

Visualization in hyperspace: making visual inferences for multivariate data

VOTech/University of Leeds
Richard Holbrey

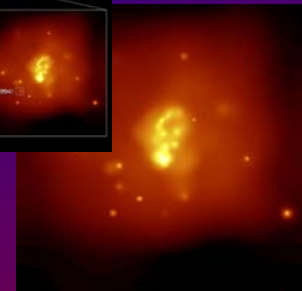
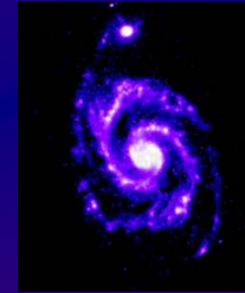
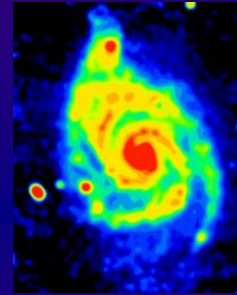
Data mining ...



- Opportunity to develop large RDBs in 90s
- Commercial push to gather customer data
 - Huge 2D tables

Tesco's had one

- Astros had to have one too
- Aim to do more multi- λ astronomy
 - Radio, IR, optical, X-ray ...
- Result
 - More huge 2D tables



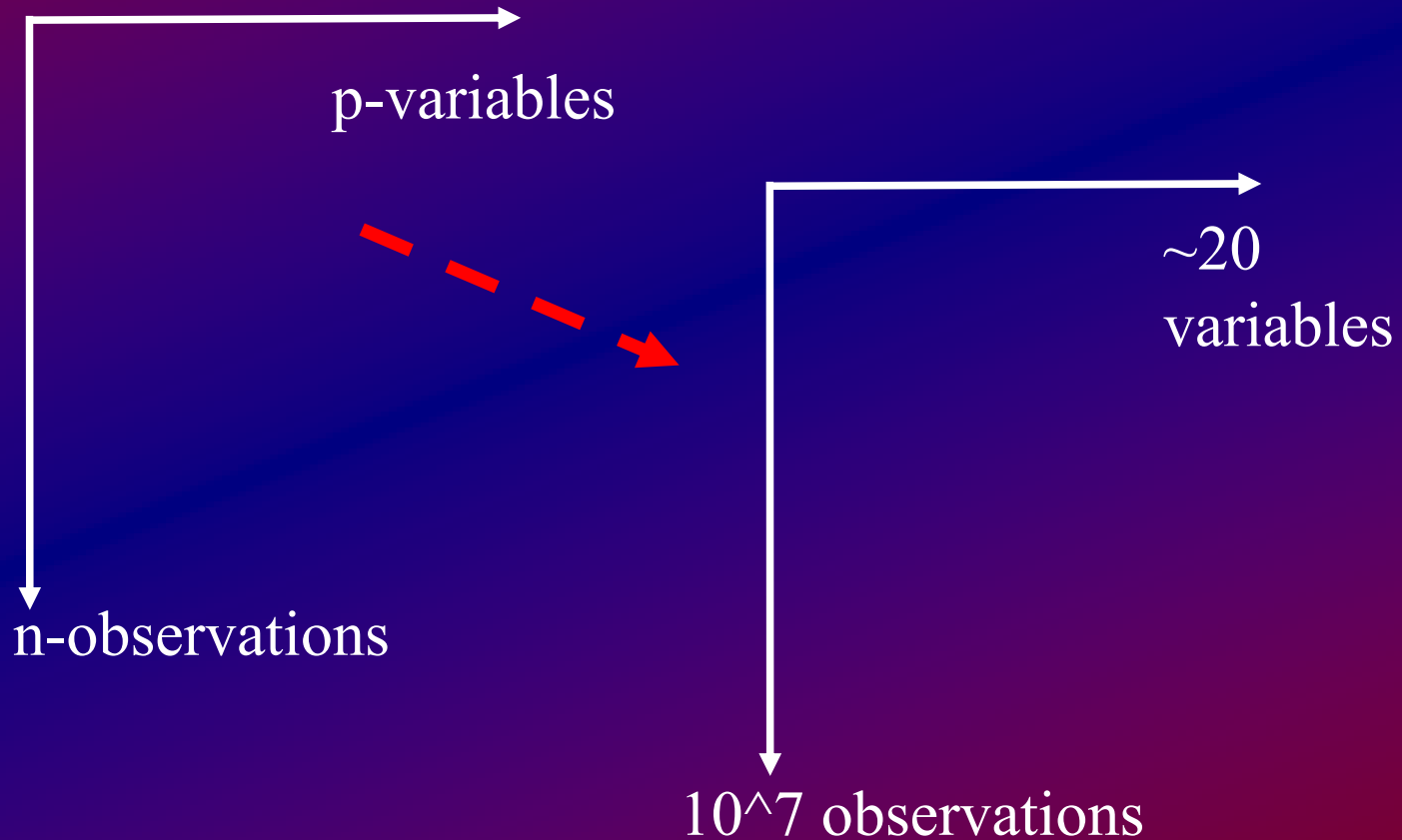
in other words, hyperspace



Um, not really...

- Multivariate data forms hyper-sphere/cube
- Hyperspace very counter-intuitive
 - diagonals disappear
 - most data appears in a thin shell at the boundary of the hypersphere
 - projections to 2D/3D could be misleading
 - data density will be skewed

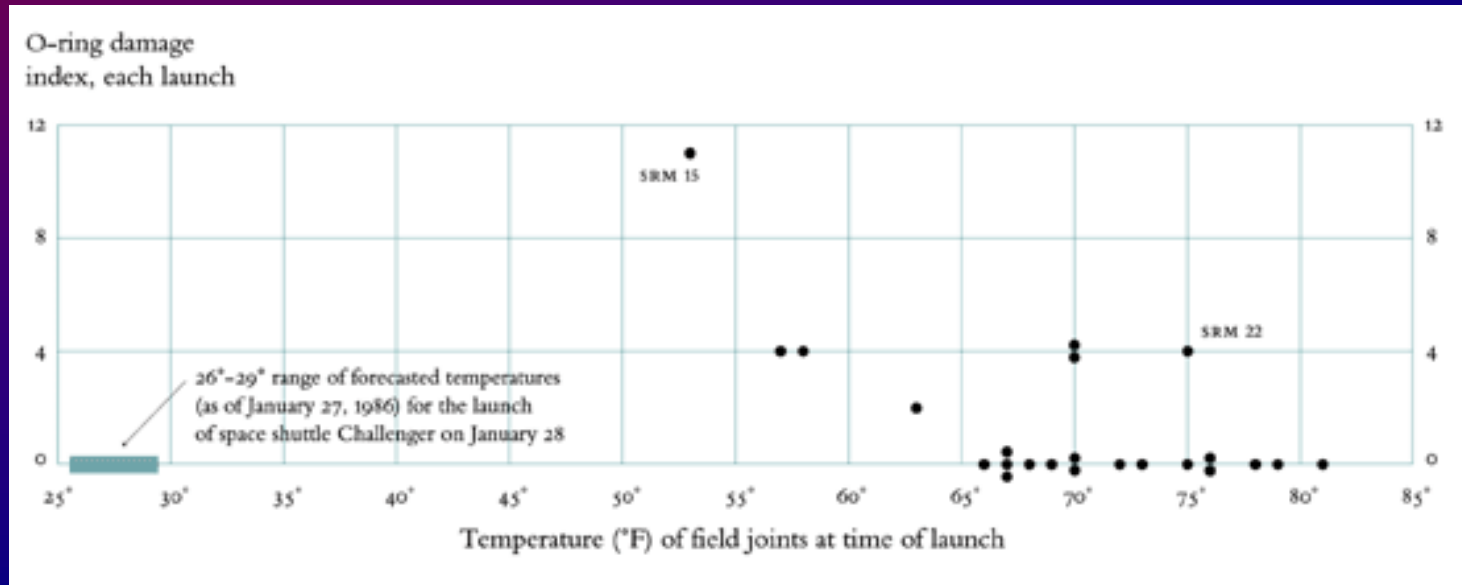
Problem definition



Looking for inspiration

- Data graphics field developed within the realm of statistics
 - Parallel coordinates, glyphs
- Fisher:
 - “diagrams prove nothing, but bring outstanding features readily to the eye ... they are no substitute for critical tests, but are valuable in suggesting such tests and in explaining the conclusions founded upon them”
- Bertin: A 2D table is defined problem
 - needs organisation to see detail

Tufte's guidelines



– Juxtaposition

Causality

– Data-ink maximisation

Clear thinking & integrity

...and visualization ?

