Techniques for visualization of high dimensional data

From data to visual representations

M. Cristina

SCC5836/0252 Visualização Computacional

High-dimensional data

- Now, what if you want to visualize more than 4, 5 or 6 attributes simultaneously?

E.g., a data table with several numerical and/or categorical atributes

$$\mathbf{x_i} = (x_{i1}, x_{i2}, ..., x_{im}),\ m >> 4\ (?)$$

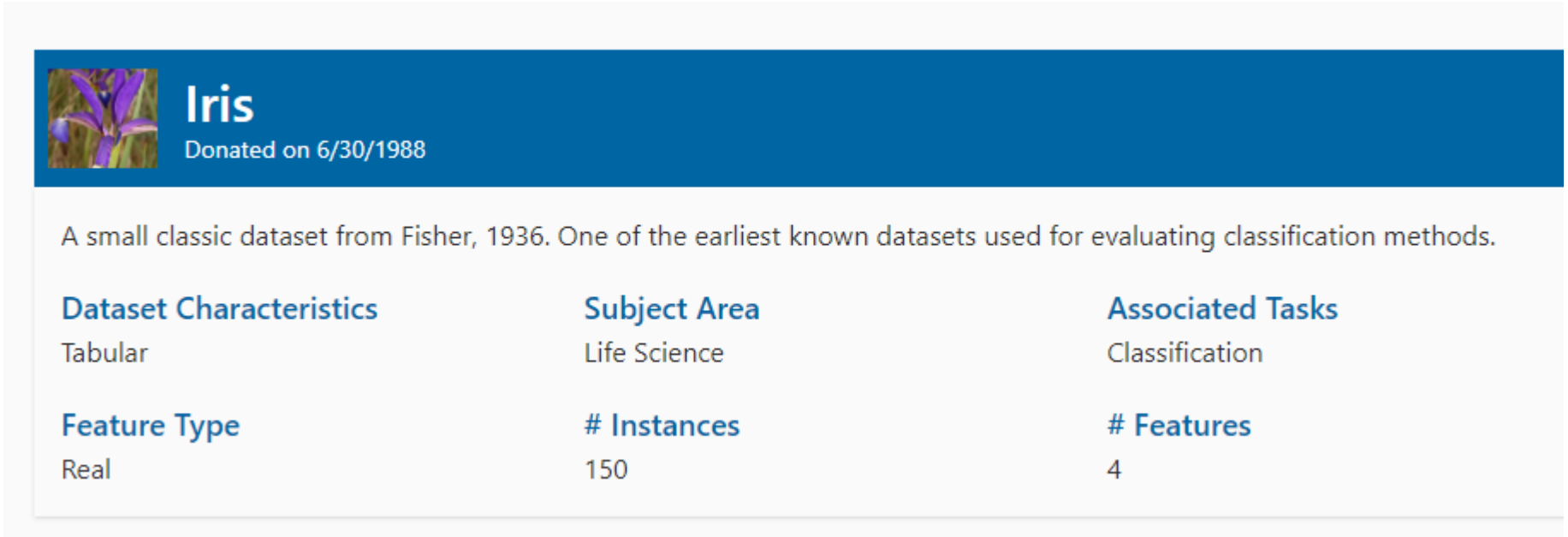
Summary

- SPLOM
- Parallel coordinates
- Radar plots
- Heatmaps
- RadViz

High-dimensional data

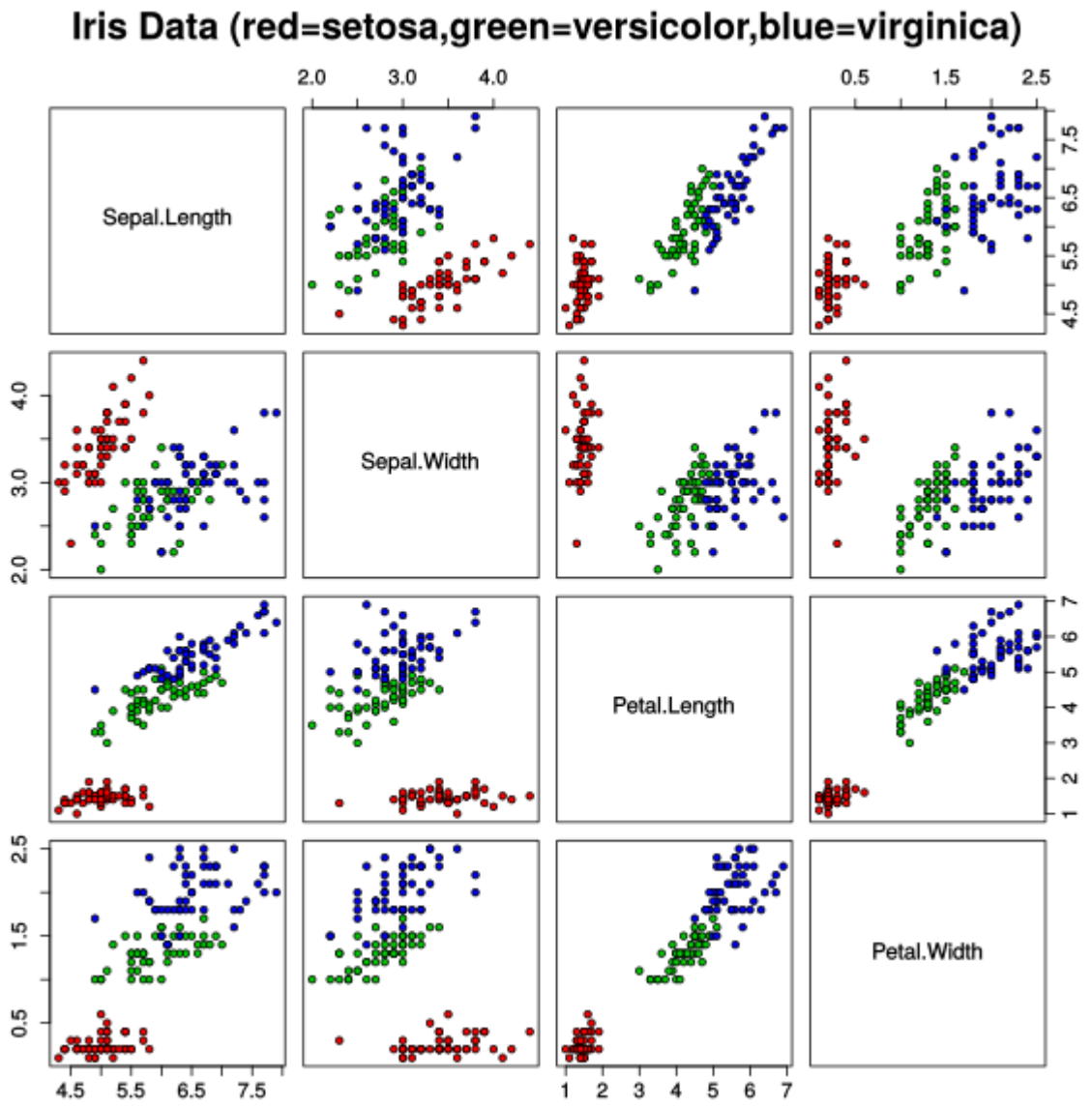
Ex. Iris data set, 4 numerical atributes + class

