A Taxonomy of Clutter Reduction for Information Visualisation

Geoffrey Ellis and Alan Dix

Abstract — Information visualisation is about gaining insight into data through a visual representation. This data is often multivariate and increasingly, the datasets are very large. To help us explore all this data, numerous visualisation applications, both commercial and research prototypes, have been designed using a variety of techniques and algorithms. Whether they are dedicated to geo-spatial data or skewed hierarchical data, most of the visualisations need to adopt strategies for dealing with overcrowded displays, brought about by too much data to fit in too small a display space. This paper analyses a large number of these clutter reduction methods, classifying them both in terms of how they deal with clutter reduction and more importantly, in terms of the benefits and losses. The aim of the resulting taxonomy is to act as a guide to match techniques to problems where different criteria may have different importance, and more importantly as a means to critique and hence develop existing and new techniques.

Index Terms— Clutter reduction, information visualisation, occlusion, large datasets, taxonomy.

✦

1 INTRODUCTION

Information visualisation is essentially about data, visual displays, people and their quest for understanding. The data is often very large, the visual displays are relatively small and hence we must explore ways, using the available computer hardware and software, to make this acquisition of knowledge as easy and enriching as possible.

In essence, too much data on too small an area of the display will result in visual clutter, which in turn diminishes the potential usefulness of the visualisation, especially when the user is exploring the data rather than posing specific questions. This problem is of course not new, but has been exacerbated by the vast amount of data generated by government and commercial organisations and no doubt in the future by ubiquitous sensing [21].

In the last two decades a large number of visualisation applications have been developed and often, due to the diversity of the knowledge domain and perhaps even more the diversity in the solutions, users and visualisation designers have found difficulty in assimilating the opportunities available to them. The problem is further compounded by the lack of comparability based on usability studies, partly due to the relatively low number of such studies but also because information visualisations are particularly hard to evaluate [5].

It is important to have a clear understanding of clutter reduction techniques in order to design visualisations that can effectively uncover patterns and trends within overcrowded displays. It should be noted that we are focussing on explorative visualisation where the user is unaware of the potential information and knowledge is locked within the mass of data.

Note that this paper is not presenting a novel visualisation technique or empirical results. Instead we have performed a systematic analysis of existing systems and literature in order to create a better understanding of the various strengths and weaknesses of different approaches. We are obviously approaching this work in the light of our own interest in sampling techniques and hence our emphasis on adopting a systematic strategy in establishing effective criteria for clutter reduction. Our aim is not to sell our own work, but, as objectively as possible, establish the means to critique and analyse work in the area. In particular we see the resulting assessments, as formative rather than summative.

- *Geoffrey Ellis is at Lancaster University, E-Mail:g.ellis@comp.lancs.ac.uk.*
- *Alan Dix is at Lancaster University, E-Mail:alan@hcibook.com.*

Section 2 provides an overview of classification schemes that have been developed for information visualisations and illustrates that only a few of them have focussed on techniques available for dealing with overcrowded displays.

Section 3 presents a set of clutter reduction techniques which were identified following a thorough review of the literature. These are classified as *appearance*, *spatial distortion* and *temporal*; a description of each technique is given. We then consider how our set of clutter reduction techniques differ from Ward's taxonomy [48].

However, a set of techniques is not particularly useful on its own unless one has a way to compare them. Section 4 thus describes a set of criteria, expressed as benefits. These criteria are based on our own experience and more importantly on an analytical review of the literature to uncover how researchers describe the benefits of their visualisation in terms of clutter reduction. The establishment of these criteria is at the heart of this work.

In Section 5, we present our taxonomy of clutter reduction in information visualisation, which shows the possible benefits of each technique, and discusses exceptions and cases which are not so clear cut.

Section 6 reflects critically on the process of creating our taxonomy and evaluating the criteria. In addition, we discuss ways in which we have used the taxonomy ourselves and provide an insight into some possible uses.

2 INFORMATION VISUALISATION CLASSIFICATION SCHEMES

Whilst taxonomies have been proposed covering various aspects of information visualisations few mention clutter reduction. This section provides a brief review of existing taxonomies or surveys, presented in chronological order. Note that we have not looked into graph drawing as there is already a large body of work which deals with the effective layout of graphs.

In the early nineties Leung et al [35] presented a comprehensive review of distortion-oriented visualisations. Although it does not mention clutter directly, in the conclusion it states that "other non-distortion techniques, such as information suppression, should be investigated further since they are potentially powerful".

Shneiderman's 'type by task' taxonomy [43] illustrates how high level tasks (overview, zoom, filter, details-on-demand, relate, history and extract) can be applied to some basic data types (1, 2 & 3-dimensional data, temporal and multi-dimensional data, tree and network data). In terms of clutter reduction he points out that, zooming reduces the amount of data as does filtering which "filters out uninteresting items". Wong et al [51] review the developments in multidimensional multivariate visualization over the years. Although clutter is not raised when discussing various techniques, in the conclusion it states that "scientists have to deal with data that is many thousand times bigger than the number of pixels on display".

A decade later we now have more pixels to play with, but dataset sizes have increased at a significantly greater rate!

Keim [28] presents a classification of data visualisation techniques (geometric, icon-based, pixel-oriented, hierarchical, graph-based and hybrid), interaction techniques (mapping, projection, filtering, link&brush, zoom, detail on demand) and distortion techniques (simple, complex). He mentions sampling, querying, segmentation and aggregation as methods to reduce the amount of data displayed. In addition he gives techniques for dimensional reduction including multidimensional scaling. However, we will not be considering this as clutter reduction as reducing dimensions does not reduce the number of data points, except in the special cases where, for example, it would reduce the elements in a scatterplot matrix.

Card et al [9] present a data-centric approach to classifying visualisations based on succinct mappings between the data and presentational properties. In a later paper, Card et al [10] proposed the Reference model for Visualisation which introduces human interaction into the process of mapping the data to a visual form. In so doing, it categorises techniques such as zooming, focus+context, magic lens and dynamic queries which are forms of clutter reduction.

Chi's taxonomy [13] takes a more detailed, processing-centric approach to classifying visualisation techniques going from 'value' to 'view' via analytical and visualisation abstractions. The transformations used to change state are also detailed for a large number of example visualisations. In doing this, Chi shows how techniques can be built using a modest number of operators. It is interesting to note that dynamic value filtering occurs at the analytical, visualisation stages as well as in the user view stage. Other clutter reduction operations such as zoom, view filtering, level-filtering and change distortion focus are included at the view stage.

