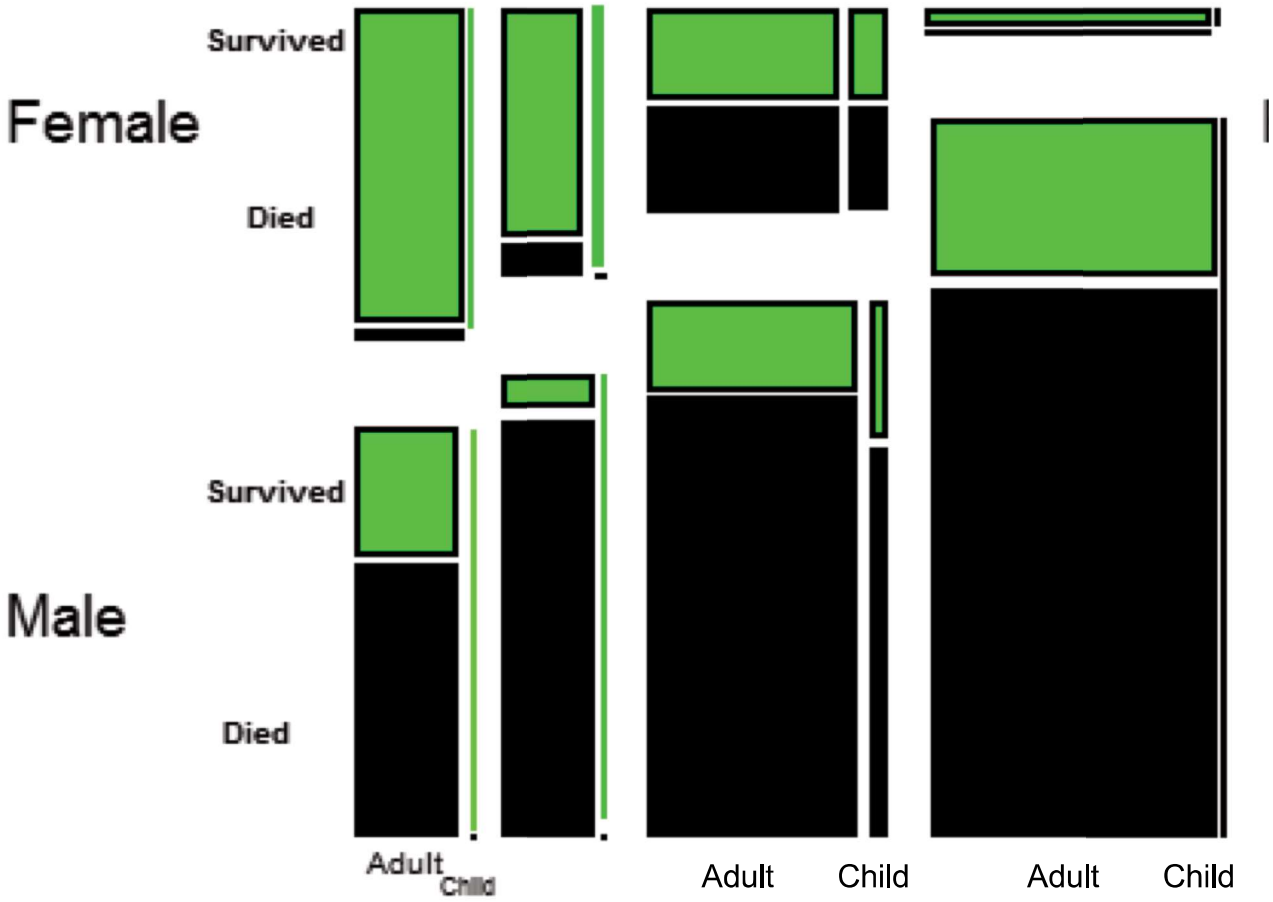
Survived	Age	Gender	Class			
			1st	2nd	3rd	Crew
No	Adult	Male	118	154	387	670
Yes			57	14	75	192
No	Child		0	0	35	0
Yes			5	11	13	0
No	Adult	Female	4	13	89	3
Yes			140	80	76	20
No	Child		0	0	17	0
Yes			1	13	14	0

Details of the Titanic disaster

Mosaic plots



Steps in the creation of a mosaic plot representing the Titanic disaster


Survived
Female
Died
Survived
Male
Died
Adult
Child
Adult
Child
Adult
Child
First
Second
Third
Crew

(d)

Outline

- Data types & data complexity
- Encoding of values
  - Univariate data
  - Bivariate data
  - Trivariate data
  - Multidimensional data
    - scatter plots + color /size...
    - scatter plot matrixes
    - parallel coordinates
    - radviz
    - star plots
    - mosaic plots
    - icons
- Encoding of relations
- Lines
- Map & Diagrams
- Trees
- Support for design

Icons

- Object visibility : representing single objects in a way that its attributes (or a subset) can easily be assimilated
- Some studies exist on icons, among them
  - Chernoff ' s faces
  - Multidimensional icons

Chernoff's face

- We are very good in distinguishing faces
- Studies show that a stylized face can bare till 18 attributes

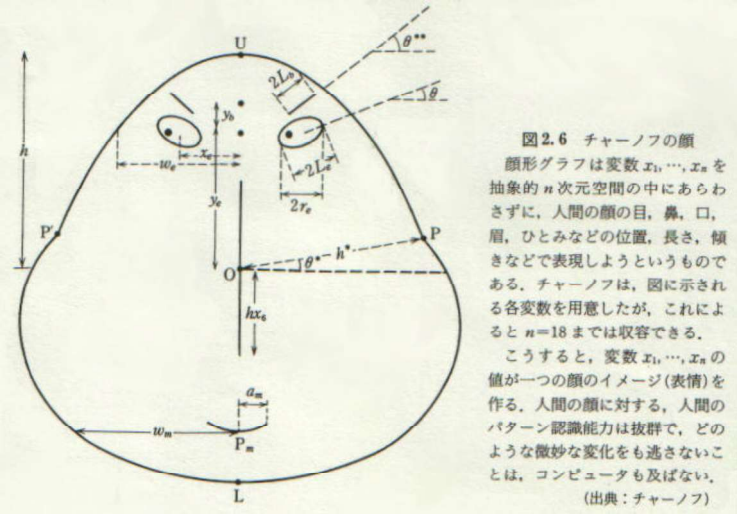
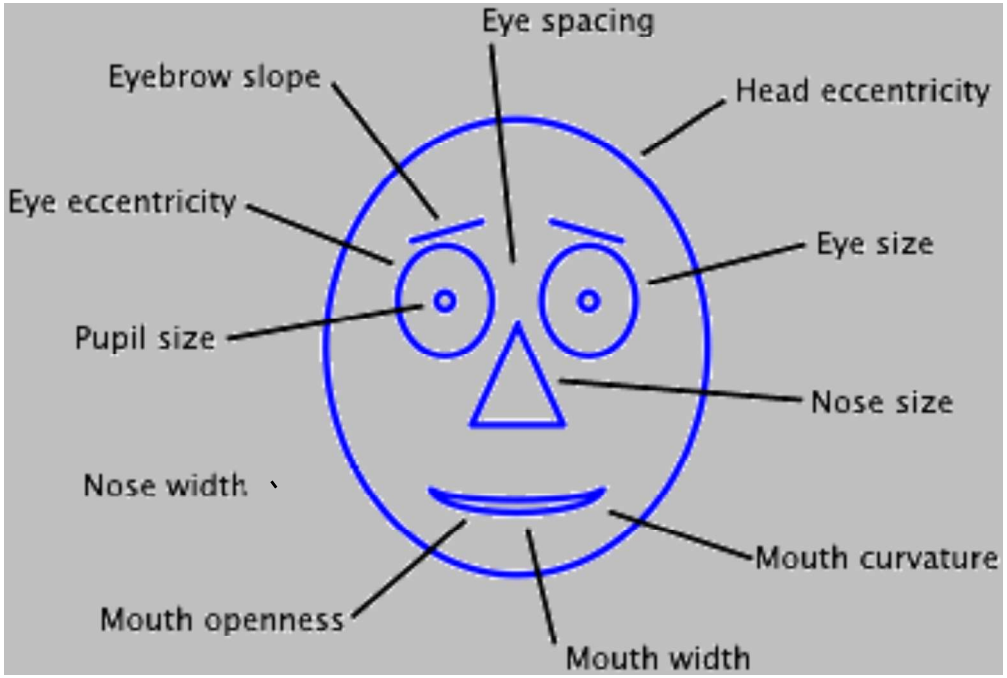
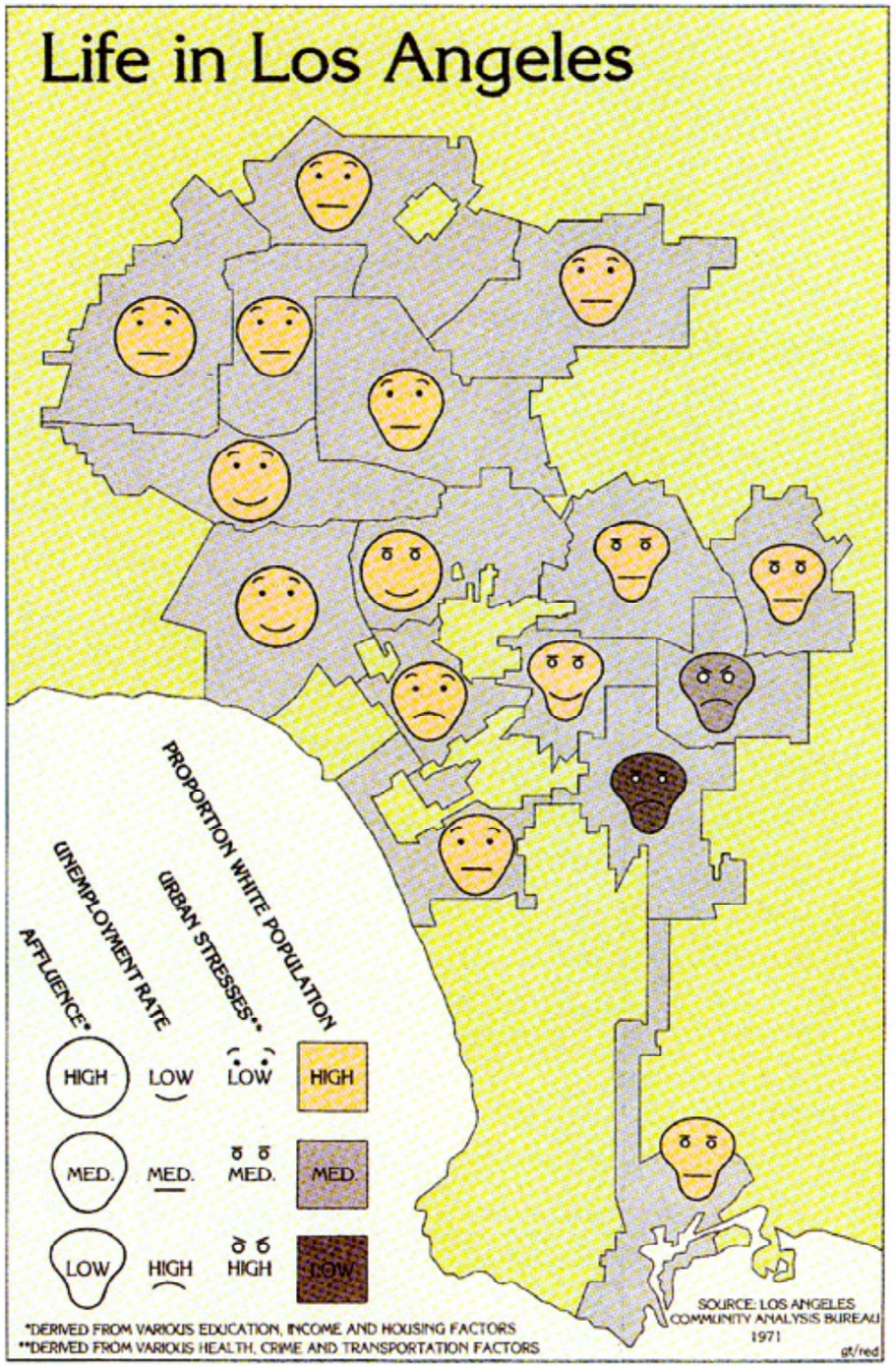


図2.6 チャーノフの顔
顔形グラフは変数 $x_1, \cdots, x_n$ を抽象的 n 次元空間の中にあらわさずに，人間の顔の目，鼻，口，眉，ひとみなどの位置，長さ，傾きなどで表現しようというものである．チャーノフは，図に示される各変数を用意したが，これによると $n=18$ までは収容できる．
こうすると，変数 $x_1, \cdots, x_n$ の値が一つの顔のイメージ(表情)を作る．人間の顔に対する，人間のパターン認識能力は抜群で，どのような微妙な変化をも逃さないことは，コンピュータも及ばない．
(出典：チャーノフ)



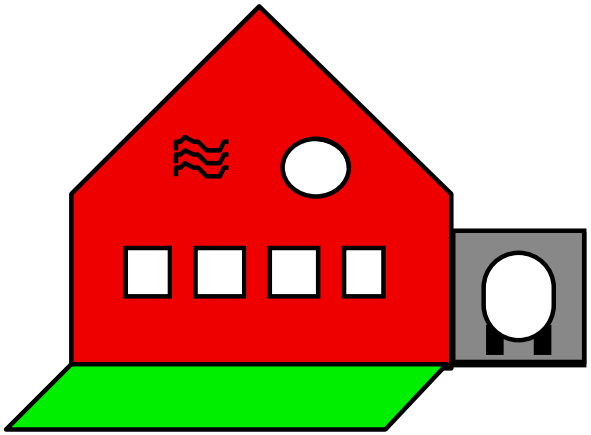
Chernoff's faces

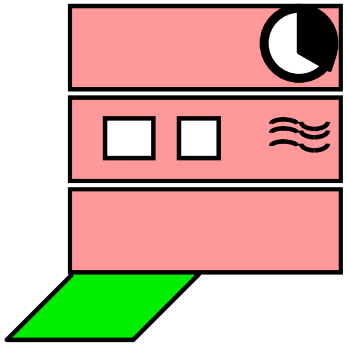
Affluence= benessere

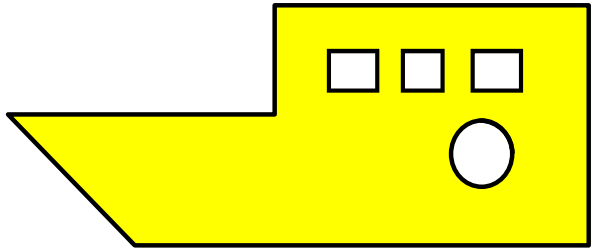


Multidimensional icons

- Looking for a place in which live ?
- Encoding eight attributes through an icon
- Searching on 56 dwellings was 200 % faster using icons..

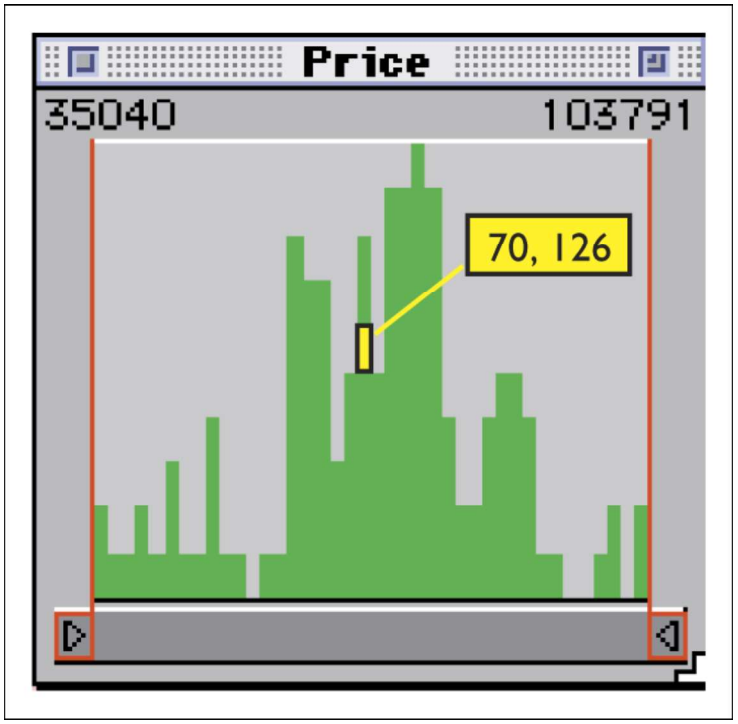




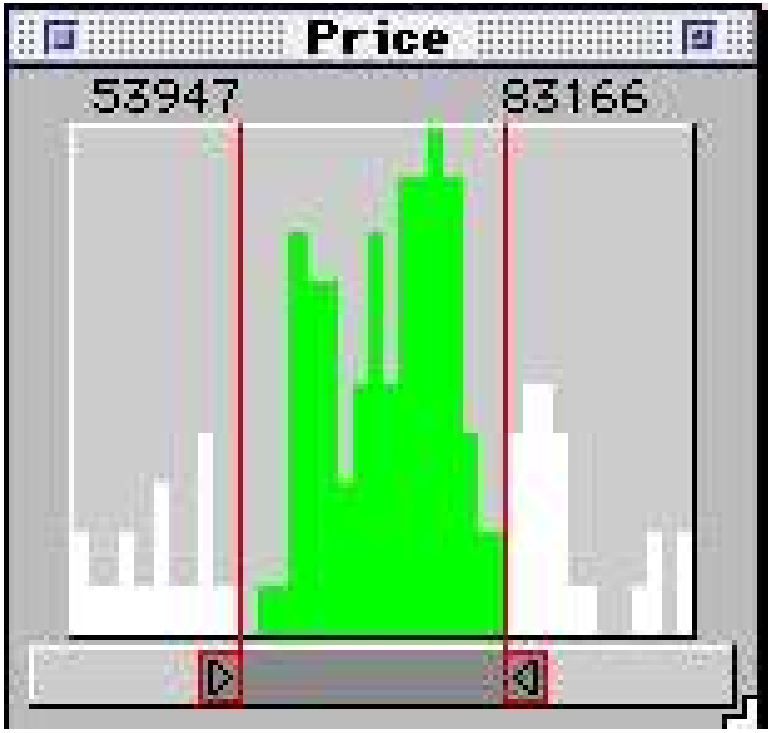


house	flat	houseboat
£400,000	£300,000	£200,000
garage	no garage	no garage
central heating	central heating	no central heating
four bedrooms	two bedrooms	three bedrooms
good repair	poor repair	good repair
large garden	small garden	no garden
Victoria 15 mins	Victoria 20 mins	Victoria 15 mins

Multiple coordinated views

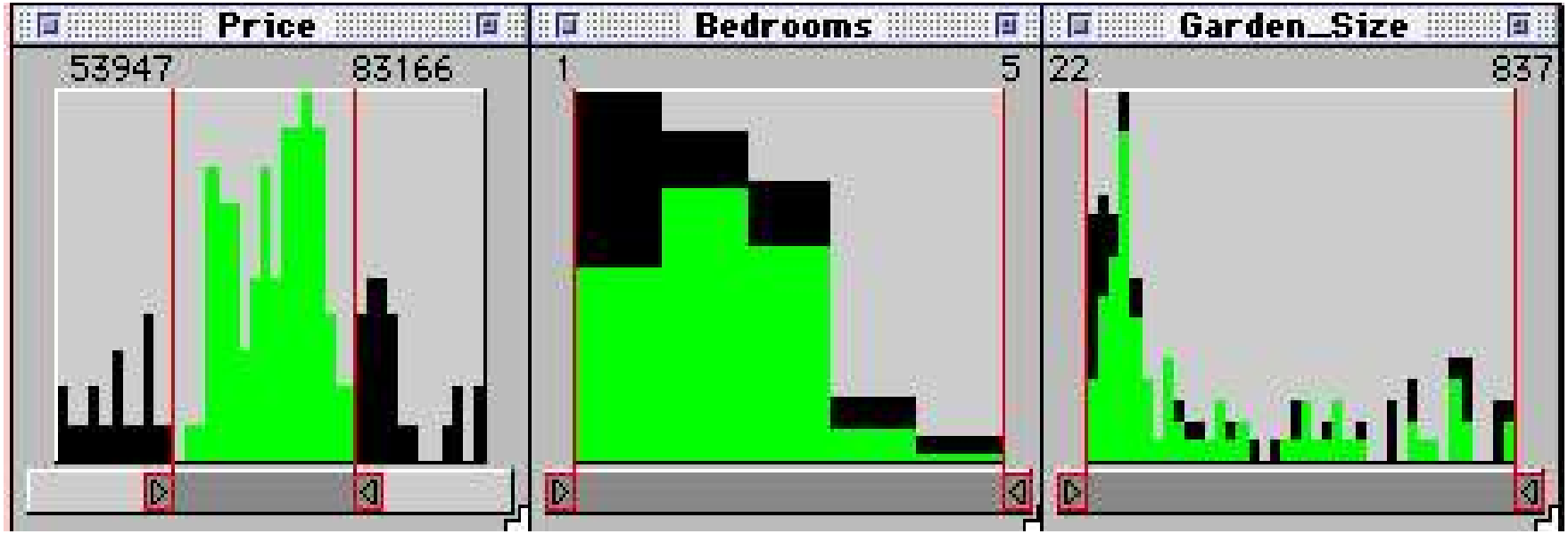


A histogram representing the prices of a collection of houses. The contribution of one house is shown in yellow