- Visualization was born in 1987 with McCormick NSF report
- Huge fillip from engineering developments
 - fluid dynamics, medical brain scanning
- Seemingly unrelated to statistics, but also justified as exploratory
- Difficult to get accurate visualizations – but some attempts

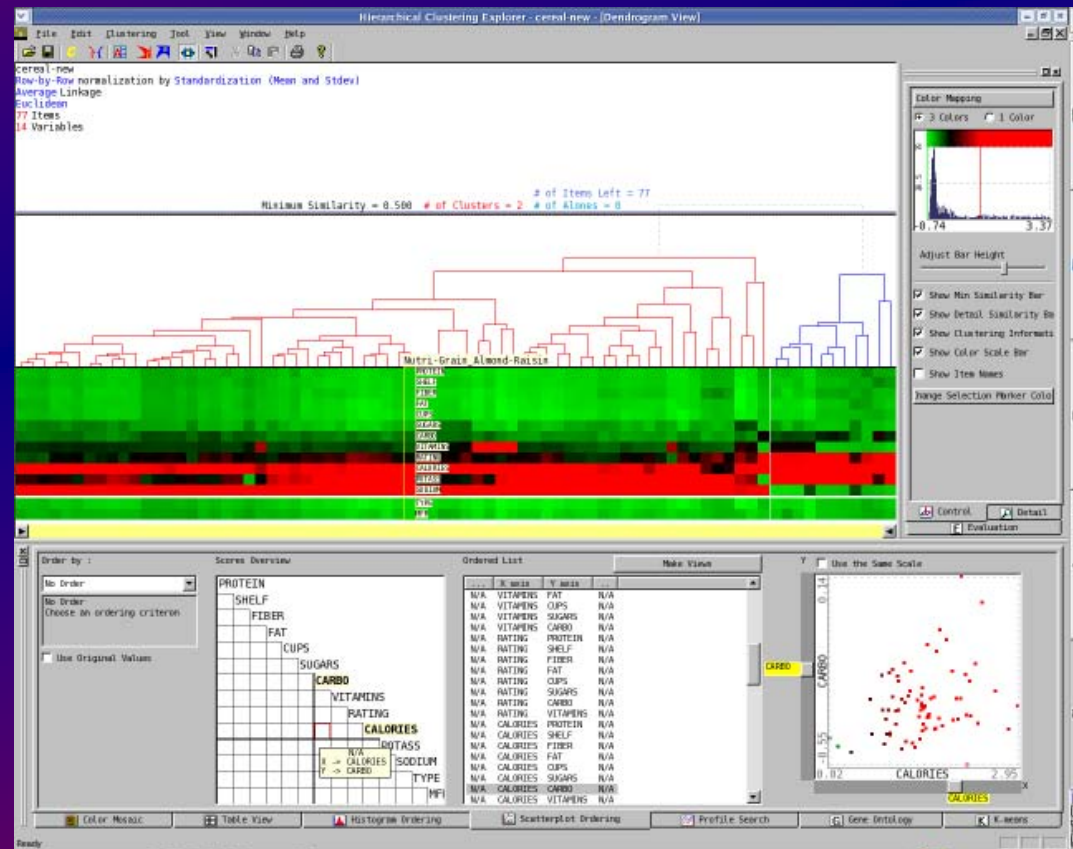
Task definition

| Table size | entries | $O(n)$ eg. μ , σ^2 etc. | $n \log(n)$ eg. quicksort | $O(n^2)$ eg. clustering |
|------------|-----------|---------------------------------------|------------------------------|----------------------------|
| small | 10^4 | 10^{-5} | 4×10^{-5} | 0.1 |
| medium | 10^6 | 10^{-3} | 6×10^{-3} | 1004.4 |
| large | 10^8 | 0.1 | 0.78 | 116 <i>days</i> |
| huge | 10^{10} | 10.02 | 100.1 | 3170 <i>years</i> |

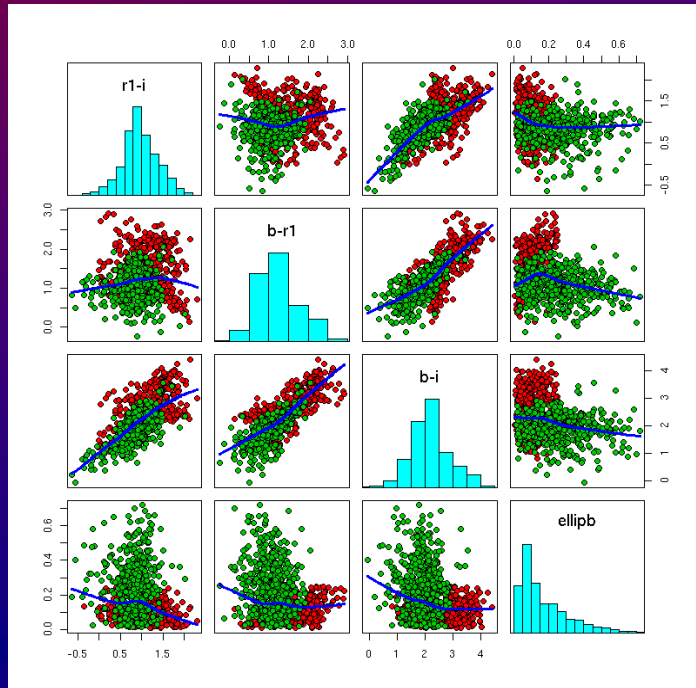
- Wegman and Solka's table (execution time in seconds)
- Combine these strands
 - huge data sets, data mining, effective visualization
 - bound interactivity and accuracy?

Small datasets

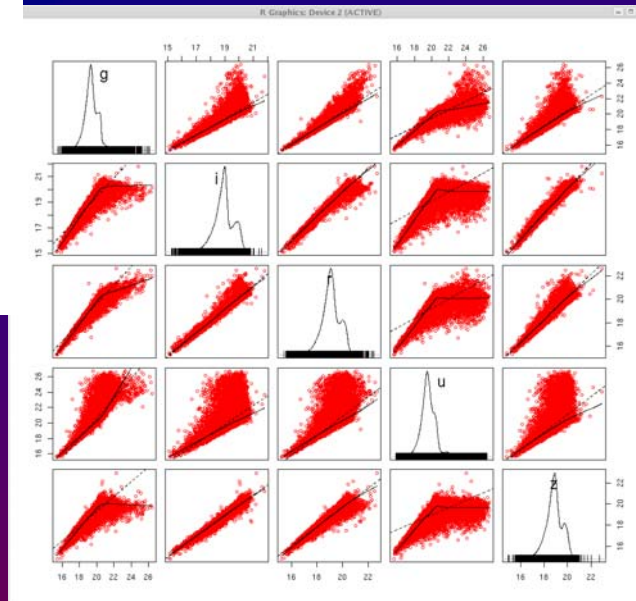
- 10^4 entries
- ~1MB
- Most apps fine
 - HCE, Mirage, xmdvtool, R ...
- HCE attempts to recommend variables
 - kurtosis, skew
 - 1D/2D histograms