A comprehensive review of data clustering algorithms is given by [25] and [36], the latter focusing on massive datasets. These do not deal directly with display clutter but they provide a means of pre-processing the data into a more manageable set.

Finally, Ward's taxonomy [48] explicitly deals with glyph placement strategies which have a significant influence on display clutter. He considers four aspects of placement, (a) position calculated from the data or determined by the structured representation of the data, (b) degree of overlap allowed, (c) screen utilisation and (d) localised displacement to improve visibility. In addition, he discusses a variety of distortion techniques to reduce clutter and overlap. He goes on to suggest that user control of distortion is necessary and recommends smooth animation between the original and distorted views. Ward states that placement techniques are a tradeoff between efficient display use, amount of occlusion and distortion, and recommends that the user should be able to alter these dynamically. Furthermore, he proposes that the system could analyse congestion and do this automatically and notes that this is a relatively unexplored area. He raises many issues pertinent to clutter reduction and presents some general guidelines on choosing a particular data or structure driven placement strategy based on dataset size and task.

3 CLUTTER REDUCTION TECHNIQUES

In our work on clutter reduction [16, 17, 18, 19] we have investigated the application of sampling and one reason for starting this classification was to compare sampling with other techniques. From our survey of the literature related principally to 2D visualisation applications (80 papers), a set of clutter reduction techniques were identified. Table 1 lists the clutter reduction techniques divided into three groups. The first group, appearance, lists those techniques which tend to affect the look of the data item. Change point size, change opacity are self-explanatory; sampling and filtering are included as they have a dramatic affect on the appearance – the item disappears! Clustering often results in a different representation of the group of individual lines or points so

Table 1. Clutter Reduction Techniques

	clutter reduction technique	examples
appearance	sampling	[15, 16, 18, 40]
	filtering	[1, 3, 8, 39, 45]
	change point size	[4, 15, 39, 52]
	change opacity	[22, 27, 32, 49]
	clustering	[12, 23, 33, 54]
spatial distortion	point/line displacement	[3, 46, 47, 50]
	topological distortion	[1, 11, 31, 34, 42]
	space-filling	[4, 44]
	pixel-plotting	[26, 30, 41]
	dimensional reordering	[38]
temporal	animation	[14, 27, 47]

is also included in the appearance group. Note that there are other appearance attributes such as colour, blurriness and texture which have been used at part of a clutter reduction strategy and these will be included in the discussion of techniques which discriminate points/lines in Section 5.

Spatial distortion includes techniques that displace the point or line in some way. The first technique in this group, point/line displacement adjusts the position of each data item, whereas the next technique, topological distortion stretches the background, either non-uniformly (e.g. Fish-eye) or uniformly (e.g. zoom) taking the data items with it. Space-filling is essentially a non-overlapping rearrangement of large, rectangular points, often driven by a particular structure (e.g. Treemaps). Pixel-plotting techniques plot data items as single pixels in order to pack as much data as possible into the available screen space. Dimensional reordering is perhaps a special case in which the attribute axes of parallel coordinate plots are rearranged and hence can be thought of as a spatial distortion. Lastly, animation techniques which are mentioned in the literature in relation to clutter reduction fall under temporal techniques.

At this point we should explain the difference between sampling and filtering. Sampling is the random selection of a subset of the data whereas filtering is the selection of a subset of data that satisfies a given criteria (e.g. year >= 2000 and filmtype = humour). If the user needs to look at a specific set of data or has an idea of what might be 'interesting' then filtering is ideal, otherwise sampling provides an alternative way of exploring the data without preconceptions.

As mentioned in Section 2, Ward's taxonomy deals explicitly with techniques which could influence the amount of clutter on the display. However, his classification differs in that he sub-divides the strategies into data-driven and structure-driven placement strategies based on distortion. Data-driven strategies include random jitter which we classify as *point/line displacement*, GridFit's [29] point reallocation which we classify as *topological distortion* and Woodruff's replacement which we classify as *change point size*. Under structure-driven placement we regard fish-eye, pliable surfaces and hyperbolic views as *topological distortion*, whereas jitter, position distortion and space padding fall under *point/line displacement*. We feel *point displacement* and *topological distortion* as being different, despite having a similar effect on the display screen due to the underlying users mental model. This is most clear when the visualisation is overlaid on some other relevant graphic such as a map: in the case of *point displacement* the map would not change, but points may be drawn in slightly different positions than their actual geo-coordinates; in the case of *topological distortion* the underlying map would also be distorted – points are in the 'right positions, but space has changed. This is discussed further in Section 5. Finally, we regard Ward's aggregation and deletion as variants of

spatial clustering (in Section 5 we see that clustering on spatial coordinates has some differences from non-spatial clustering).

4 CLUTTER REDUCTION CRITERIA

In order to compare the clutter reduction techniques mentioned in the last section, we have come up with a set of criteria, which have been reduced to eight major benefits to the system and subsequently the user. Some of these are based on what we consider to be desirable features and others have been added after carrying out a literature survey of around 50 papers, from which some 68 benefits stated or implied by their authors, were extracted. An example of a shortened criteria search records is presented in Table 2. In addition to the four columns shown, data was kept on any disadvantages noted by the authors and each visualisation was classified according to clutter reduction techniques employed.

This highly systematic methodology for establishing the criteria we deemed essential to avoid simply reproducing the criteria that we personally consider, and would naturally be drawn from our own research and concerns. Thinking explicitly about criteria is, we believe, important but is not always easy. This was evident as many papers were not explicit about the criteria that drove the work or were used to determine success; in such cases the criteria would often be buried deep in the text.