DY
150
4
Sep_lenght;Sep_width;Pet_lenght;Pet_width
0;5.1;3.5;1.4;0.2;1
1;4.4;3;1.3;0.2;1
2;6.5;2.8;4.6;1.5;2
3;6.4;2.9;4.3;1.3;2
4;6.8;2.8;4.8;1.4;2
5;5.5;2.4;3.8;1.1;2
6;6.4;3.2;5.3;2.3;3
7;6.3;2.7;4.9;1.8;3
8;7.7;3;6.1;2.3;3
9;4.6;3.1;1.5;0.2;1
10;5.4;3.9;1.3;0.4;1
…



SPLOM Scatterplot matrix

https://en.m.wikipedia.org/wiki/File:Iris_dataset_scatterplot.svg



SPLOM Scatterplot matrix

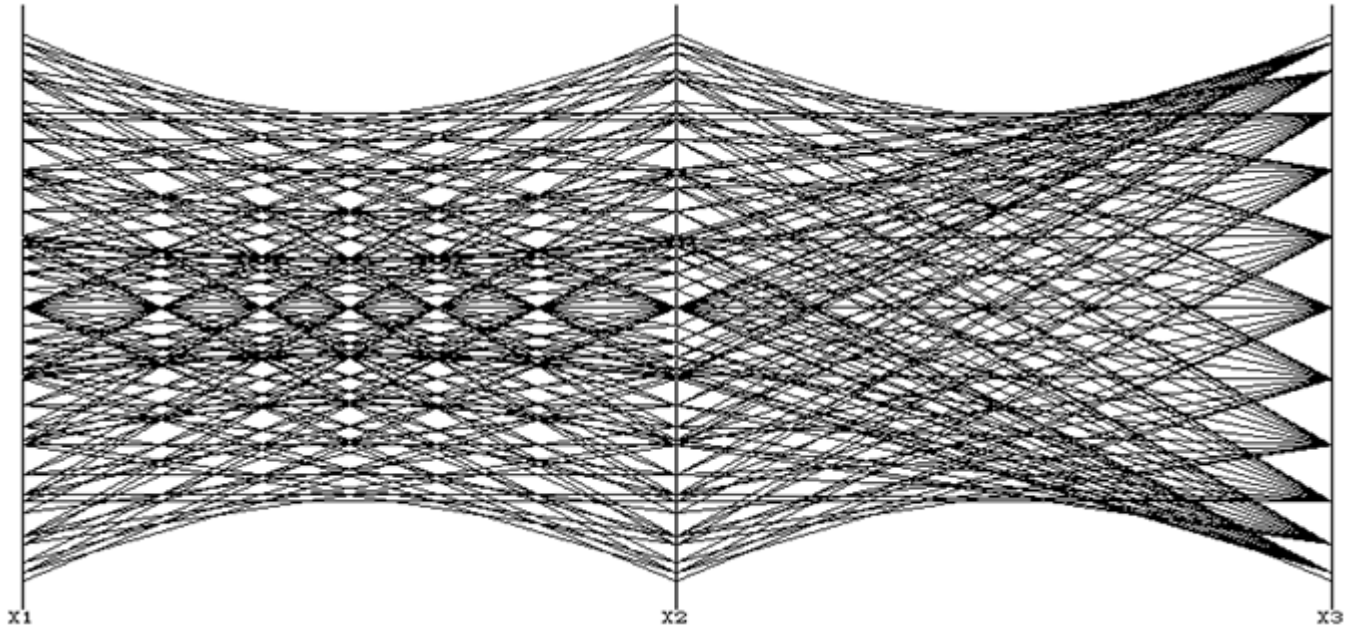
- https://plotly.com/python/line-and-scatter/
- https://plotly.com/python/splom/ (see https://www.kaggle.com/datasets/mathchi/diabetes-data-set)
- Also known as correlogram, or correlation matrix: https://www.data-to-viz.com/graph/correlogram.html

Parallel Coordinates

- Displays high-dimensional data with information about atribute behavior
  - Each atribute is mapped to one of multiple equally spaced parallel axes
  - Each data item is represented as a polyline
    - A multidimensional `data point´ is mapped as a series of line segments
  - First introduced as a solution for visualizing high-dimensional geometrical entities: Inselberg, A. *The Plane with Parallel Coordinates, Visual Computer*. 1985.

Parallel coordinates

Points on the surface of a 3D sphere



Inselberg, A. *Parallel Coordinates, Visual Multidimensional Geometry and its Applications*. Springer 2009.

Wegman, E.J. Hyperdimensional data analysis using Parallel Coordinates. *J. American Statistical Association* 1990 85(411).