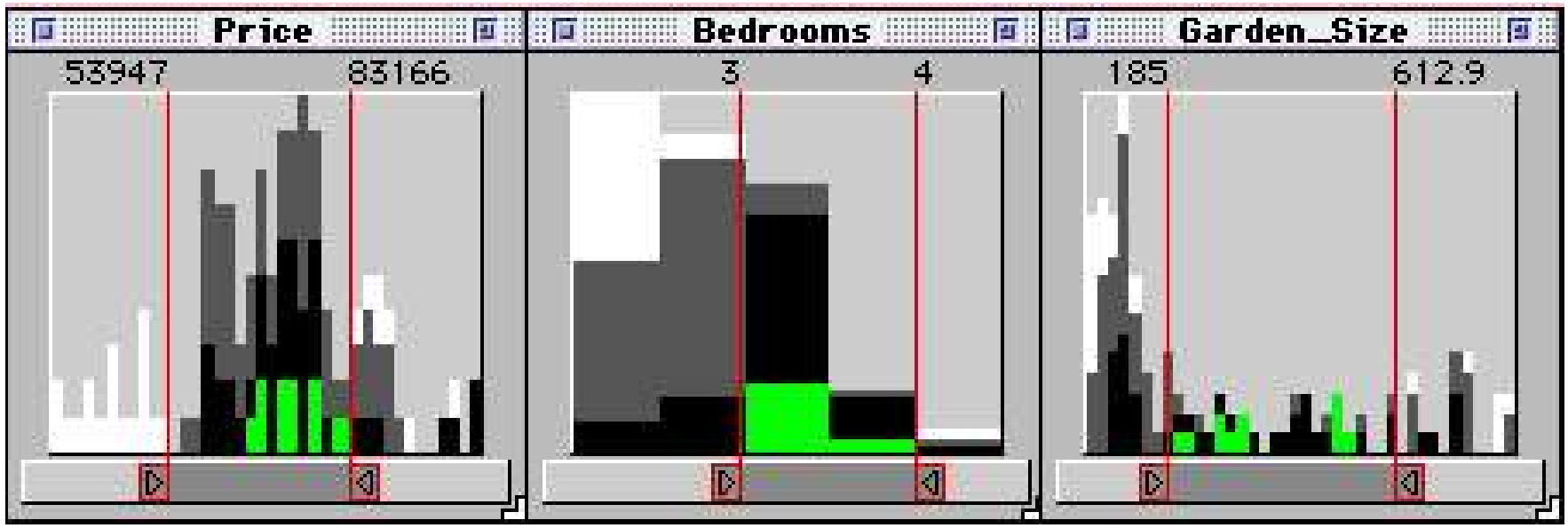
Limits on *Price* identify a subset of houses, coded green

Multiple coordinated views



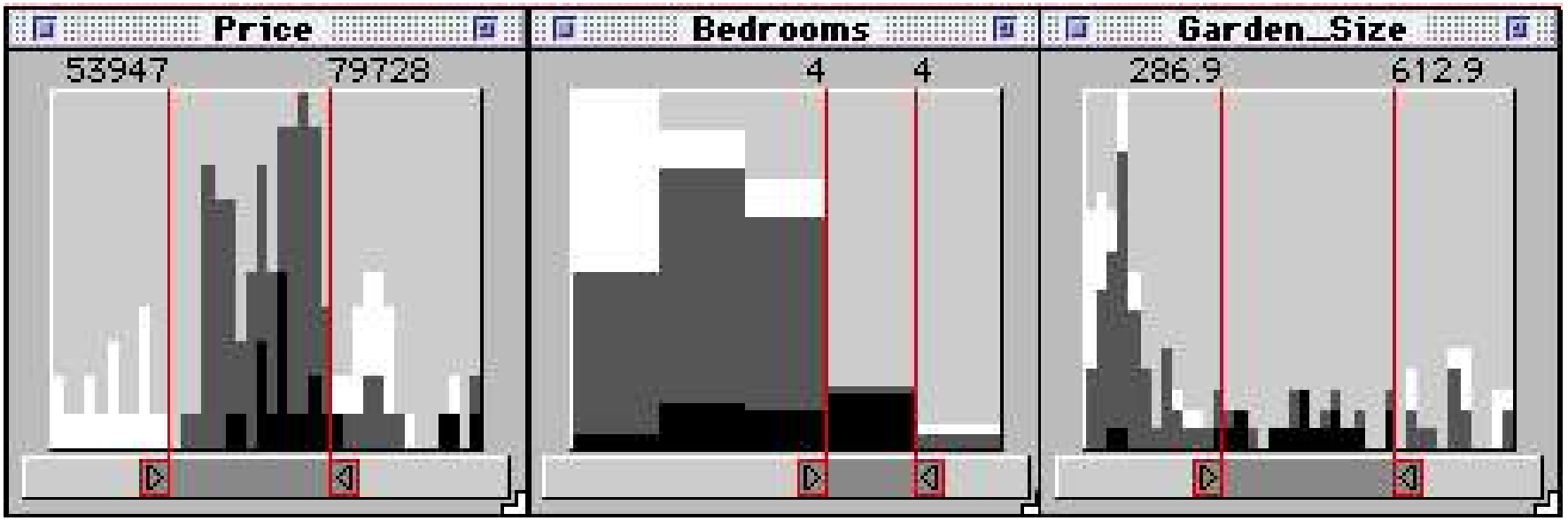
Houses defined by the limits on *Price* are coded green in other attribute histograms

Multiple coordinated views



Green coding applies only to houses which satisfy all attribute limits. Houses which fail **one limit** are coded **black**, so if a black house is positioned **outside a limit** it will turn green if the limit is extended to include it

Multiple coordinated views



Even if no houses satisfy all attribute limits, black houses, which fail only one limit, provide guidance as to the effect of relaxing limits

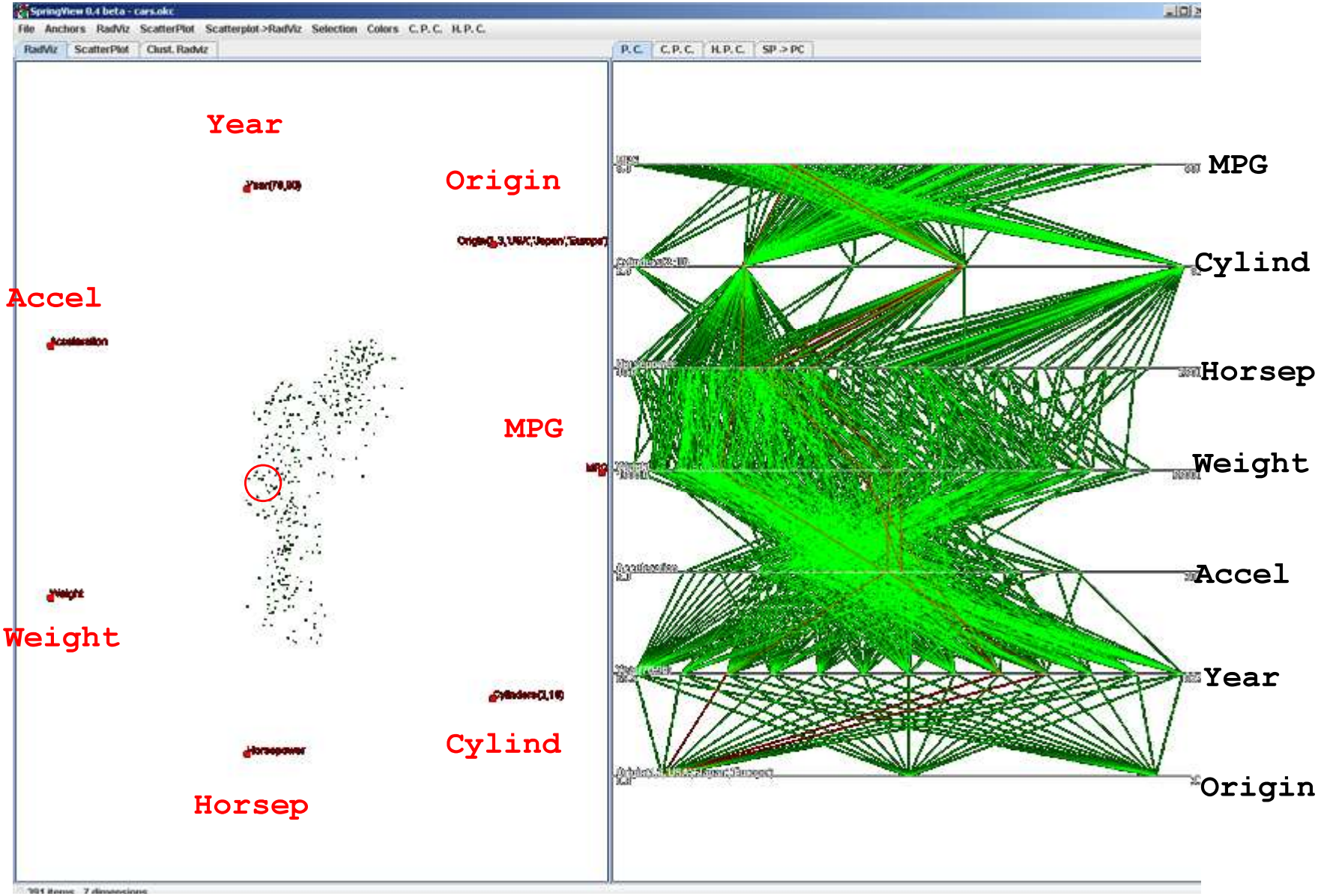
Radviz + Parallel coordinates

- Radviz Pros
  - Traditional representation (scatter plot)
  - Not data crossing
  - Easy 2D interaction through a pointer (direct manipulation)
- Radviz Cons
  - No details about dimensions' values
  - Not unique correspondence between data and screen points

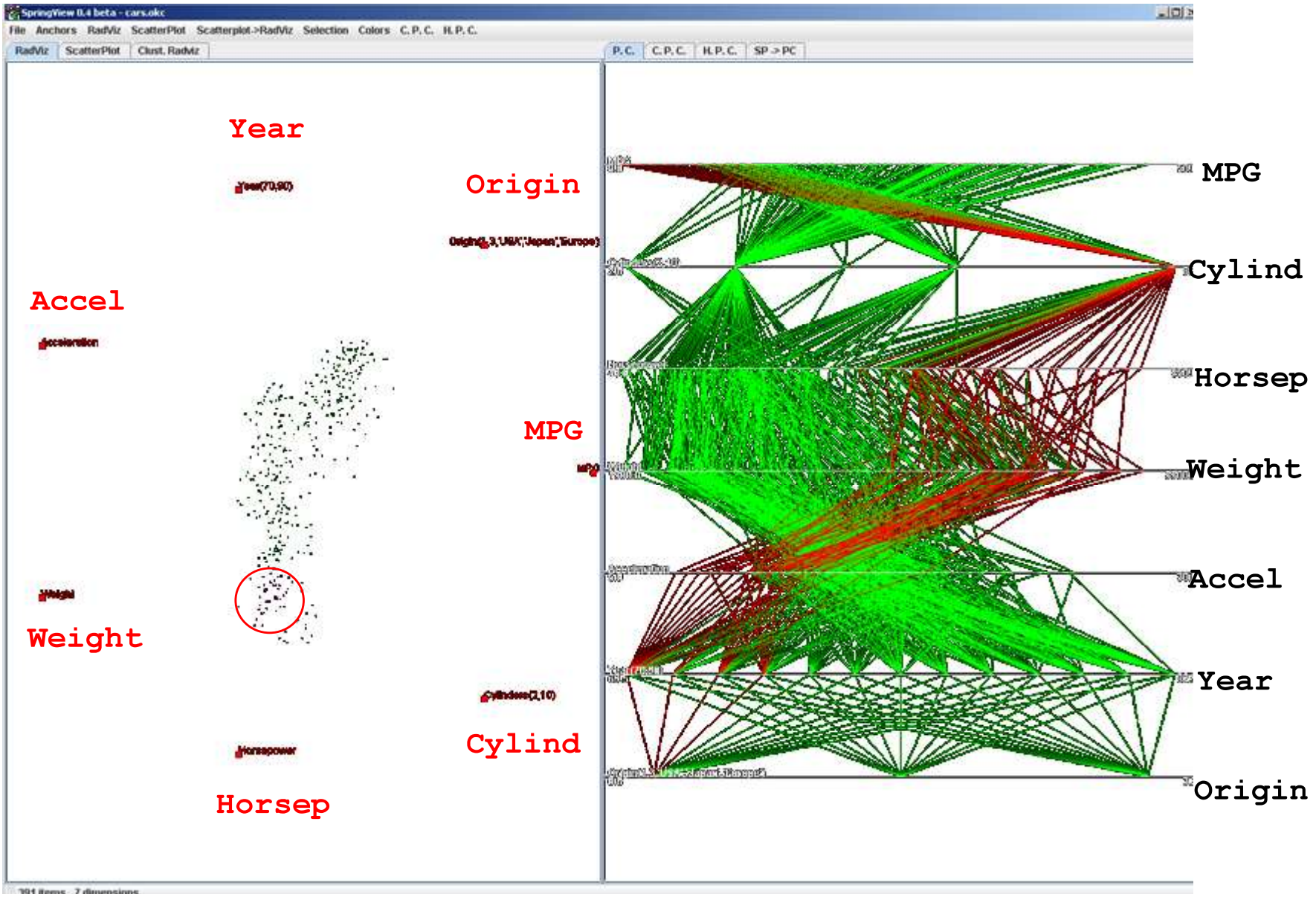
Radviz + Parallel coordinates

- Parallel coordinates Cons
  - Non traditional representation
  - Data crossing
  - Not easy interaction through a pointer
    - e.g., selecting a car subsets requires SEVEN range selections
- Parallel coordinates Pros
  - Details about dimensions' values
  - Unique correspondence between data and screen polylines

Understanding Radviz



Selecting data subsets

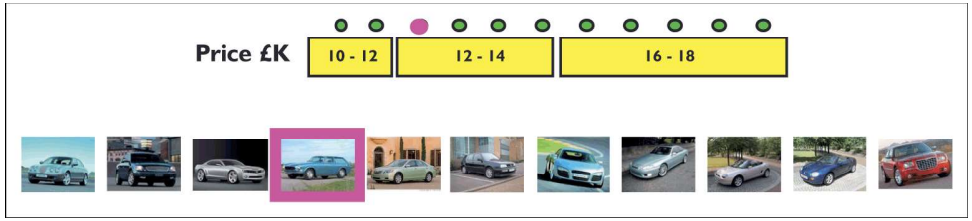


Outline

- Data types & data complexity
- Encoding of values
  - Univariate data
  - Bivariate data
  - Trivariate data
  - Multidimensional data
- Encoding of relations & relationships
- Lines
- Map & Diagrams
- Trees
- Support for design

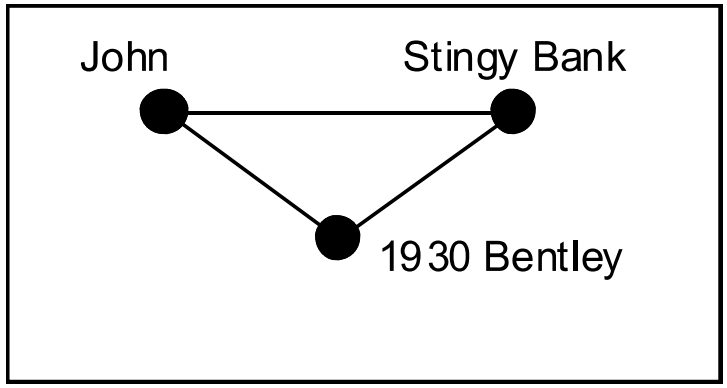
Encoding of relations and relationships

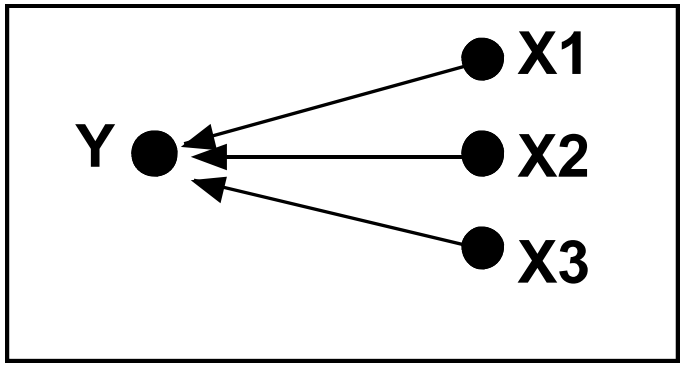
- Relation: logical or natural association between **two or more <u>values</u>**



Make	Price (£)	MPG	Rating	Age (yrs)
Ford	15,450	31	*****	3
Chevy	12,450	27	***	4

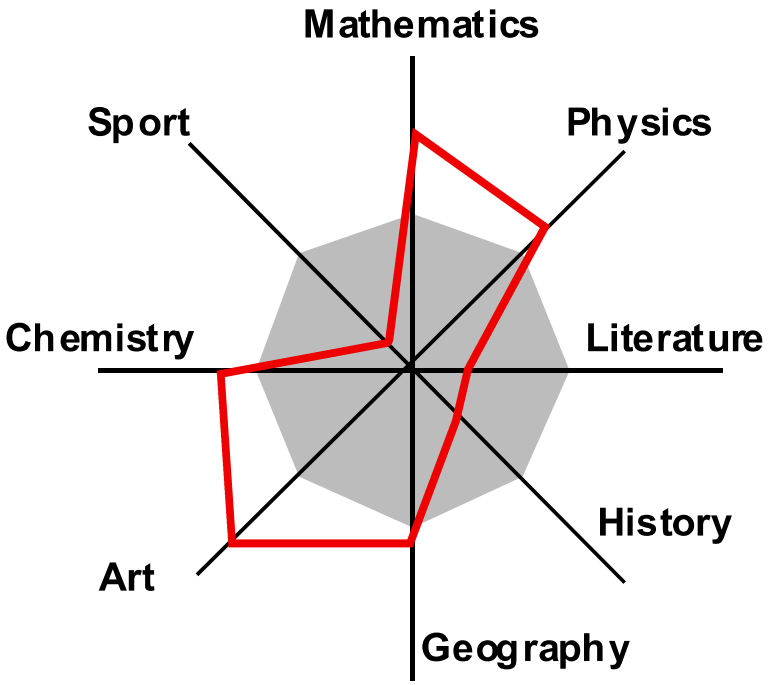
Relationship: connection between **two or more <u>data items</u>**

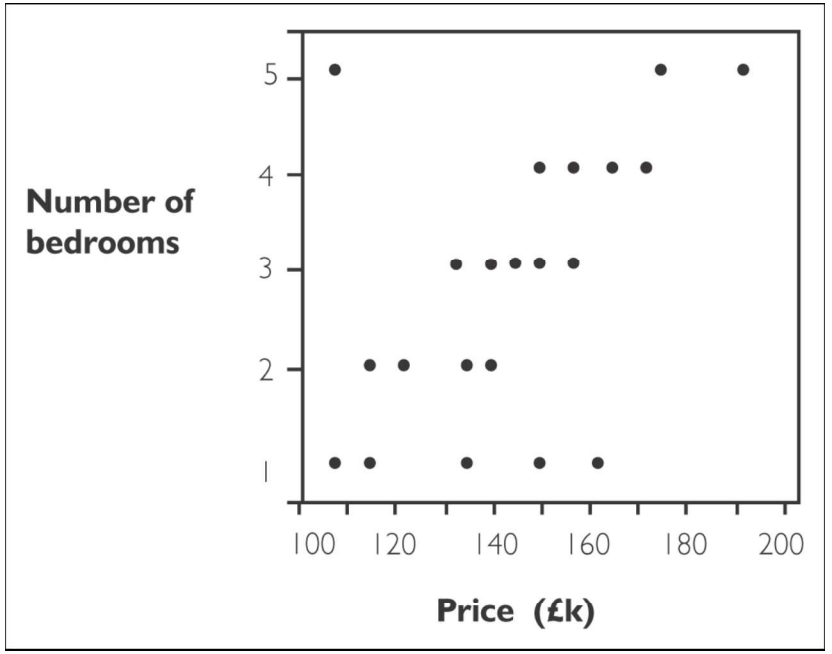




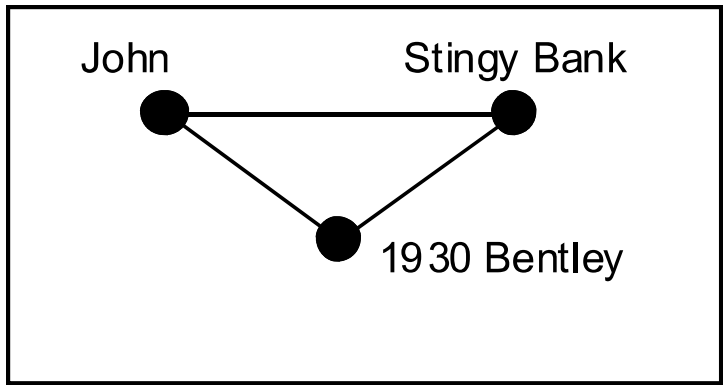
What is the **visual** difference between the two?