This process led to eight high-level criteria. These are listed below together with an explanation of why these particular ones were chosen. Examples of each are presented in the relevant figures at the end of the paper.

avoids overlap
This is a major benefit cited by many researchers and perhaps an obvious one. These include *reduce clutter* [e.g. 45, 53, 52], *ability to see/identify patterns* [2, 18, 31], *gain a better understanding of the network graph* [40], *have less hidden data* [16], *avoid the problem of losing information* [31], *improve a messy display* [54], *provide efficient browsing on small displays* [14], *give more display space to points/nodes* [33, 3], *see more detail* [23, 34] and *see the number of points* [46, 31]. See Figures 1, 2 and 7 for examples.

keeps spatial information
Obviously for geo-spatial data, the x-y position of a point is significant. However, the accuracy to which the user can measure the absolute position of a point is questionable and other factors such as landmarks (e.g. representations of physical or political boundaries) may have a greater influence on our spatial awareness. It could be argued that when searching for patterns within the data, only relative positions are important, namely those which essentially define the clusters of points or lines so absolute positions hold less importance. As a result, the criteria would also pertain to keeping relative spatial information. Similarly, for some scale-less visualisations (e.g. Treemaps, graphs, self-organising maps) it is the relative position of data items which conveys information to the user as the absolute position is not quantifiable. See Figures 3 and 4 for examples.

can be localised
Localised in this context means a specific region or regions of the display. The most common manifestation of this is in focus and context techniques, although non-linear distortion or sampling would also meet this criteria. Examples of benefits of localisation include, *reducing the clutter in localised regions to reveal information underneath* [50], *providing an overview and detail in a temporal dataset* [39], *allowing the user to examine small items in detail whilst keeping in context* [44], *avoid losing information in low density areas when reducing overplotting in high density areas* [7]. See Figures 1, 3, 4 and 5 for examples.

is scalable
Clutter reduction technique which have the ability to cope with very large datasets would seem to be a desirable characteristic. Several papers simply refer to large datasets and few quantify this such as *this method is only limited by the number of available pixels* [26]. See Figures 1, 6 and 7 for examples.

is adjustable
Most visualisations are interactive in that they allow the user to control some aspect of the visual display. For example, from the early days of dynamic queries [1], users could reduce the size of the results set by adjusting a slider control. This criteria is concerned with the ability to adjust some parameter of the system that influences the degree of display clutter. The benefits given in the literature include, the *ability to adjust the sampling rate to an appropriate level to see patterns* [16], *interactive adjustment helps the user understand the cluster distribution* [12], *users can set visual characteristics and hence tune the clutter reduction* [6] and the *ability to highlight different aspects of data visualized with parallel coordinates* [27]. See Figures 1, 3, 6 and 8 for examples.

can show point/line attribute
It is often useful, especially when displaying multivariate data, to map the value of one or more attributes to the colour, shape, opacity of the displayed points or lines. See Figures 2, 5 and 7 for examples.

can discriminate points/lines
It is desirable to distinguish between individual points or lines so they can be easily identified in what may be a crowded display. This is backed up by benefits such as *help to differentiate overlapping points* [8], *distinguish between individual points in a crowded display* [22], *reduce apparent clutter by making each cluster of lines distinct* [23] and *the user can quickly see a particular set of points within a crowded display* [32]. See Figures 8 and 9 for examples.

can see overlap density
If overplotting is present in the display, then users ought to be made aware of this, otherwise they may not realise that data is hidden from their view. We may also want to see where the higher density regions are and gauge the amount of overplotting. Fekete mentions that *it would be useful to see the amount of overplotting and hence the distribution of data* [22] and Wegman argues that *a density plot shows the degree of overplotting and also helps to discriminate individual lines such as outliers* [49]. See Figures 8 and 9 for examples.

A number of authors comment on the efficiency or speed of their algorithms or graphical techniques. Although this is important in terms of improving the interactivity of the visualisation and/or the ability to cope with large datasets, this is more of an implementation issue. Some other benefits we came across such as *avoid overwhelming the whole view by only showing additional detail within the lens* [45], *reduce distracting outliers which often confuse the user* [38] and *automatic clutter reduction reduces the need for user interaction* [19, 6] were specific to one or two applications and hence have not been included. Finally we did consider 'provides a good mental model for the user' as a useful criteria, but found it difficult to classify all the clutter reduction techniques as it is more an assessment of a visualisation system as a whole, so this has also been left out.

Table 2. Example of Clutter Reduction Criteria Search Records

paper	criteria found	classified as	comments
Fekete 2002 – Million items	closely packed points often merge so a good idea to be able to distinguish individual points in a crowded display	can discriminate points/lines	use smooth shaded rectangle (tilt + fog) implemented in graphics hardware
Keim 2004 - PixelMaps	can see all the points so as not to lose point attribute information	avoids overlap can show point/line attribute	densest regions get the space required to place all the data points close to each other

Table 3. Clutter Reduction Taxonomy

Key: ✓ satisfies criterion; ✗ does not satisfy criterion; $^{+}$ some exception/ special cases (discussed below)
More complex cases: possibly/partly (see explanations below)

		1 sampling	2 filtering	3 point size	4 opacity	5 clustering	6 point/line displacement	7 topological distortion	8 space-filling	9 pixel-plotting	10 dimensional reordering	11 animation
A	avoids overlap	possibly	possibly	possibly	partly	possibly	✓$^{+}$	possibly	✓$^{+}$	✓$^{+}$	partly	✓$^{+}$
B	keeps spatial information	✓	✓	✓	✓	partly	✗$^{+}$	possibly	✓$^{+}$	possibly	✓	✓
C	can be localised	✓	✓	✓	✓	✗$^{+}$	✓	✓	✗	✗	✗	✓$^{+}$
D	is scalable	✓	✓	✗	✗$^{+}$	✓	✗	✗	✗	✗	✗	✓$^{+}$
E	is adjustable	✓	✓	✓	✓	✓	possibly	✓	✗$^{+}$	✗$^{+}$	✓	✓$^{+}$
F	can show point/line attribute	✓	✓	✓	✗$^{+}$	partly	✓	✓	✓	✓	✓	✓
G	can discriminate points/lines	✗	✗	possibly	✓$^{+}$	✓$^{+}$	possibly	✗	✗	✗	✗	✗
H	can see overlap density	✗	✗	✗	✓$^{+}$	possibly	✗	✗$^{+}$	✗$^{+}$	✗$^{+}$	✗	✗$^{+}$

5 TAXONOMY

In this section we take our set of clutter reduction techniques defined in Section 3 and classify them in terms of the criteria defined in the previous section. The resulting taxonomy is given in Table 3. We have placed a ✓ against those techniques that meet the given criteria and a ✗ against those that we feel do not meet the criteria. However, some technique-benefit combinations have limitations or are special cases or warrant a mention and these are marked with a $^{+}$ or some text and are discussed below. Note that 'possibly' indicates that the criteria is met in some situation but not in others; whereas 'partly' indicates that the criteria is only partly met in some situations. Some of the cases discussed are illustrated by figures at the end of the paper.