A. Telea, Data visualization principles and practice. CRC Press 2015.

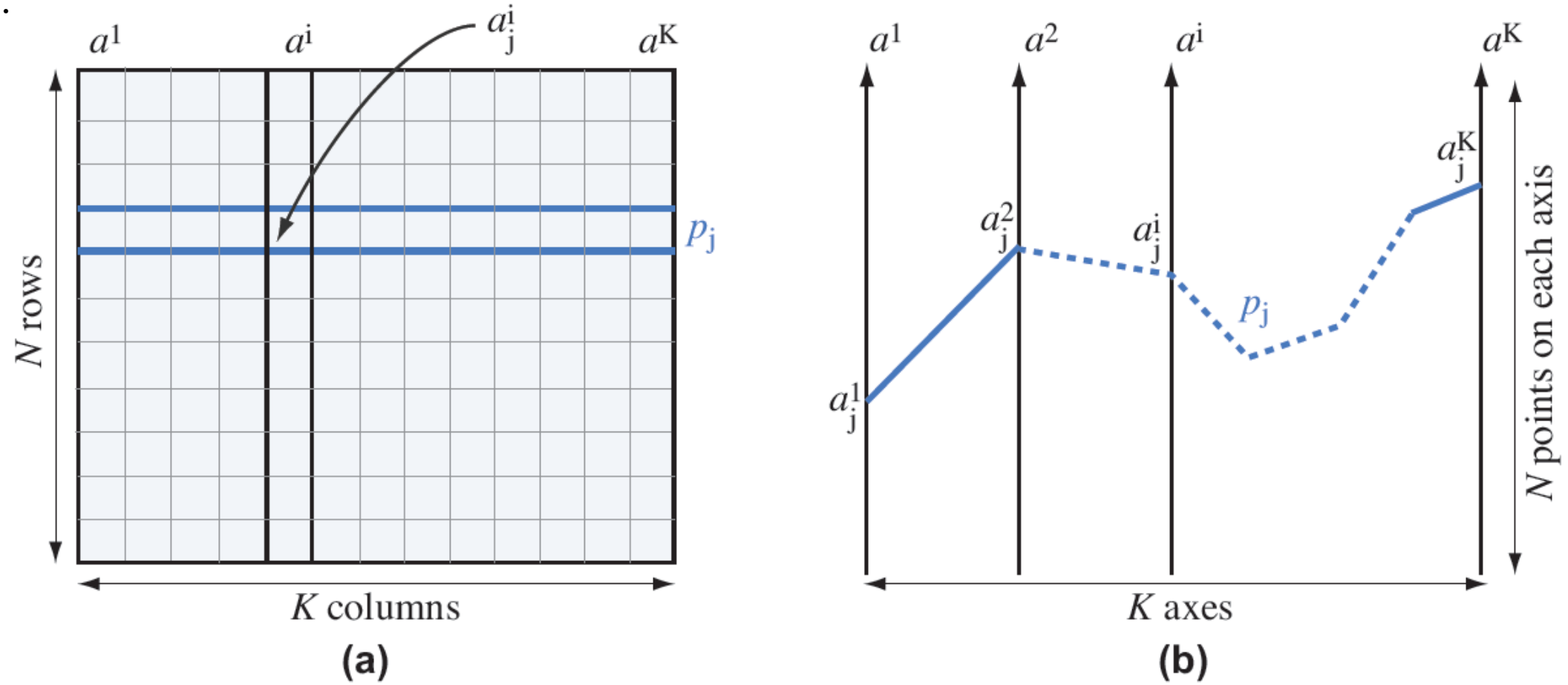
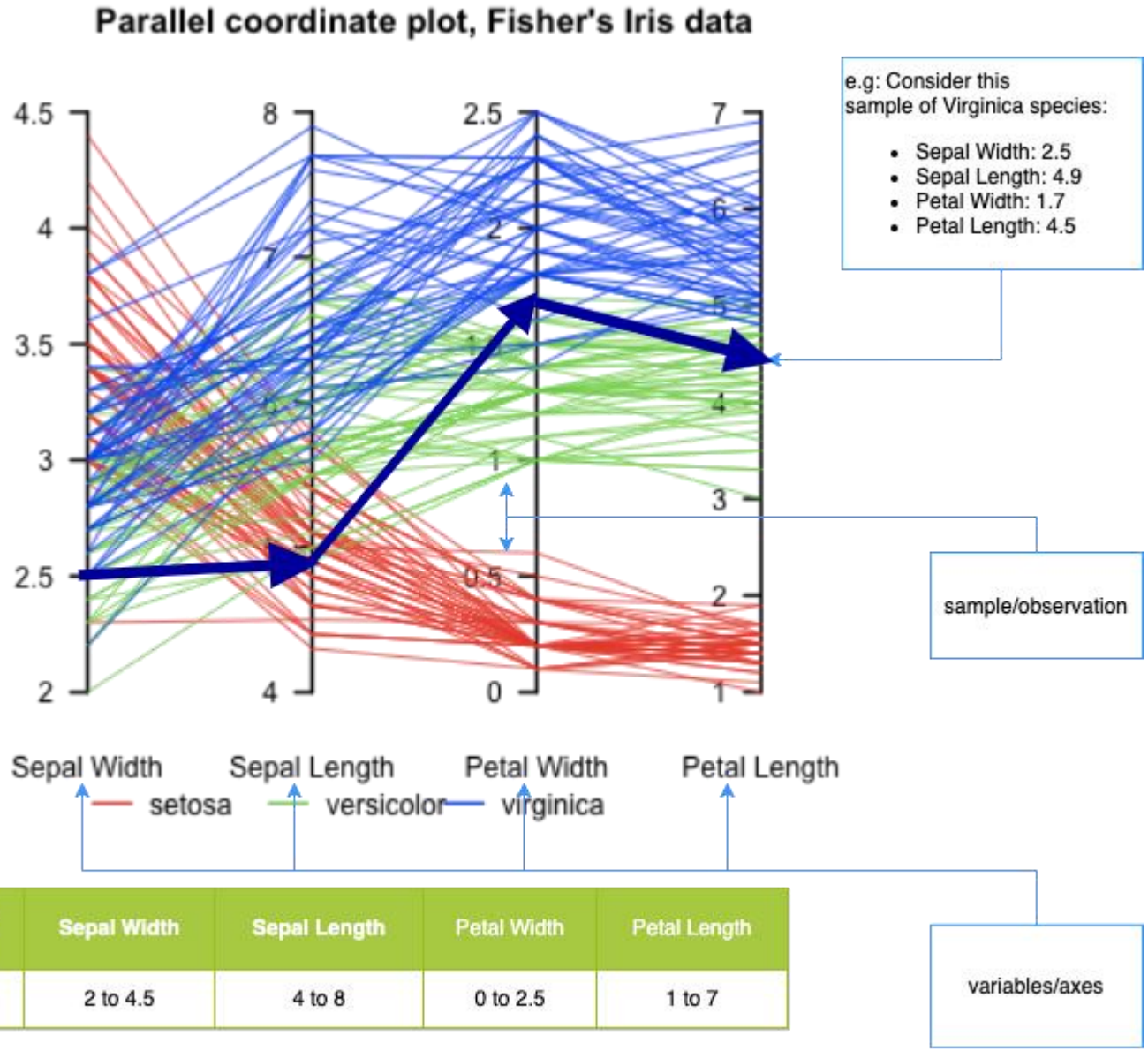


Figure 11.27. Schematic description of (a) table visualization vs. (b) parallel coordinate plots. A K-dimensional point p_j is shown in blue in both plots.

