Analysis Case Studies: Notes

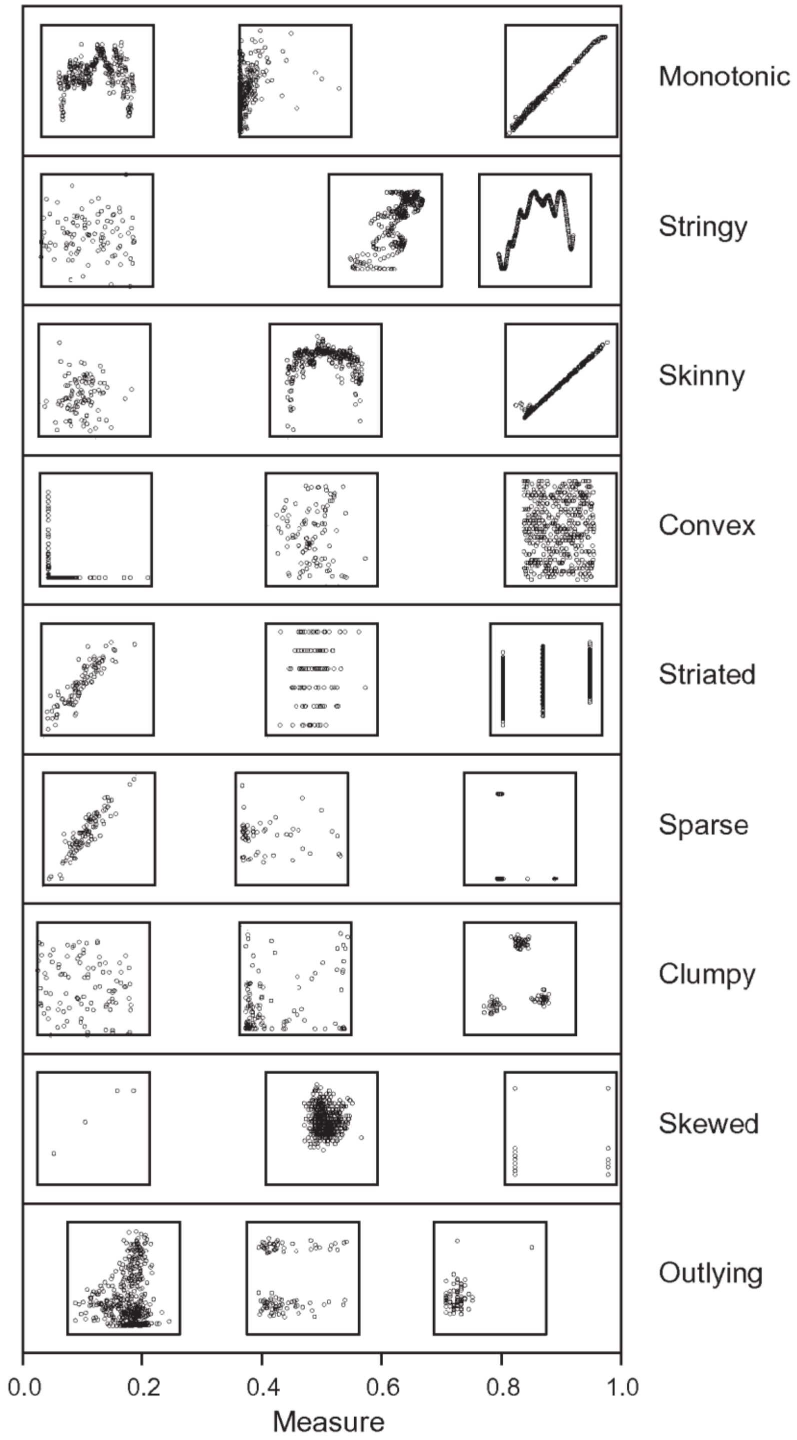
Advanced Data Analytics Week 5

Slide 5

The following analyses are descriptive examinations of the final design of each system, not prescriptive statements that the particular choices made by these designers are the only good solution that fits the requirements. As you read through this chapter, it is a useful exercise to generate a set of alternatives for each choice made by these designers and to consider the pros and cons of each