The assignment of ticks and crosses on Table 3 is based on *our own* assessment of *existing* systems and using the self assessment of authors within the literature where present. As with any assessment, another person might rate things differently. In order to mitigate this we discuss any cases that we feel are problematic. Because it is based on what systems actually do, we have not distinguished whether, for example, the presence of a cross means that it is fundamentally impossible for the given technique, or just not found. We should note that the purpose of this table is not to assess the techniques against each other, still less to justify our own. The purpose is instead to act as a guide to match techniques to problems where different criteria may have different importance, and more importantly a means to critique and hence develop existing and new techniques.

avoids overlap

As row A of Table 3 shows, not all clutter reduction techniques avoid overlap.

A1 Sampling cannot avoid overlap altogether as data items are not displaced, but it can be used successfully to reveal hidden patterns. In our experience and as suggested by [48], an acceptable amount of overlap can be tolerated and allowing the user to adjust the sampling rate (or overlap amount with auto-sampling [19]) is an optimum solution. (See Figure 1).

A2 Like sampling, filtering cannot necessarily avoid overplotting altogether, yet it can reduce the results set sufficiently to reveal the desired relationships for the chosen data range. (See Figure 3).

A3 Large points may conceal or partially conceal any points underneath (plotted earlier), so reducing their size may be beneficial. However, there are trade-offs between overlap and points size. If the point colour represents some attribute value, then too small a point makes the colour difficult to discern. Also, a glyph representing multiple attributes may need simplifying when reduced in size, resulting in a loss of data. Furthermore, images need to be large enough to see the required detail [15].

A4 Although a change in opacity cannot avoid overlap, it can reveal a small number of underlying or partially overlapping points. (See Figures 6 and 9).

A5 Clustering here means reducing the number of data items in order to simplify the plot. So it can be used to avoid overplotting by either representing a group of points by a single point (the size of which represents the number of original points [52]) or a group of lines by a single line or band (as in some parallel coordinate clutter visualisations [e.g. 53, 27]). Zhang et al [54] instead uses 'zip zooming' to combine several adjacent axes on parallel coordinate plots to reduce overlap. (See Figure 6).

A6, A8, A9 Some visualisations that employ point displacement (e.g. smart jitter [46]), space-filling (e.g. Treemaps [4]) and pixel plotting (e.g. TableLens [41], Information Mural [26]) are specifically designed to avoid overlap. However, they are limited by the number of pixels on the display and human visual acuity. Other methods (e.g. Mobile 2D scatter space [47], EdgeLens [50]) displace points or lines locally to reduce but not necessarily avoid overlap. (See Figure 5).

A7 Topological distortion involves stretching the virtual drawing surface either uniformly (zooming) or non-uniformly to give extra display space to the data point. Some visualisations avoid overlap by virtue of their design (e.g. PixelMaps [31]) while others such as pliable surface/rubber sheet interfaces [42, 11] do not. The size of the points do not change apart from semantic zoom visualisation [52] and Fisheye lens type distortion, if used as a magnifier. (See Figures 2 and 4).

A10 Dimensional reordering as applied by Peng et al [38] arranges the scatter matrix dimensions or parallel coordinate order with a view to minimizing their clutter measure.

A11 Animation used in Rapid Serial Visual Presentation [14] avoids overlap by showing a stack of images to the user in quick succession and in 'Cenimation' [20] overlapping data bubbles 'float' to the surface in quick succession, hence no data item is ever hidden completely.

keeps spatial information

Sampling, filtering, changes to opacity and point size all preserve spatial information.

B5 Clustering, by default, loses individual spatial information but as a group (or cluster) it can show aggregate values, typically via colour/shading/opacity. Good examples can be found in [27].

B6 The actual point displacement depends on density of the data. The higher the number of overlapping points, the greater the spatial distortion to accommodate the points without overlap. However, as discussed in Section 4 the amount of information actually lost is not necessarily directly dependent on the displacement as relative positions are perhaps a more determining factor.

B7 Topological distortion is similar to **B6** in that the amount of distortion from the normal x, y position is increased with the degree of overlap (density) but because the x-y space is being stretched or squashed it could be argued that the viewer does not lose spatial information, as long as sufficient landmarks are available. Keim et al [29] claim that in their visualisation the points are plotted as close to their 'original' position as possible. However, this perception of little or no spatial distortion must rely on providing appropriate spatial cues to the user. For example, Carpendale et al [11] attempt to inform the user of the amount of distortion by either superimposed grid or by shading. (See Figure 4).

B8 Space-filling techniques such as TreeMaps [4] and Sunburst [44] plot hierarchical data that is not geo-spatial in the first instance. So it is questionable whether this criteria can be applied. The level in the hierarchy is significant and as space-filling does keep this information, one could say that this criteria is met.

B9 Pixel-oriented techniques such as Information Mural, Table Lens and Pixel bar charts do retain the spatial information of the original data. However pixel-spirals [30] plots data based on one of a set of packing algorithms and is not based on any attribute of the data apart from order. As with **B8**, it is questionable whether this criteria can be applied.

can be localised
Sampling can be localised to a particular region of the display by using a lens metaphor [18] and also through non-uniform sampling across a scatterplot [7]. Likewise EdgeLens restricts line displacement to a lens [50] to clarify node and edge relationships. By adopting a 'Magic Lens' approach [45], filtering could also be localised as well as changes to point size, opacity and animation effects. Topological distortion is localised with Fish-eye lenses and also with the pliable surface interfaces. However, it is difficult to see how clustering, pixel-oriented techniques or dimensional reordering could be restricted to a particular region of the display. (See Figures 3 and 5).

C5 We have placed a cross against this as clustering based on arbitrary similarity measures does not necessarily maintain any spatial locality, hence passing a lens over a set of points to see clusters would be largely meaningless. However, if the clustering is based on spatial attributes alone, then it would make sense if the lens was large compared with a typical cluster diameter.