Parallel Coordinates

Notice patterns!
What can we tell about this data?



Variable Axis	Sepal Width	Sepal Length	Petal Width	Petal Length
Range	2 to 4.5	4 to 8	0 to 2.5	1 to 7

Source: https://www.analyticsvidhya.com/blog/2021/11/visualize-data-using-parallel-coordinates-plot/

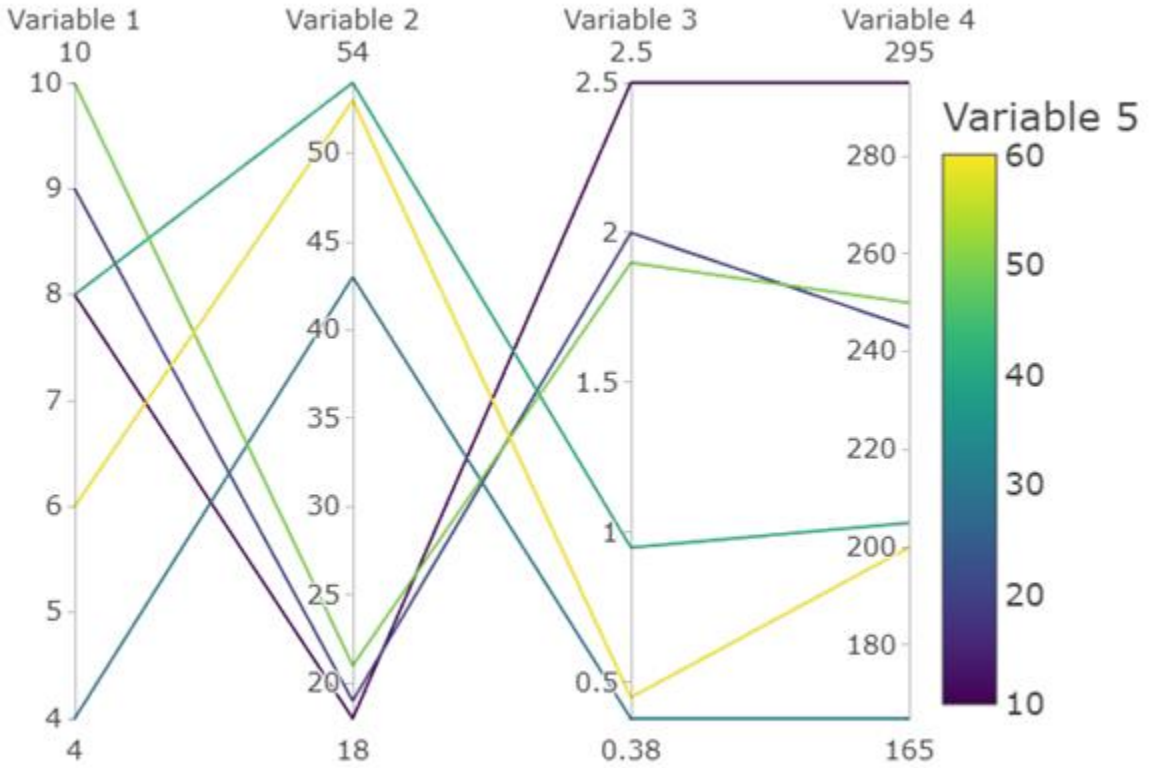
Parallel Coordinates

- http://mbostock.github.io/d3/talk/20111116/iris-parallel.html
- https://plotly.com/python/parallel-coordinates-plot/
- https://www.data-to-viz.com/graph/parallel.html

Parallel Coordinates

- Issues
  - Scale/Normalization
  - Axes reordering. Why?
  - Interaction: brushing
  - Overplotting
  - Many many enhancements...

Parallel Coordinates



Source: https://towardsdatascience.com/parallel-coordinates-plots-with-plotly-dffe3f526c6b

Example: axes reordering

Data from a Kaggle Competition related to evaluating how some chemical properties affect the quality of red variants of the Portuguese “Vinho Verde” wine [Cortez et al., 2009]. Quality scores between 0 (Bad) and 10 (Excellent) provided for each wine in the dataset

