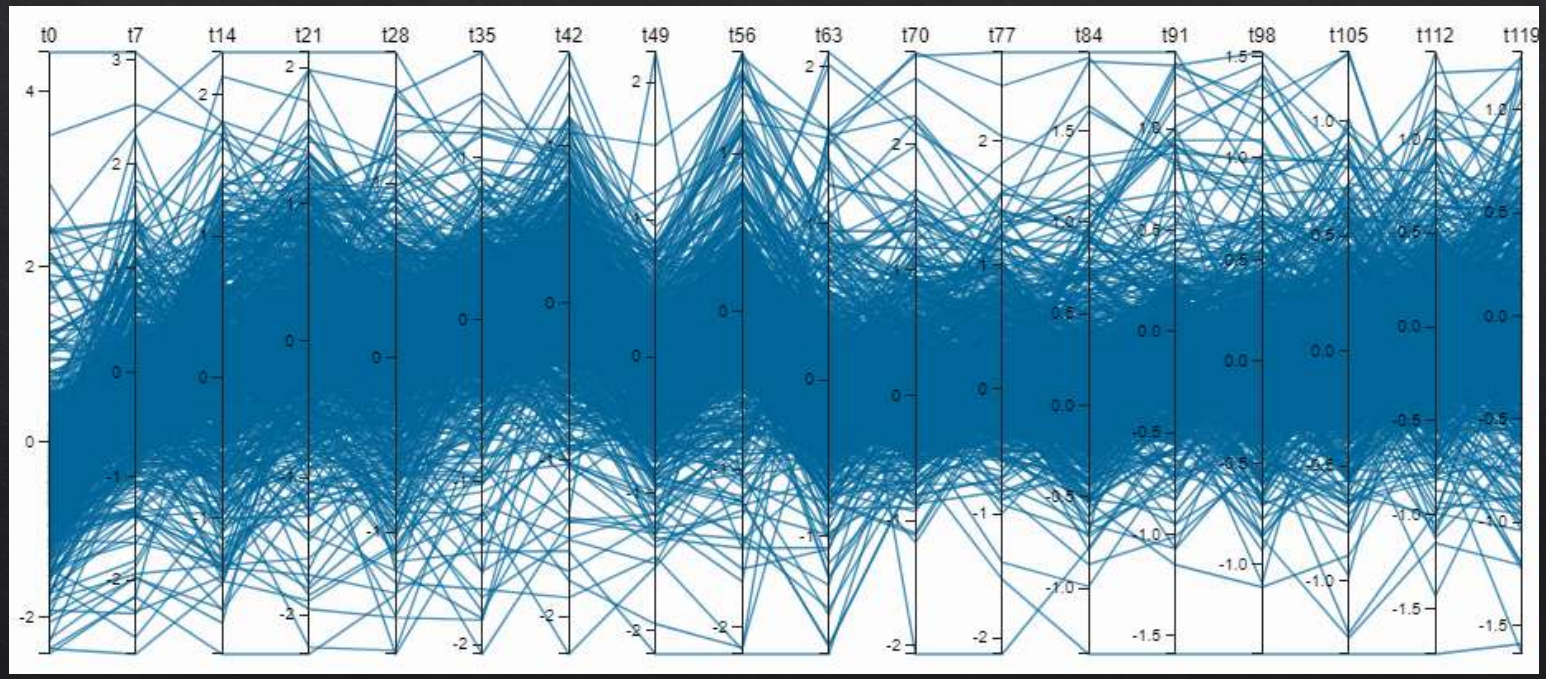
How to deal with over-plotting?

Visualizing Gene Expression Data

800 genes involved in cell-cycle regulation (Spellman et al., 1998),
measured every 7 minutes over 2 hours - interpreting time points as dimensions

'profile plot'

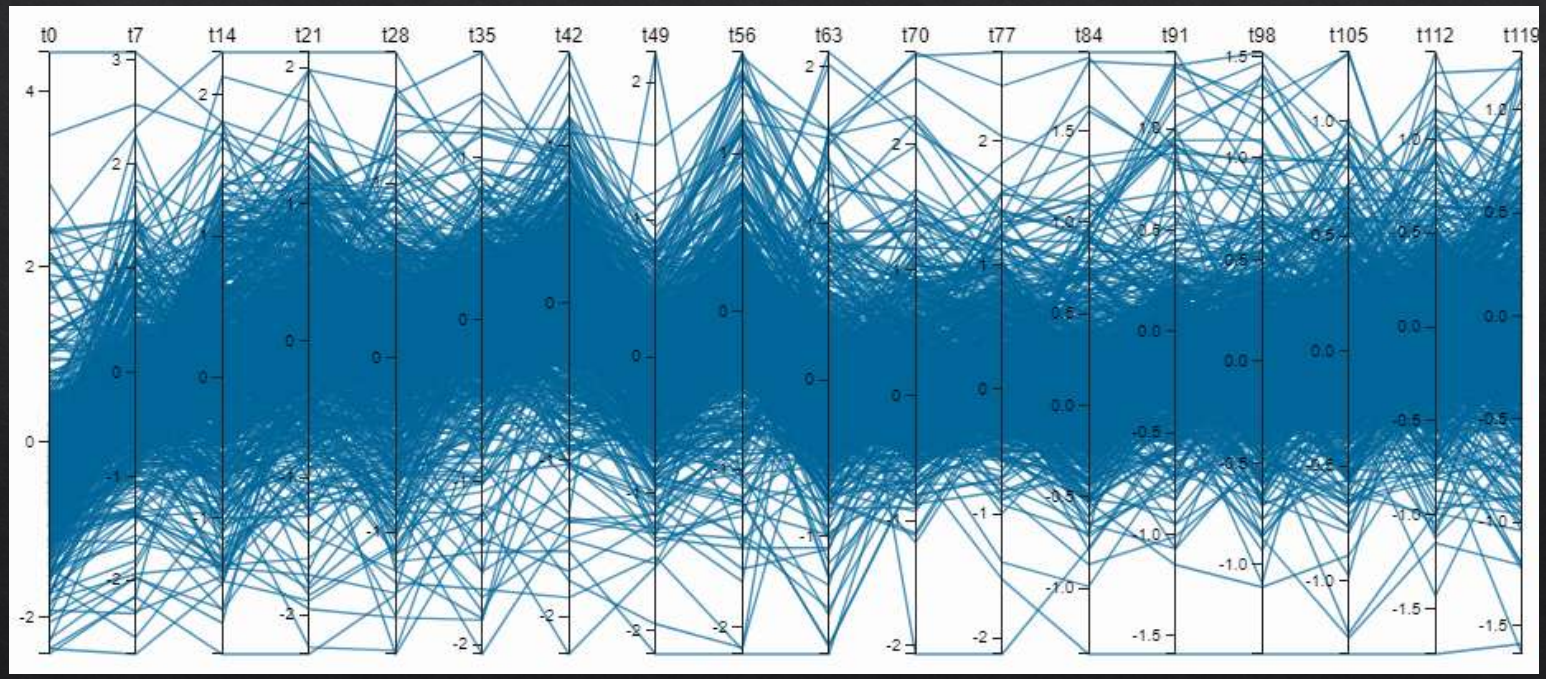


values greater than zero signify that the corresponding gene is being expressed or 'on', while values below zero denote genes that are 'off'

Visualizing Gene Expression Data

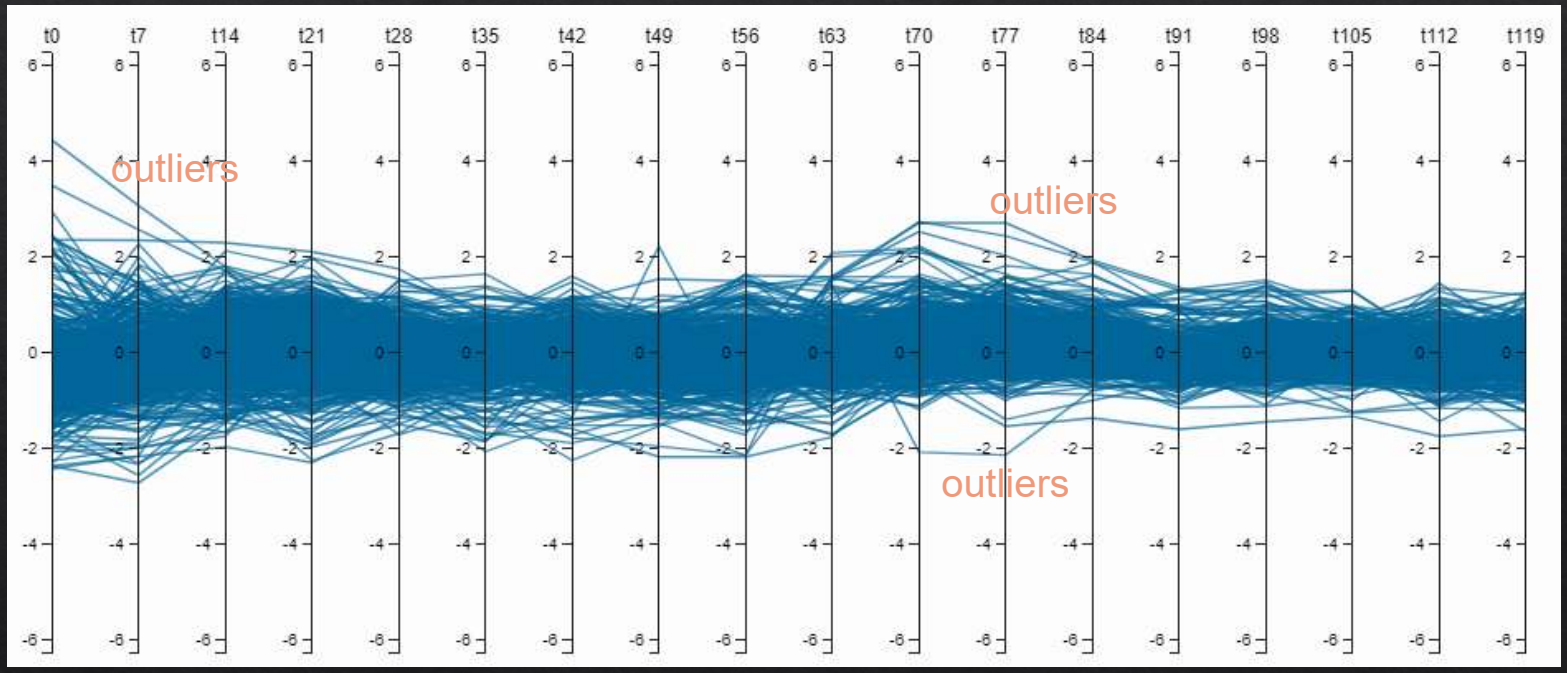
800 genes involved in cell-cycle regulation (Spellman et al., 1998),
measured every 7 minutes over 2 hours - interpreting time points as dimensions

'profile plot'



values greater than zero signify that the corresponding gene is being expressed or 'on', while values below zero denote genes that are 'off'