C11 Animation can be used to show examples of the data in an overplotted area or in the case of RSVP [14], cycle through the images on a particular stack on the display. (See Figure 7).

is scalable
Sampling, filtering and clustering can all be scaled up to deal with very large datasets as they all inherently reduce the number of plotted points. The limiting factor is the computational resource available. Animation techniques [e.g. 14, 20] which show data items in sequence can deal with large data sets, however the time to show all the data would need to be taken into account. Reducing the point size is limited by the resolution of the display and visual acuity All the other techniques are ultimately limited by the number of pixels available on the display. With topological distortion, a very large number of data items within one area would lead to a significant spread of points (or stretching of the underlying topology), hence unmanageable distortion for the viewer. (See Figure 2).

D4 Healey [24] suggests that opacity is only useful when up to five items overlap.

D11 Some animation techniques [14, 20] which show data items in sequence can deal with very large datasets, however the time to show all the data would need to taken into account.

is adjustable
This criteria is looking at whether the amount or degree of clutter reduction can be adjusted interactively. Sampling rate, dynamic query range, point size, opacity and to some extent cluster size can all be adjusted. Often dimensional scaling algorithms can be adjusted by the user (e.g. [53]) to control the dimensional reduction. The magnification factor of localised topological distortion techniques (e.g. Fish-eye lens, pliable surfaces) can be changed dynamically and in some of the dimensional reordering visualisations implemented by Peng et al [39], the user can adjust the cluster width threshold. (See Figures 1 and 3).

E6 Only a few applications (e.g. Mobile 2D scatter space [47]) appear to allow the user to control the amount of displacement. In some other applications, the displacement will depend on the data density.

E8, E9 Although users can navigate through some space-filling visualisations (e.g. Sunburst [44]), they cannot necessarily control the amount of clutter reduction. This also seems to apply for pixel-plotting visualisations.

E11 The animation rate can easily be adjusted for temporal techniques and this affects how the user perceives the data, but may have little effect on the clutter reduction.

can show point/line attribute
All techniques apart from opacity and clustering do not affect the use of the physical attribute of the point/line (e.g. colour, shape) to represent another attribute, however one should be aware of the perception problems associated with packing pixels tightly. As mentioned earlier, reducing the point size can affect the perception of colour as well. It should be noted that if there is overplotting, the point/line attributes of the 'top' data item will be on view and hence the display is dependent on the order in which the items are plotted. Placing the items in a random order should also be considered [31].

F4 Reducing the opacity will diminish the significance of the data point, especially colour.

F5 Clustering shows aggregate values.

can discriminate points/lines
It seems desirable to distinguish between individual points or lines so they can easily be identified in a crowded display.

G3 An interesting side effect of geo-spatial semantic zoom [52] is that outliers by definition tend to be in sparse areas of the map and hence are given additional prominence if increased in size. This is ideal if one wants to identify these particular points, otherwise it may present a distorted view of the data distribution.

G4 Opacity is used with good effect in parallel coordinate plots to discriminate overlapping lines, especially in association with clustering [e.g. 27, 23]. There are other appearance attributes such as colour [23], blurriness [32], and texture [27] which would be effective as part of a clutter reduction strategy. (See Figure 6).

G5 Clustering algorithms can be used to detect outliers as well as create groups. This is used to good effect by Novotny et al [37] in their outlier-preserving visualisation of parallel coordinates.

G6 Wong et al [50] claim that curving lines in their EdgeLens technique helps disambiguate the connected nodes of a graph.

can see overlap density
H4 With careful adjustment, opacity can lead to helpful density maps [22, 49]. Aggregation can be used and the result represented as size, colour, opacity etc. But opacity is the only clutter reduction technique which promotes a visual indication of overlap density. (See Figure 9).

H5 Clustering does not inherently show the overlap density, however the aggregate value can be displayed, as discussed in **H4**.

H7 Topological distortion achieves this indirectly. Carpendale's pliable surfaces [11] can indicate the amount of overlap by showing the distortion necessary to spread out the points, but this is not really

quantifiable. Topological distortion does not however separate coincident points. (See Figure 4).

H8, H9 Obviously space-filling and pixel-plotting are designed to avoid overlap, thus there is no provision to show the overlap density.

H11 It is interesting to note that animation-based image browsing applications [e.g. 14] can, by virtue of the height of stack of images, show the number to be browsed. But this is making use of a pseudo-3D effect which is not part of this classification.

6 EVALUATING THE CRITERIA

In building the taxonomy of techniques and criteria summarised in Table 3, we have two aims: validity and utility.

We have attempted to ensure the *validity* of the chosen criteria by adopting a systematic and inductive approach – that is through the methodological rigour of the *process*. However, the inductive approach is limited by what we, and those we have studied, can conceive of now. New, techniques are being and will continue be developed. Many will fall under the broad scope of one or other of the identified classes of technique, but others will be radically new; so we would expect Table 3 to grow new columns occasionally. Criteria tend to be more stable than technology. However it may well be the case that tacit criteria are, by their nature, not stated explicitly in existing literature and so were missed in this work.

One way to increase confidence on the validity of this taxonomy would be to produce a more formal theoretical model of the interactions between information, perception and cognition in visualisation. Stated thus, this sounds more like a Grand Challenge for the whole area – however, it may be that in a restricted area, such as clutter reduction, this may be possible. Such a model would also act as a way to establish more soundly whether the assessments of Table 3 are about fundamental features of the techniques or just the way they happen to have been used to-date. It may also prompt new techniques or criteria. We have been considering ways to relate the criteria to high and low-level visualisation tasks or goals. However, for the moment, we regard the development of a substantially more grounded model as desirable, but very challenging, future work.

In terms of utility, we have tried to produce categories of techniques and criteria that are broad enough to apply easily. As we have emphasised several times, the purpose of this work is not to produce a closed categorisation and 'scoring' of techniques, but rather as a tool for thought. Certainly, as the authors discussed the construction of Table 3 this has been the case, for example, in clarifying the different properties of spatial and non-spatial clustering for **C5**.

One of the ways we have been using Table 3 ourselves is to see whether techniques can be combined in order to allow weaker aspects of one to compliment stronger aspects of another. In fact, simple combinations do not always work, and this often has led us to more deeply understand the techniques and their interactions. For example, clustering is strong in terms of scalability, but weak in terms of ability to show attributes unless those attributes can be reduced to a summary statistic. At first glance, pixel plotting or space filling would seem to be complimentary, but actually it is very hard to imagine combinations that would preserve clustering's scalability. Most often, as an initial observation, complementary techniques seem to be best combined through overlays, alternative views or drill downs. For example, one could imagine a pop-up space-filling representation of an individual cluster.