You can encode a relation using values





You **cannot** encode a relationship using values



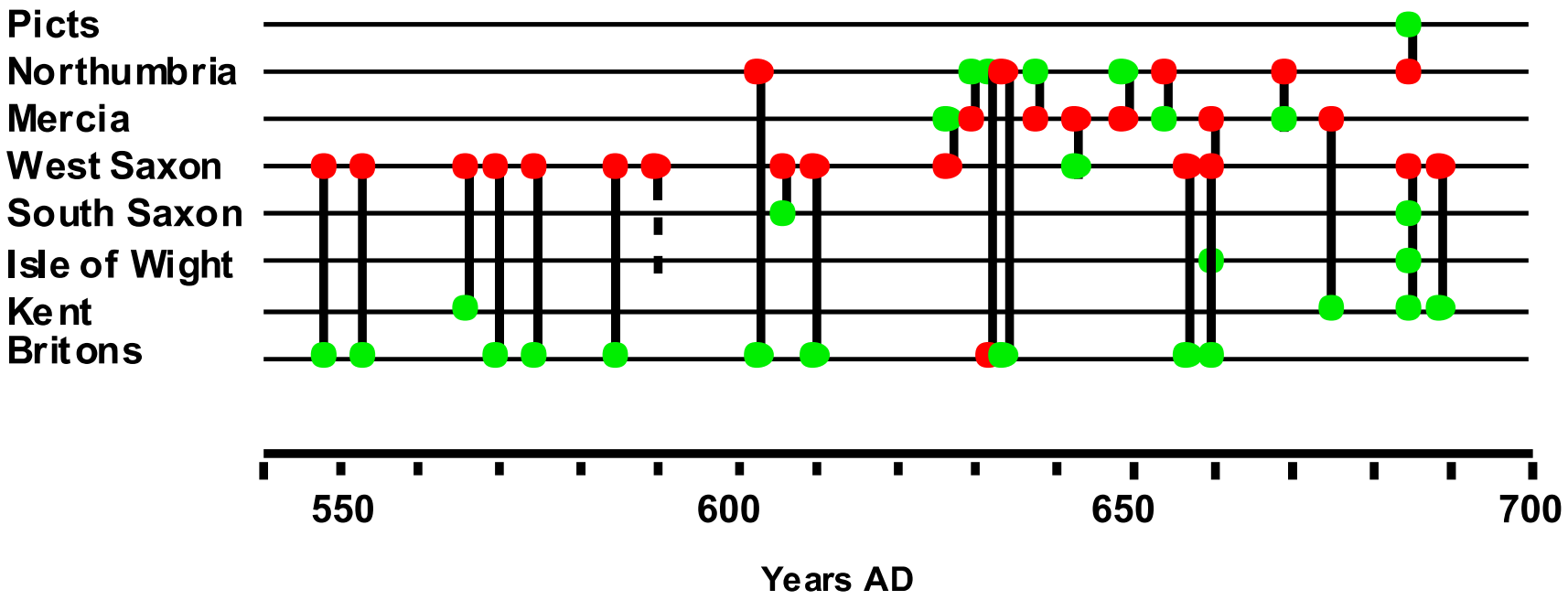
Line
Position
Brushing
~~Values~~

But line, position, and brushing,
are also fine with relationships !

Outline

- Data types & data complexity
- Encoding of values
  - Univariate data
  - Bivariate data
  - Trivariate data
  - Multidimensional data
- Encoding of relations
  - Lines
- Map & Diagrams
- Trees
- Support for design

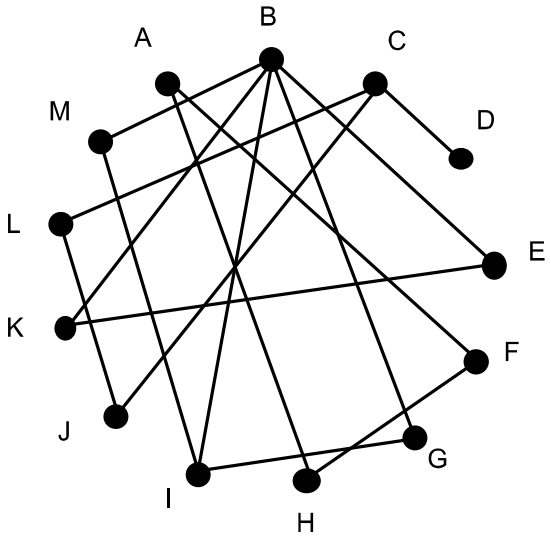
Lines! (+ color + time)



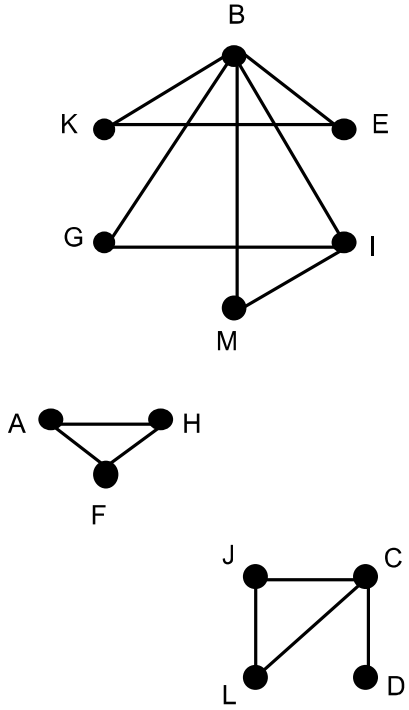
The incidence of warfare in early Anglo-Saxon England between 550 AD and 700 AD. Red indicates the aggressor, green the attacked

Lines !

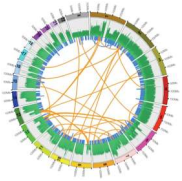
Originator	Receiver
A	H
C	L
I	M
B	E
F	H
G	I
I	B
B	M
K	B
G	B
K	E
C	J
D	C

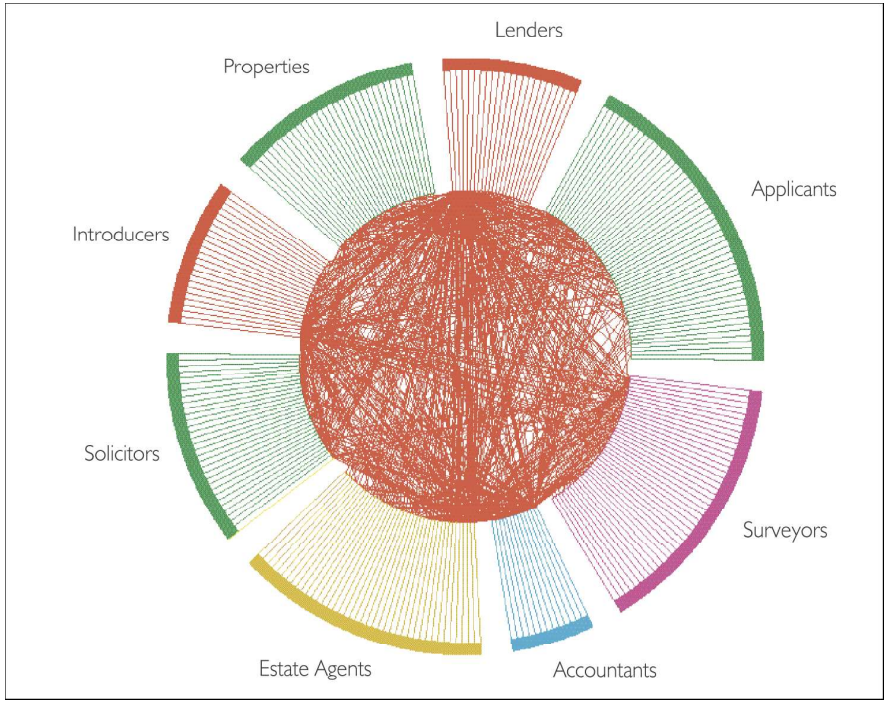


Phone calls ?

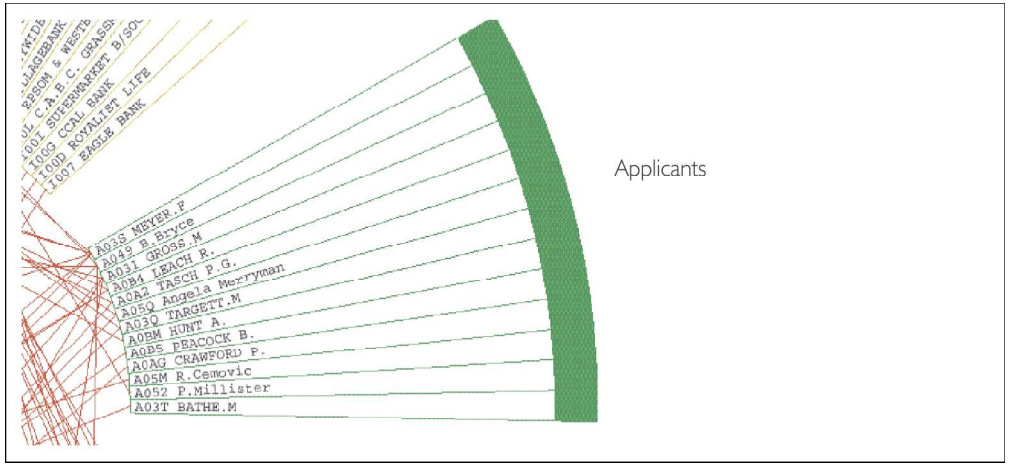


Radial graph (ancestor of chord diagrams)
Lines + color + partitioning + semantic zoom





(a)

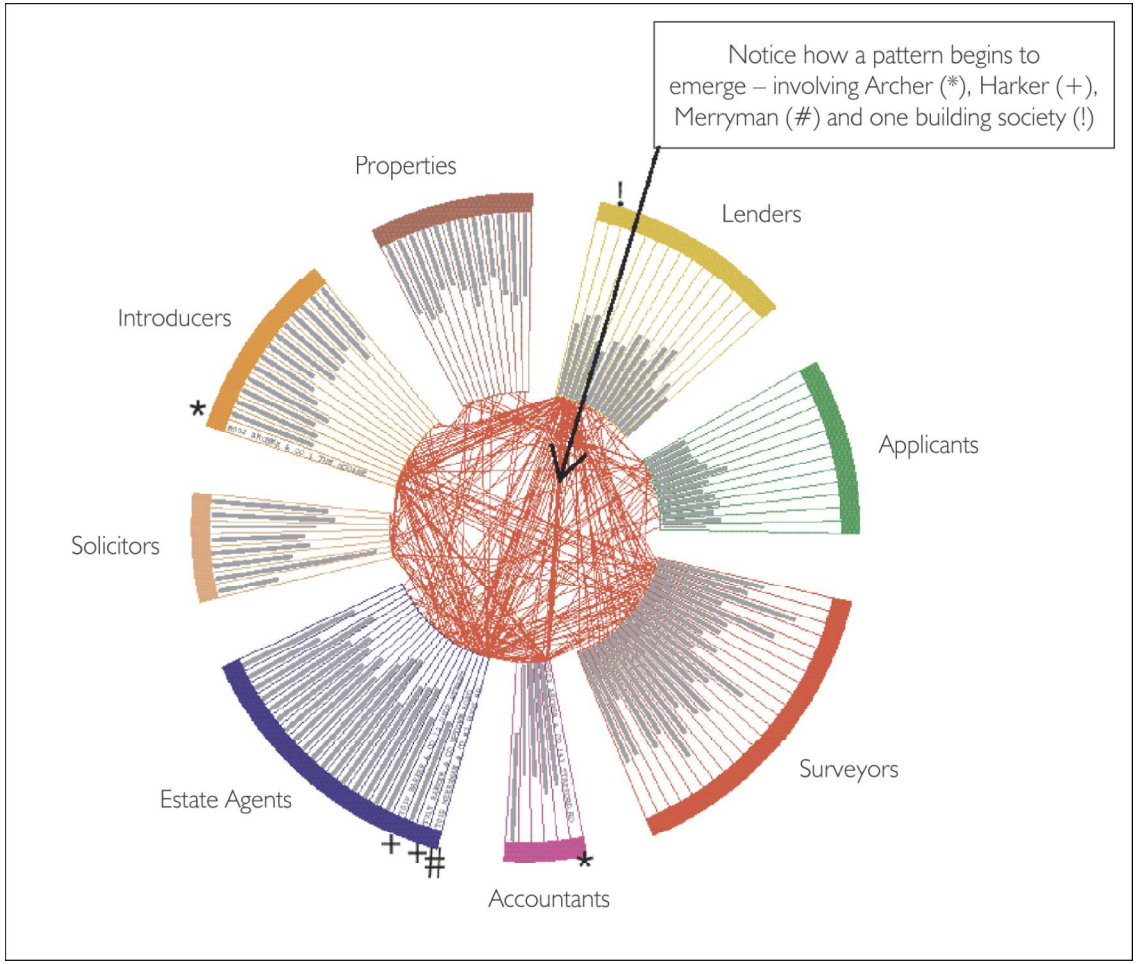


(b)

mortgage = ipoteca
lender = prestatore
surveyor=ispettore
applicant= richiedente
accountant=contabile

A representation of mortgage activity (a). Lenders, properties (houses), buyers, etc. are represented by small radial segments of an annulus as shown in (b), and their relationships denoted by straight lines

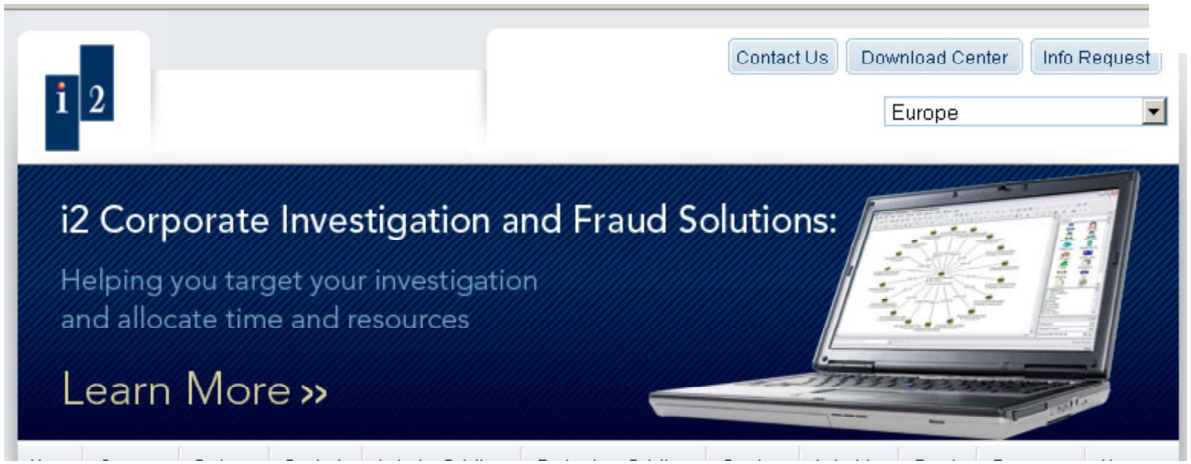
A mortgage fraud visualized & discovered



A threshold has been imposed to suppress the display of **normal** behavior. As a result, **unusual behavior** is revealed by the patterns formed by the lines

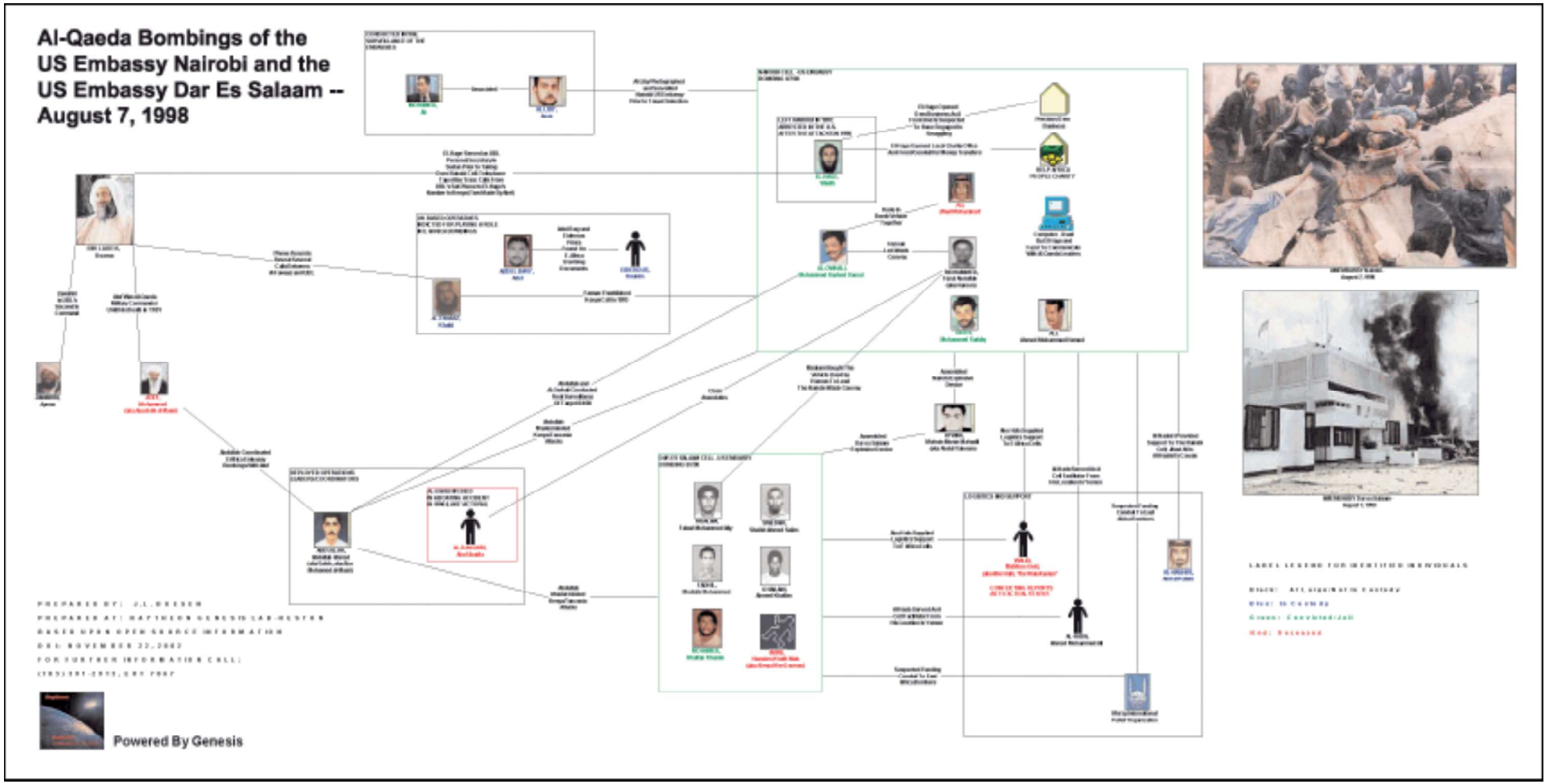
Representing connection between people

- Increasing interest about the matter
- intelligence analysis
  - associated with chart
  - timeline chart
  - for annotations
  - for explaining
- social networks
  - facebook visualization