Visualizing Gene Expression Data

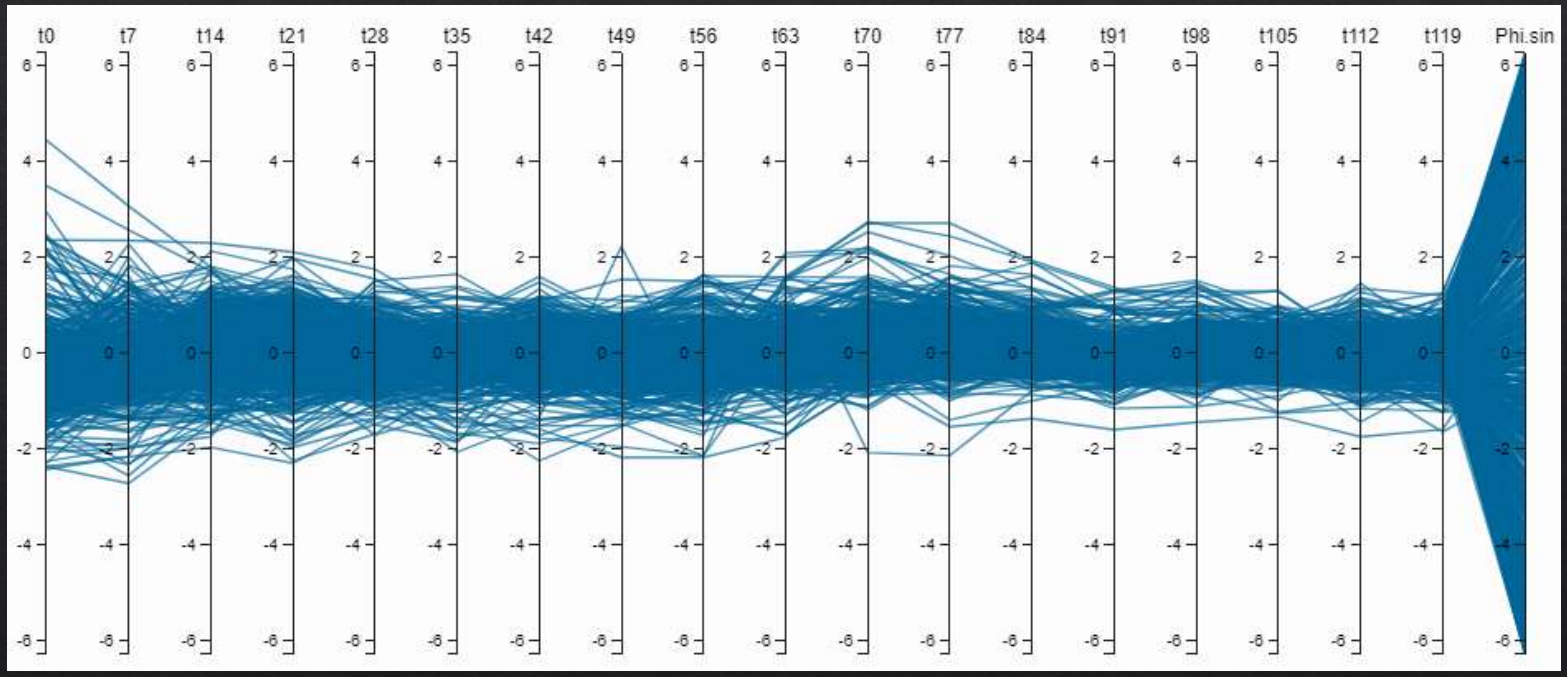


put a common scale in place

Visualizing Gene Expression Data

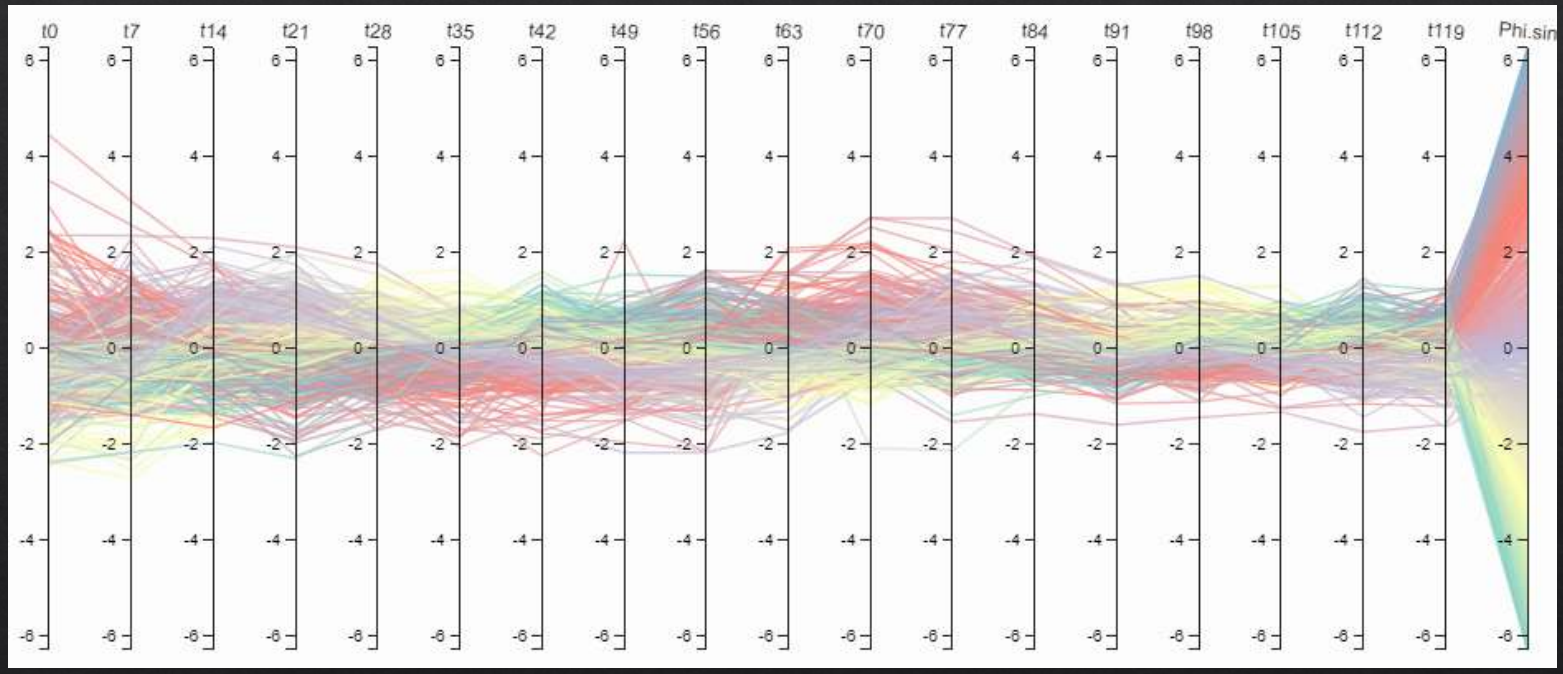
Add another dimension to the data

statistic that was derived from the time-series



Applied to every gene: genes with similar values for this dimension share a similar activation pattern over time.

Visualizing Gene Expression Data



To visualise this, we apply a colormap to the new dimension:

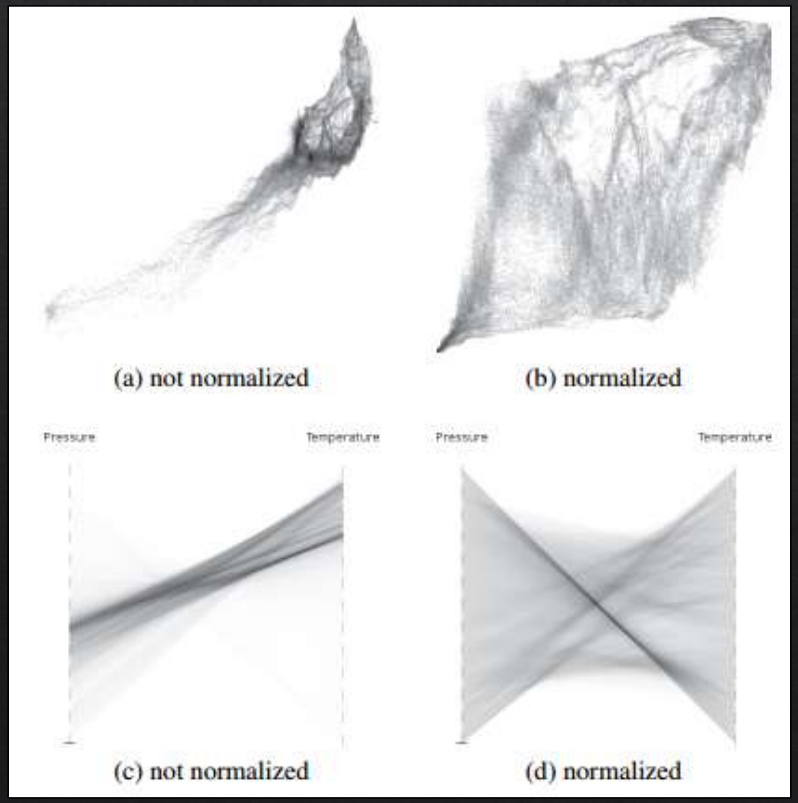
Now, we can see cyclic patterns emerge from the colored polylines.

Handling Big Data

Sampling – Distribution – Continuous Domain

Continuous Model [Blaas & Botha, 2008]

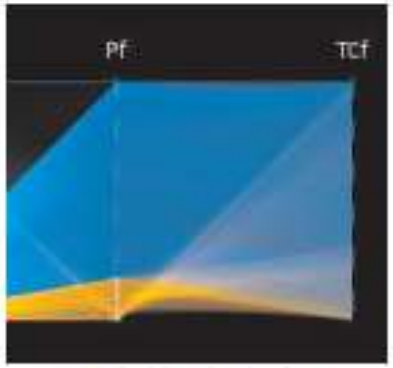
(special processing, 25 million data about 10 dimensions)



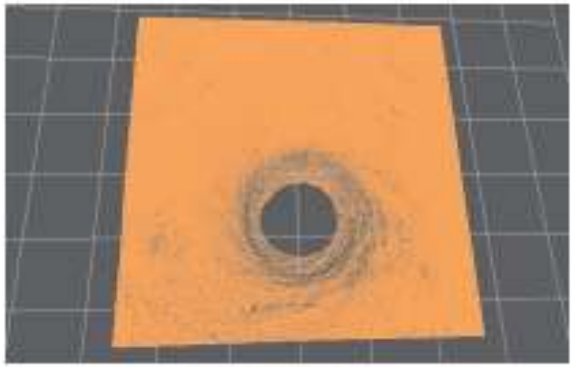
LZO (Lempel-Ziv-Oberhumer) compression,
- a public implementation is available
- realtime decompression

- Don't forget to Normalize the data

Continuous Model
[Blaas & Botha, 2008]
(special processing, 25 million data about 10 dimensions)