This said, we continued the above line of investigation, looking for techniques to compliment the weakness of clustering in displaying point/line attributes, such as colour, that cannot easily be aggregated. Animation techniques often do satisfy this criterion and this suggested incorporating animation that would cycle the attribute displayed for a cluster through the various specific values of its constituent members. In retrospect, this seems an obvious solution, and we would not be surprised to find it used elsewhere; the crucial thing is that for us, in the situation, it prompted new ideas.

Of course, the ultimate test of utility is whether other people, not just the authors, can use the taxonomy in their own research or practice, and we hope this paper will allow precisely that.

7 CONCLUDING REMARKS

We have come a long way since starting to evaluate our sampling-enabled visualisations. Having experienced many of the problems of conducting effective user studies, as highlighted at a recent workshop [5], we have adopted a more analytical approach to assess where sampling fits into the gamut of clutter reduction techniques. From a detailed review of the literature we have systematically selected a set of techniques and a set of criteria which can be used to compare them. Based on our own assessment of existing systems and using the self-assessment of authors where present, we have constructed a taxonomy which indicates the extent that each technique satisfies each criterion. In addition, we annotate the taxonomy with a discussion of any cases that we feel are problematic.

The purpose of our classification is to act as a guide to match techniques to problems where different criteria may have different importance, and more importantly a means to critique and hence develop existing and new techniques. In this sense it is more a tool for thinking about the appropriateness of cutter reduction techniques than a ranking system. We have given some examples of how we have used the taxonomy to gain an insight into existing and new techniques and we hope that researchers and practitioners will find it useful for such tasks.

REFERENCES

[1] C. Ahlberg, C. Williamson and B. Shneiderman, "Dynamic Queries for Information Exploration: An Implementation and Evaluation", *Proc. CHI'92*, Monterey, California, pp. 619-626, 1992, ACM Press

[2] C. Ahlberg and B. Shneiderman, "Visual Information Seeking: Tight Coupling of Dynamic Query Filters with Starfield Displays", *Proc. CHI'94*, Boston, pp. 313-317, 1994, ACM Press

[3] C. Ahlberg, "Spotfire: an information exploration environment", ACM SIGMOD, 25(4), pp. 25-29, Dec 1996

[4] B.B. Bederson, B. Shneiderman and M. Wattenberg, "Ordered and Quantum Treemaps: Making Effective Use of 2D Space to Display Hierarchies", *ACM Trans. on Graphics, 21(4)*, pp. 833-854, Oct 2002

[5] BELIV '06 , *Proc. AVI workshop on BEyond time and errors*, Eds. E. Bertini, C. Plaisant, G. Santucci, Venice, Italy, 2006, ACM Press

[6] E. Bertini and G. Santucci, "Quality metrics for 2D scatterplot graphics: automatically reducing visual clutter", *Proc. Smart Graphics'04*, Banff, Canada, pp. 77-89, 2004, Springer Verlag

[7] E. Bertini and G. Santucci, "Give chance a chance - modeling density to enhance scatter plot quality through random data sampling", *Information Visualisation, 5(2)*, pp. 95-110, June 2006

[8] D. Brodbeck, M. Chalmers, A. Lunzer and P. Cotture, "Domesticating Bead: Adapting an Information Visualization System to a Financial Institution", *Proc. InfoVis'97*, Phoenix, pp. 73-80, 1997, IEEE

[9] S.K. Card and J. Mackinlay, "The structure of the information visualization design space", *Proc. InfoVis'97*, pp. 92-100, 1997, IEEE

[10] S.K. Card, J.D. Mackinlay and B. Shneiderman, *Readings in Information Visualization: Using Vision to Think*, Chapter 1&2, 1999, Morgan Kaufmann

[11] M.S.T. Carpendale, D.J. Cowperthwaite and F.D. Fracchia, "3-Dimensional Pliable Surfaces: For the Effective Presentation of Visual information", *Proc. UIST'95*, 1995, ACM Press

[12] K. Chen and L. Liu, "A Visual Framework Invites Human into the Clustering Process", *Proc. Int. Conf. Scientific and Statistical Database Management*, pp. 97-106, 2003, IEEE

[13] E.H Chi, "A Taxonomy of Visualization Techniques using the Data State Reference Model", *Proc. InfoVis 2000*, pp. 69-75, 2000, IEEE

[14] O. de Bruijn and R. Spence, "Rapid serial visual presentation: a space–time trade-off in information presentation", *Proc. AVI 2000*, Trento, Italy, pp. 51-60, 2000, ACM Press

[15] M. Derthick, M.G. Christel, A.G. Hauptmann and H.D. Wactlar, "Constant Density Displays Using Diversity Sampling", *Proc. InfoVis'03*, Seattle, pp. 137-144, 2003, IEEE

[16] A. Dix and G.P. Ellis, "by chance: enhancing interaction with large data sets through statistical sampling", *Proc. AVI'02*, L'Aquila, Italy, pp. 167-176, 2002, ACM Press

[17] G.P. Ellis and A. Dix, "Density control through random sampling : an architectural perspective", *Proc. Information Visualisation 2002*, London, pp. 82-90, 2002, IEEE

[18] G.P. Ellis, E. Bertini and A. Dix, "The Sampling Lens:Making Sense of Saturated Visualisations ", *CHI'05 Extended Abstracts*, Portland, USA, pp. 1351-1354, 2005, ACM Press

[19] G.P. Ellis and A. Dix, "Enabling Automatic Clutter Reduction in Parallel Coordinate Plots", *IEEE Trans. Visualization and Computer Graphics (Proc. InfoVis'06), 12(5)*, pp. 717-723, Sept 2006

[20] S. Engle, J. Shearer, M. Ogawa, S. Haroz and K-L. Ma, "Free Your Data! Cenimation: Visualization for Constrained Displays", *InfoVis'06 Contest*, Baltimore, 2006, ACM Press

[21] I.A. Essa, "Ubiquitous Sensing for Smart and Aware Environments ", *IEEE Personal Communications*, 2000, IEEE

[22] J-D. Fekete and C. Plaisant, "Interactive Information Visualization of a Million Items", *Proc. InfoVis'02*, pp. 117-124, 2002, IEEE