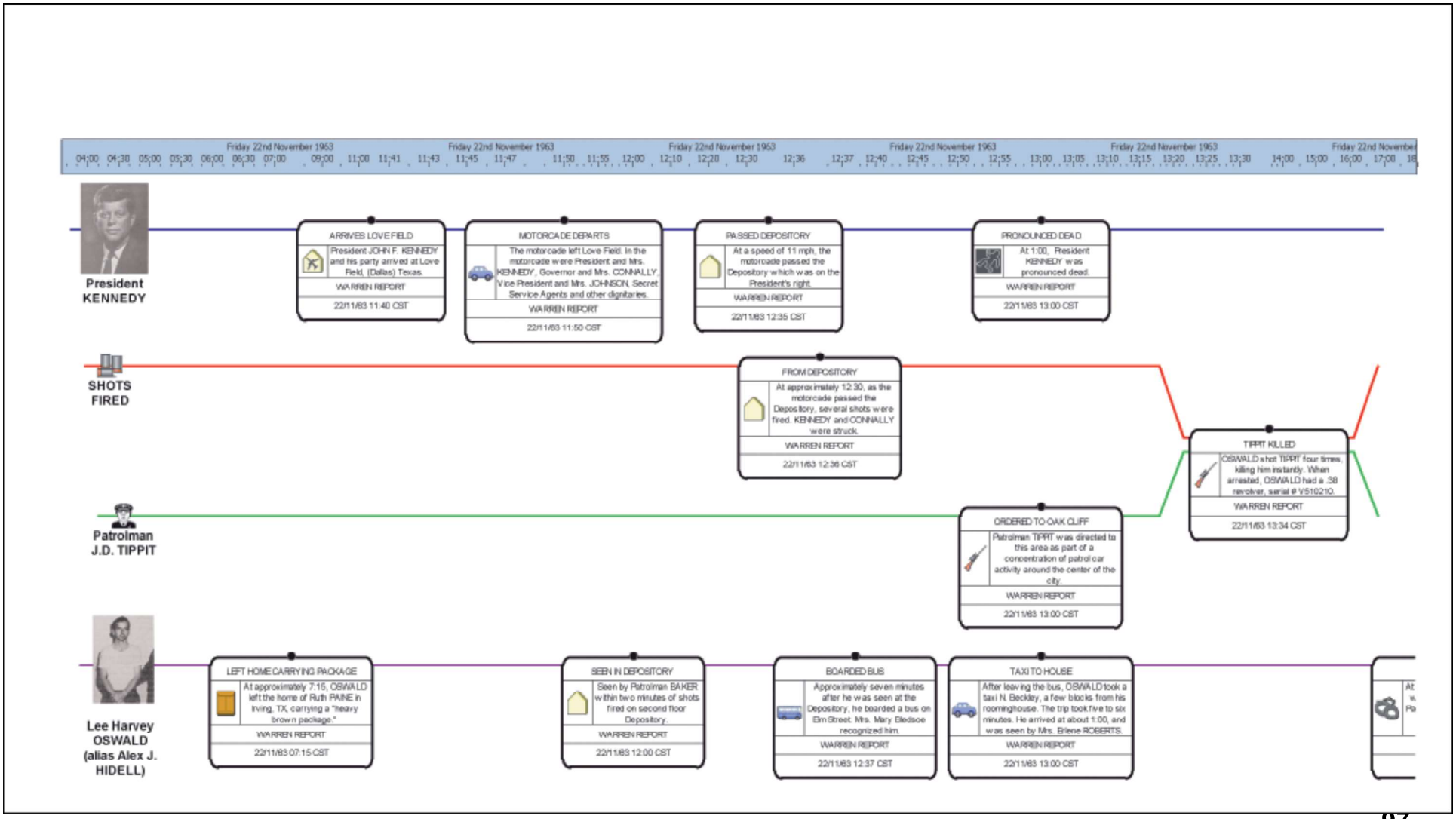


i2 Limited (2006)
a successful story
(now IBM)

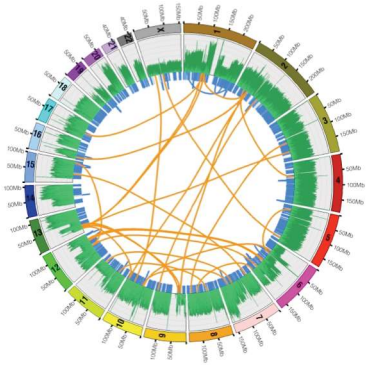
Association chart about African bombing relationships

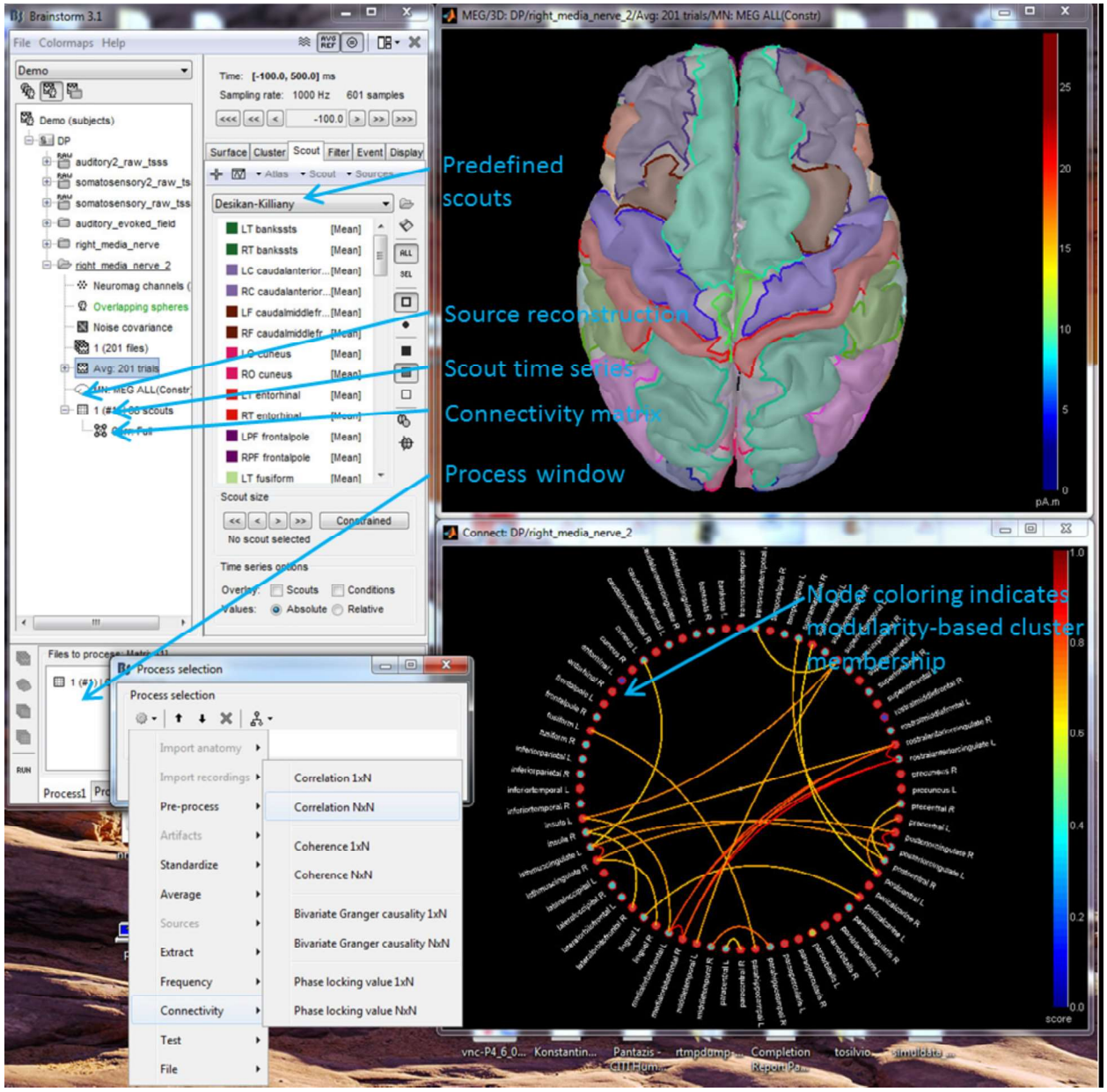


Timeline chart about Kennedy assassination (relationships + time)

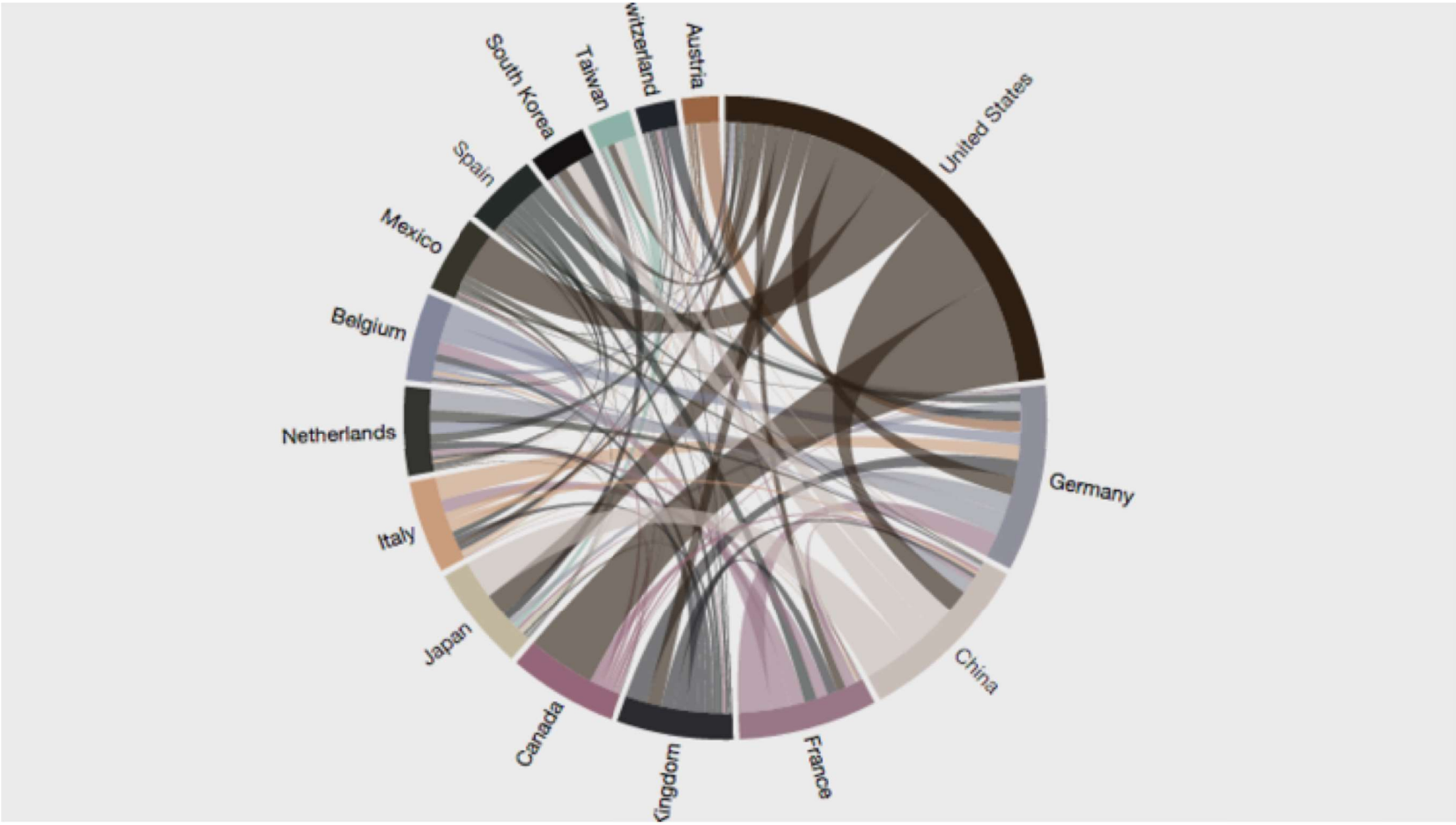


Chord diagrams (relationships)

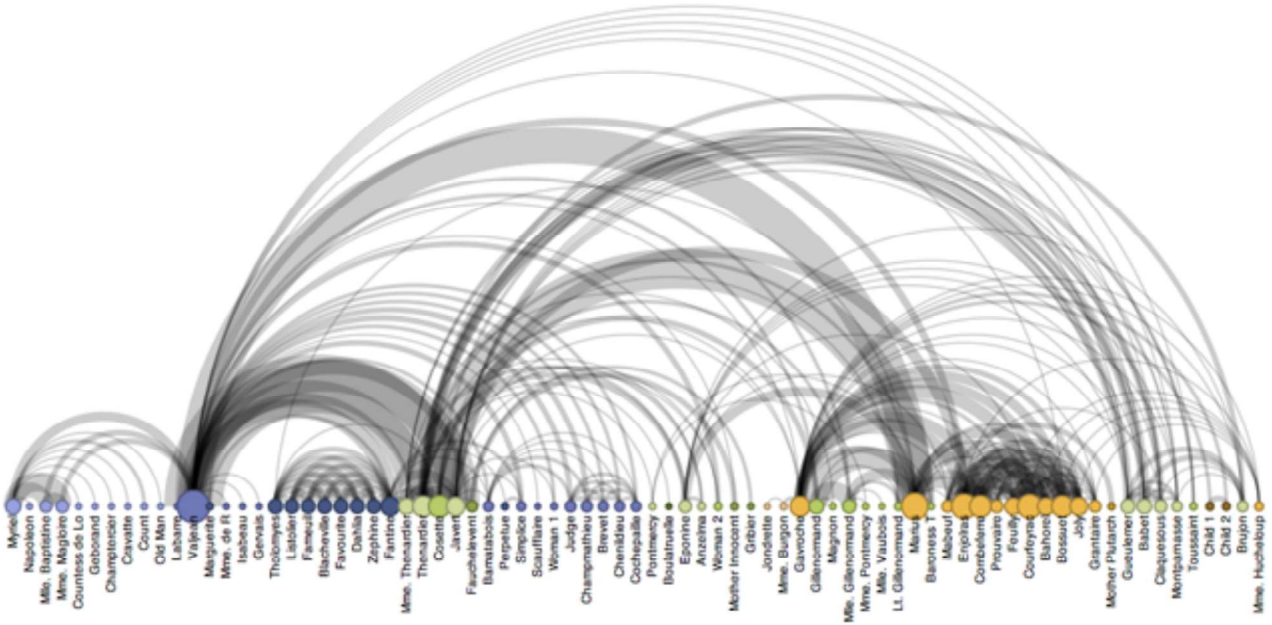




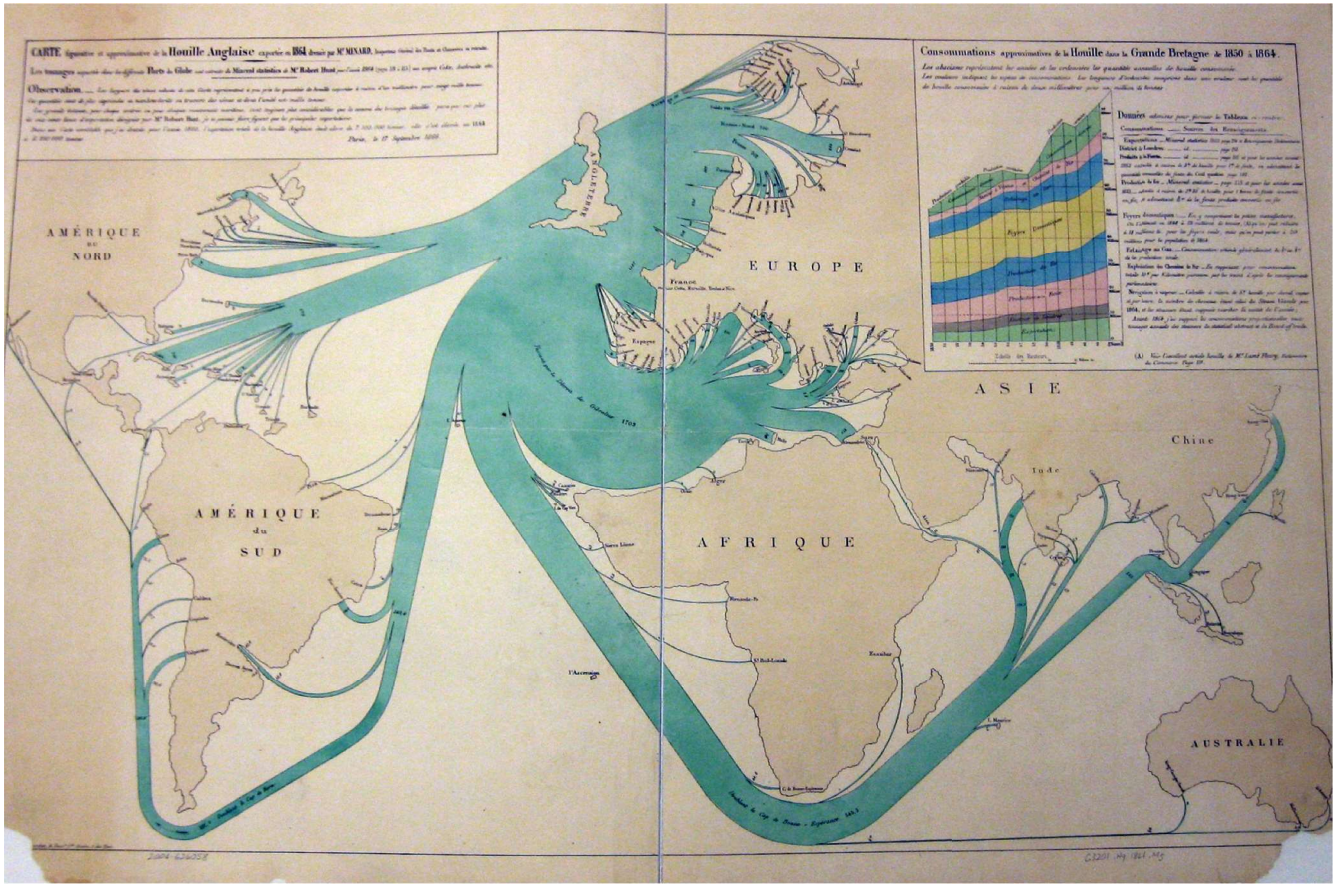
Chord diagrams (relationships + numerical values)



'Opened' chord diagram...

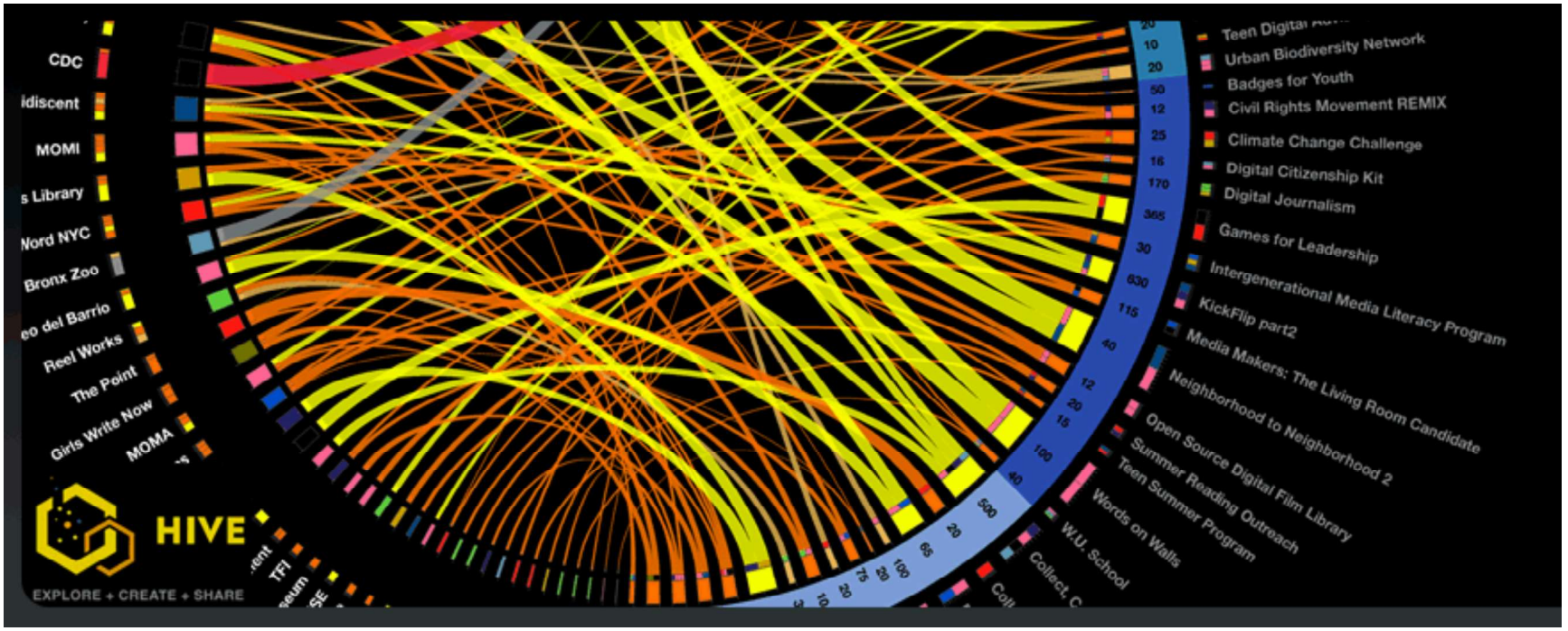


Minard idea (and ideas for home works!)