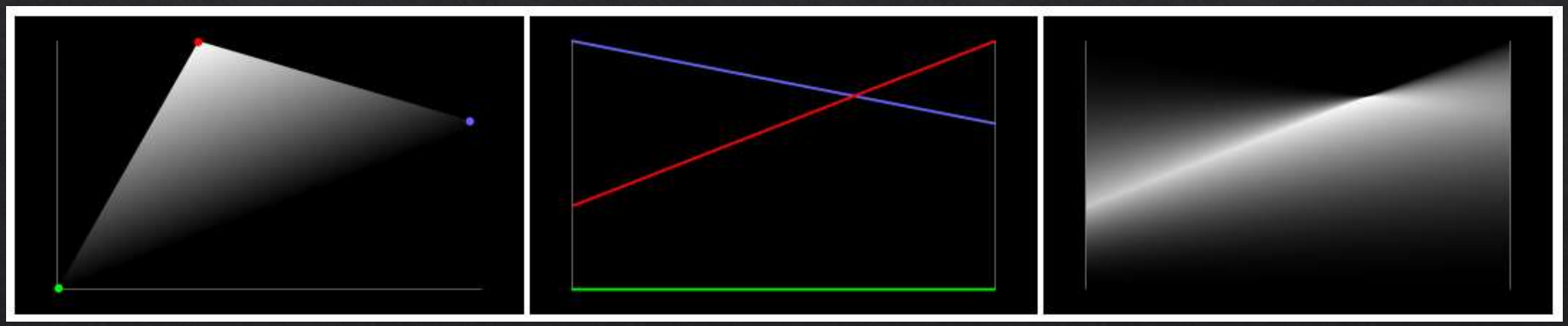
(a) PCP selection



(b) 3D view

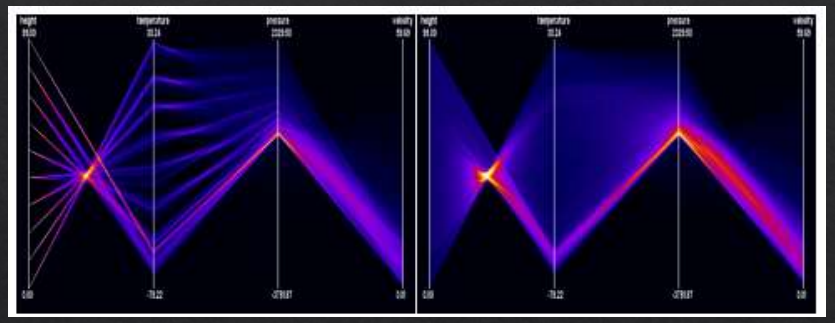
Fig. 9: Selection of low pressure areas (shown in orange) reveals the area of low-temperature near the eye of the hurricane. The compact horizontal shape of the orange band in the PCP reveals that low-pressure areas mostly have a low-temperature as well.

Continuous Model
[Heinrich & Weiskopf, 2009]

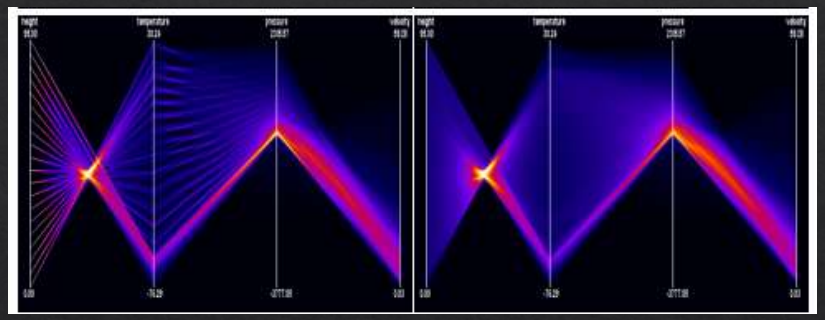


Continuous Model [Heinrich & Weiskopf, 2009]

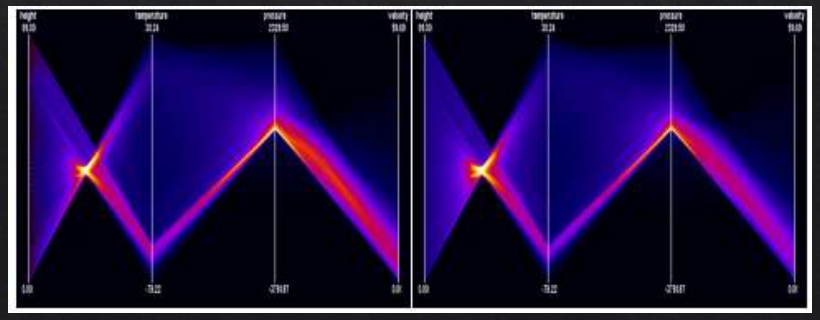
Discrete Continuous



Discrete Continuous



Discrete Continuous

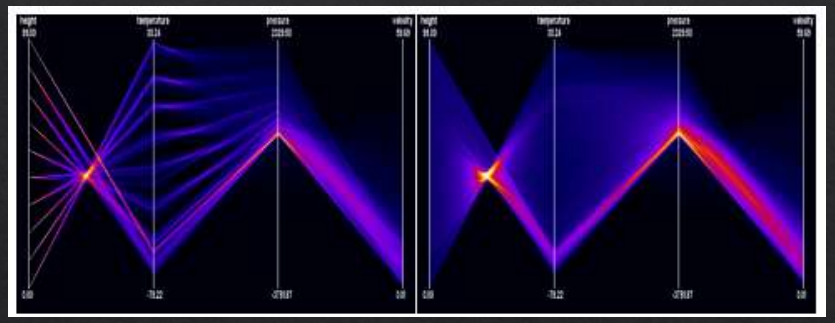


The “hurricane Isabel” dataset at different spatial resolutions
50×50×10,
100×100×20,
500×500×100

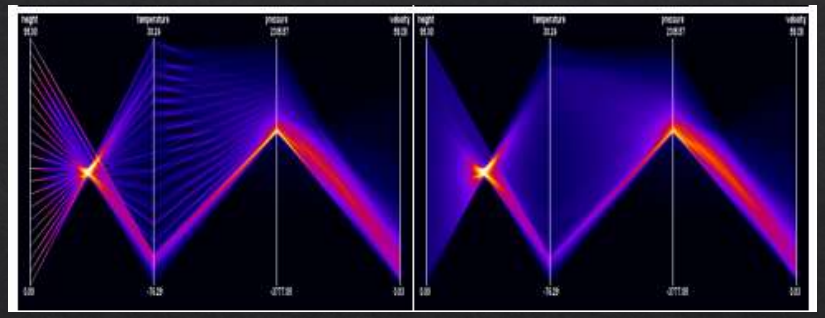
Continuous Model
[Heinrich & Weiskopf, 2009]

Sampling artifacts

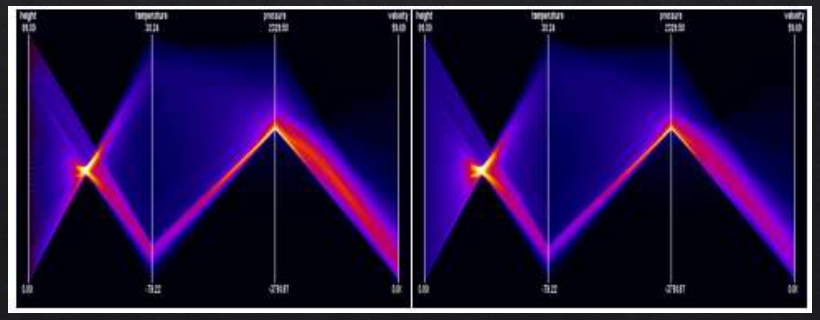
Better Visualization



50×50×10



100×100×20



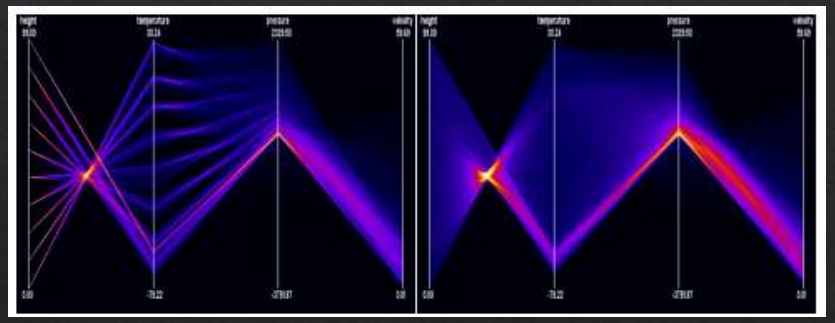
500×500×100

Sampling artifacts stemming from the discrete mapping lead to misrepresentation of key information in discrete PCP.

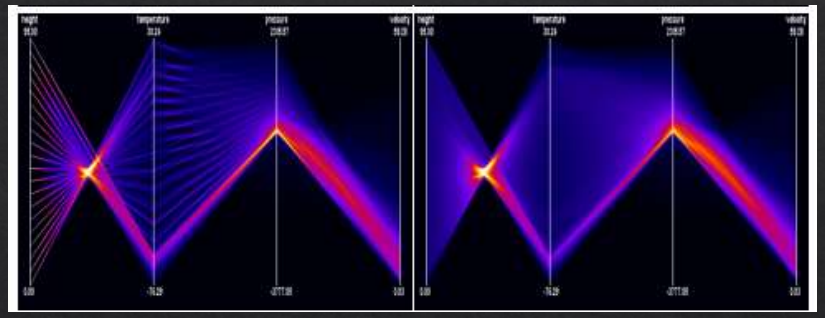
Continuous Model
[Heinrich & Weiskopf, 2009]

Sampling artifacts

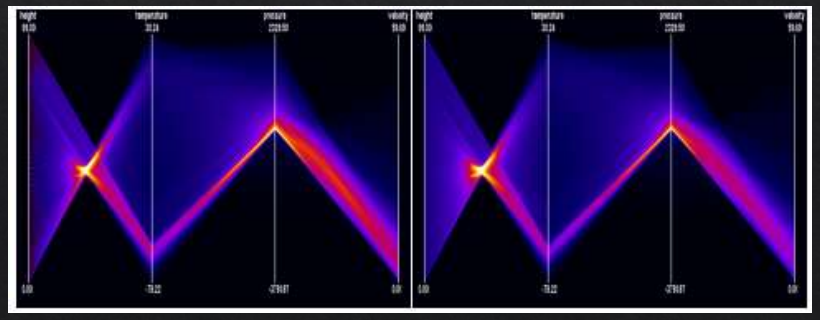
Better Visualization



50×50×10



100×100×20



500×500×100

Apart from differences regarding the sampling of the data, however, continuous parallel coordinates share most of the advantages and problems of discrete parallel coordinates.

Axis-Ordering Problem

Comparing All Pairs
Each Pair Exactly Once!