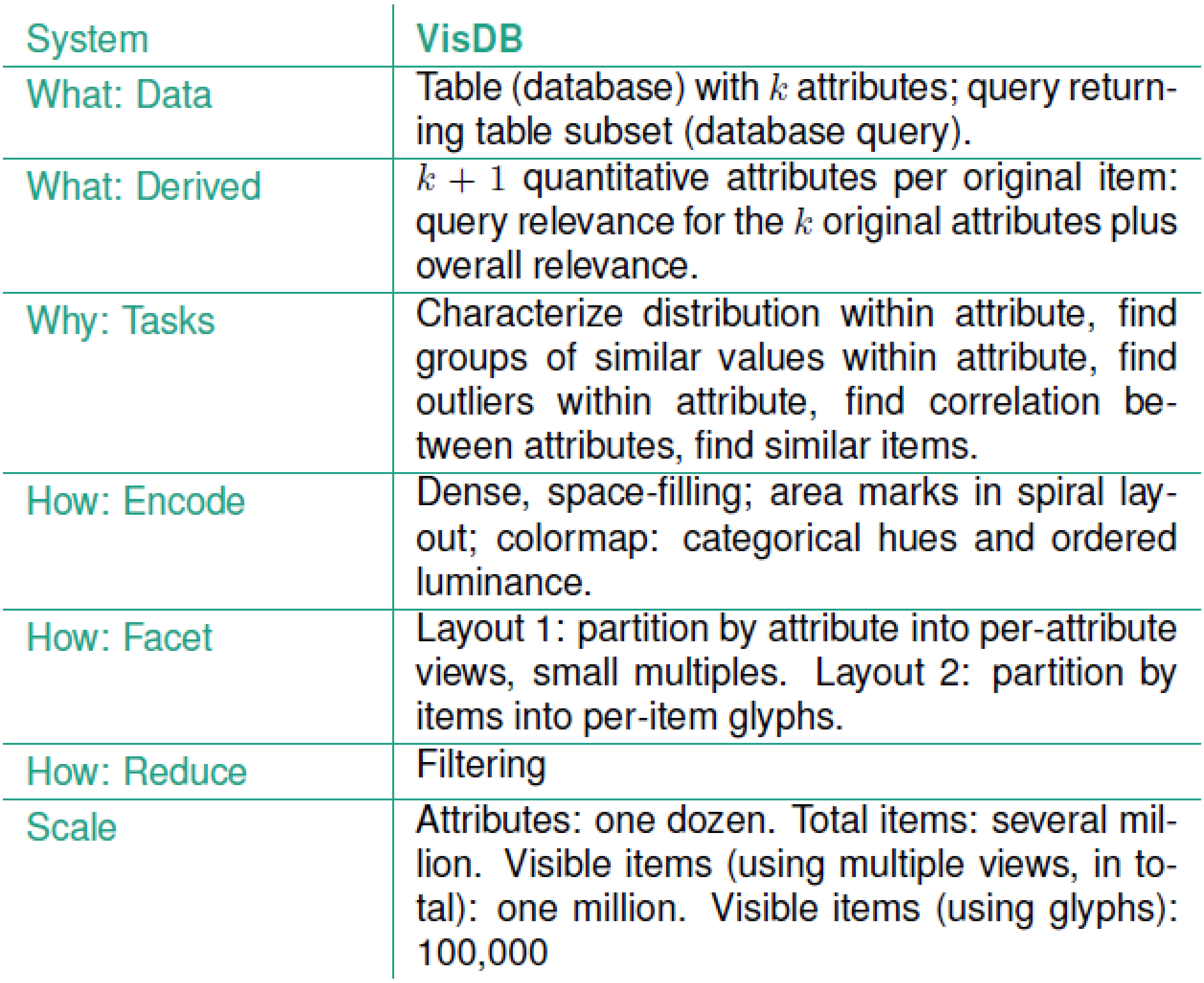
Slide 8

The nine scagnostics measures that describe scatterplot shape, with examples of real-world datasets rated low, medium, and high for each of the nine measures.