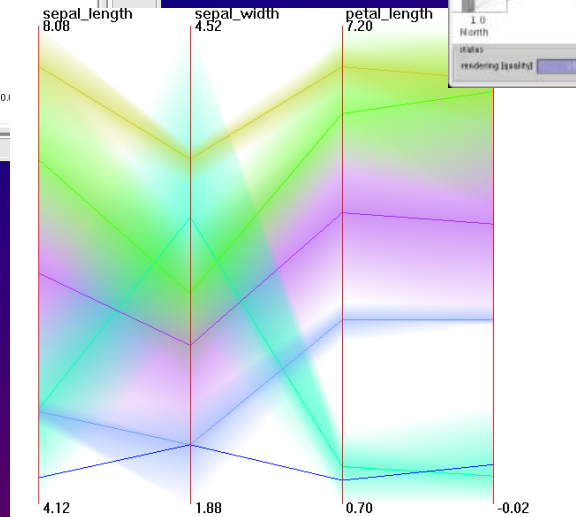
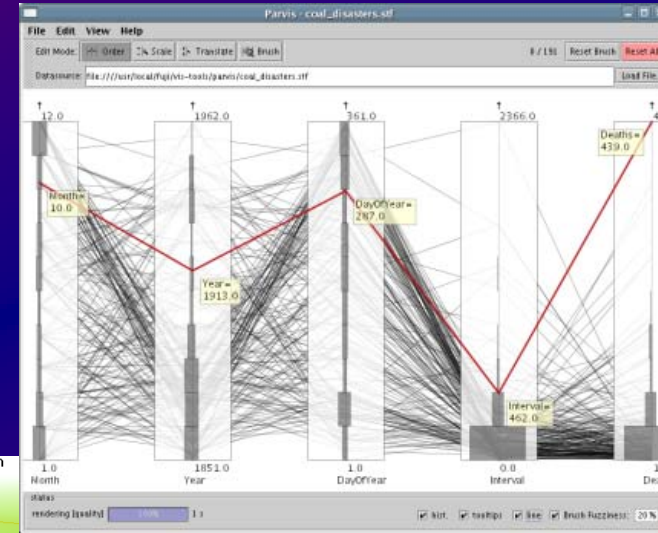
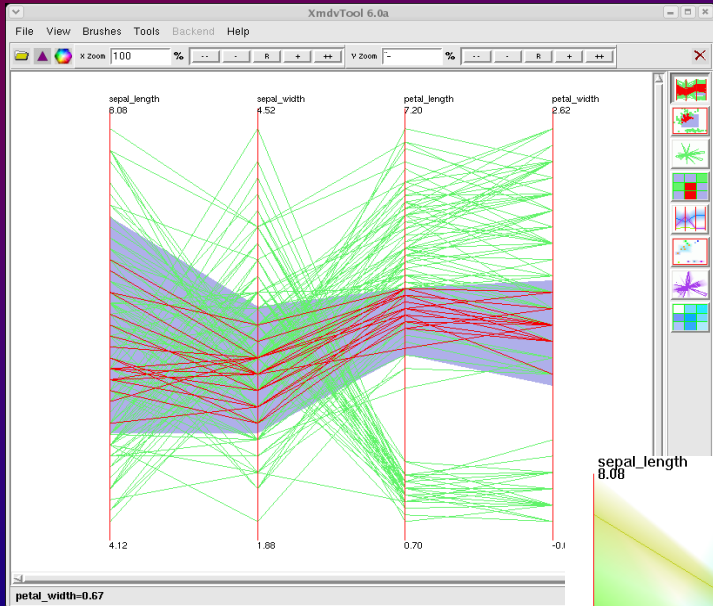
Tools & Techniques



- Scatterplots
 - colour, regression lines
 - density plots



Parallel coordinates



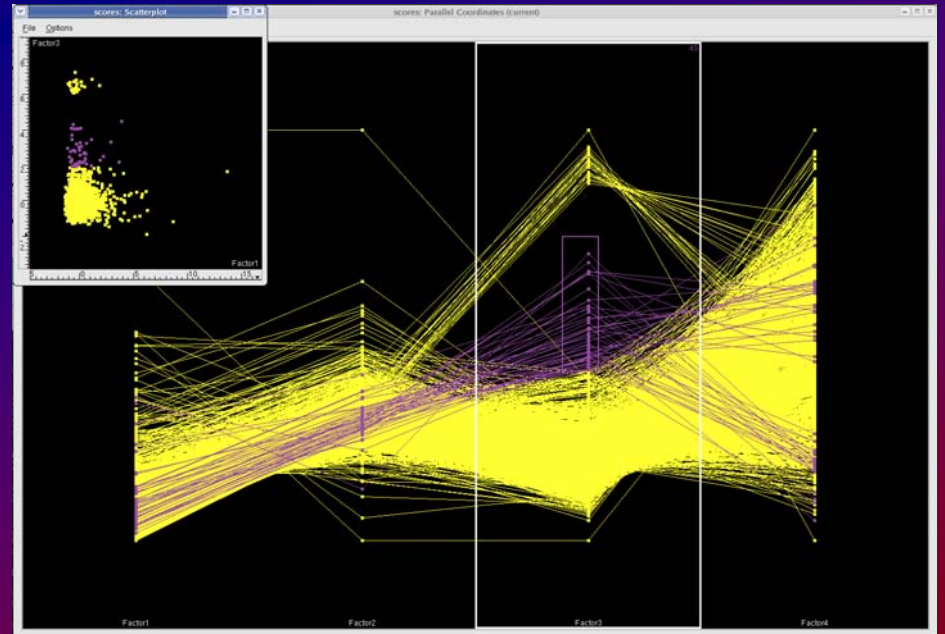
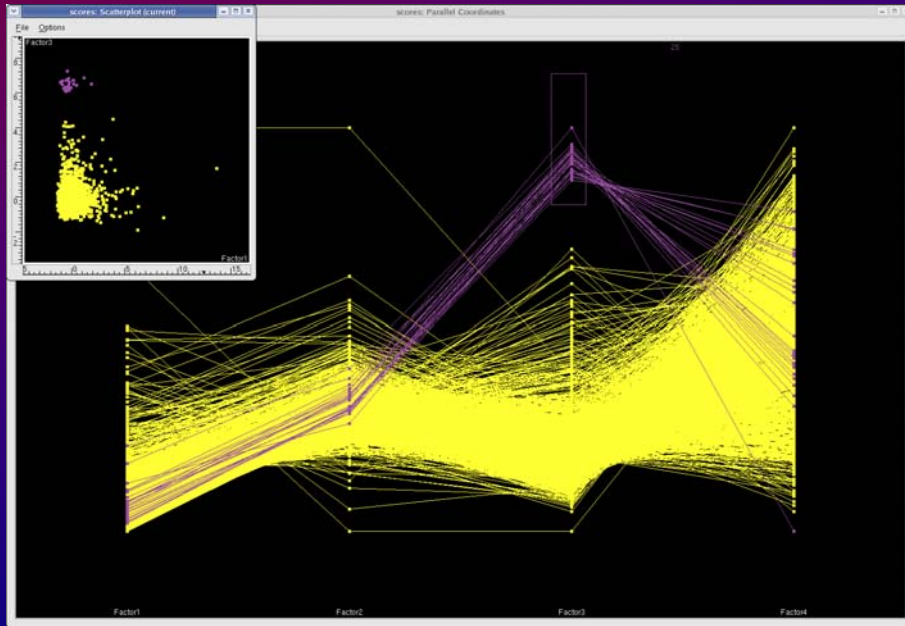
- axes are parallel
 - brushing
 - hierarchical clustering
 - even histograms