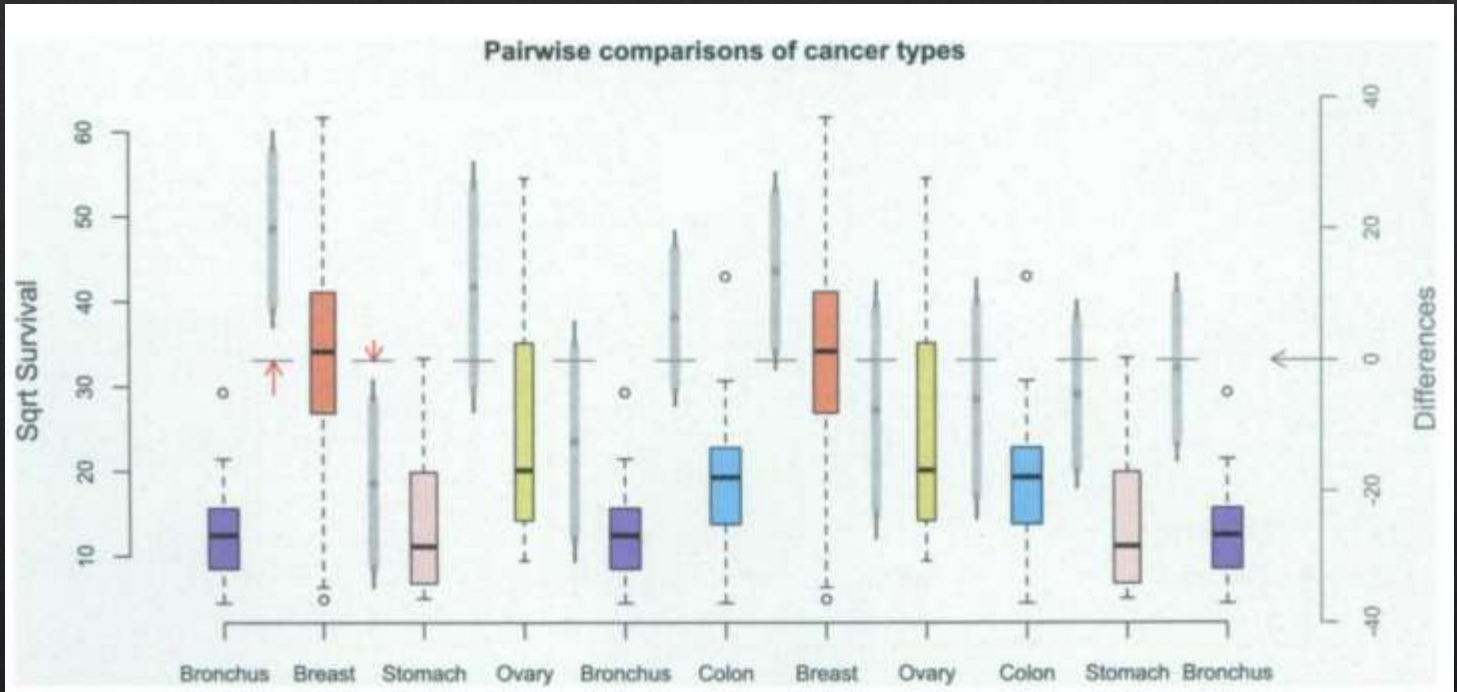


Figure 11. Boxplots and pairwise comparisons of vitamin-C treated cancer patients. The left axis and boxplots refer to square-root transformed survival times; the right axis and gray scale vertical bars refer to confidence intervals for pairwise differences of means. Red arrows indicate comparisons significantly different from 0.

Graphs of Even Degree → Eulerian Tour!

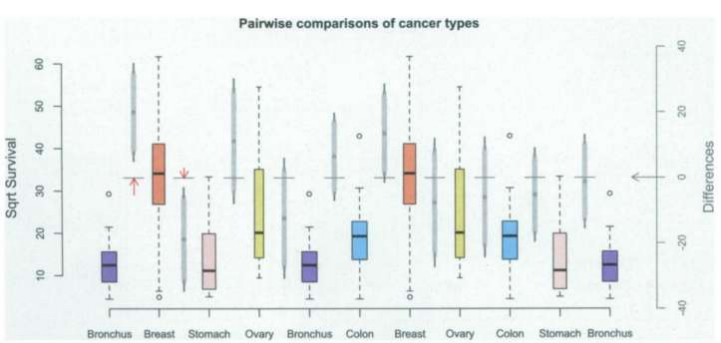
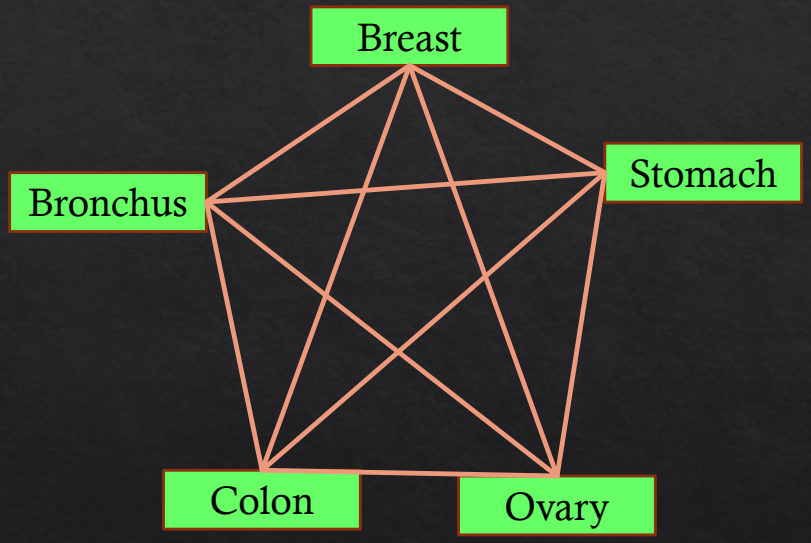


Figure 11. Boxplots and pairwise comparisons of vitamin-C treated cancer patients. The left axis and boxplots refer to square-root transformed survival times; the right axis and gray scale vertical bars refer to confidence intervals for pairwise differences of means. Red arrows indicate comparisons significantly different from 0.



Graphs of Even Degree → Eulerian Tour!

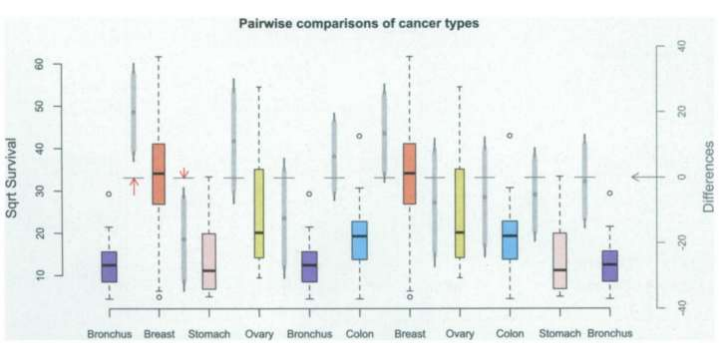
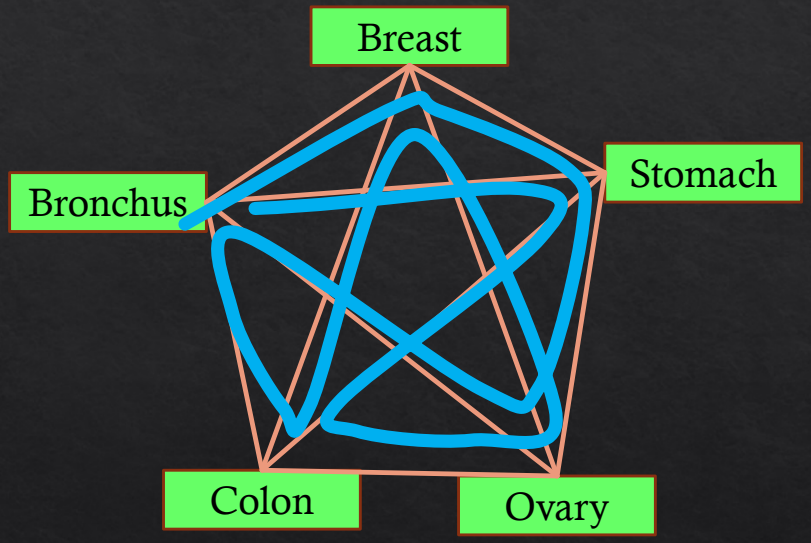


Figure 11. Boxplots and pairwise comparisons of vitamin-C treated cancer patients. The left axis and boxplots refer to square-root transformed survival times; the right axis and gray scale vertical bars refer to confidence intervals for pairwise differences of means. Red arrows indicate comparisons significantly different from 0.



Graphs of Even Degree → Eulerian Tour!

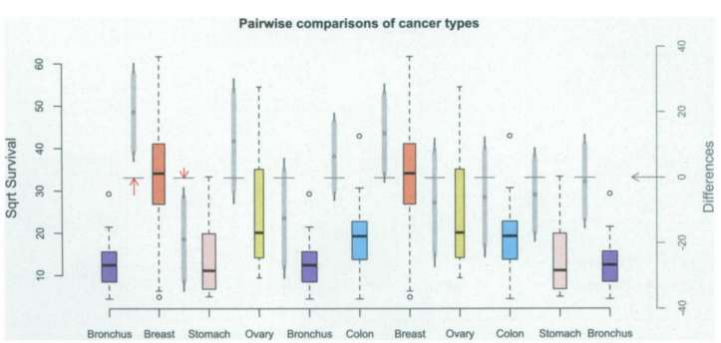
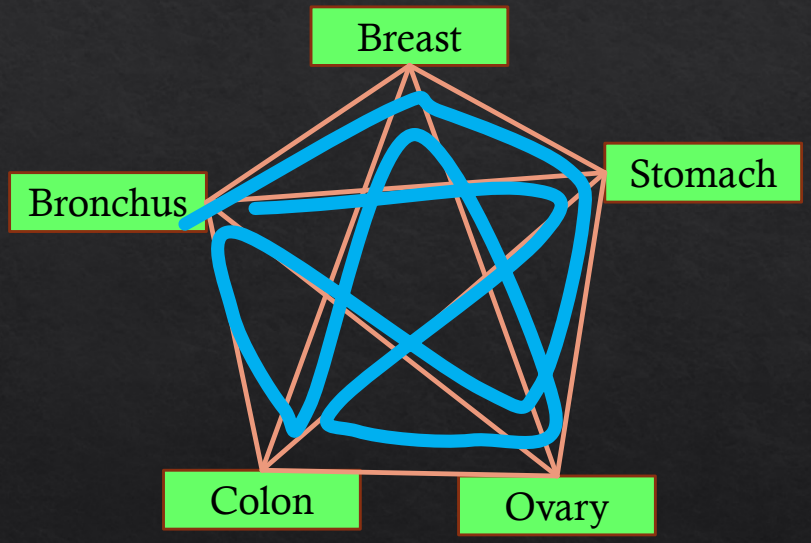


Figure 11. Boxplots and pairwise comparisons of vitamin-C treated cancer patients. The left axis and boxplots refer to square-root transformed survival times; the right axis and gray scale vertical bars refer to confidence intervals for pairwise differences of means. Red arrows indicate comparisons significantly different from 0.



Graphs of Even Degree → Eulerian Tour!

Algorithm 1 Hierholzer (1873) (adapted to find an Eulerian tour or open Eulerian path)

Require: A connected graph G that is even or that has exactly two odd vertices.

1: Choose a vertex v. If G is even, v can be any vertex; otherwise v is one of the two odd vertices.
2: Starting at v construct a path T in G, stopping when a vertex is reached without an unused edge.
3: **while** there are edges of G not already in path T **do**
4: Choose *any* vertex w in T that is incident on an unused edge.
5: Starting at w, construct a path D of unused edges stopping when a node is reached without any unused edges.
6: Enlarge T by splicing path D into T at vertex w.
7: **end while**
8: **return** T

Graphs of Even Degree → Eulerian Tour!