Slide 14

Summary of VisDB analysis



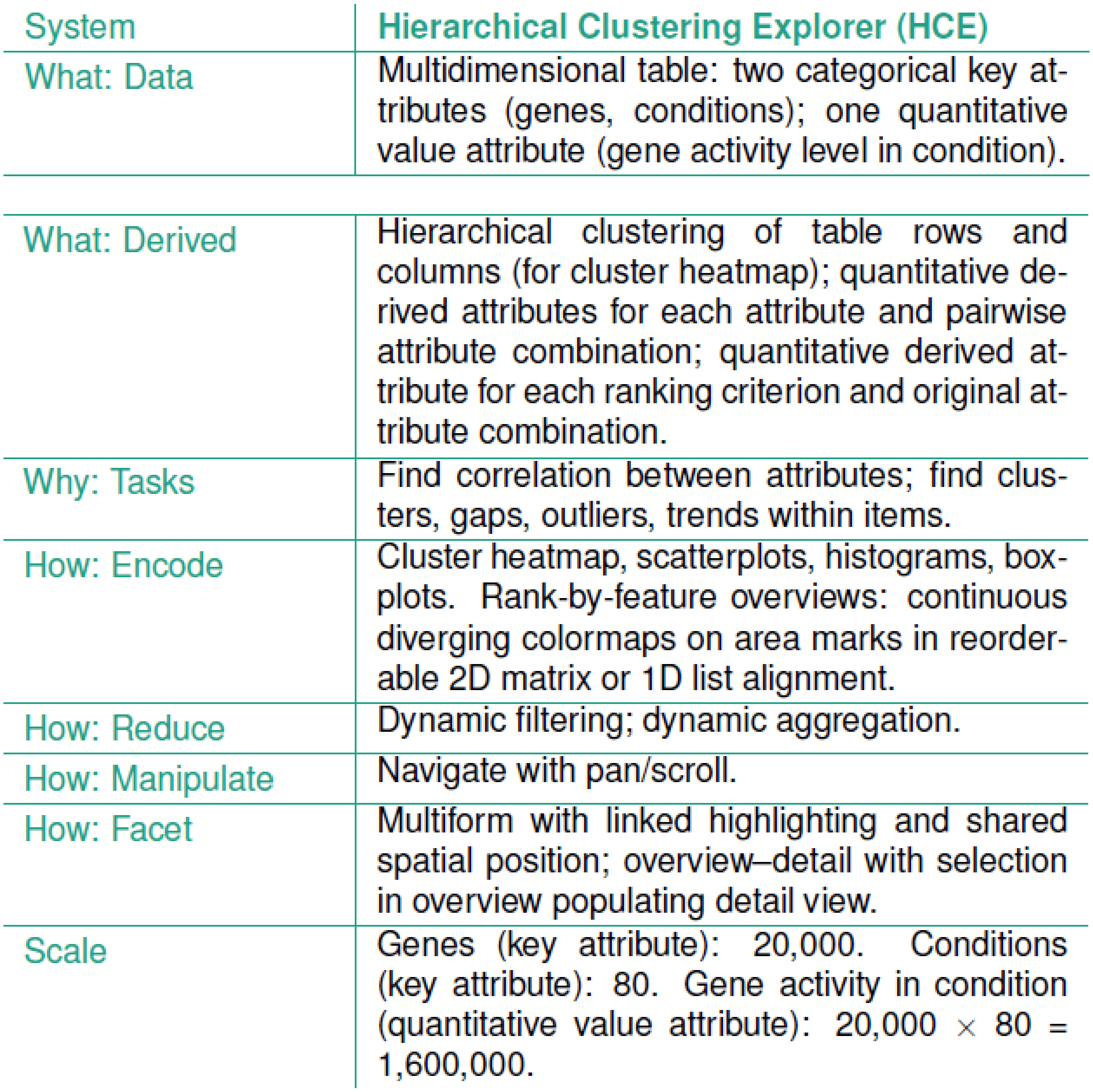
System	VisDB
What: Data	Table (database) with k attributes; query returning table subset (database query).
What: Derived	$k + 1$ quantitative attributes per original item: query relevance for the k original attributes plus overall relevance.
Why: Tasks	Characterize distribution within attribute, find groups of similar values within attribute, find outliers within attribute, find correlation between attributes, find similar items.
How: Encode	Dense, space-filling; area marks in spiral layout; colormap: categorical hues and ordered luminance.
How: Facet	Layout 1: partition by attribute into per-attribute views, small multiples. Layout 2: partition by items into per-item glyphs.
How: Reduce	Filtering
Scale	Attributes: one dozen. Total items: several million. Visible items (using multiple views, in total): one million. Visible items (using glyphs): 100,000