Column reduction

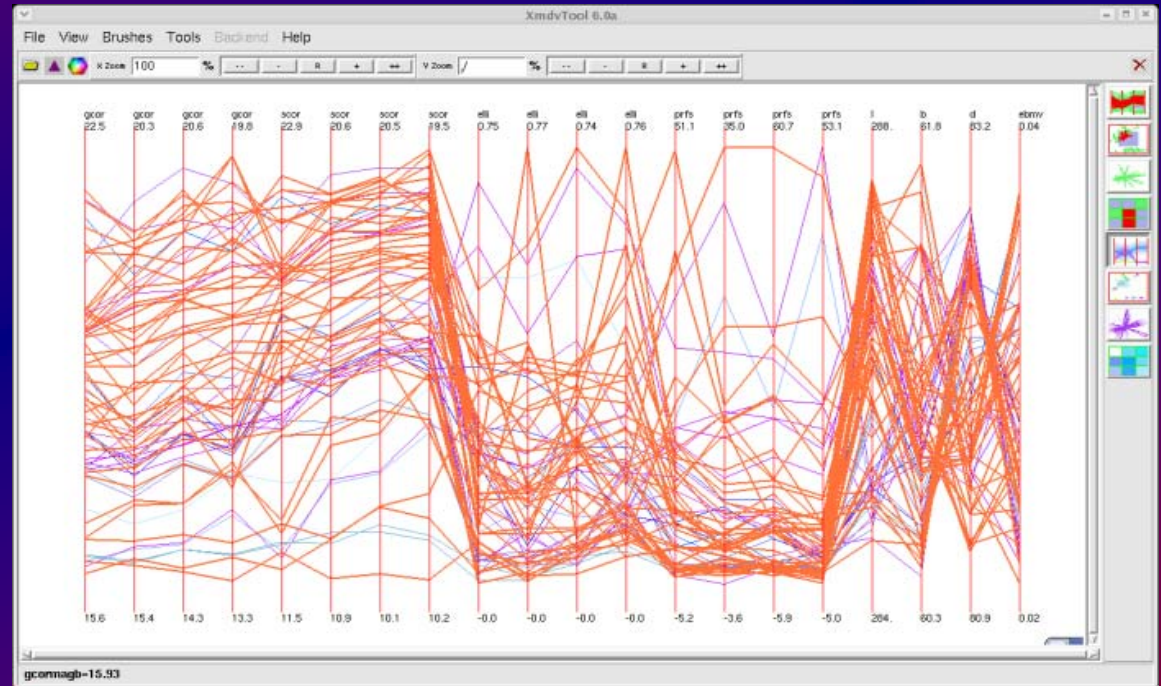
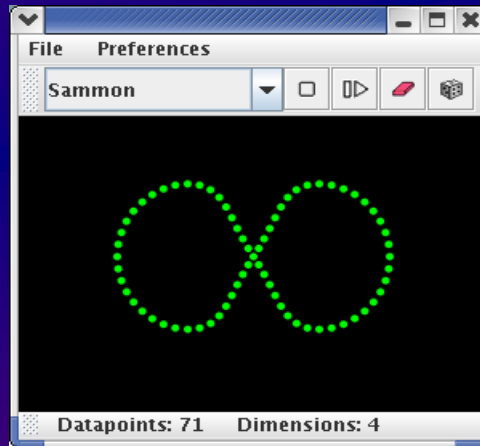
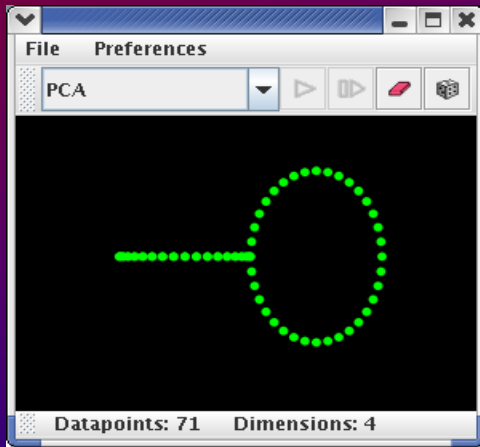
- Analysis of covariance
 - PCA
 - Factor analysis
 - MDS ...
- PCA
 - difficult to interpret
- FA with rotation is easier

| Variable | FA | | | | PCA | | | |
|-------------|-------|--------|------|--------|--------|--------|--------|--------|
| | fa1 | fa2 | fa3 | fa4 | pc1 | pc2 | pc3 | pc4 |
| dt | – | – | – | 0.459 | -0.222 | -0.172 | 0.682 | – |
| latitude | – | – | – | -0.23 | – | – | -0.339 | -0.132 |
| longitude | – | 0.446 | – | – | -0.196 | 0.72 | – | 0.107 |
| long_carr | – | – | – | – | – | – | – | 0.98 |
| area | 0.743 | -0.161 | – | – | -0.428 | -0.499 | 0.14 | – |
| long_extent | 0.763 | 0.237 | – | – | -0.584 | – | -0.264 | – |
| n_spots | 0.878 | 0.9 | – | – | -0.581 | 0.119 | -0.288 | – |
| mindistar | – | – | 0.95 | -0.117 | 0.223 | -0.407 | -0.489 | – |
| n_flares | 0.501 | – | – | – | na | na | na | na |