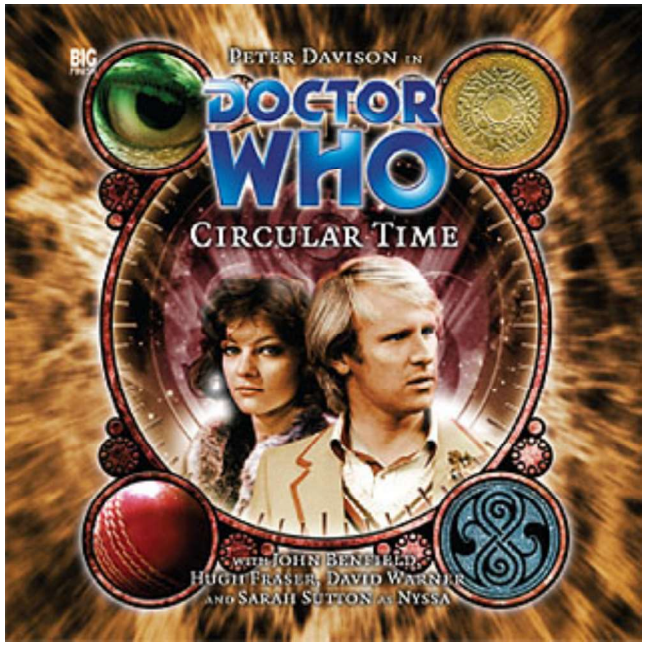


Rephrase historical drawing in modern viz!

And radial on radial...

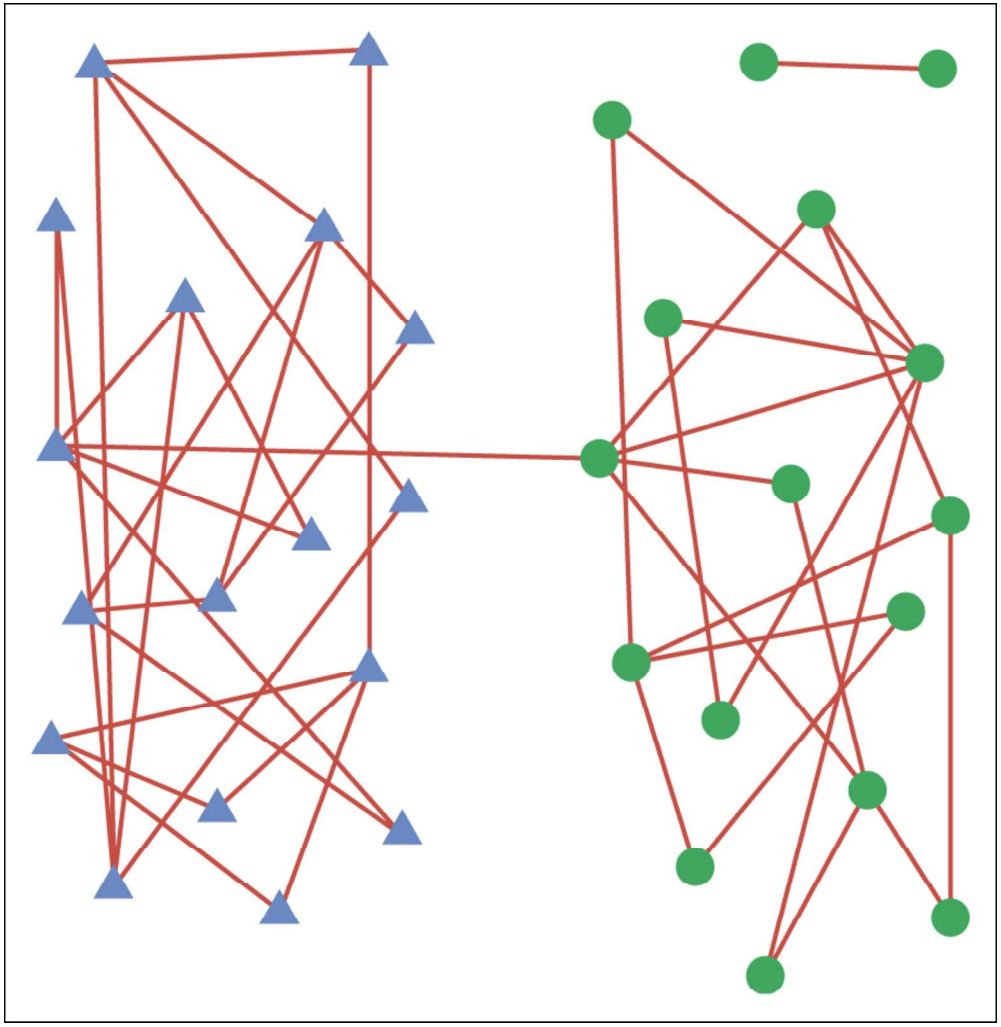


Please do not abuse of radial chart



Think of the data!

Social networks



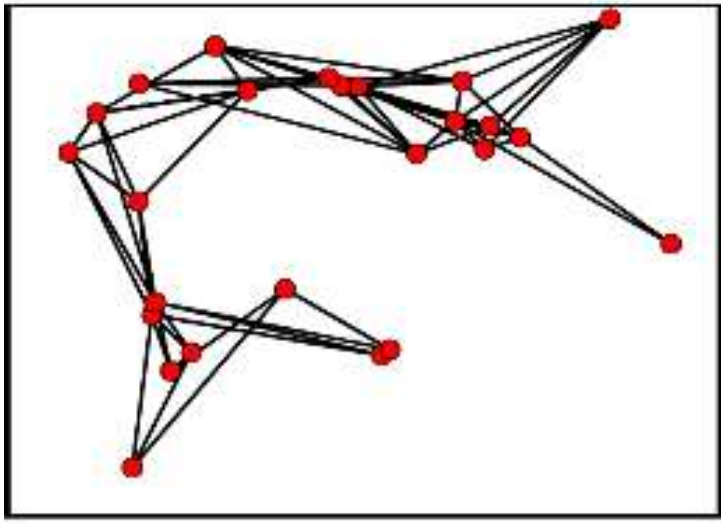
Social choices of fourth grade students in a school

boys chose boys

girls chose girls

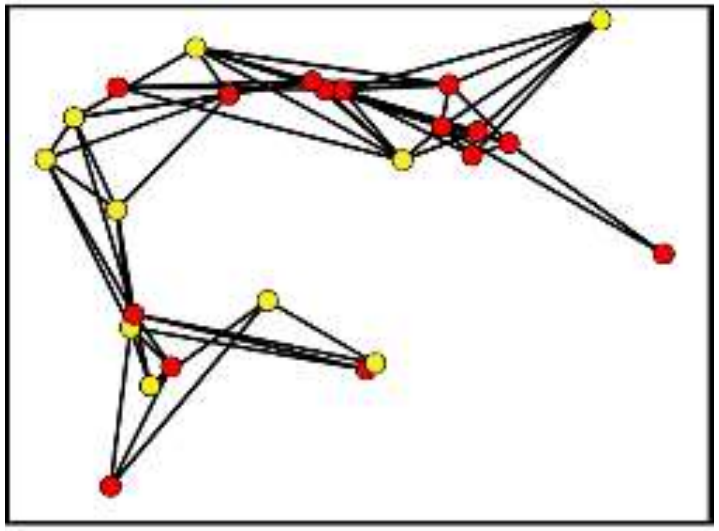
....

Social networks



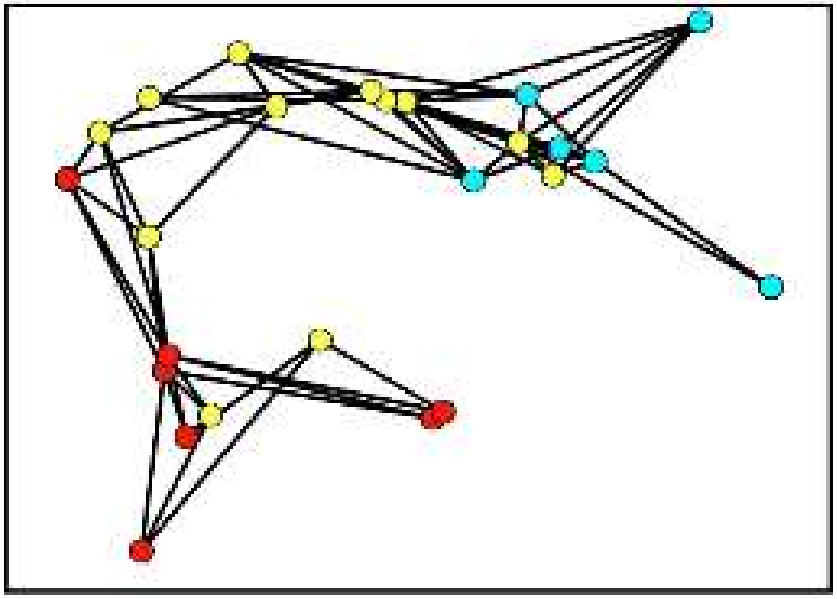
Social choices of
employees in
a department store
during recreational activities
(coffee break)
Data is collected using electronic
badges or videos

Reason for interacting: social networks + color



gender,
ethnicity,
marital status ?
no...

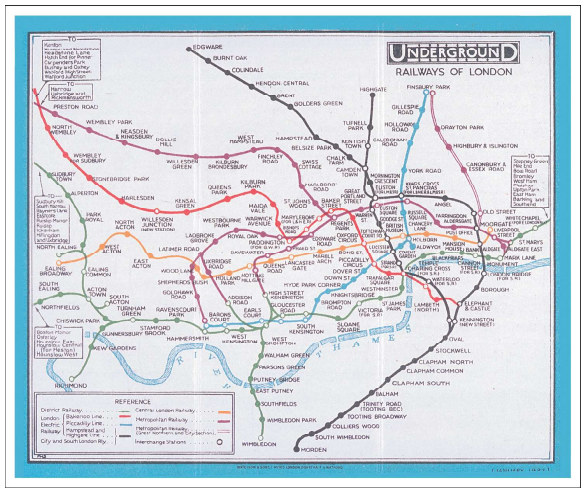
Social networks + color



Age!

(blue <30, 30 <yellow <40, red >40)

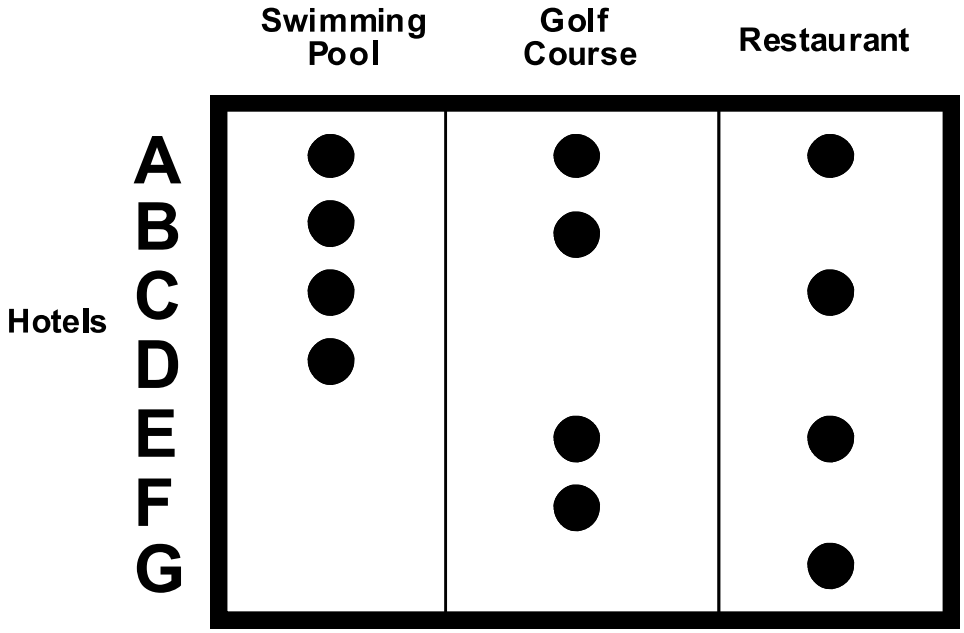
Do not forget physical connection...

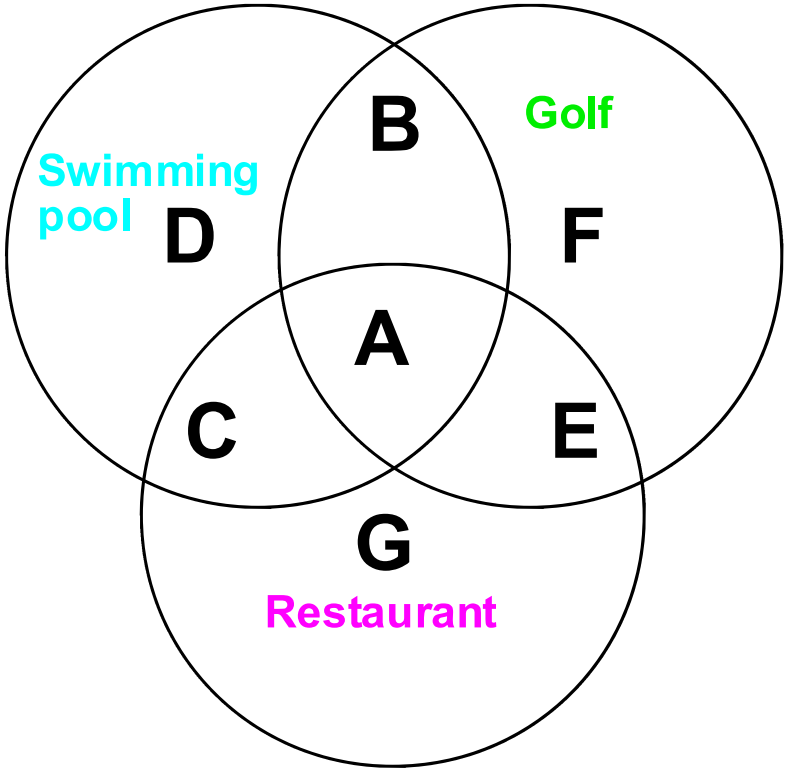


Outline

- Data types & data complexity
- Encoding of values
  - Univariate data
  - Bivariate data
  - Trivariate data
  - Multidimensional data
- Encoding of relations
- Lines
- Map & Diagrams
- Trees
- Support for design