[23] Y-H. Fua, M.O. Ward and E.A. Rundensteiner, "Hierarchical Parallel Coordinates for Exploration of Large Datasets", *Proc. Visualization'99*, Los Alamitos, CA, pp. 43-50, 1999, IEEE

[24] C.G. Healey, K.S. Booth and J. Enns, "Visualizing Real-Time Multivariate Data Using Preattentive Processing", *Trans. Modeling and Computer Simulation, 5(3)*, pp. 190-221, 1995

[25] A.K. Jain, M.N. Murty and P.J. Flynn, "Data Clustering: A Review", *ACM Computing Surveys, 31(3)*, pp. 264-323, Sept 1999

[26] D.F. Jerding and J.T. Stasko, "The Information Mural: A Technique for Displaying and Navigating Large Information Spaces", *IEEE Trans. Visualization and Computer Graphics, 4(3)*, pp. 257-271, 1998

[27] J. Johansson, P. Ljung, M. Jern and M. Cooper, "Revealing Structure in Visualizations of Dense 2D and 3D Parallel Coordinates", *Information Visualization, 5*, pp. 125-136, 2006

[28] D.A. Keim, "Visual Techniques for Exploring Databases", Invited tutorial KDD'97, Newport Beach, CA, 1997

[29] D.A. Keim and A. Herrmann, "The Gridfit Algorithm: An Efficient and Effective Approach to Visualizing Large Amounts of Spatial Data", *Proc. Visualization'98*, Research Triangle Park, NC, pp. 181-188, 1998, IEEE

[30] D.A. Keim, "Designing Pixel-Oriented Visualization Techniques: Theory and Applications", *IEEE Trans. Visualization and Computer Graphics, 6(1)*, pp. 1-20, Mar 2000

[31] D.A. Keim, S C North, C Panse and C P M Sips, "Pixel Based Visual Mining of Geospatial Data", *Computers and Graphics, 28(3)*, pp. 327-344, June 2004

[32] R. Kosara, S. Miksch and H. Hauser, "Focus+Context Taken Literally", *Computer Graphics & Applications, 22(1)*, pp. 22-29, Jan 2002

[33] M. Kreuseler and H. Schumann, "Information visualization using a new Focus+Context Technique in combination with dynamic clustering of information space", *Proc. NPIV'99*, Missouri, pp. 1-5, 1999, ACM Press

[34] J. Lamping and R. Rao, "Visualizing Large Trees Using the Hyperbolic Browser", *Proc. CHI'96*, Vancouver, pp. 388-389, 1996, ACM Press

[35] Y.K. Leung and M.D. Apperley, "A Review and Taxonomy of Distortion-Oriented Presentation Techniques", *ACM Trans. Computer-Human Interaction, 1(2)*, pp. 126-160, June 1994

[36] F. Murtagh, "Clustering in Massive Data Sets", Chemical Data Analysis in the Large, *Proc. Beilstein-Institut Workshop*, May, 2000, Bozen, Italy

[37] M. Novotny and H. Hauser, "Outlier-Preserving Focus+Context Visualization in Parallel Coordinates", *IEEE Trans. Visualization and Computer Graphics, 12(5)*, pp. 893-900, Sept 2006

[38] W. Peng, M.O. Ward and E.A. Rundensteiner, "Clutter Reduction in Multi-Dimensional Data Visualization Using Dimension Reordering", *Proc. Infovis'04*, Austin, Texas, 2004, IEEE

[39] C. Plaisant, B. Milash, A. Rose, S. Widoff and B. Shneiderman, "LifeLines: Visualizing Personal Histories", *Proc. CHI'96*, pp. 221-227, 1996, ACM Press

[40] D. Rafiei and S. Curial, "Effectively Visualizing Large Networks Through Sampling", *Proc. Visualization'05*, pp. 48-55, 2005, IEEE

[41] R. Rao and S. Card, "The Table Lens: Merging graphical and symbolic representations in an interactive focus + context visualization for tabular information", *Proc. CHI'94*, Boston, pp. 111–117, 1994, ACM Press

[42] M. Sarkar, S.S. Snibbe, O.J. Tversky and S.P. Reiss, "Stretching the rubber sheet: a metaphor for viewing large layouts on small screens", *Proc. UIST'93*, Atlanta, Georgia, pp. 81-91, 1993, ACM Press

[43] B. Shneiderman, "The Eyes Have It; A Task by Data Type Taxonomy for Information Visualization", Univ. of Maryland, TR-96-66, 1996

[44] J. Stasko and E. Zhang, "Focus+Context Display and Navigation Techniques for Enhancing Radial, Space-Filling Hierarchy Visusalization", *Proc. InfoVis 2000*, 2000, IEEE

[45] M. Stone, K. Fishkin and E.A. Bier, "The Movable Filter as a User Interface Tool", *Proc. CHI'94*, pp. 306-312, 1994, ACM Press

[46] M. Trutschl, G. Grinstein and U. Cvek, "Intelligently Resolving Point Occlusion", *Proc. InfoVis'03*, pp. 131-136, 2003, IEEE

[47] C. Waldeck and D. Balfanz, "Mobile liquid 2D scatter space (ML2DSS)", *Proc. Information Visualisation 2004*, London, pp. 494-498, 2004, IEEE

[48] M.O. Ward, "A taxonomy of glyph placement strategies for multidimensional data visualization", *Information Visualization, 1*, pp. 194–210, 2002

[49] E.J. Wegman and Q. Luo, "High Dimensional Clustering Using Parallel Coordinates and the Grand Tour", *Computing Science and Statistics, 28*, pp. 352-360, July 1996

[50] N. Wong, S. Carpendale and S. Greenberg, "EdgeLens: An Interactive Method for Managing Edge Congestion in Graphs", *Proc. InfoVis'03*, pp. 51-58, 2003, IEEE

[51] P.C. Wong and R.D. Bergeron, "30 Years of Multidimensional Multivariate Visualization", *Scientific Visualization: Overviews, Methodologies & Techniques*, 1997

[52] A. Woodruff, J. Landay and M. Stonebraker, "Constant Density Visualizations of Non-Uniform Distributions of Data", *Proc. UIST'98*, San Francisco, pp. 19-28, 1998, ACM Press

[53] J. Yang, M.O. Ward and E.A. Rundensteiner, "Visual hierarchical dimension reduction for exploration of high dimensional datasets", *Proc. Sym. Data visualisation '03* , 2003, Eurographics