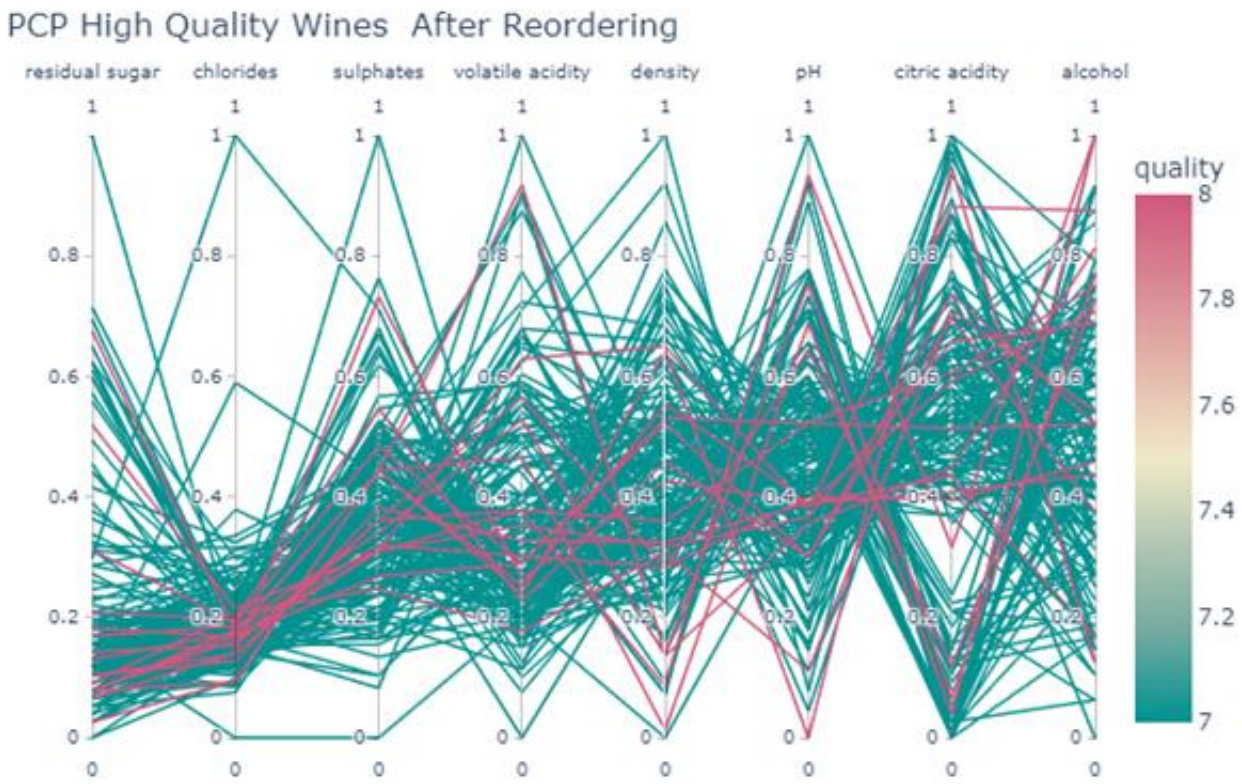
Indice	fixed acidity	volatile acidity	citric acidity	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality
0	11.2	0.28	0.56	1.9	0.075	17	60	0.998	3.16	0.58	9.8	6
1	7.8	0.88	0	2.6	0.098	25	67	0.9968	3.2	0.68	9.8	5
2	7.8	0.76	0.04	2.3	0.092	15	54	0.997	3.26	0.65	9.8	5
3	7.9	0.6	0.06	1.6	0.069	15	59	0.9964	3.3	0.46	9.4	5
4	7.5	0.5	0.36	6.1	0.071	17	102	0.9978	3.35	0.8	10.5	5
5	7.8	0.58	0.02	2	0.073	9	18	0.9968	3.36	0.57	9.5	7
6	7.3	0.65	0	1.2	0.065	15	21	0.9946	3.39	0.47	10	7
7	7.4	0.7	0	1.9	0.076	11	34	0.9978	3.51	0.56	9.4	5
8	7.4	0.7	0	1.9	0.076	11	34	0.9978	3.51	0.56	9.4	5
9	7.4	0.66	0	1.8	0.075	13	40	0.9978	3.51	0.56	9.4	5

Source: https://towardsdatascience.com/parallel-coordinates-plots-with-plotly-dffe3f526c6b

Example: axes reordering

Source: https://towardsdatascience.com/parallel-coordinates-plots-with-plotly-dffe3f526c6b