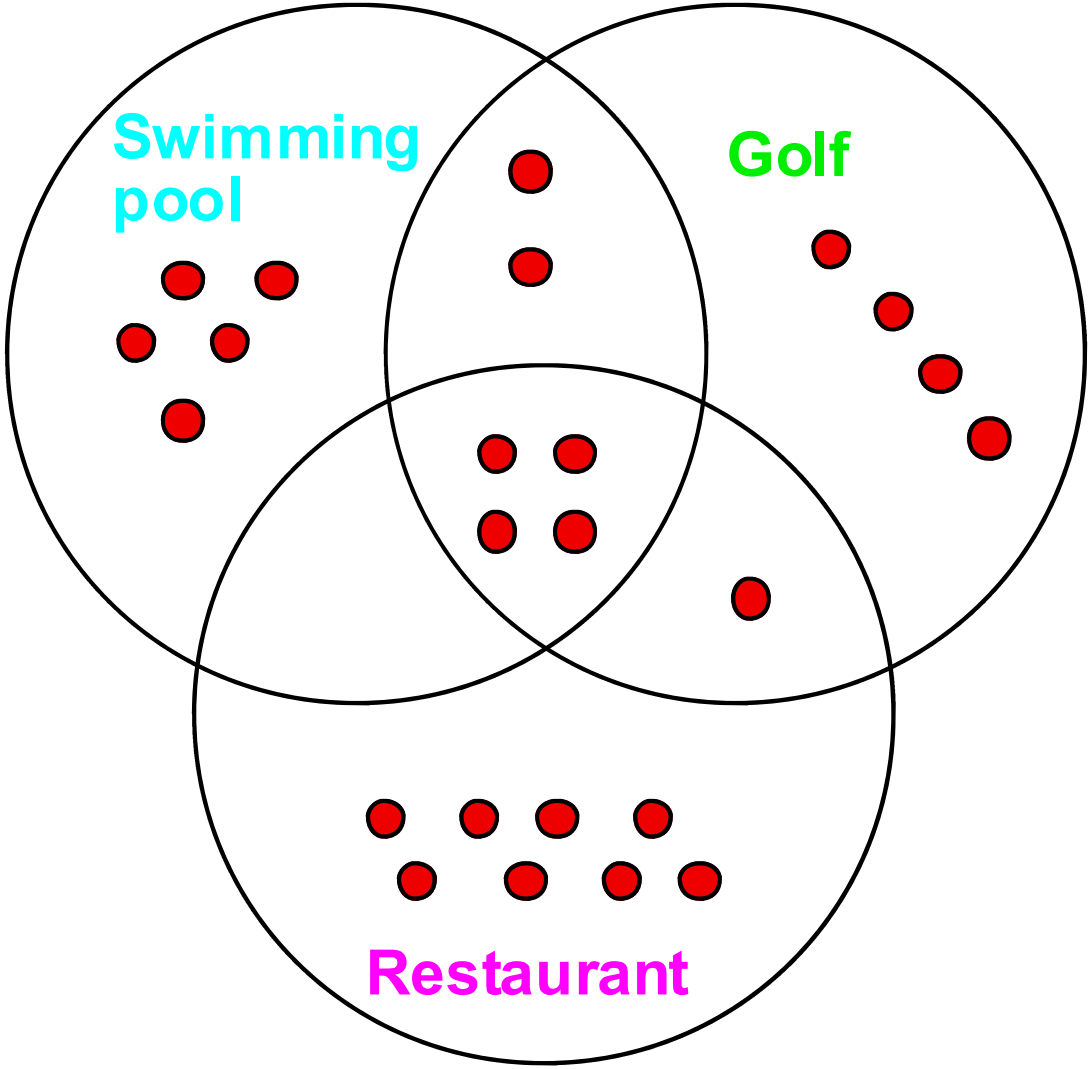
Maps & diagrams



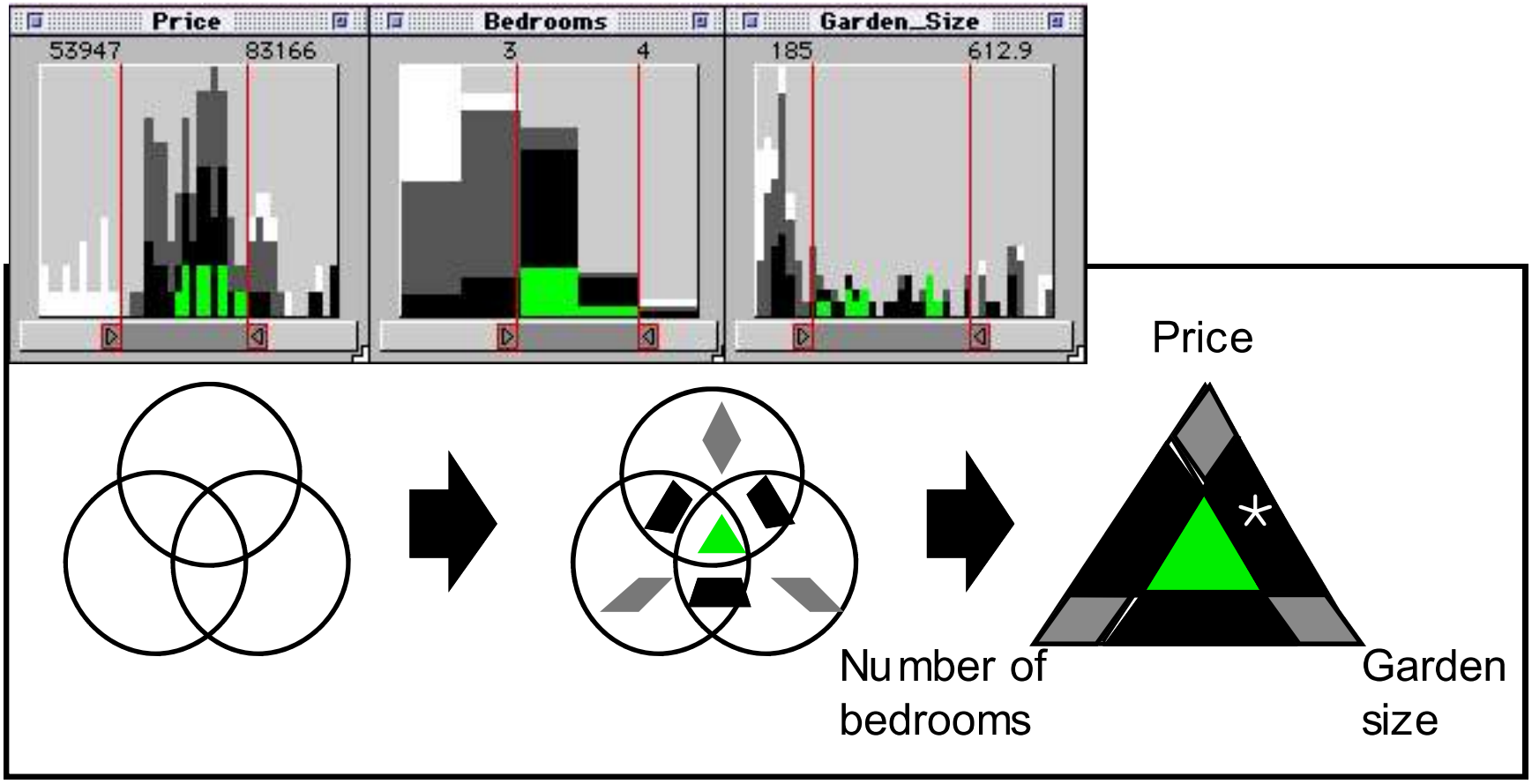


A Venn diagram might help

More hotels

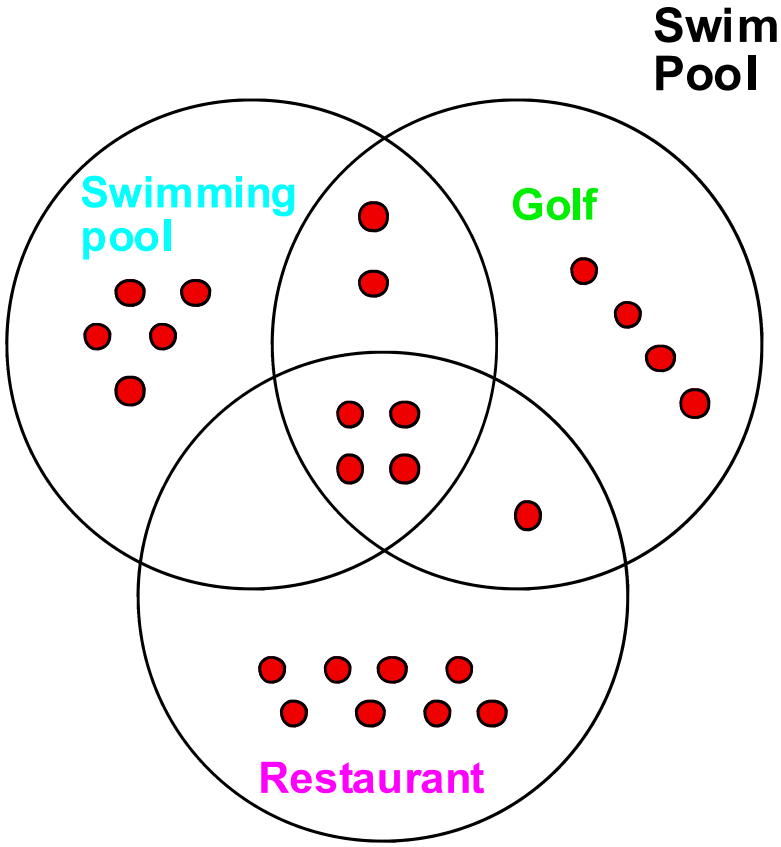


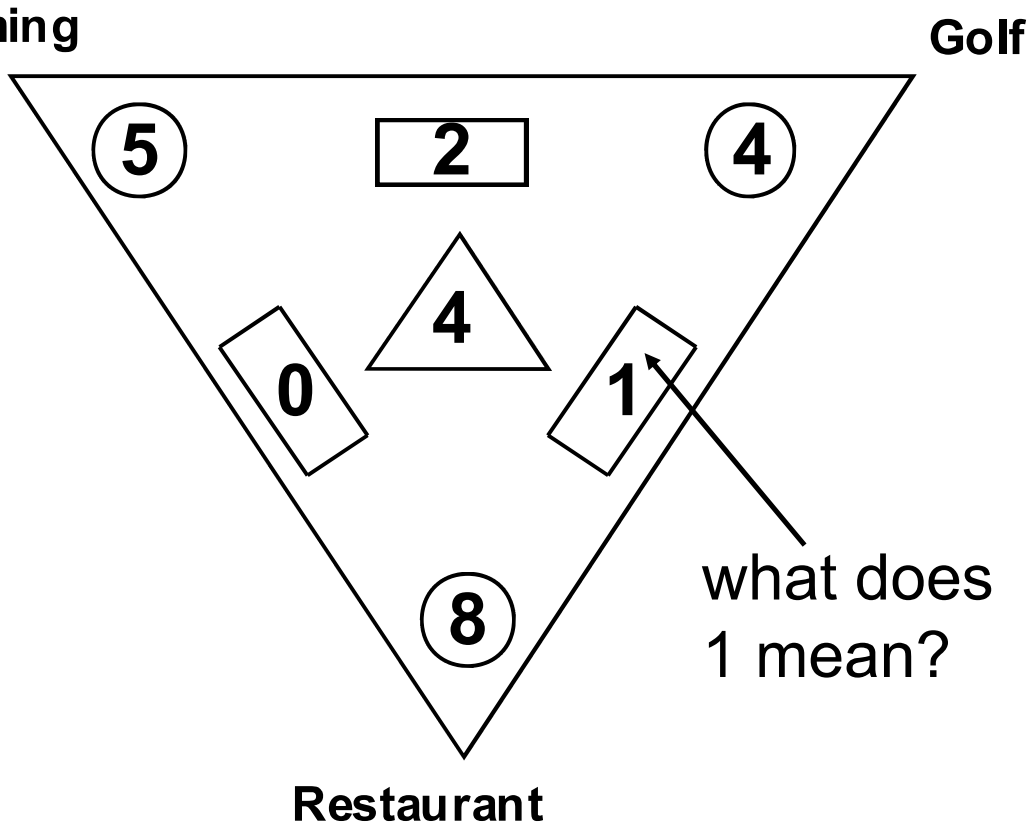
Representing Venn diagrams: InfoCrystal



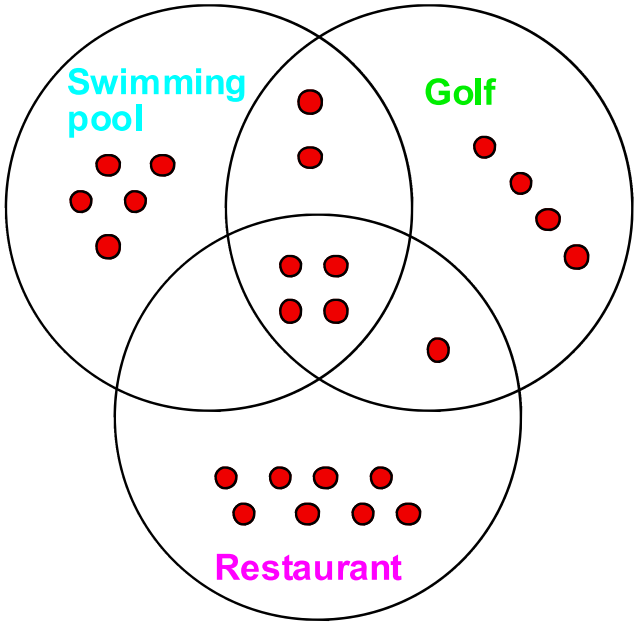
InfoCrystal allows visual queries to be made concerning price, garden size and number of bedrooms. The asterisk represents houses satisfying criteria on price and garden size but not number of bedrooms

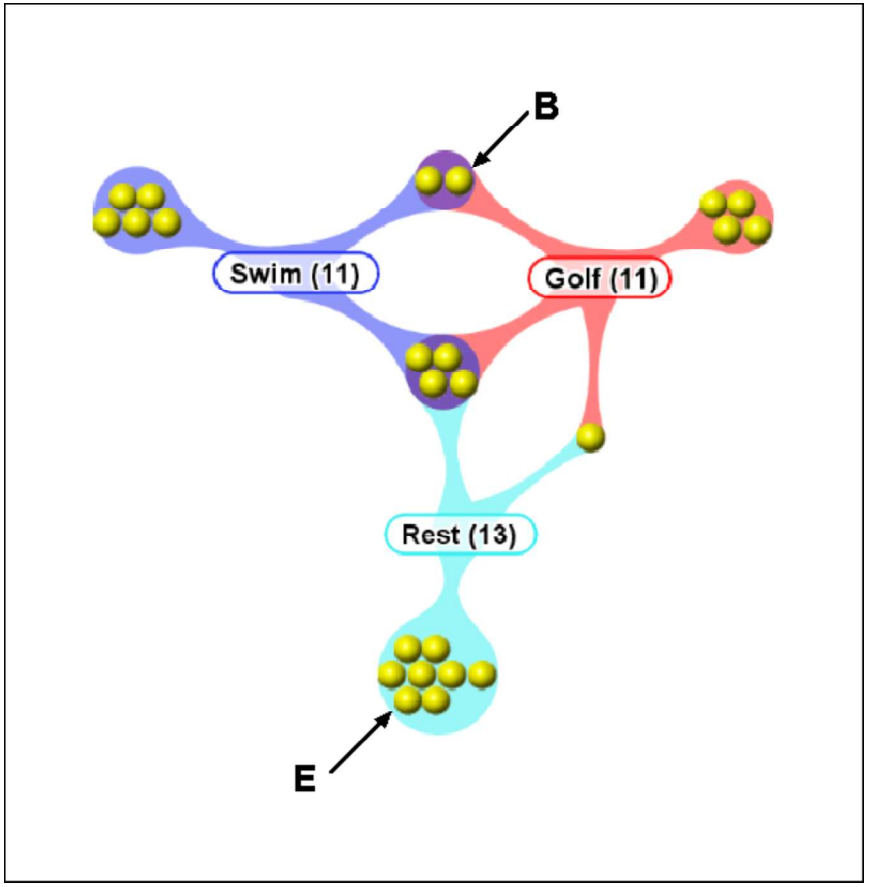
Back to the hotels



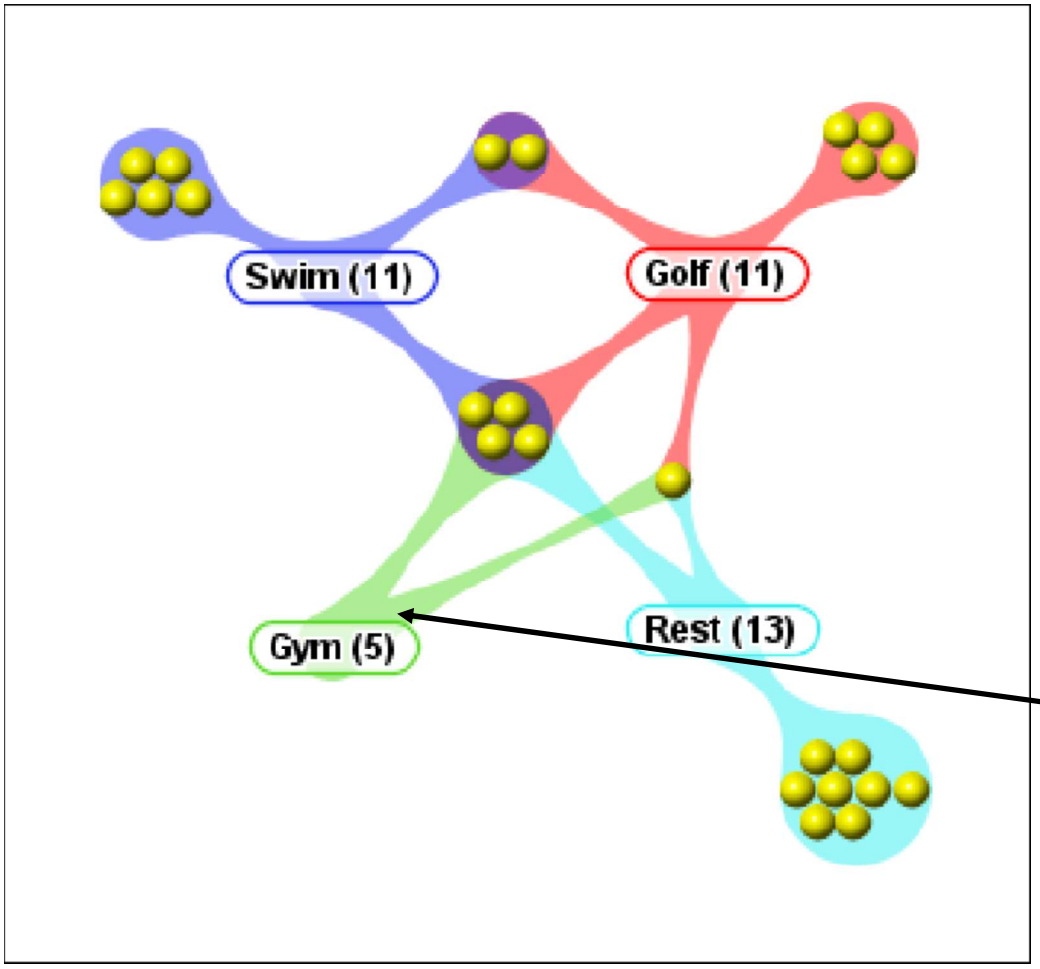


Cluster maps





Cluster maps



While Venn diagrams and InfoCrystal do not scale Cluster Maps do!

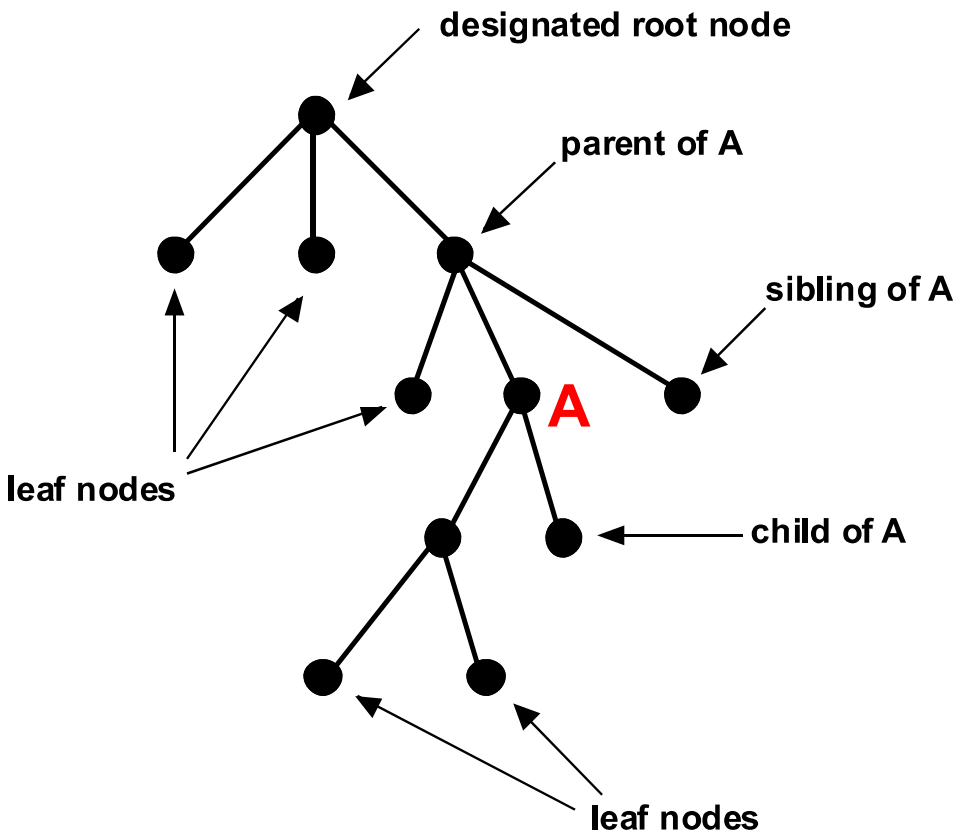
Empty?
If a hotel has Gym it always has another facility

Outline

- Data types & data complexity
- Encoding of values
  - Univariate data
  - Bivariate data
  - Trivariate data
  - Multidimensional data
- Encoding of relations
- Lines
- Map & Diagrams
- Trees
- Support for design

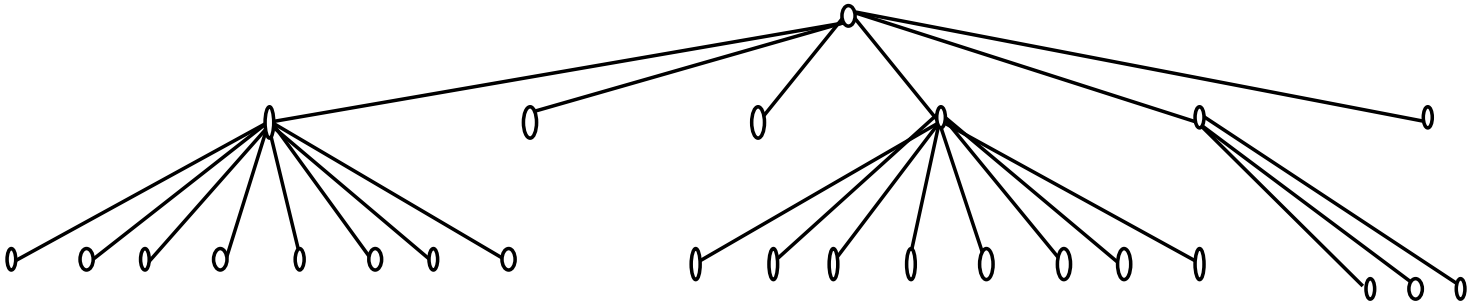
Tree representation

- Tree are a particularly interesting kind of graphs (a lot of data relationships have a hierarchical structure)



Trees are hard to draw

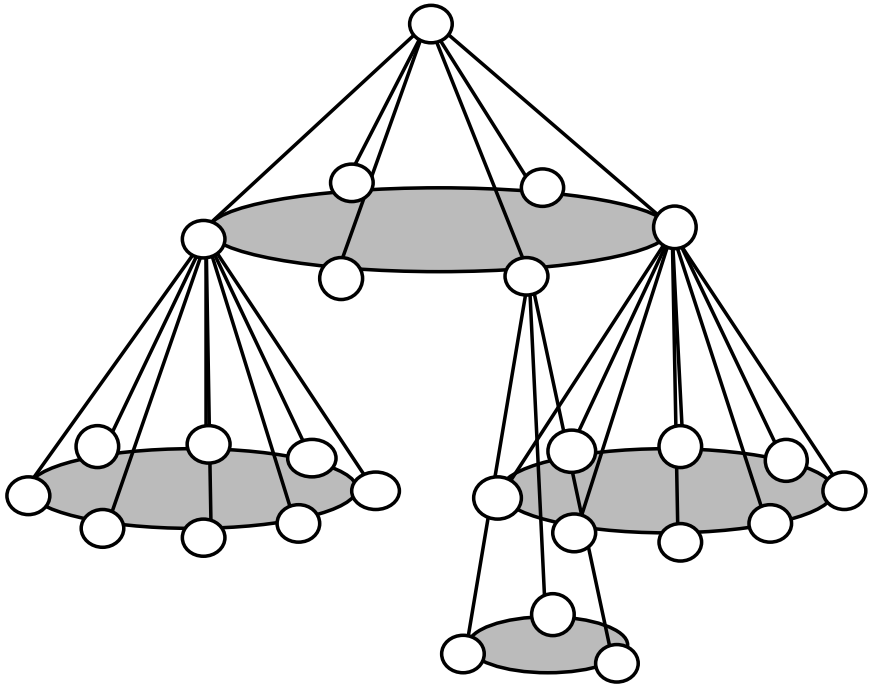
- They require more horizontal then vertical space