Slide 16

The overview cluster heatmap at the top uses an aggregated representation where an entire dataset of 3614 genes is shown with fewer than 1500 pixels by replacing individual leaves with the average values of adjacent leaves. The density level of the overview can be interactively changed by the user, for a tradeoff between an aggregate view where some detail is lost but the entire display is visible at once, and a more zoomed-in view where only some of the columns are visible simultaneously and navigation by horizontal panning is required to see the rest. The horizontal line through the dendrogram labelled Minimum Similarity is an interactive filtering control. Dragging it down vertically dynamically filters out the columns in the heatmap that correspond to the parts of the dendrogram above the bar and partitions the heatmap into pieces that correspond to the number of clusters just below the bar.

The detail view at the bottom shows a heatmap of the cluster selected in the top overview. It also shows the second dendrogram for hierarchical clustering of the rows on the side; this dendrogram is not shown above in order to maximize the number of columns that can fit within the overview. The background of the selected cluster is highlighted in yellow in the overview, and the correspondence between the views is emphasized by colouring the column labels along the top of the detail view yellow as well, for linked highlighting.

Slide 17



System	Hierarchical Clustering Explorer (HCE)
What: Data	Multidimensional table: two categorical key attributes (genes, conditions); one quantitative value attribute (gene activity level in condition).
What: Derived	Hierarchical clustering of table rows and columns (for cluster heatmap); quantitative derived attributes for each attribute and pairwise attribute combination; quantitative derived attribute for each ranking criterion and original attribute combination.
Why: Tasks	Find correlation between attributes; find clusters, gaps, outliers, trends within items.
How: Encode	Cluster heatmap, scatterplots, histograms, boxplots. Rank-by-feature overviews: continuous diverging colormaps on area marks in reorderable 2D matrix or 1D list alignment.
How: Reduce	Dynamic filtering; dynamic aggregation.
How: Manipulate	Navigate with pan/scroll.
How: Facet	Multiform with linked highlighting and shared spatial position; overview–detail with selection in overview populating detail view.
Scale	Genes (key attribute): 20,000. Conditions (key attribute): 80. Gene activity in condition (quantitative value attribute): 20,000 × 80 = 1,600,000.

References

Case Study 1: Scagnostics

L. Wilkinson, A. Anand, and R. Grossman, “Graph-Theoretic Scagnostics”, In *Proceedings of the IEEE Symposium on Information Visualization (InfoVis)*, (2005), 157-164.

Case Study 2: VisDB

D. A. Keim and H-P. Kriegel, VisDB: Database Exploration Using Multidimensional Visualization.” *IEEE Computer Graphics and Applications*, **14** (1994), 40-49.

Case Study 3: HCE

J. Seo and B. Shneiderman, “Interactively Exploring Hierarchical Clustering Results”, *IEEE Computer* **35**, (2002), 80-86/