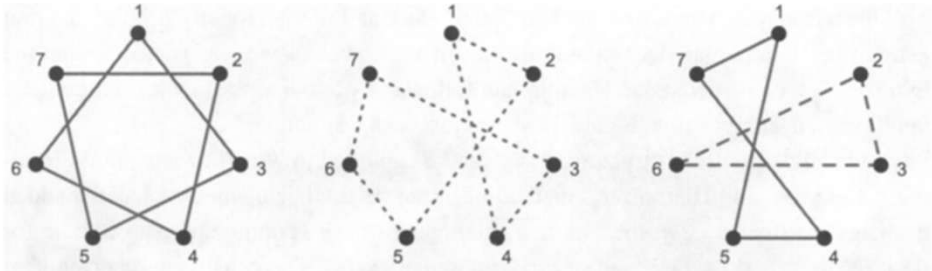
Algorithm 1 Hierholzer (1873) (adapted to find an Eulerian tour or open Eulerian path)

Require: A connected graph G that is even or that has exactly two odd vertices.

1: Choose a vertex v. If G is even, v can be any vertex; otherwise v is one of the two odd vertices.
2: Starting at v construct a path T in G, stopping when a vertex is reached without an unused edge.
3: **while** there are edges of G not already in path T **do**
4: Choose *any* vertex w in T that is incident on an unused edge.
5: Starting at w, construct a path D of unused edges stopping when a node is reached without any unused edges.
6: Enlarge T by splicing path D into T at vertex w.
7: **end while**
8: **return** T

Figure 15 shows how an application of Hierholzer's method might create an Eulerian tour for K_7. Starting at node 1 the selection of edges is such as to produce the Hamiltonian cycle 13572461 of Figure 15(a), followed by a second Hamiltonian cycle 12567341 of Figure 15(b), and finally by the short cycle 15471 of Figure 15(c). At this point, node 1 has no further unused edges and $T = 1357246\ 1\ 256734\ 1\ 547\ 1$. Path D (of Algorithm 1 line 5) is the dashed cycle 2362 of Figure 15(c), which line 6 of the algorithm allows to be spliced into T at node 2. The resulting Eulerian tour can be either 1357 2362 46125673415471 or 13572461 2362 5673415471.

Hierholzer's method applies to the graph K_n^e for all n. When $n = 2m + 1$, $K_{2m+1}^e = K_{2m+1}$ is even and it yields an Eulerian tour. When $n = 2m$, K_{2m}^e is an augmented version of K_{2m} with exactly two odd nodes, and the result is an open Eulerian path.



(a) First Hamiltonian cycle (b) Second Hamiltonian cycle (c) Two non-Hamiltonian cycles

Figure 15. An application of Hierholzer's method to K_7 which happens to follow one Hamiltonian cycle after another.

Application: Analyzing Consistency in Ratings

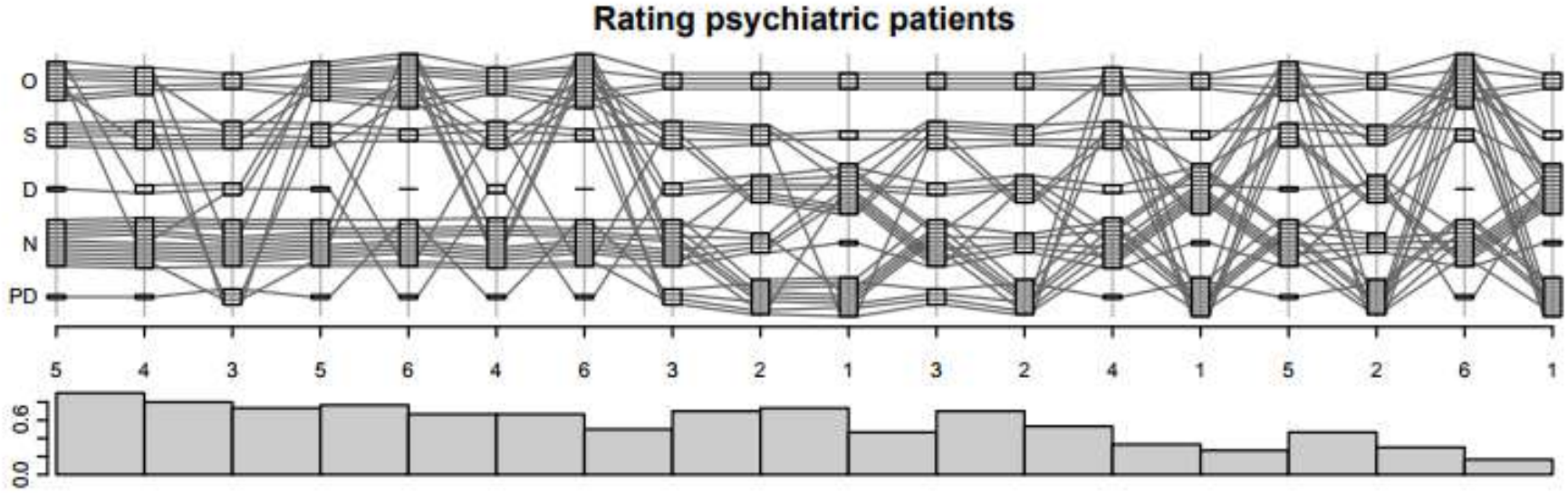
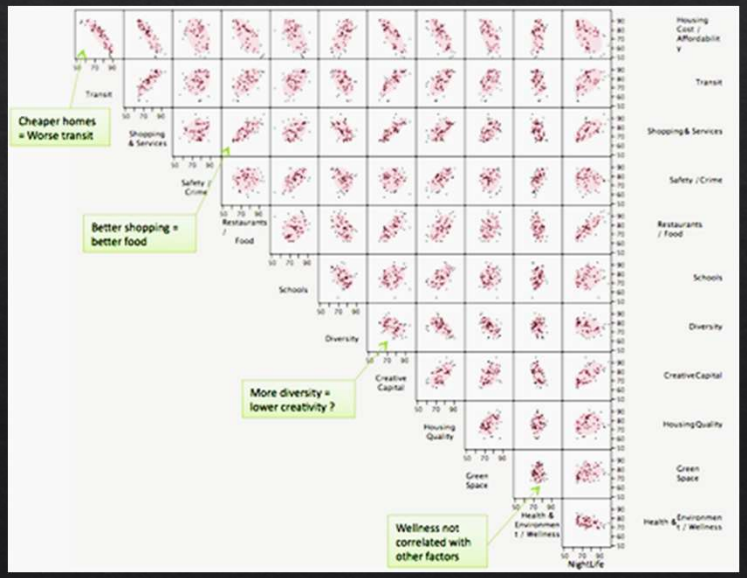


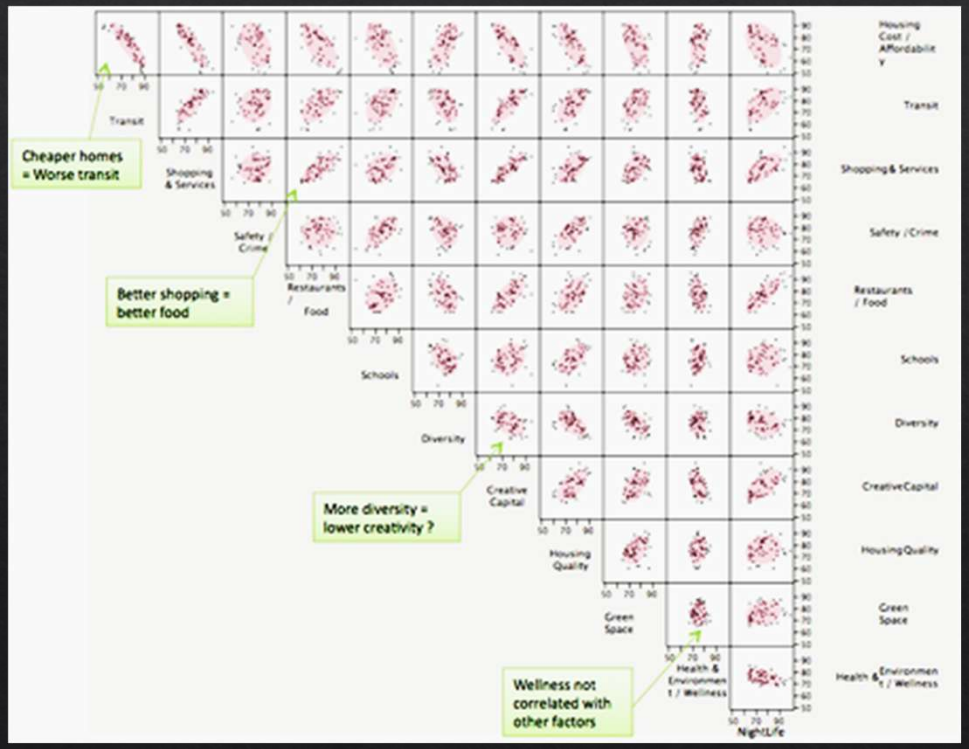
Fig. 2 Parallel coordinate plot of ratings. Each axis represents a rater, all pairs of raters appear adjacently. Overlaid rectangles show each rater's distribution of ratings. The lower bar chart shows the agreement proportion for each adjacent pair of raters.

SPLOM

But we are just comparing dimensions – why cannot we just compare all possible pairs in a traditional plot?



Scatterplot Matrix (SPLOM) (created by JMP)



A scatter-plot matrix neatly organizes all of the pairwise correlation information.

New York neighborhoods

12 factors (housing affordability, transit, green space, nightlife, etc.)