Can combine techniques

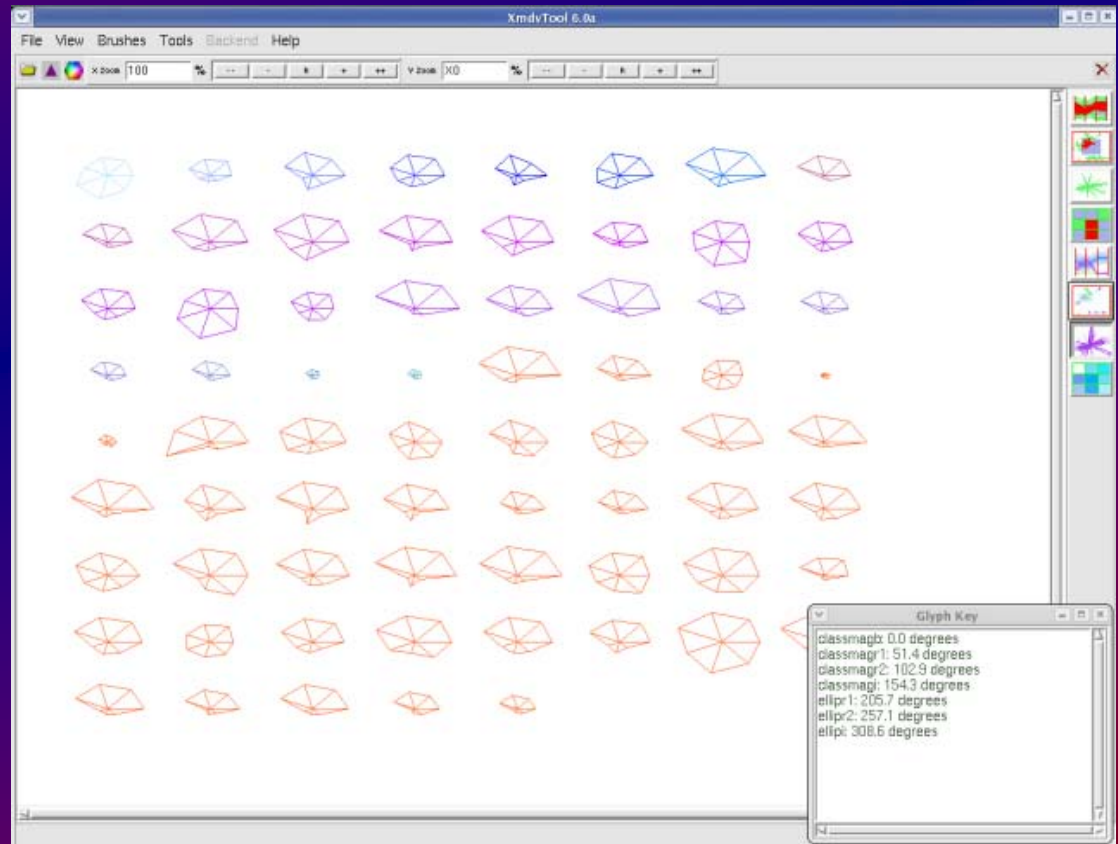


but some signs of trouble



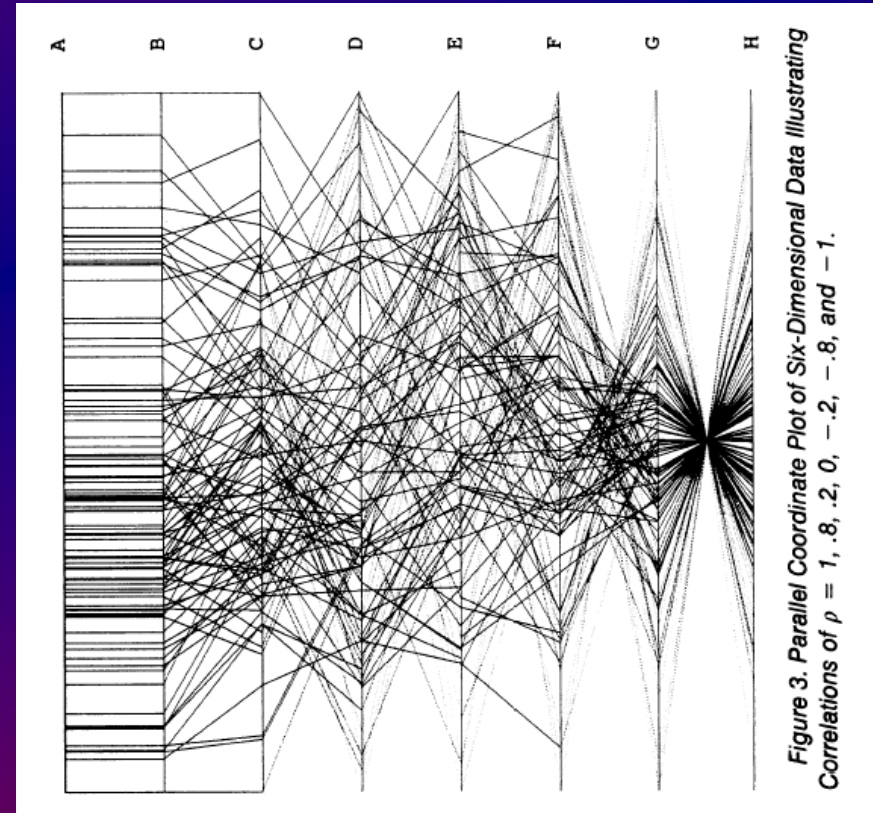
...also apparent in the plot key

- Glyph drawings
 - radial coord
 - not labelled
- Clustering also applied
- How easy is it to read?

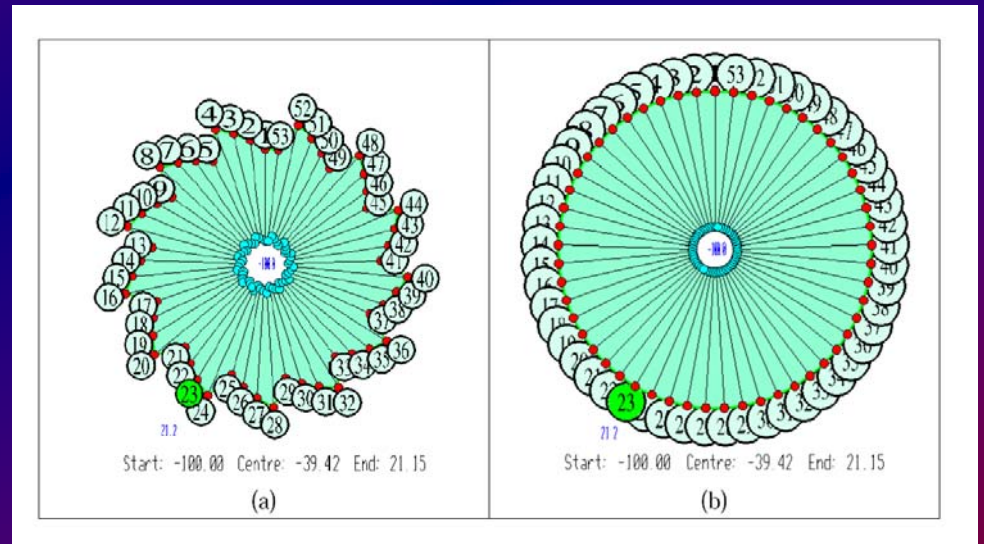
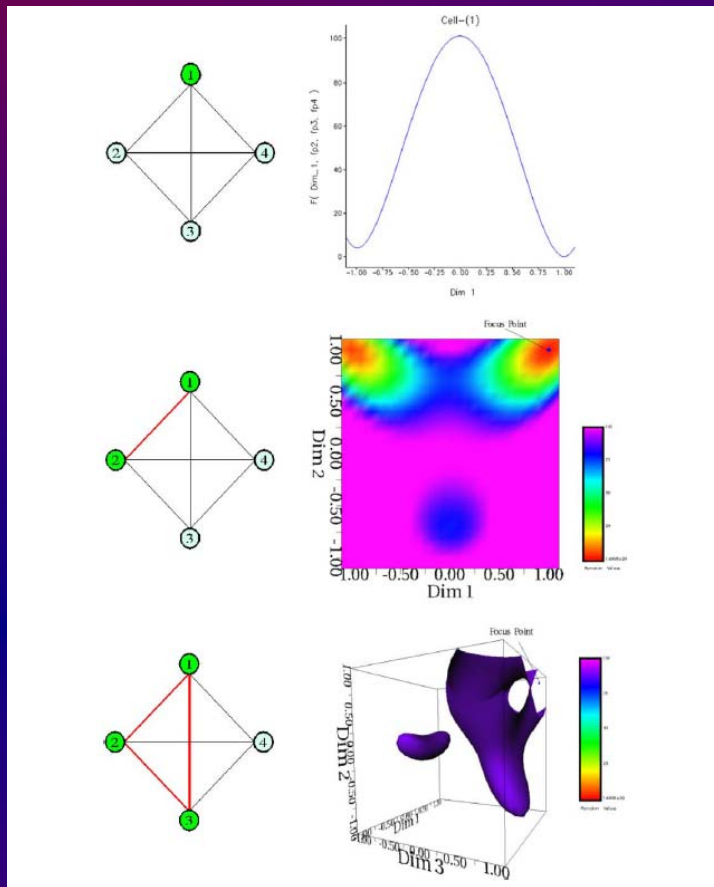


also in parallel coord space

- Correlation in parallel coords
- Even more difficult if several clusters are present
- 10^3 rows * ~ 20 variables for most displays
 - small table
- perhaps more a cognitive limit than physical