- There exist a lot of proposals for automatically draw trees (and graphs)

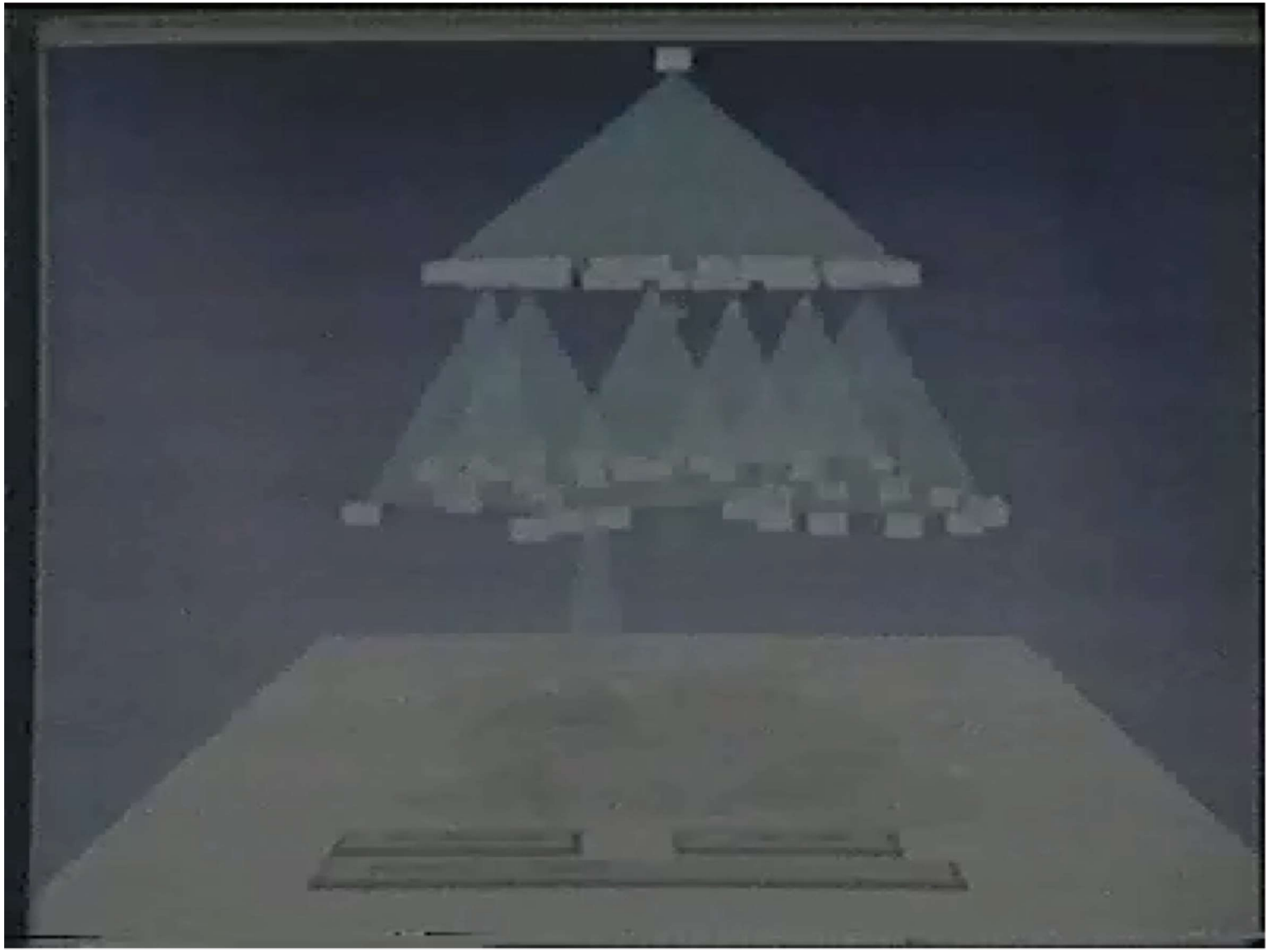
A look to the past: 3D trees

- All nodes subordinate to a given node are arranged into a 3D cone
- More compact
- Occlusions: it requires strong interaction support

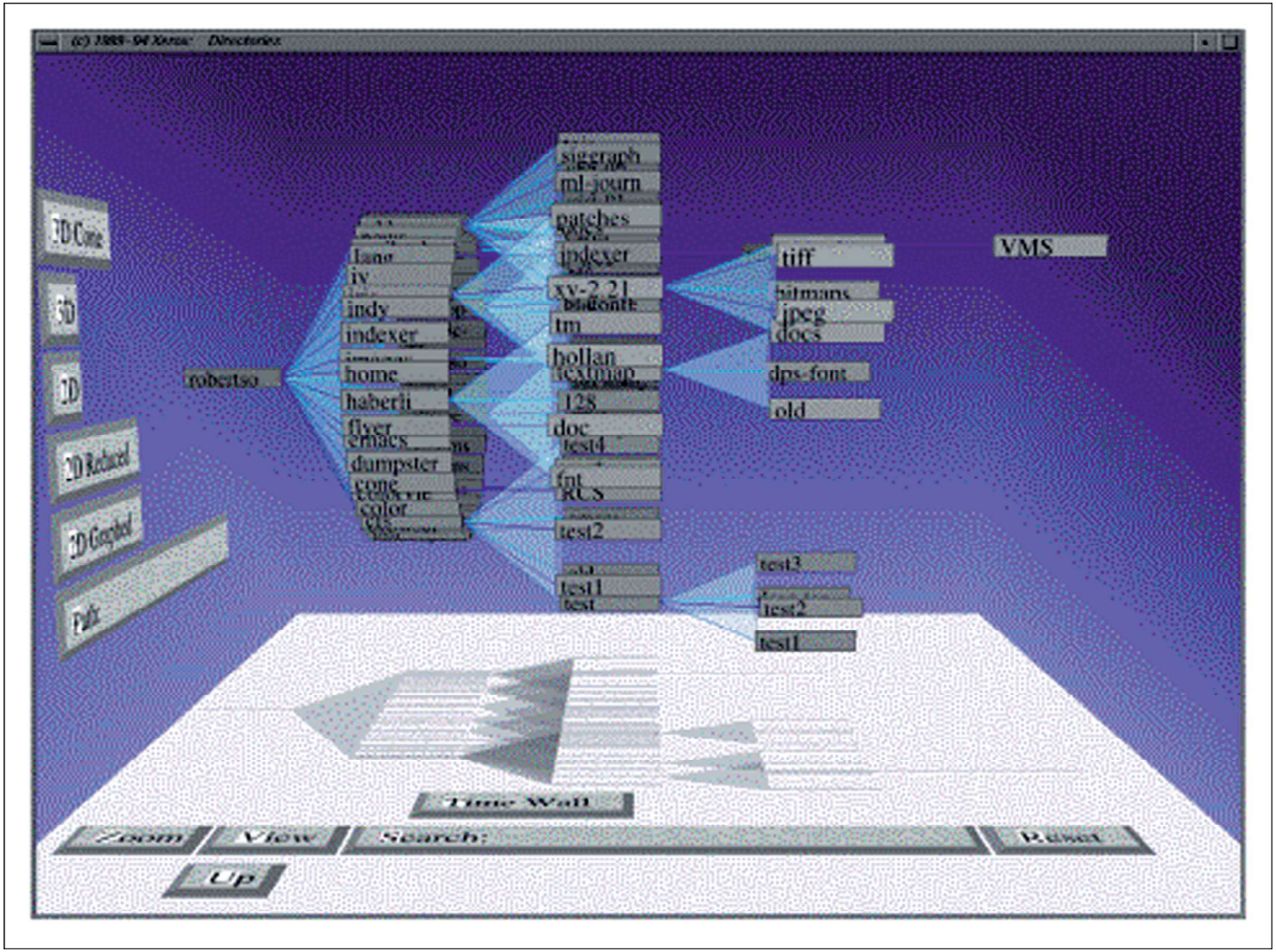


Note the animation

A look to the past: 3D trees

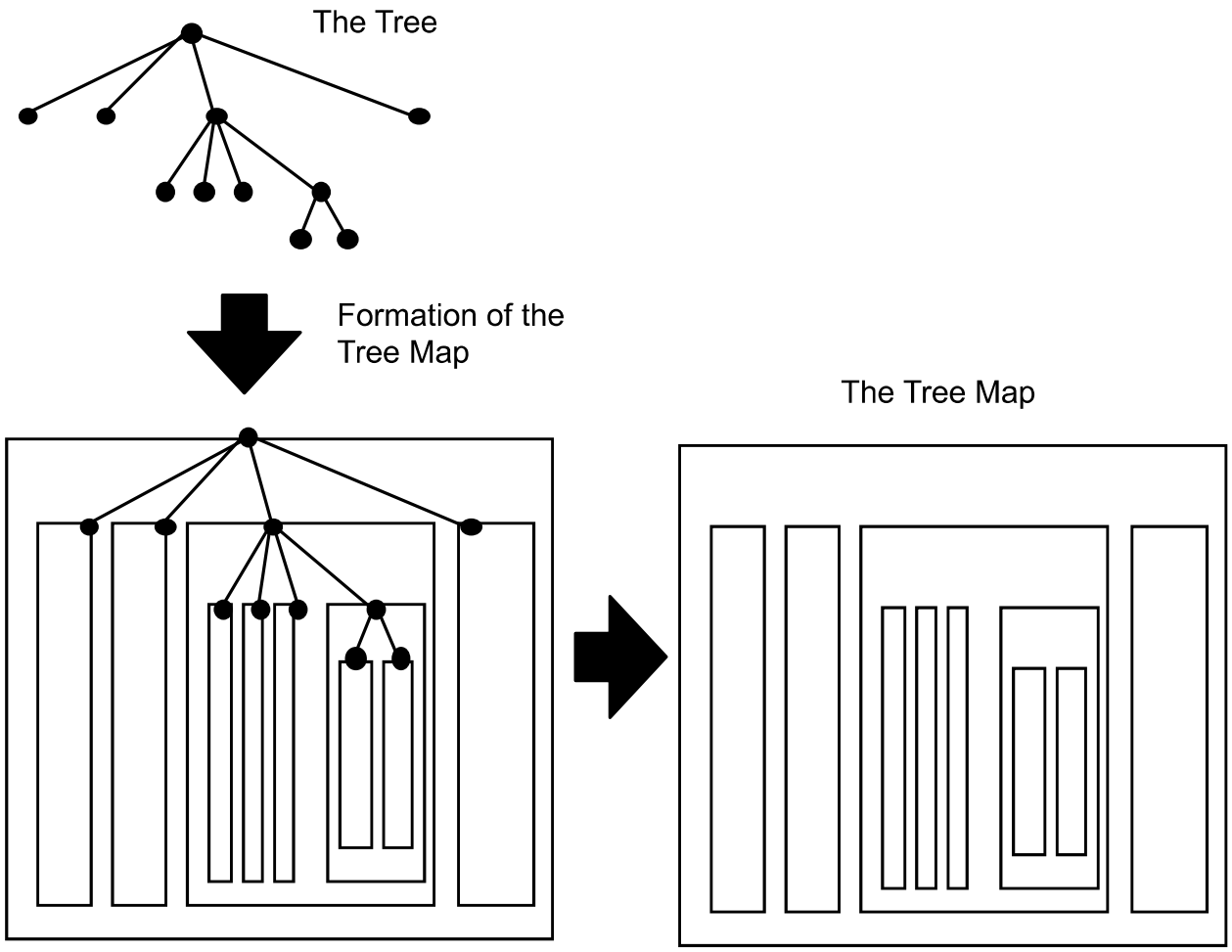


A look to the past: 3D trees



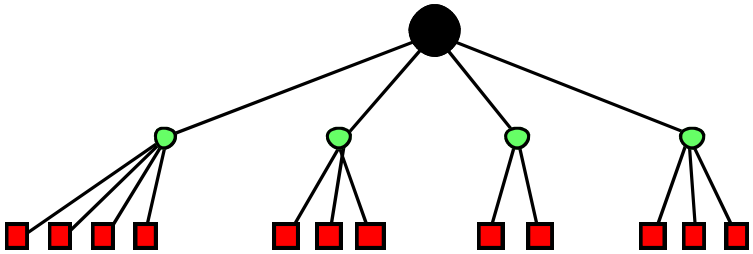
A reorientation, more convenient for the textual labelling

Tree maps



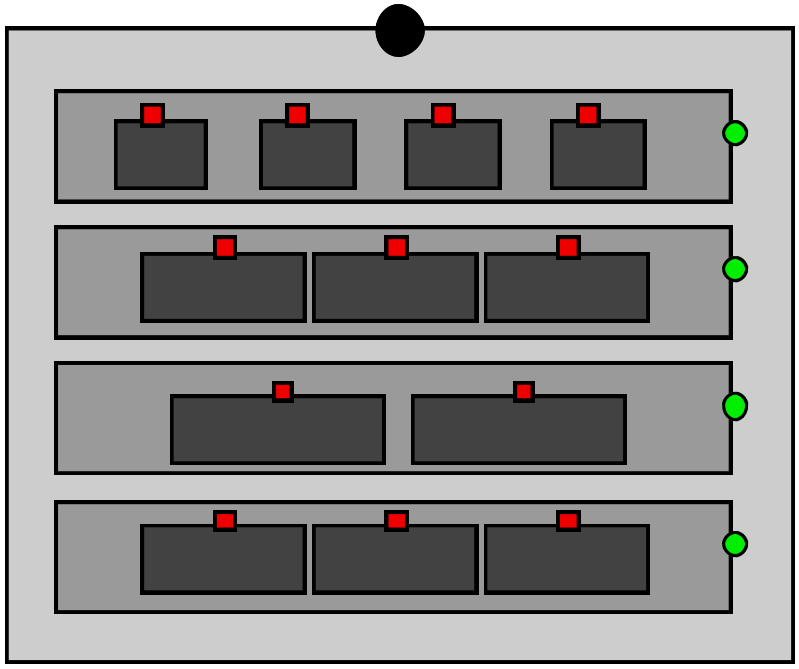
Tree maps

- Slice-and-dice , more suited for including text and images



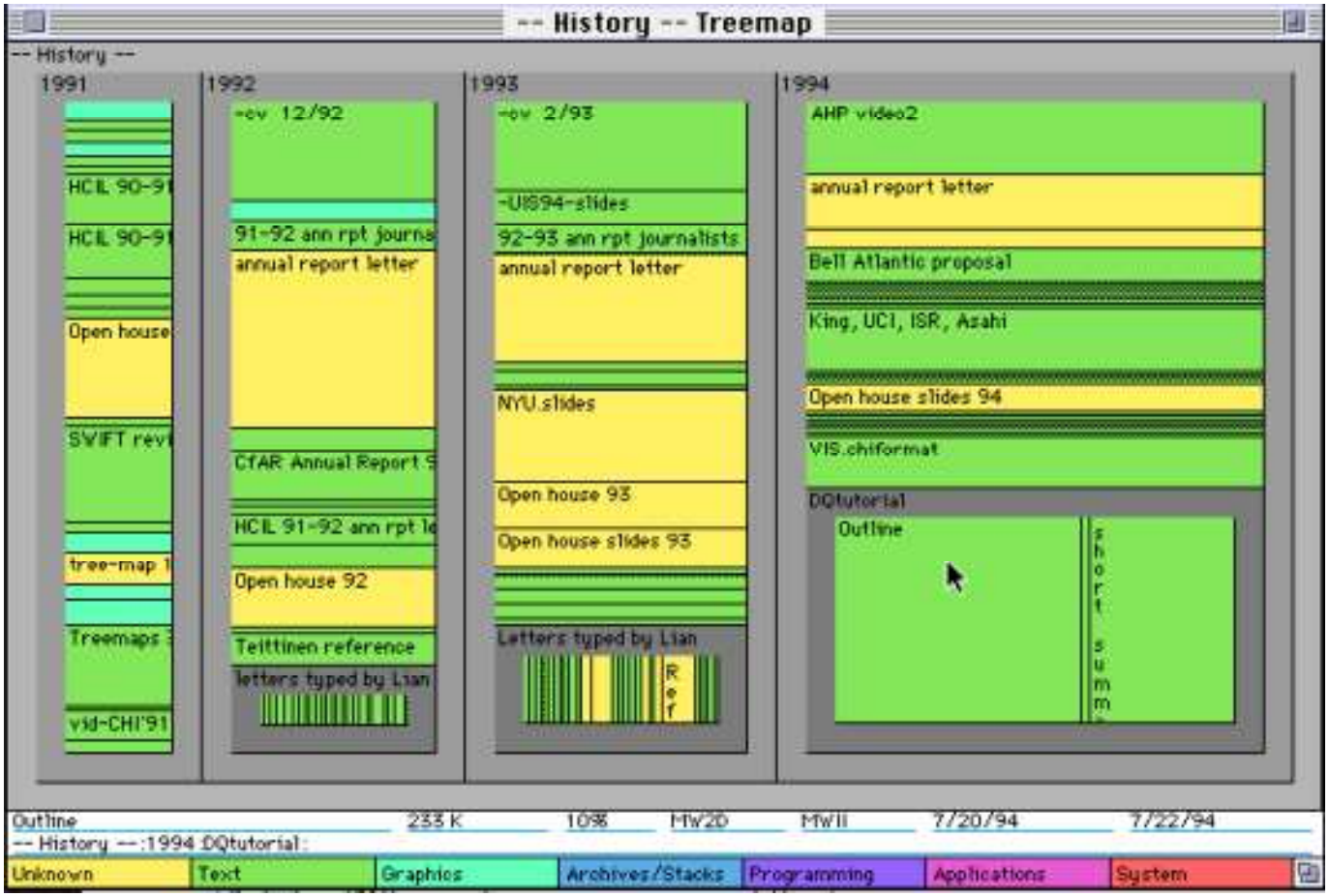
Tree

Tree Map

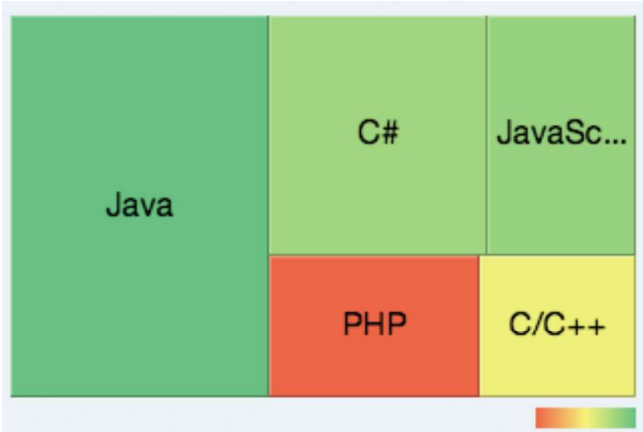


Tree maps

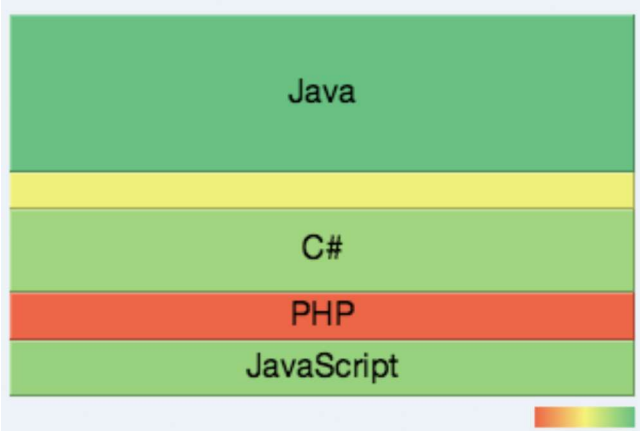
- Vertical and horizontal alternation
- Hierarchy is not easy to discern