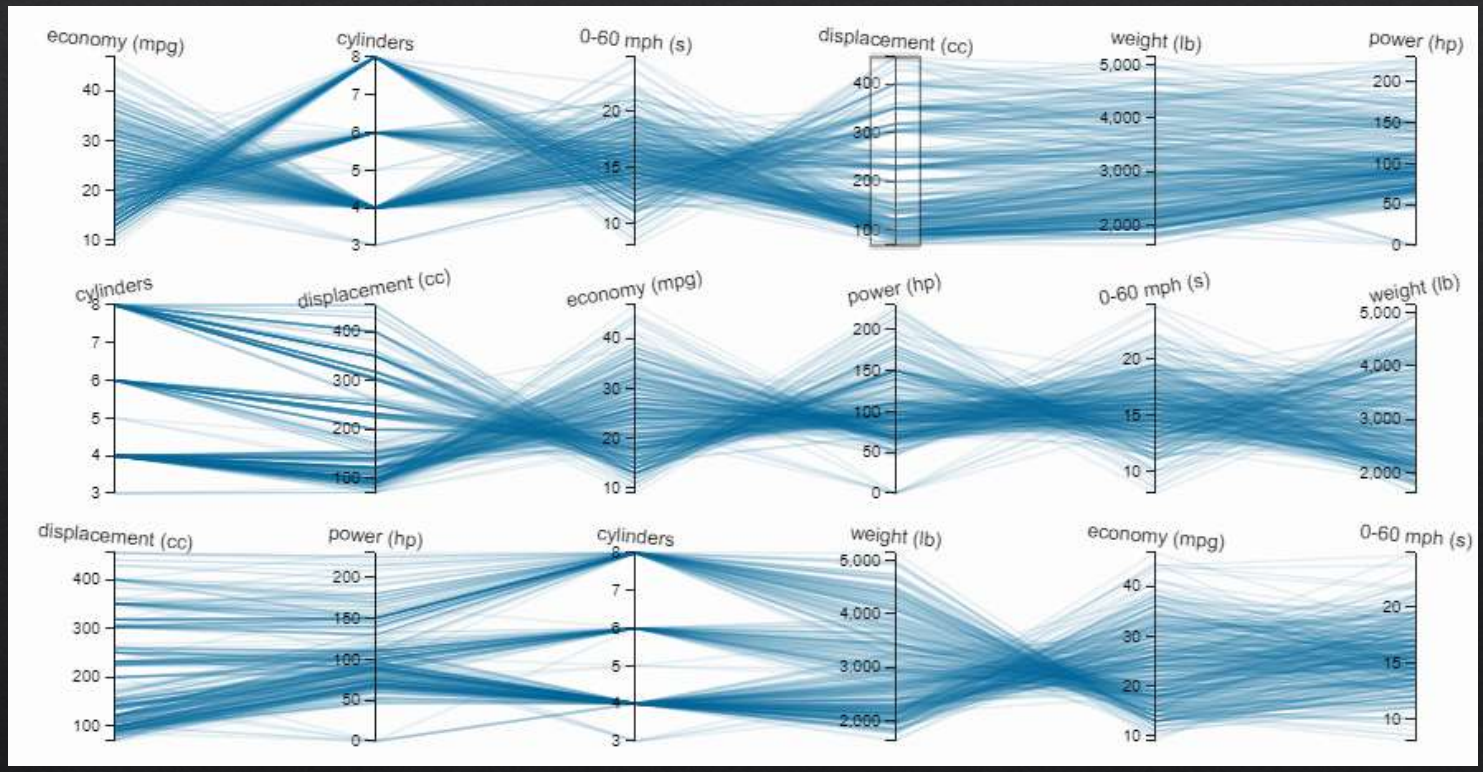
Dots represent the neighborhoods. The pink patch contains the "middle 75%" of the nieghborhoods

SPLOM is probably not Scalable – but it really shows all possible pairs!

Is there anything in the realm of Parallel-Coordinates which is similar to SPLOM?

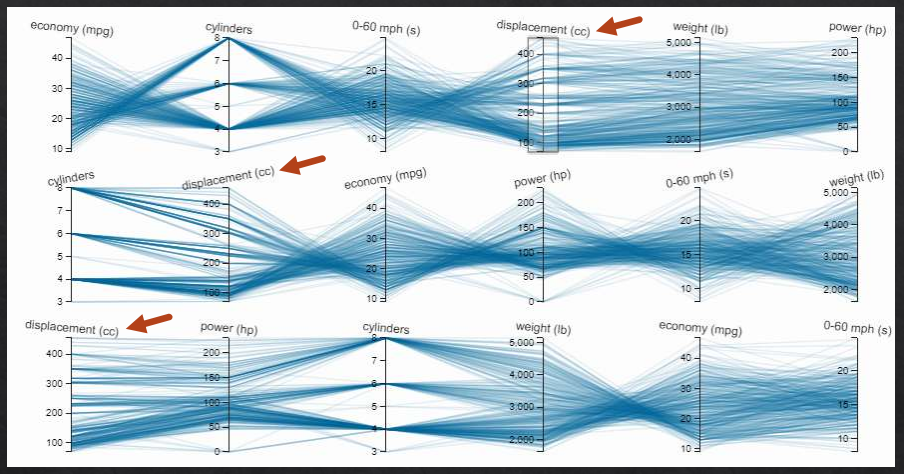
Parallel-coordinates matrix (PCM)

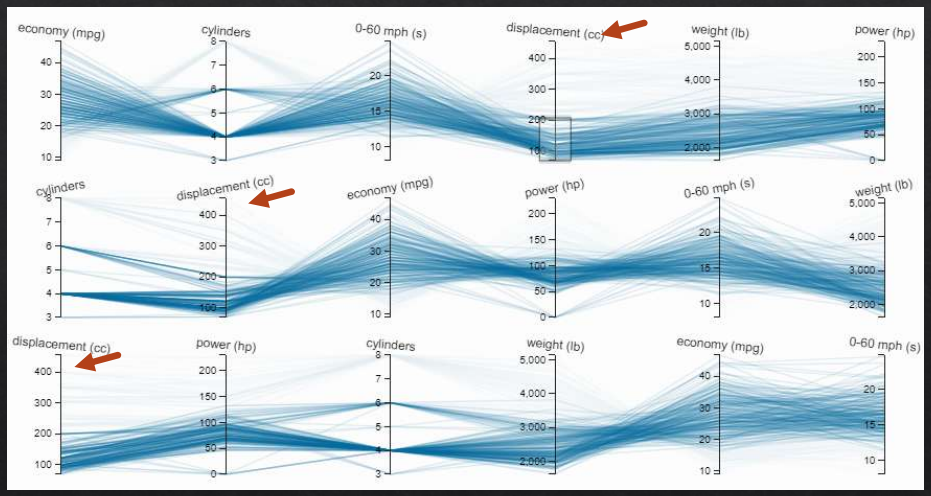
Heinrich, Stasko & Weiskopf, 2012



Parallel-coordinates matrix (PCM)

Heinrich, Stasko & Weiskopf, 2012



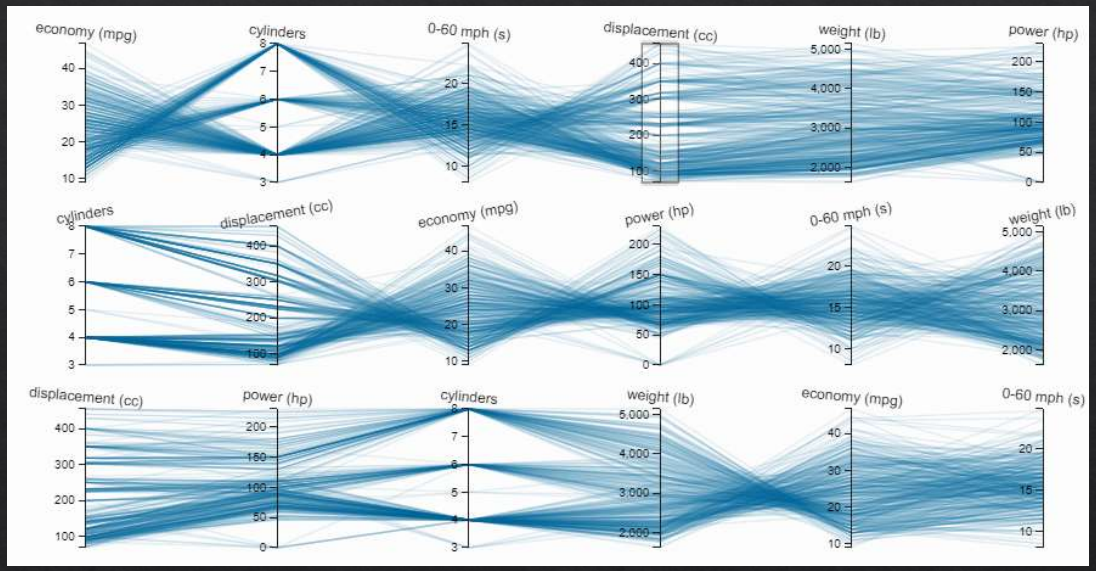


Parallel-coordinates matrix (PCM)

Model the set of dimensions as a graph with N nodes, where
each node represents a dimension and
an edge represents a pairwise relation.

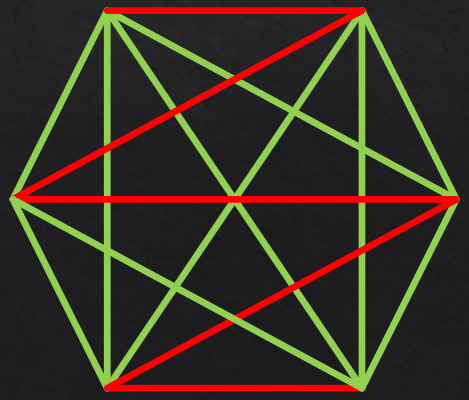
Our goal to see all pairwise relations is then represented as the complete graph K_N. Finding a set of axes orders that satisfy our requirements from above is then simply a matter of finding an appropriate Hamiltonian path decomposition of K_N.

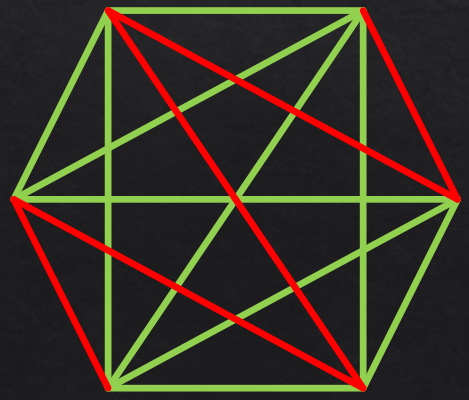
How to get PCM?
Compute a Hamiltonian Path Decomposition

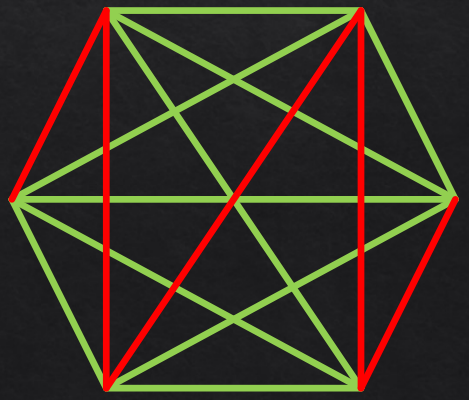


How to get PCM?
Compute a Hamiltonian Path Decomposition

(6 choose 2) = 15 = (3*5)



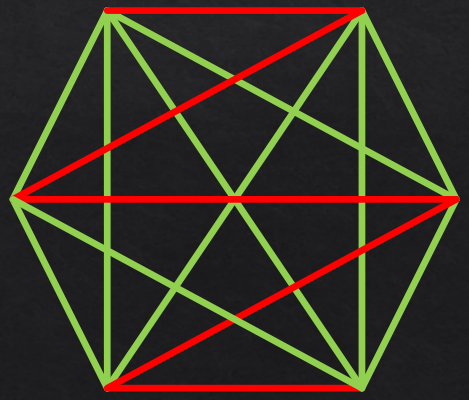


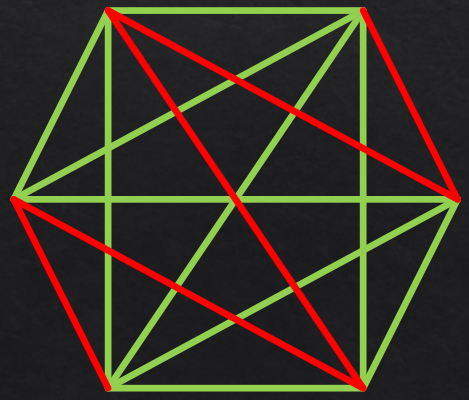


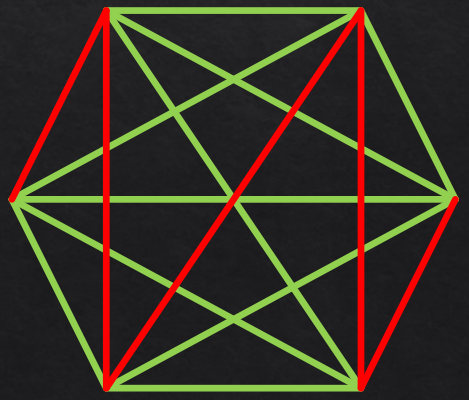
How to get PCM?
Compute a Hamiltonian Path Decomposition

(6 choose 2) = 15 = (3*5)

Think: why all the paths are covered?