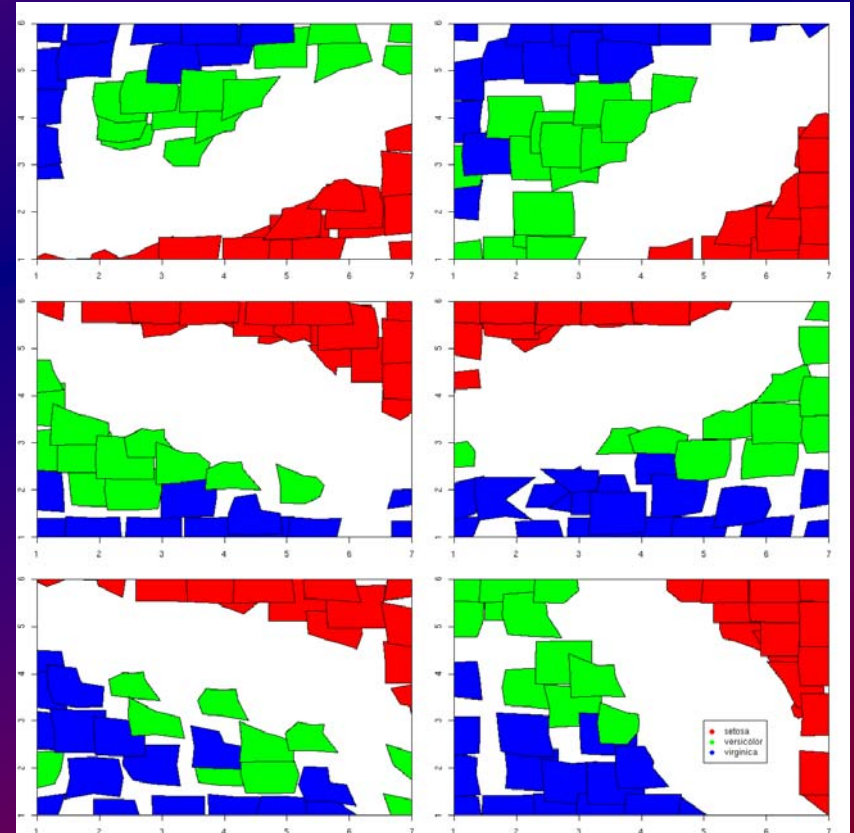


Loss of overview if p high



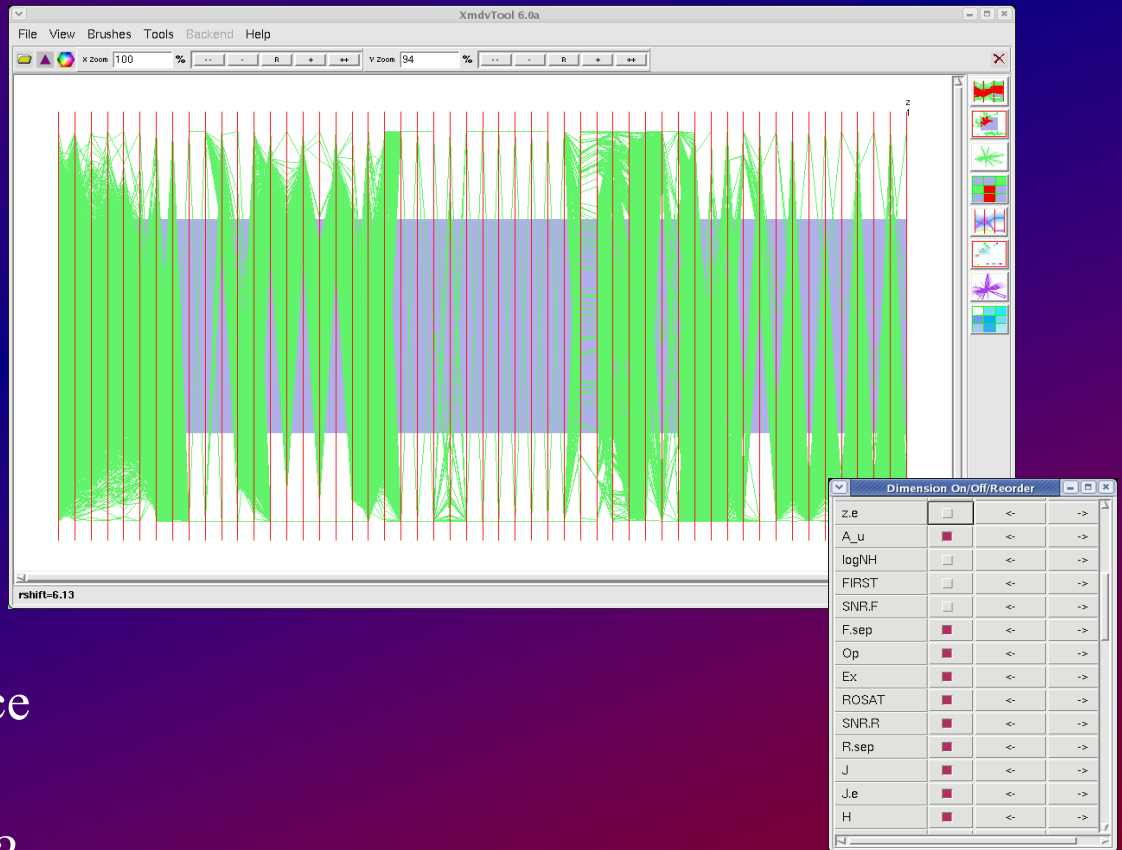
Row reduction

- Clustering
 - Kmeans etc
 - maximum likelihood
 - 'spectral'
- Sampling
 - bootstrap, MCMC
 - density
- SOM/klAR

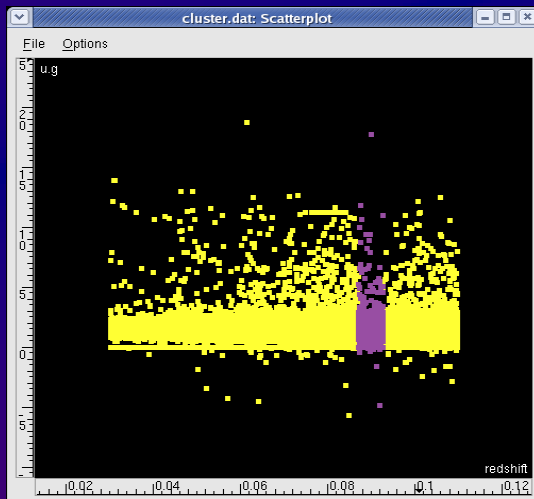
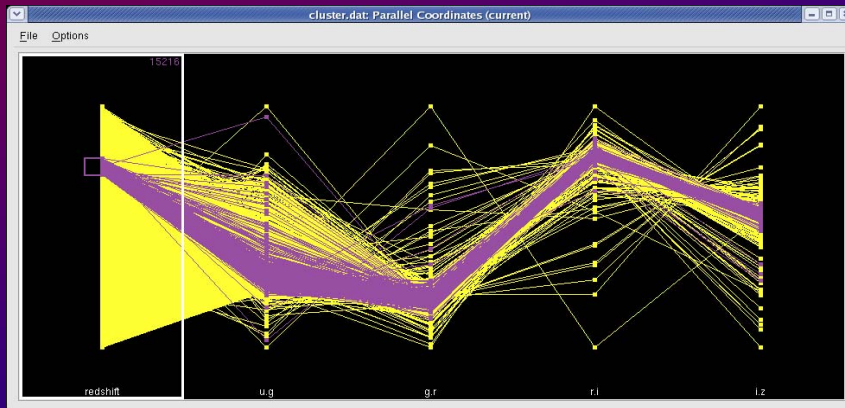


Medium datasets

- 10^6 entries
- ~10-15 MB
 - clustering ✗
 - scatterplots ✓
 - parallel coords ✓
 - PCA etc ✓ ?lapack
- Few apps will try
 - Xmdvtool/R with patience
 - Astroneural $\sim 5.6 \times 10^5$
 - GOODS data ...but time?



And another problem...



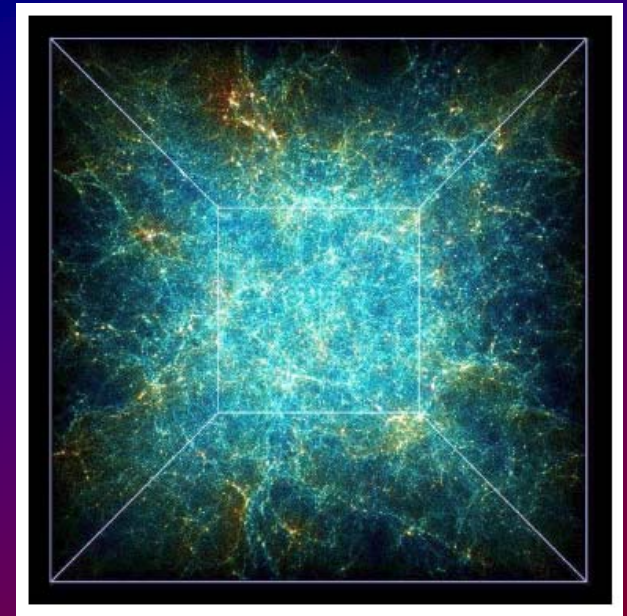
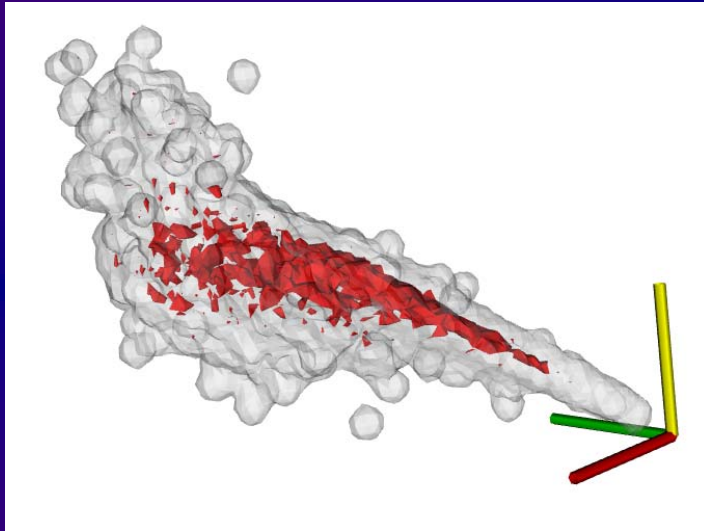
- Can interact, but slow
- brushing also works
 - but slow
- Can't really see wood for trees
 - eye is drawn to outliers
 - data density issue