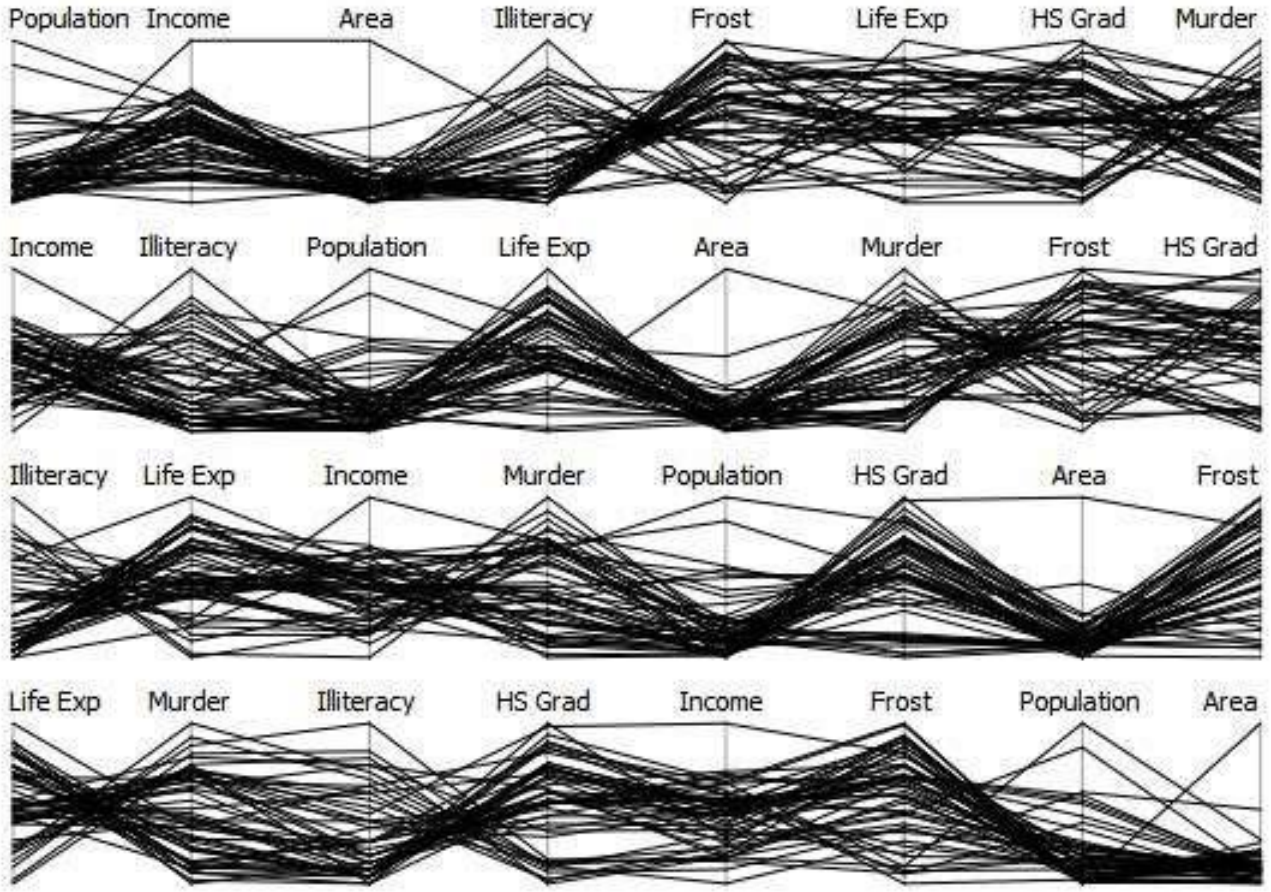
Example: axes reordering



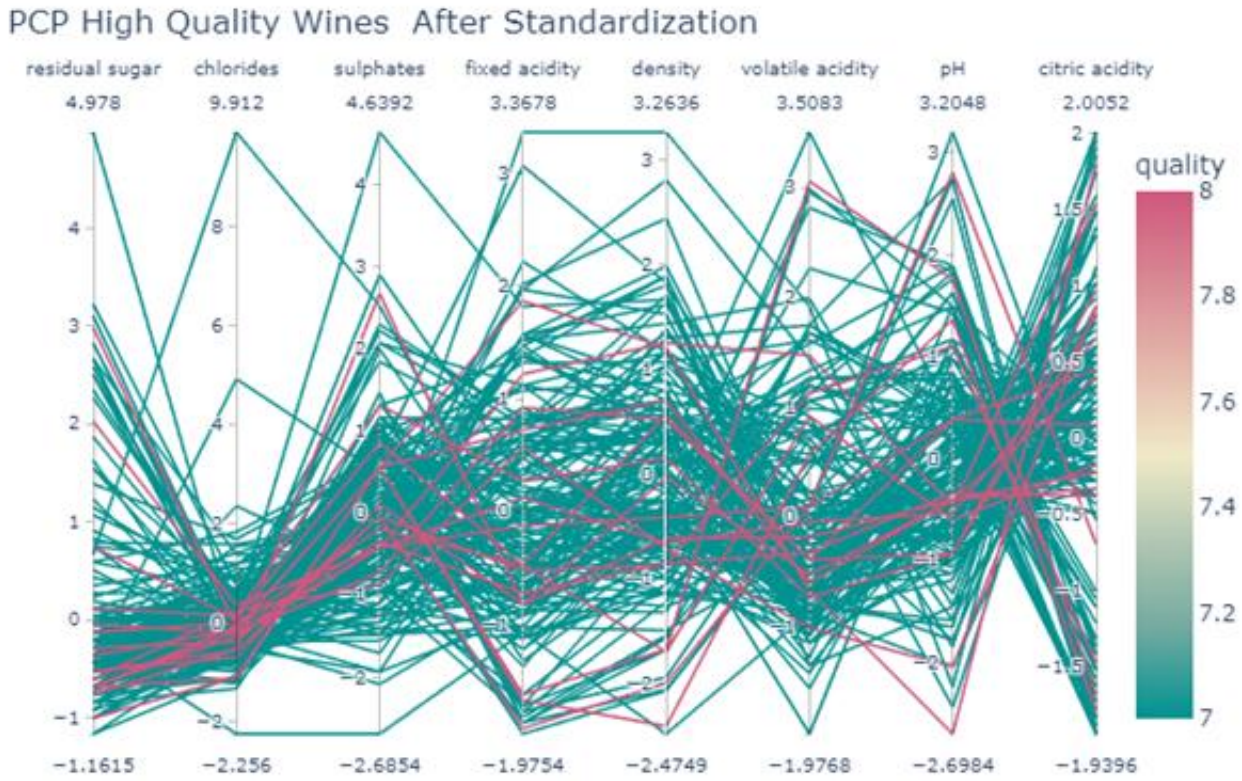
Source: https://towardsdatascience.com/parallel-coordinates-plots-with-plotly-dffe3f526c6b

Example: axes order matters

Source:
https://jeheonpark93.medium.com/vc-parallel-coordinates-1ae8c119e062



Example: standardization (not a good choice)



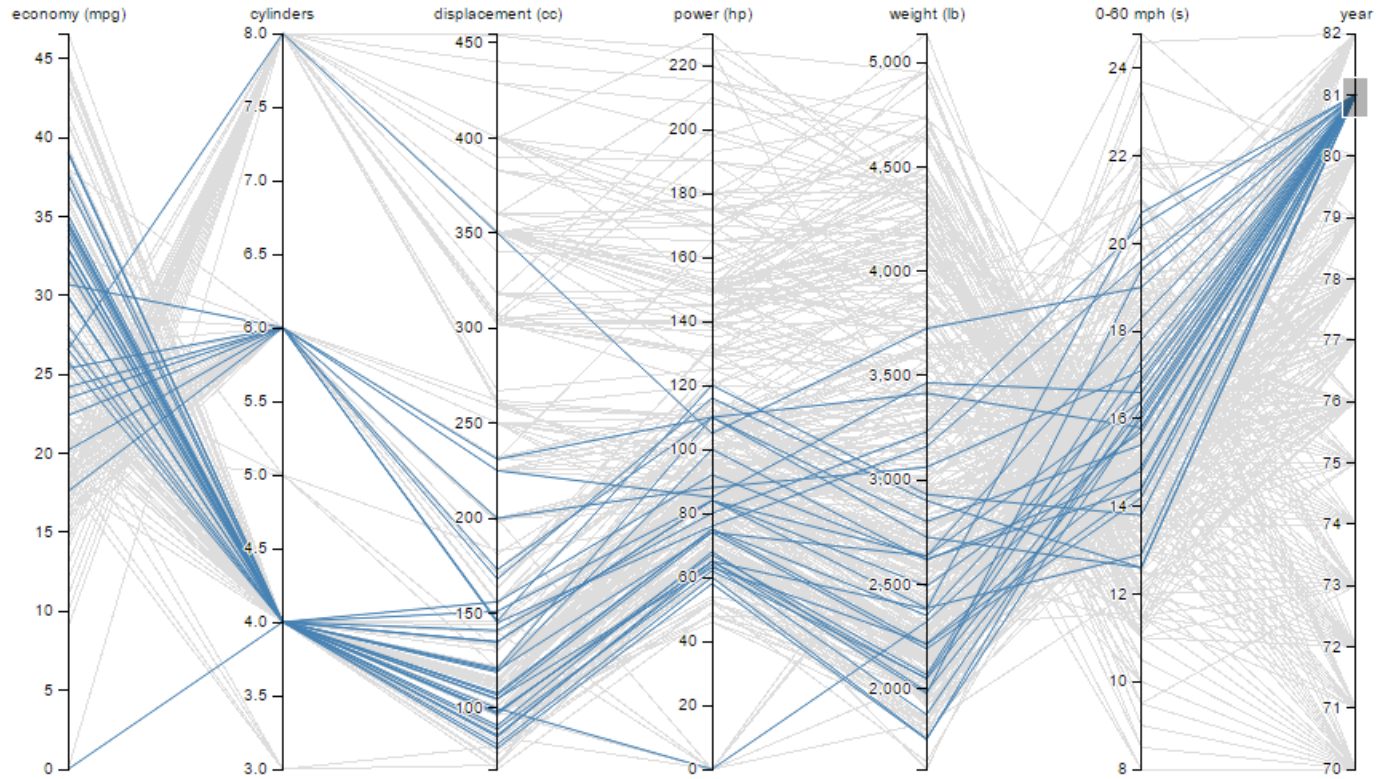
Source: https://towardsdatascience.com/parallel-coordinates-plots-with-plotly-dffe3f526c6b