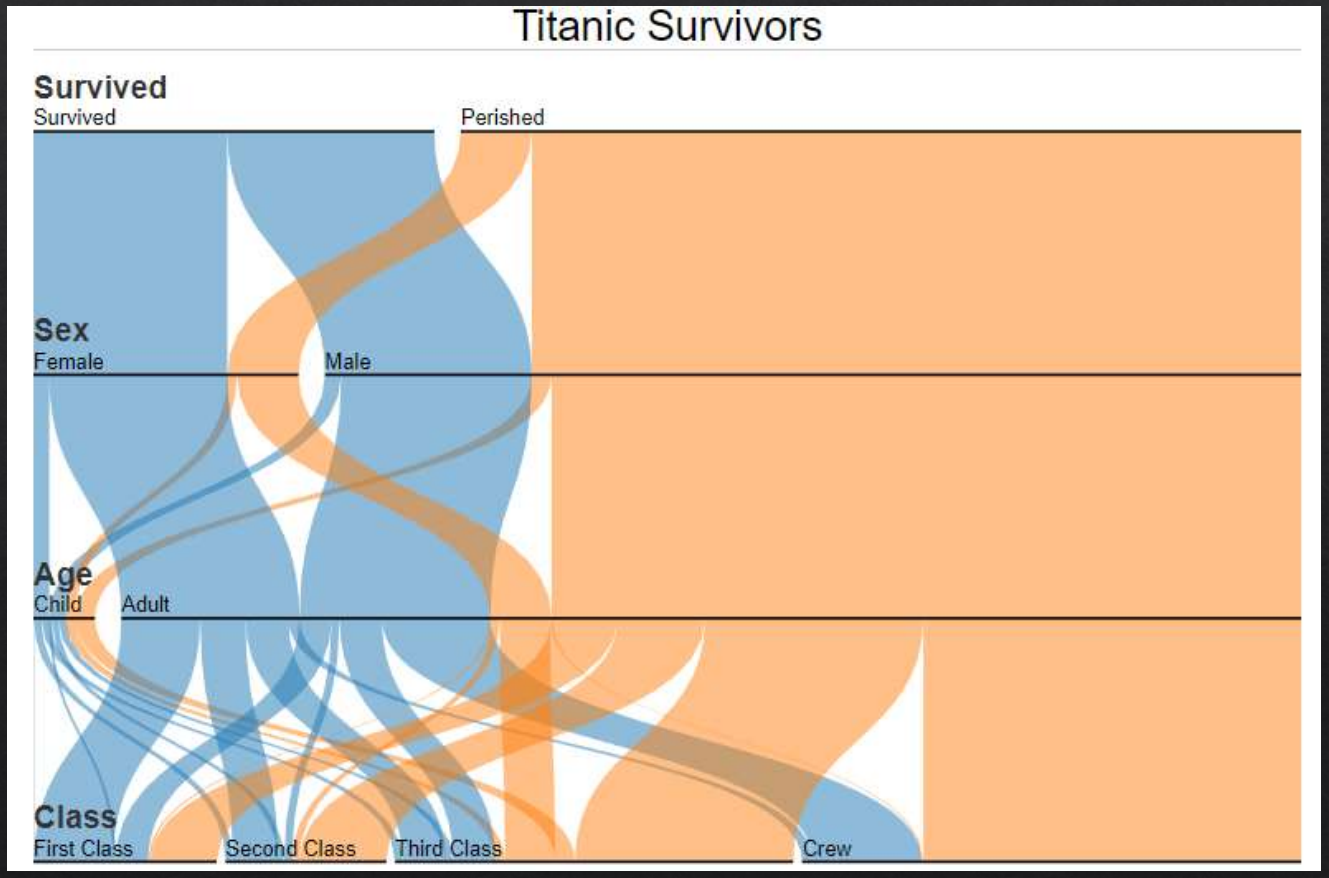




Dealing with Categorical Dimensions

Parallel Sets

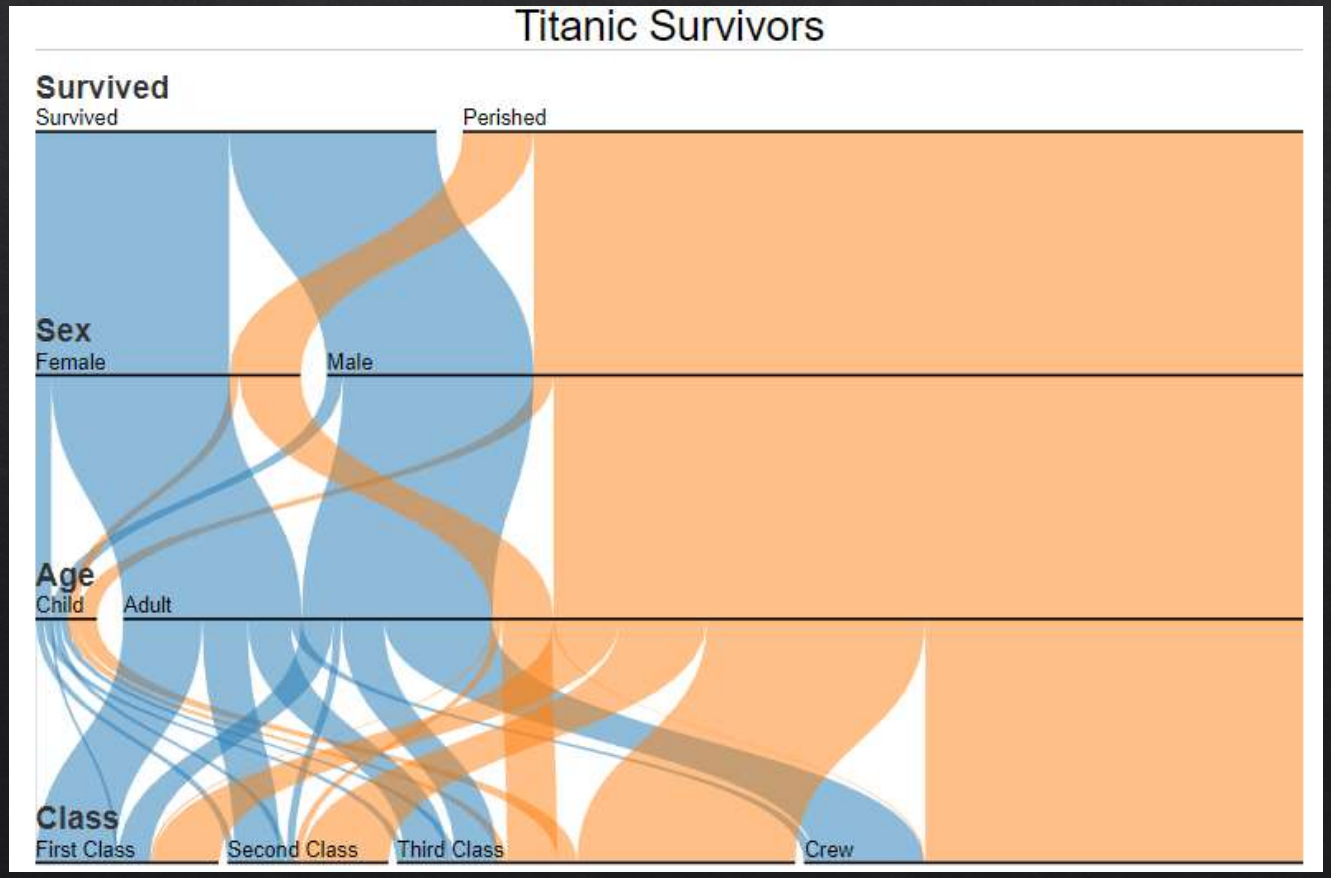
https://www.jasondavies.com/parallel-sets/



Dealing with Categorical Dimensions

Parallel Sets

https://www.jasondavies.com/parallel-sets/

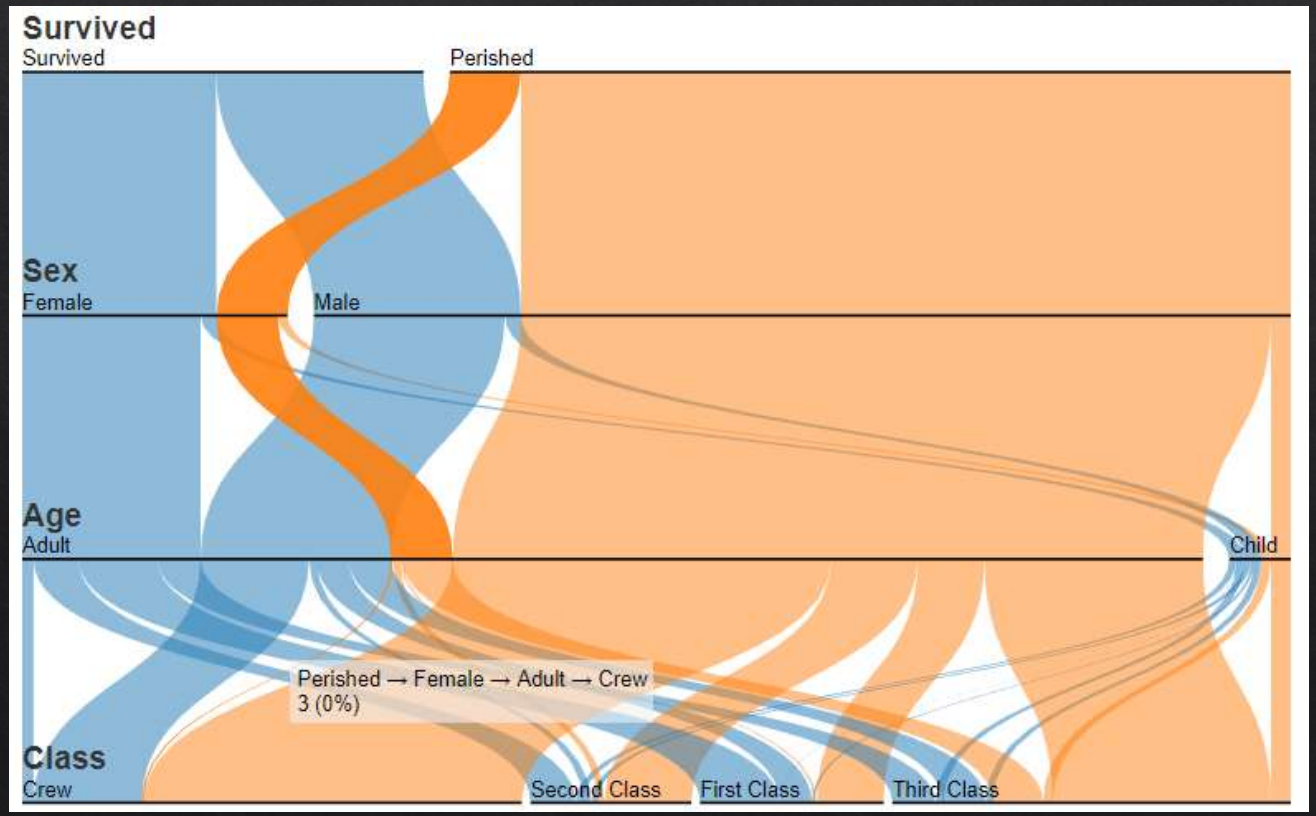


Is there any Female Crew who Perished?