Large datasets

- 10^8 entries
- ~200 MB
 - clustering ✗
 - scatterplots (single eg. Topcat) or splats ✓
 - parallel coords ✗
 - PCA ,
 - ?maybe, no rotation

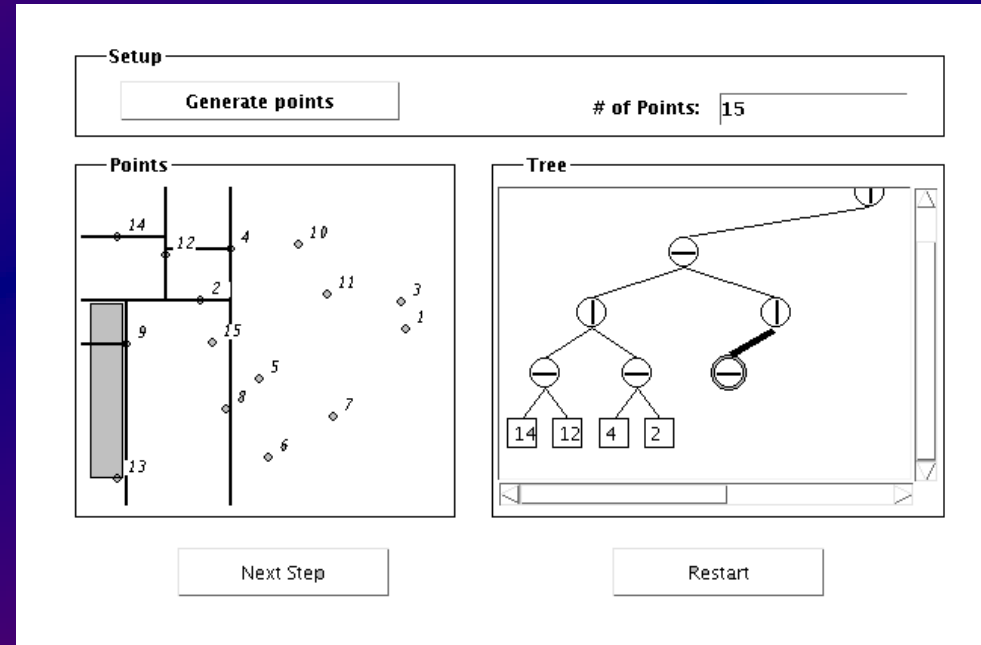
Splatting

- Composite of data space



Kd-trees

- Gray & Moore
 - can be applied to a large class of problems
 - notably density
 - Correlation, n-pt funcs

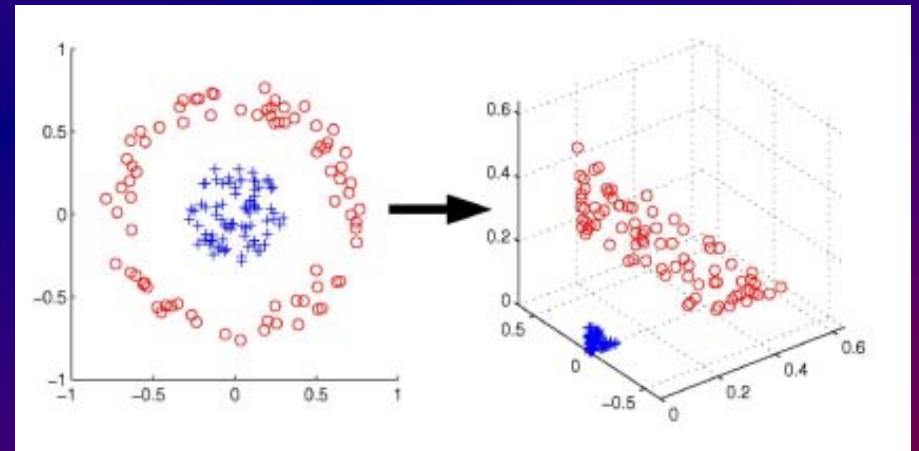


Speed-ups with kd-trees

- $O(n \log n)$ construction
- $O(\log n)$ search
- employ parallel processing + gpu kd-tree
- *Pruning* or approximation
- may be able to contour/isosurface in a fairly cheap way













Oddly, more dimensions can help

- SVM/kernel trick
- Can cast some problems into higher dimensions
- Non-linear separation



Some solutions?

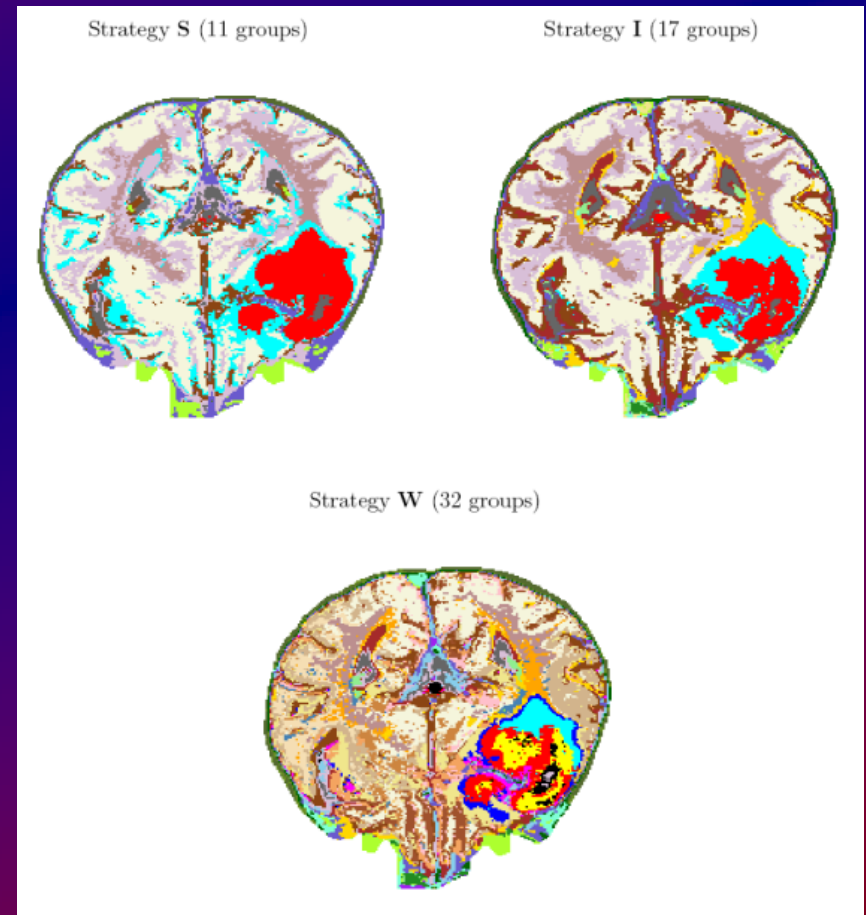
- Maintain overview
 - Hyperbolic trees
 - Sparklines
 - Explore increasing dimensions
- Statistical backup
 - R, libsvm, vq...
 - security? speed?