Example: brushing

- highlighting part of the data by delimiting a region over an axis with the mouse

Example: brushing

https://bl.ocks.org/jasondavies/1341281



highlighting part of the data by delimiting a region over an axis with the mouse (filtering)

Examples: PCP & brushing

- https://medium.com/@ilievski.vladimir/neural-networks-hyperparameter-search-the-visualized-way-9c46781bea28
- https://towardsdatascience.com/visualizing-backpropagation-in-neural-network-training-2647f5977fdb
- http://www.cs.nott.ac.uk/blaramee/research/callCenter/brushing/roberts18smart.pdf

Parallel Coordinates: discussion

- Appropriate for comparing many numerical variables simultaneously
  - Multivariate numerical data, particularly when data variables have different scales and/or different units of measurement

- A `complex´ visual mapping: must be done carefully (e.g., normalization, outlier removal, …
  - Difficult to understand by non-technical audiences
- Issues with too many data items or too many data attributes

Parallel Coordinates: discussion

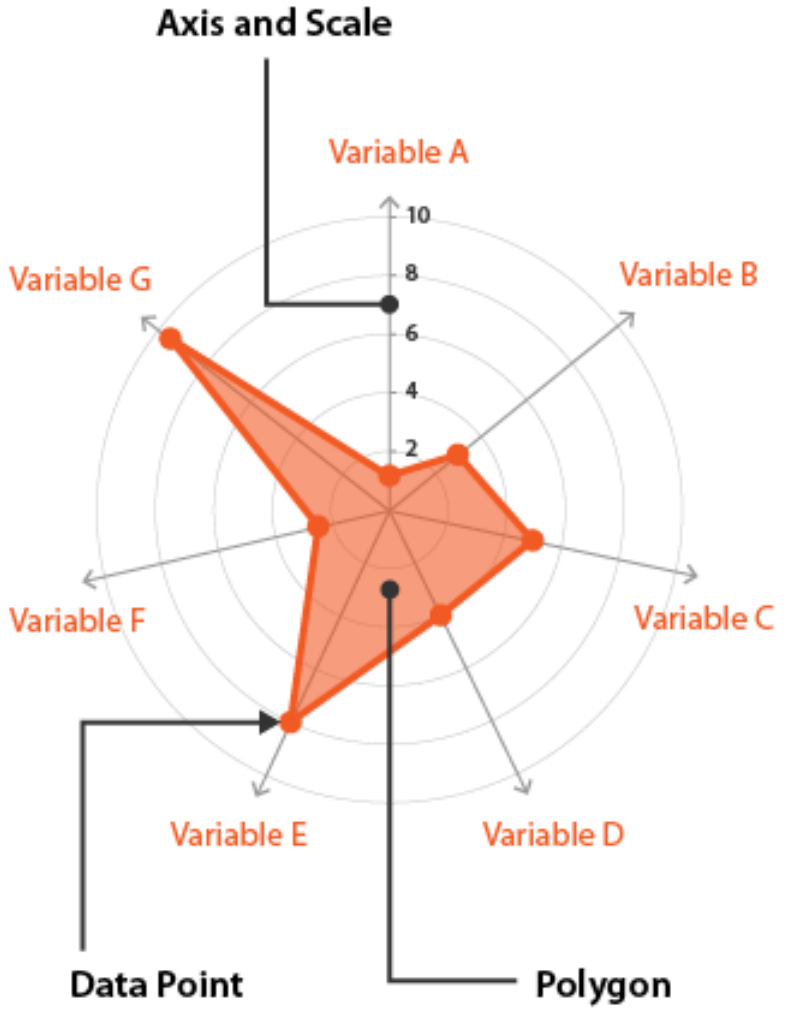
“The first time that I saw a parallel coordinates visualization, I almost laughed out loud. My initial impression was "How absurd!" I couldn’t imagine how anyone could make sense of the dense clutter caused by hundreds of overlapping lines. This certainly isn’t a chart that you would present to the board of directors or place on your Web site for the general public. In fact, the strength of parallel coordinates isn’t in their ability to communicate some truth in the data to others, but rather in their ability to bring meaningful multivariate patterns and comparisons to light when used interactively for analysis.”