Dealing with Categorical Dimensions

Parallel Sets

https://www.jasondavies.com/parallel-sets/

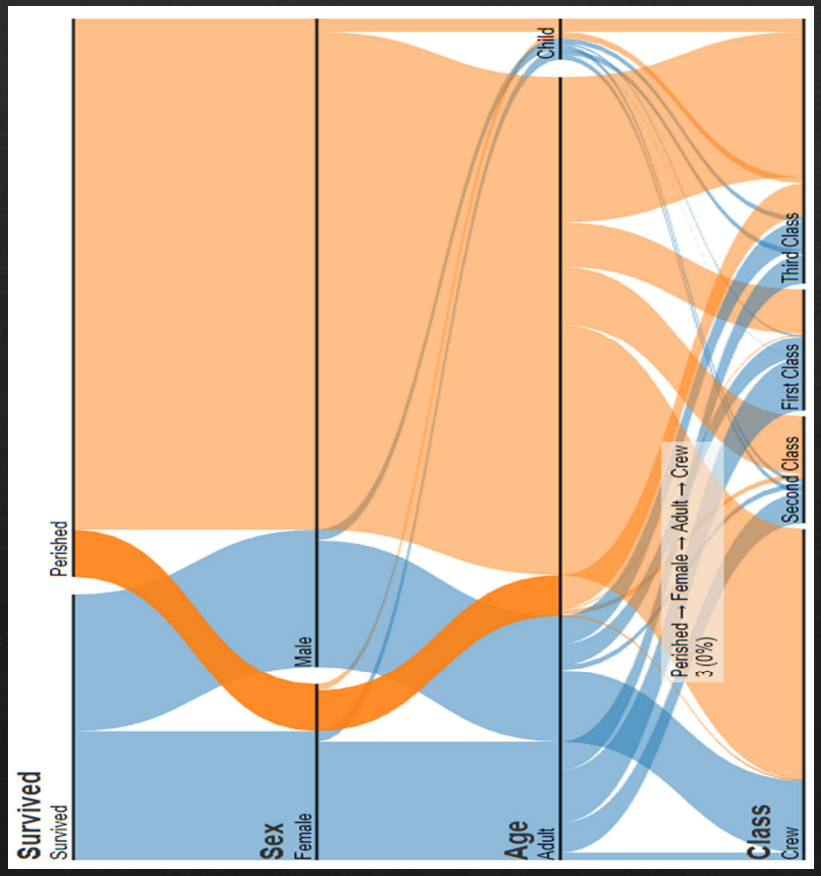


Is there any Female Crew who Perished?

Dealing with Categorical Dimensions

Parallel Sets

https://www.jasondavies.com/parallel-sets/



Parallel Coordinates !!
Really?

Motivation for **Parallel Sets**

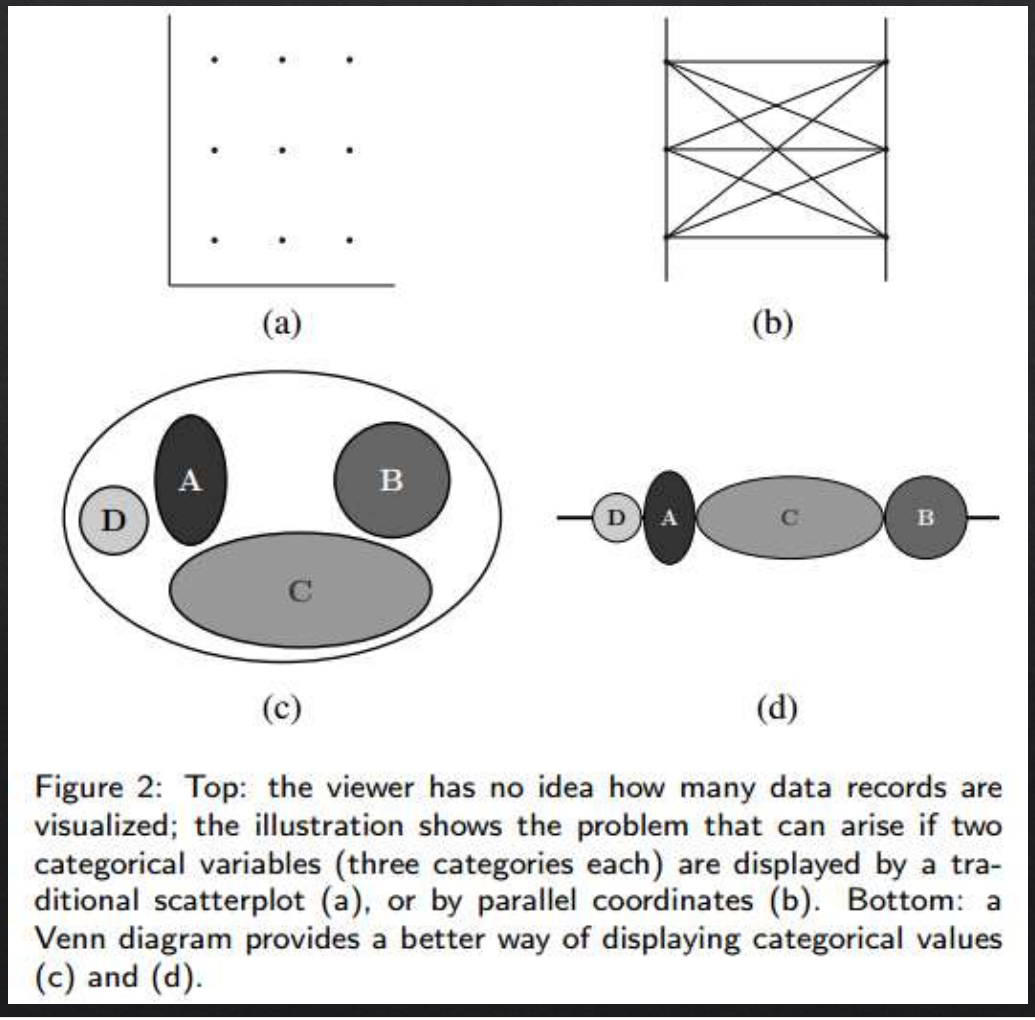


Figure 2: Top: the viewer has no idea how many data records are visualized; the illustration shows the problem that can arise if two categorical variables (three categories each) are displayed by a traditional scatterplot (a), or by parallel coordinates (b). Bottom: a Venn diagram provides a better way of displaying categorical values (c) and (d).

Motivation for **Parallel Sets**

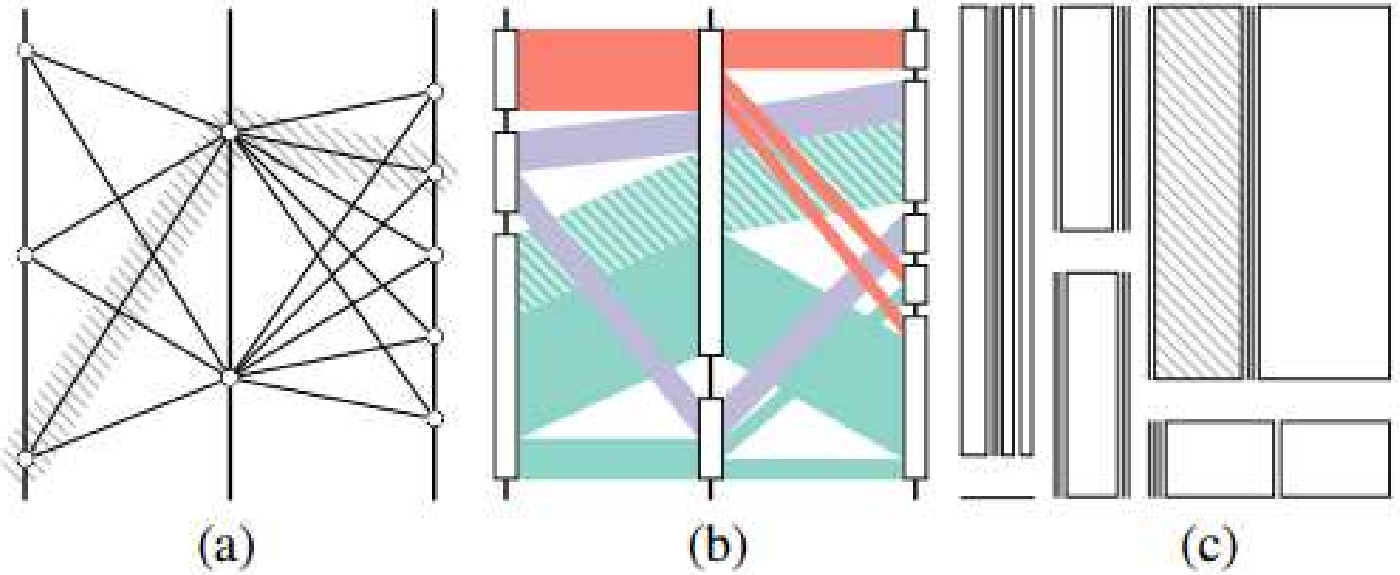


Figure 3: Three different visualization techniques displaying the same data: (a) the categories are represented by points on continuous axes in parallel coordinates, (b) Parallel Sets show the frequencies of categories and relations, and (c) a Mosaic Display provides a compressed view of the data (the hatched parts represent the same subset).

Quote from Authors
(Bendix , Kosara & Hauser, 2005)

Interactive visual analysis implies two requirements for a visualization technique: (1) an adequate visual metaphor that offers the user a comprehensible mental vision of the abstract data, and (2) a powerful, user-friendly, and user-driven interaction scheme.

Parallel Sets fulfills both requirements. It adopts the layout from parallel coordinates (that makes the displayed dimensions visually independent from each other), but uses a frequency based representation for categorical data variables, since frequency data is best represented by areas and not by individual data points (thus, the visualization becomes independent to the number of displayed data records).

For data exploration, the dynamic layout and our sophisticated interaction scheme are important: adding dimensions to the view by drag and drop, reordering dimensions and categories, dimension composition, highlighting, and so on.

Learning Goal

- Understanding the relation between scatter plot and parallel coordinate plots (PCP)
- Learn the interactions that may be used in PCPs
- Able to state real life usages of parallel coordinates
- Learn techniques to deal with the over-plotting and axis ordering problem
- Describe the basic idea behind parallel coordinate matrix (PCM) construction
- Get familiar with the ideas of visualizing large scale data using PCP
- Understanding relation between PCP and Parallel set