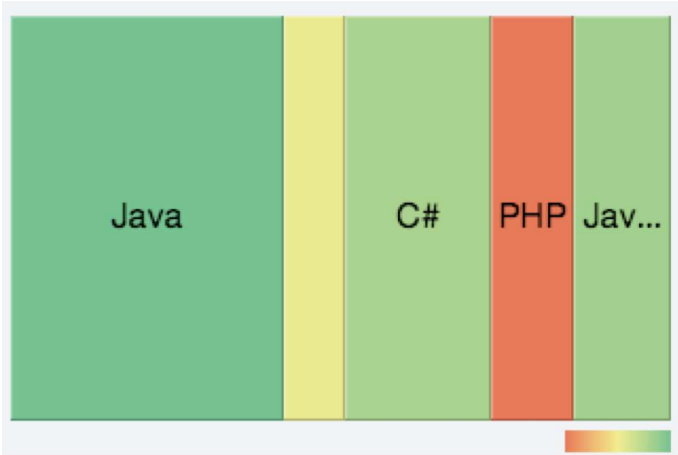
Treemaps styles



Squarified



Slice and dice vertical



Slice and dice horizontal

Paper

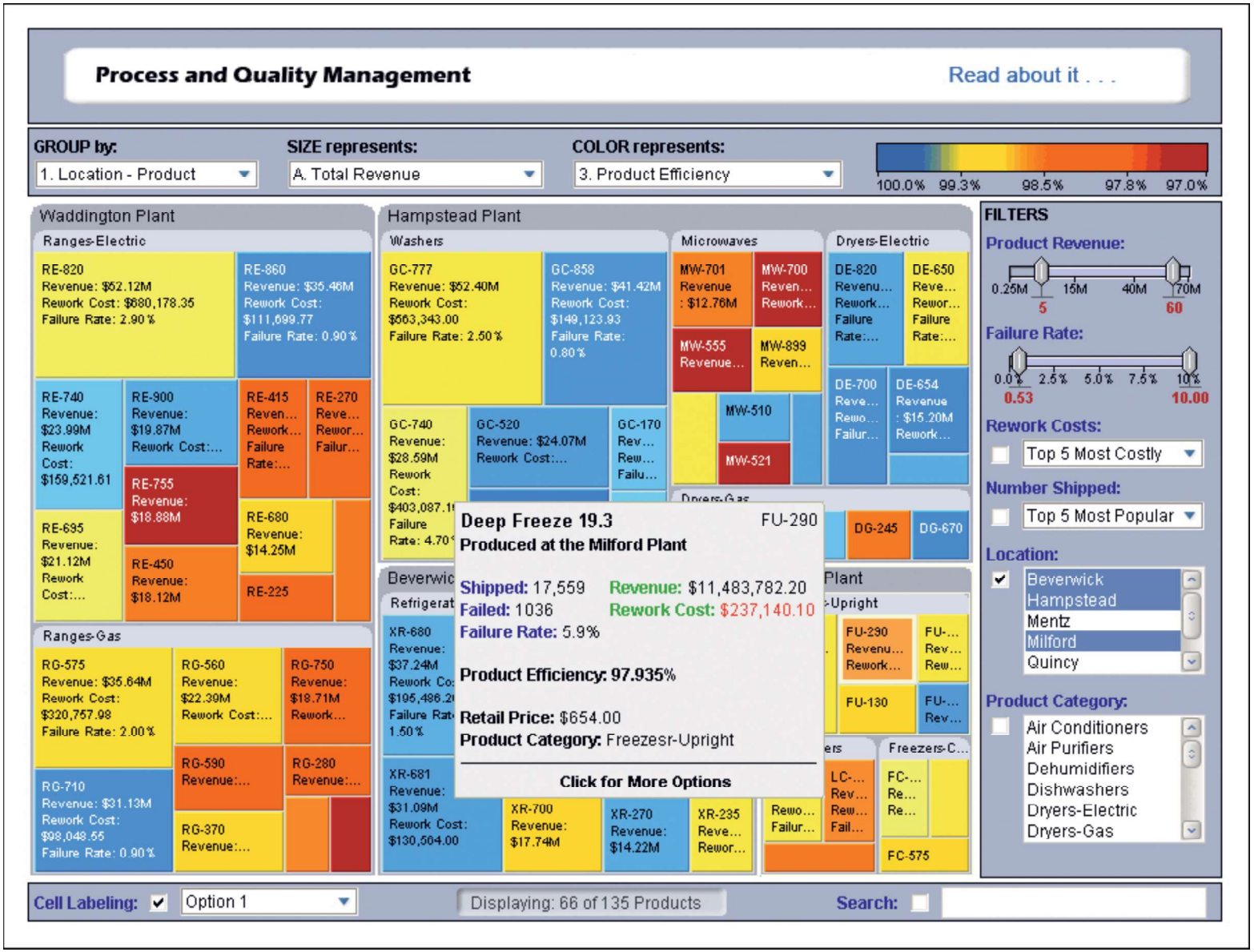
Ordered Treemap Layouts Ben Shneiderman 2001

Treemaps, a space-filling method of visualizing large hierarchical data sets, are receiving increasing attention. Several algorithms have been proposed to create more useful displays by controlling the aspect ratios of the rectangles that make up a treemap. While these algorithms do improve visibility of small items in a single layout, they introduce instability over time in the display of dynamically changing data, and fail to preserve an ordering of the underlying data. This paper introduces the ordered treemap, which addresses these two shortcomings. The ordered treemap algorithm ensures that items near each other in the given order will be near each other in the treemap layout. Using experimental evidence from Monte Carlo trials, we show that compared to other layout algorithms ordered treemaps are more stable while maintaining relatively favorable aspect ratios of the constituent rectangles. A second test set uses stock market data.

Smartmoney

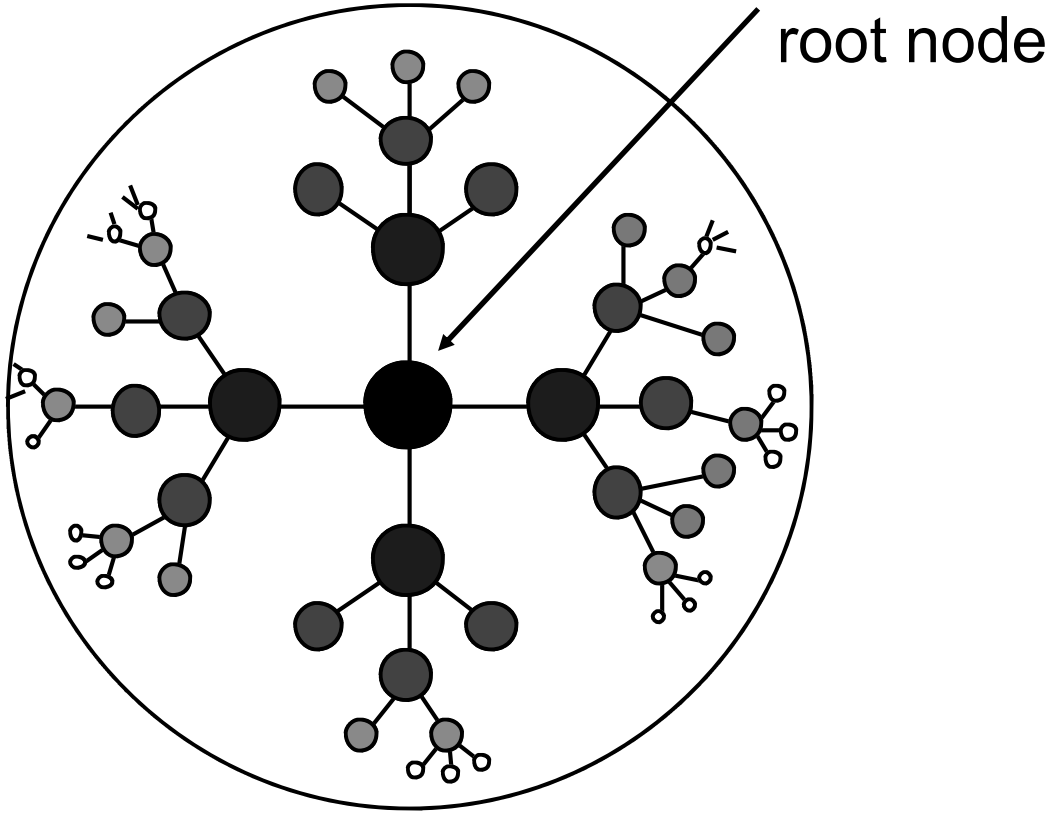

Health Care
Financial
Energy
Utilities
Basic Mat...
Capital Goods
Transport
Consumer Cyclicals
Technology
Communication
Consumer Staples

Treemap + color + filtering



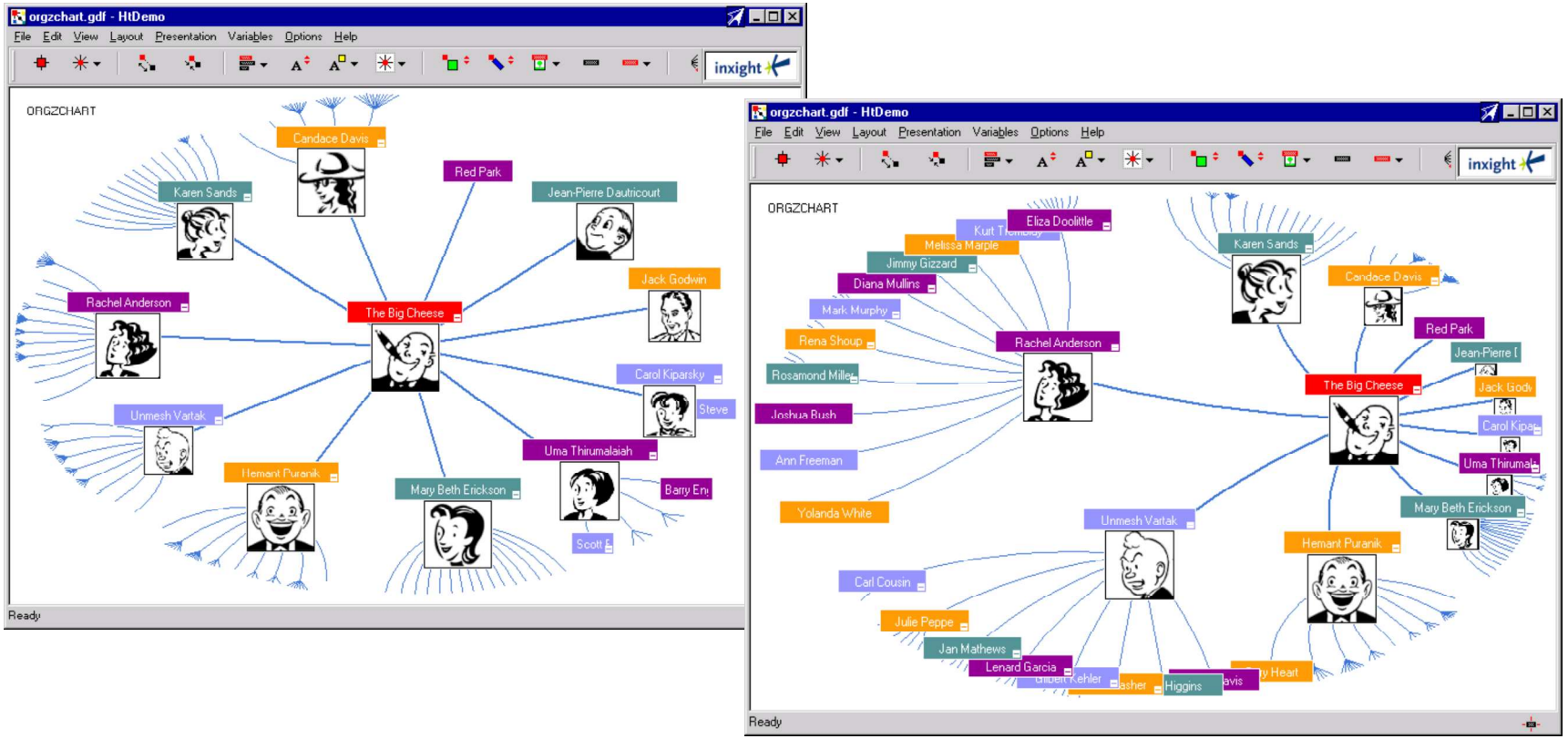
Hyperbolic browser

- Rescuing (more or less) the tree design



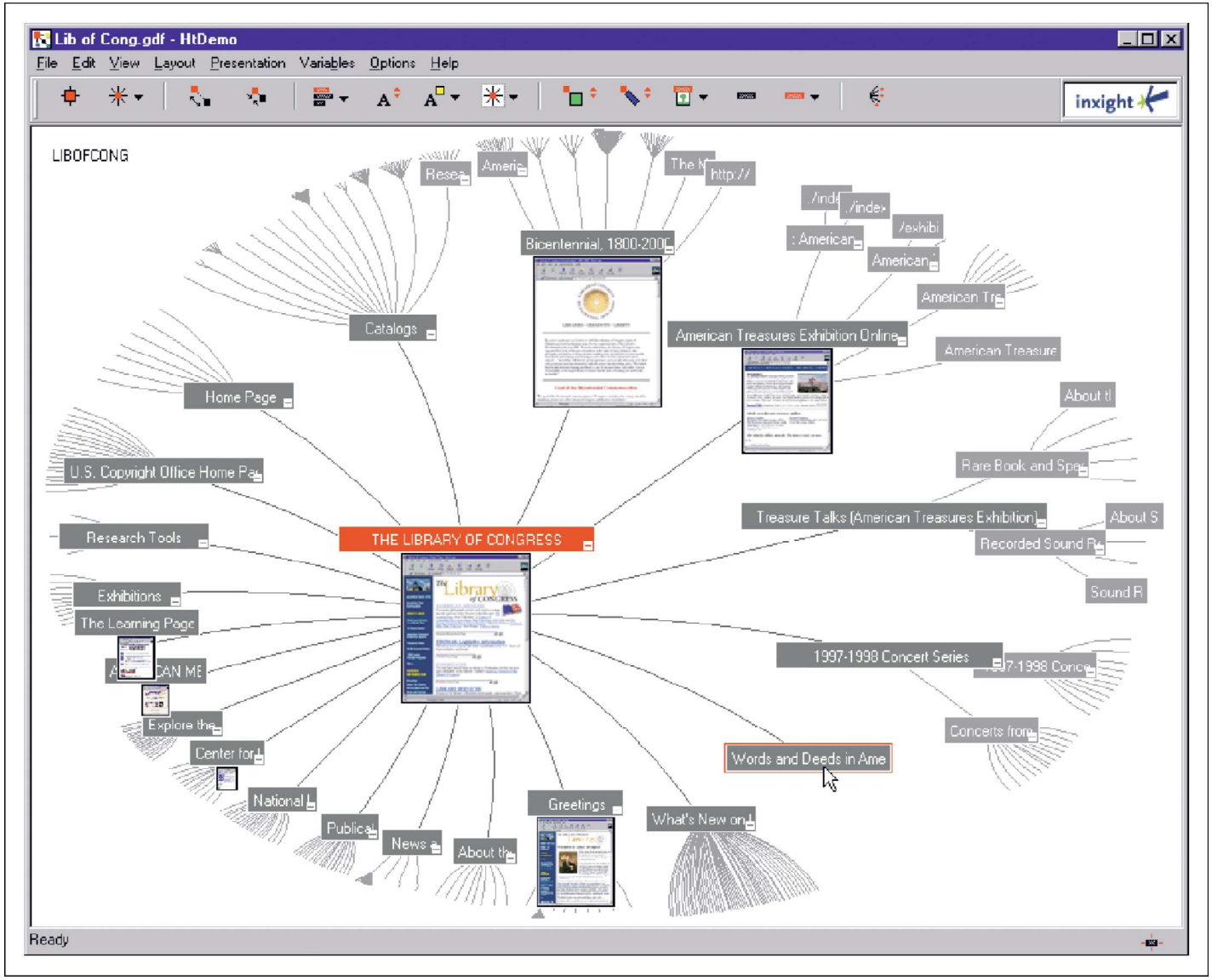
The further away a node is from the root node, the closer it is to its superordinate node, and the area it occupies decreases
The limit is the pixel...

Interaction !



(a) The reporting structure of the employees of a company. (b) One employee of interest, Rachel Anderson, has been moved towards the centre, revealing her subordinates

The Library of Congress

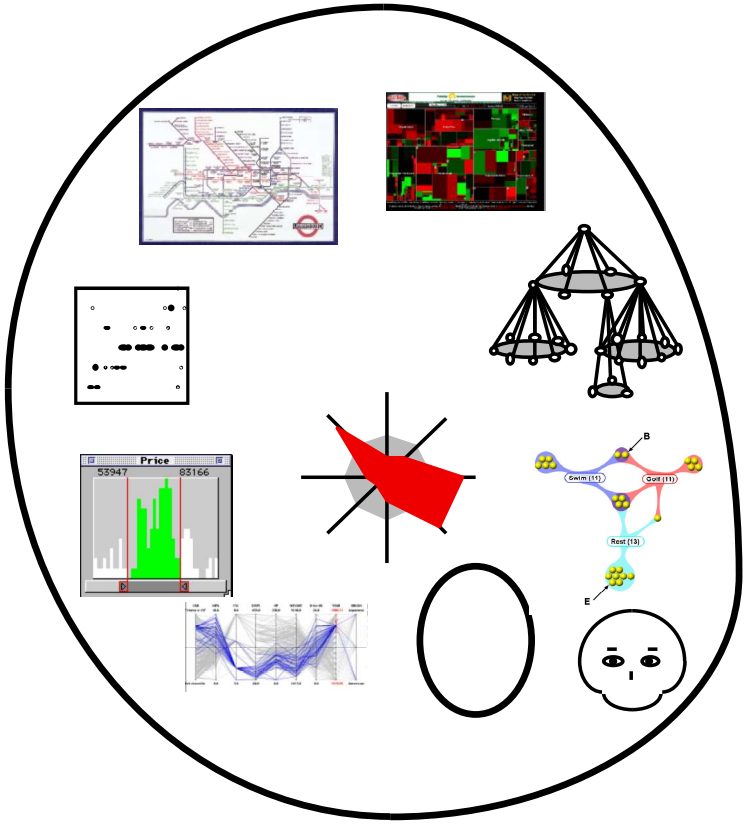


Hyperbolic browser original idea

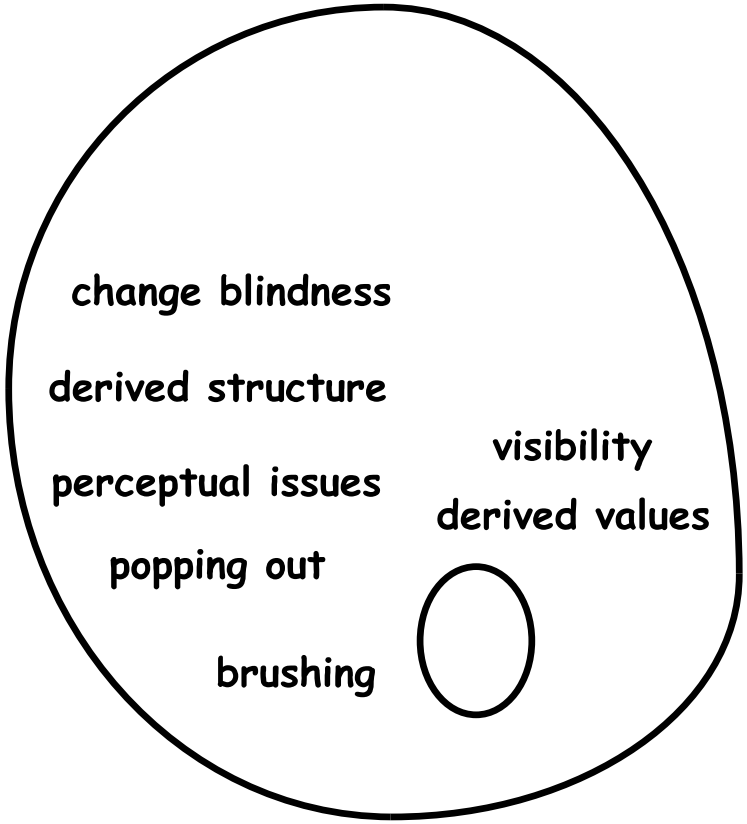
Outline

- Data types & data complexity
- Encoding of values
  - Univariate data
  - Bivariate data
  - Trivariate data
  - Multidimensional data
- Encoding of relations
- Lines
- Map & Diagrams
- Trees
- Support for design

Representation design palettes



Techniques



Concepts