[54] L. Zhang, C. Tang, Y. Shi, Y. Song, A. Zhang and M. Ramanathan, "VizCluster and Its Application on Clustering Gene Expression Data", *Distributed and Parallel Databases, 13(1)*, pp. 73 - 97, 2003
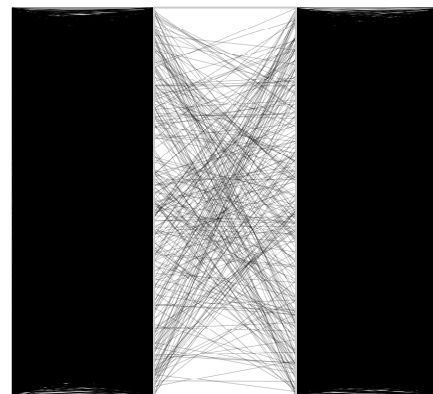
Fig. 1. Using sampling to reduce the number of overlapping lines. Example shows inter-axis sampling lens [19] on a parallel coordinate plot with 30,000 records. Sampling rate is 1%.
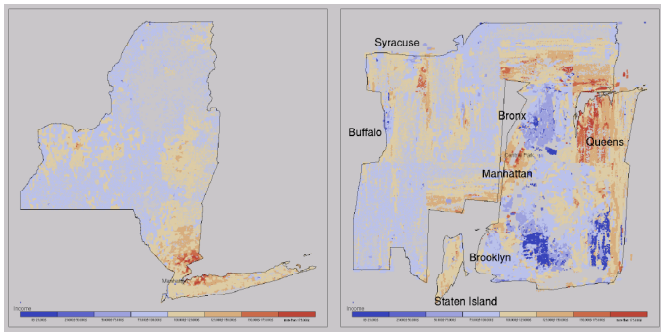
Fig. 2. PixelMap avoids overlap altogether by distorting the underlying map [31]. (Map on the left is undistorted)
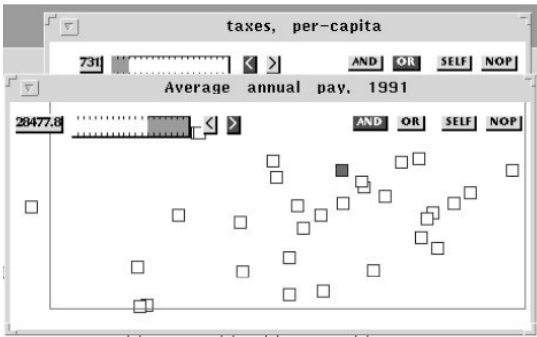

Fig. 3. An example of filtering with a Magic Lens [45]. The spatial information is retained, although one could argue that the points which are no longer visible have been displaced significantly!
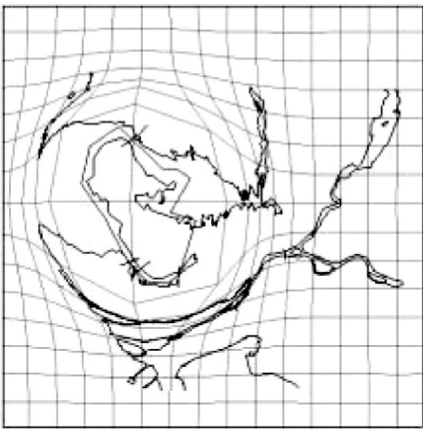

Fig. 4. The plot has undergone a topological distortion, however the overlaid grid squares do provide a reference for the user and helps to keep spatial information [11].
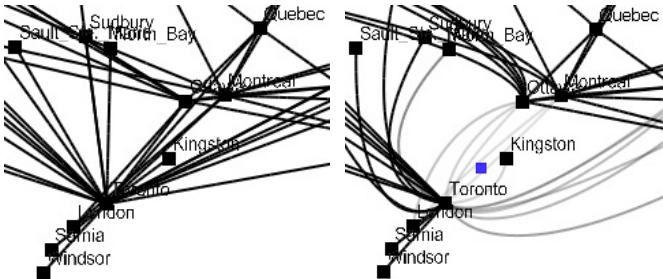

Fig. 5. EdgeLens [50] displaces lines to both reveal the data underneath and helps to disambiguate the edges and nodes.
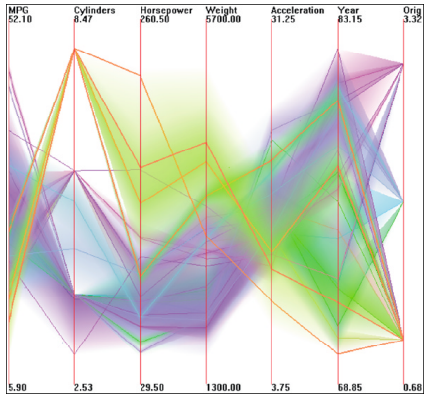

Fig. 6. The clustering used in Hierarchical parallel coordinates [53] is scaleable to very large datasets, only limited by computational resources.

Fig. 7. In this example of an RSVP carousel, the 'points' obviously retain their image attribute [14].
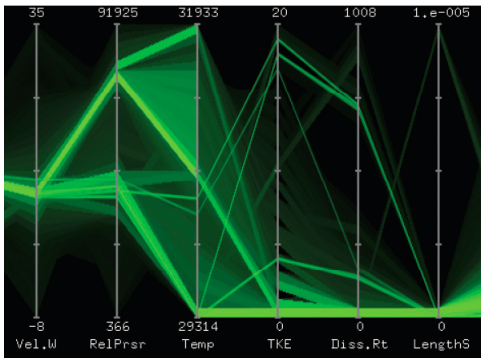

Fig. 8. Discrimination of lines in a parallel coordinate plot utilising the outlier-preserving technique of Novotny et al and opacity [37].
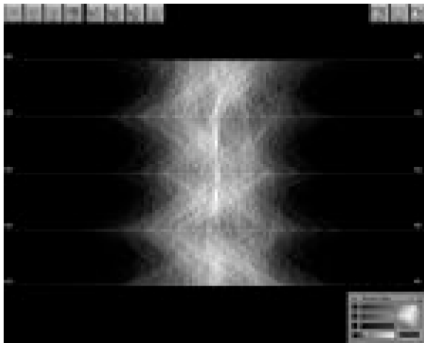

Fig. 9. Reducing the opacity of lines can indicate the density of the overlapping lines [49]