| Measures | | ... | | |
|------------|--|--------------------|------------|---------|
| Region ... | Market Share P30 | Target Group Count | Units Sold | Sales |
| + 2141 |  1,3% | 210 | 21.618 | 413.890 |
| + 2142 |  0,8% | 205 | 14.694 | 277.894 |
| + 2143 |  0,8% | 271 | 17.813 | 339.998 |
| + 2144 |  0,9% | 243 | 18.389 | 339.786 |
| + 2145 |  1,1% | 240 | 21.206 | 382.798 |
| + 2146 |  0,6% | 179 | 16.836 | 314.964 |
| + 2147 |  1,5% | 221 | 13.142 | 308.911 |
| + 2148 |  1,2% | 243 | 17.296 | 332.079 |
| + 2149 |  1,5% | 229 | 12.799 | 277.175 |
| + 2150 |  1,6% | 209 | 13.073 | 285.644 |
| + 2151 |  1,6% | 214 | 11.146 | 239.930 |
| + 2152 |  1,4% | 236 | 13.306 | 270.886 |
| | | | 11.388 | 251.572 |
| | | | 13.116 | 291.661 |

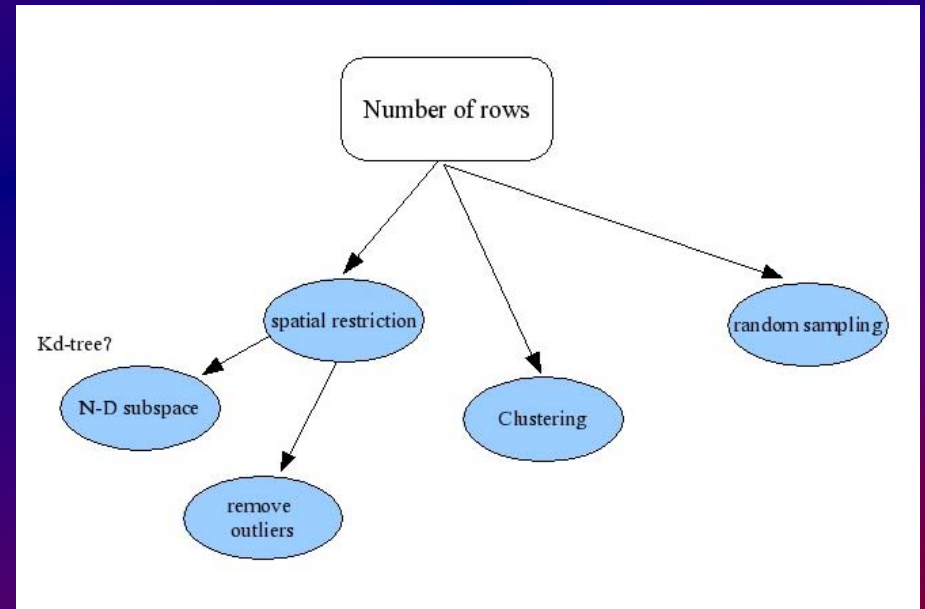
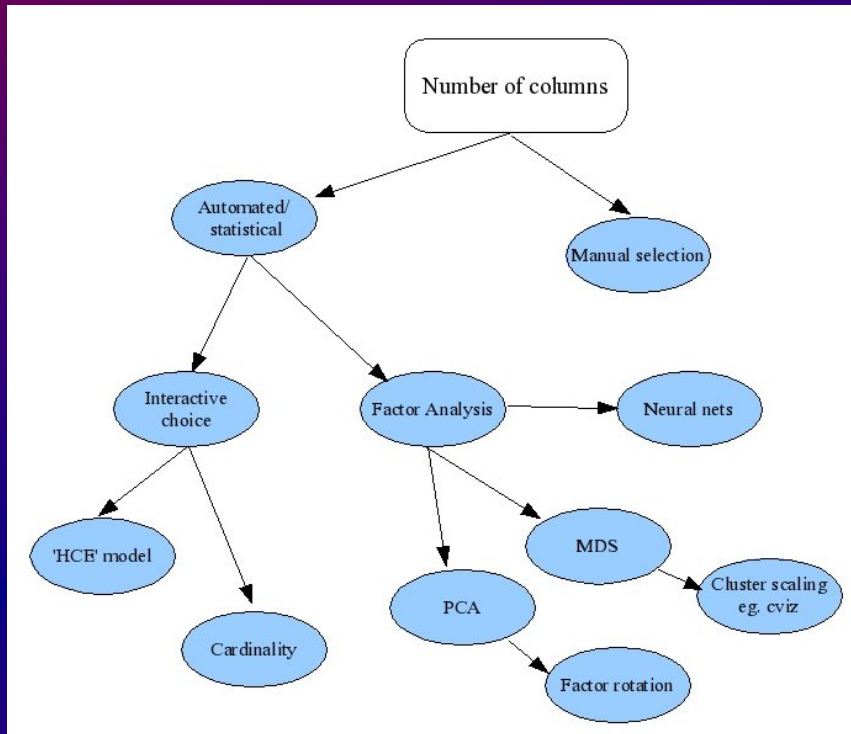


The future seems dense

- Density sampling
 - Parametric methods: normal distribution
 - simple and adaptable
 - Non-parametric: Kernel methods, K-NN
 - Semi-parametric: mixture models (Fraley et al)
 - seeded by HC, BIC



Dimension reduction



On this one Columbia slide, a PowerPoint festival of bureaucratic hyper-rationalism, 6 different levels of hierarchy are used to display, classify, and arrange 11 phrases:

- Level 1 Title of Slide
- Level 2 ● Very Big Bullet
- Level 3 — big dash
- Level 4 ◆ medium-small diamond
- Level 5 • tiny square bullet
- Level 6 () parentheses ending level 5

The analysis begins with the dreaded Executive Summary, with a conclusion presented as a headline: "Test Data Indicates Conservatism for Tile Penetration." This turns out to be unmerited reassurance. Executives, at least those who don't want to get fooled, had better read far beyond the title.

The "conservatism" concerns the *choice of models* used to predict damage. But why, after 112 flights, are foam-debris models being calibrated during a crisis? How can "conservatism" be inferred from a loose comparison of a spreadsheet model and some thin data? Divergent evidence means divergent evidence, not inferential security. Claims of analytic "conservatism" should be viewed with skepticism by presentation consumers. Such claims are often a rhetorical tactic that substitutes verbal fudge factors for quantitative assessments.

As the bullet points march on, the seemingly reassuring headline fades away. Lower-level bullets at the end of the slide undermine the executive summary. This third-level point notes that "Flight condition [that is, the debris hit on the Columbia] is significantly outside of test database." How far outside? The final bullet will tell us.

This fourth-level bullet concluding the slide reports that the debris hitting the Columbia is estimated to be $1920/3 = 640$ times larger than data used in the tests of the model! The correct headline should be "Review of Test Data Indicates Irrelevance of Two Models." This is a powerful conclusion, indicating that pre-launch safety standards no longer hold. The original optimistic headline has been eviscerated by the lower-level bullets.

Note how close readings can help consumers of presentations evaluate the presenter's reasoning and credibility.

The Very-Big-Bullet phrase fragment does not seem to make sense. No other VBB's appear in the rest of the slide, so this VBB is not necessary.

Spray On Foam Insulation, a fragment of which caused the hole in the wing

A model to estimate damage to the tiles protecting flat surfaces of the wing

Review of Test Data Indicates Conservatism for Tile Penetration

- The existing SOFI on tile test data used to create Crater was reviewed along with STS-87 Southwest Research data
 - Crater overpredicted penetration of tile coating significantly
 - ◆ Initial penetration to described by normal velocity
 - Varies with volume/mass of projectile (e.g., 200ft/sec for 3cu. In)
 - ◆ Significant energy is required for the softer SOFI particle to penetrate the relatively hard tile coating
 - Test results do show that it is possible at sufficient mass and velocity
 - ◆ Conversely, once tile is penetrated SOFI can cause significant damage
 - Minor variations in total energy (above penetration level) can cause significant tile damage
 - Flight condition is significantly outside of test database
 - ◆ Volume of ramp is 1920cu in vs 3 cu in for test

BOEING

Here "ramp" refers to foam debris (from the bipod ramp) that hit Columbia. Instead of the cryptic "Volume of ramp," say "estimated volume of foam debris that hit the wing." Such clarifying phrases, which may help upper level executives understand what is going on, are too long to fit on low-resolution bullet outline formats. PP demands the shorthand of acronyms, phrase fragments, and clipped jargon in order to get at least some information into the tight format.

(How not to visualize...)