Stephen Few

Source: http://www.perceptualedge.com/articles/b-eye/parallel_coordinates.pdf

Radar chart (spider chart)

Also useful for comparing
multiple quantitative variables



Radar chart (spider chart)

https://www.data-to-viz.com/caveat/spider.html

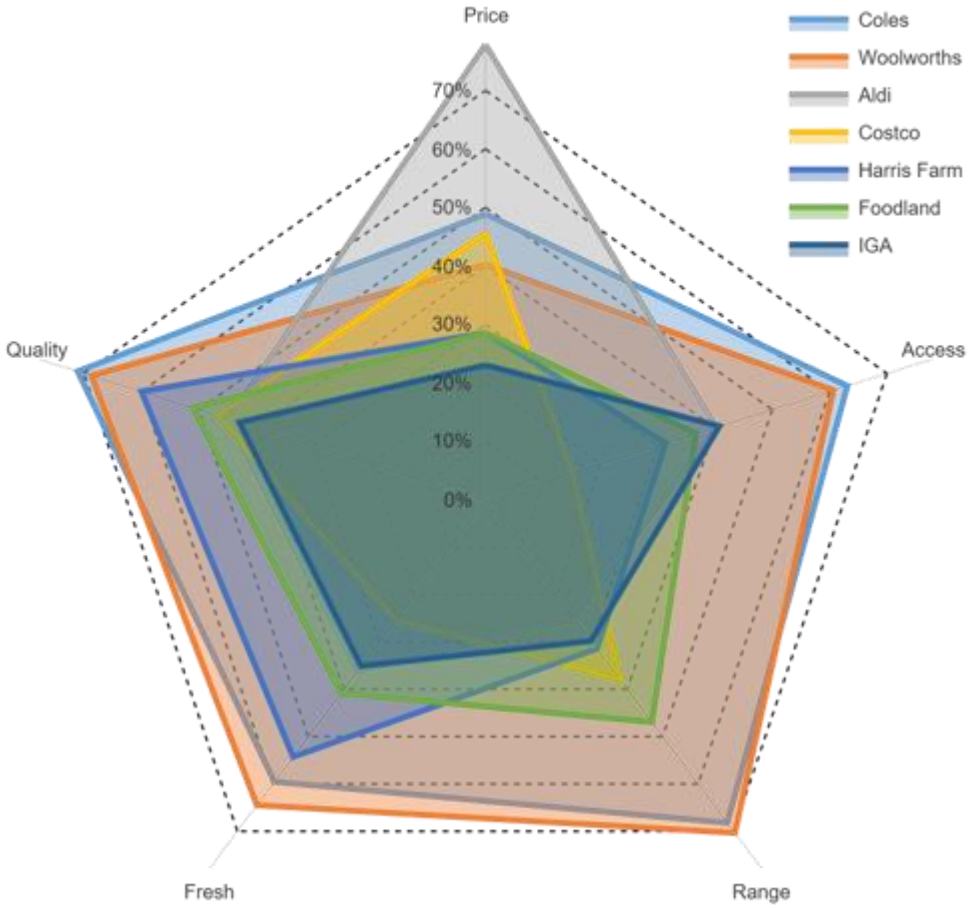
https://www.fusioncharts.com/resources/chart-primers/radar-chart

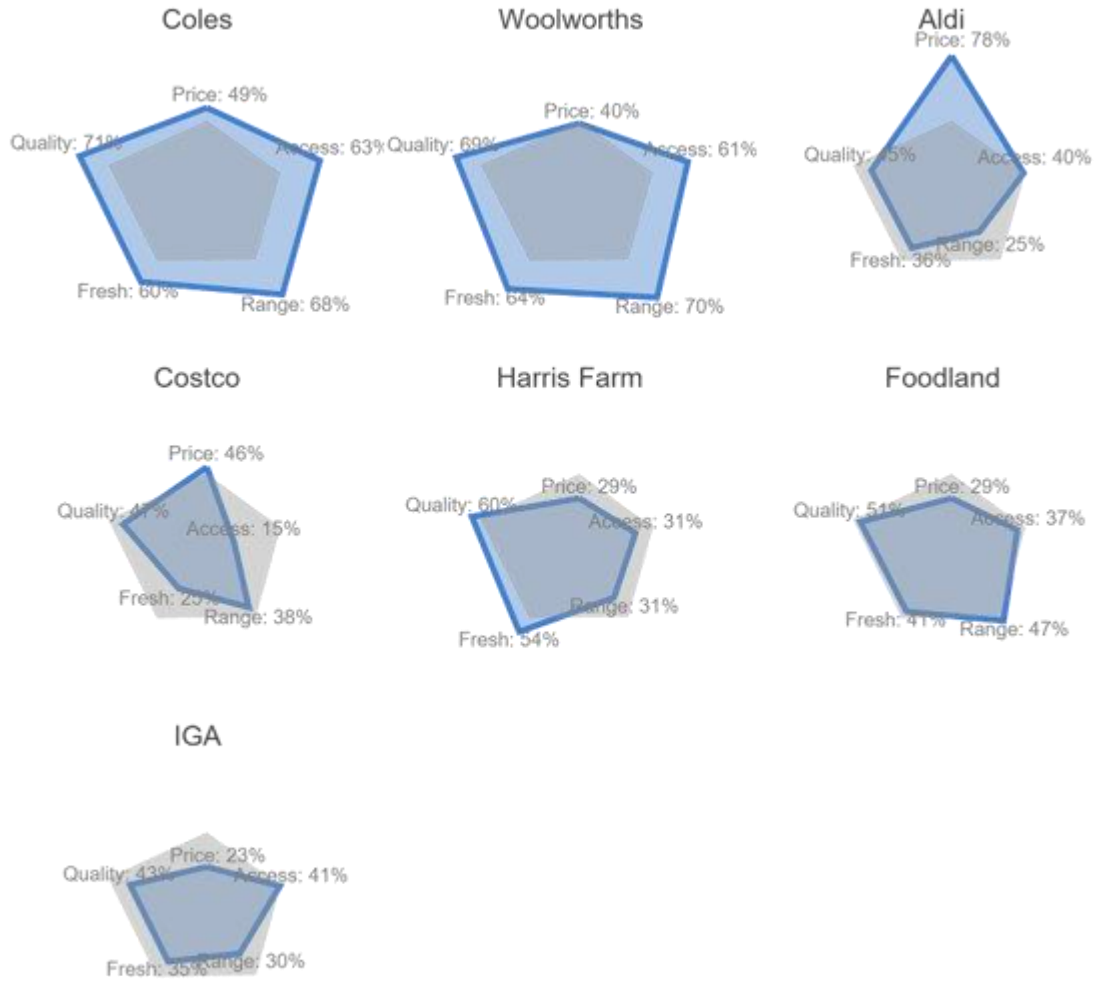
Radar chart (spider chart)

- Best for comparing the behavior (on multiple atributes) of a small numbers of items

- Often appear in small multiplex charts

Radar chart (spider chart)



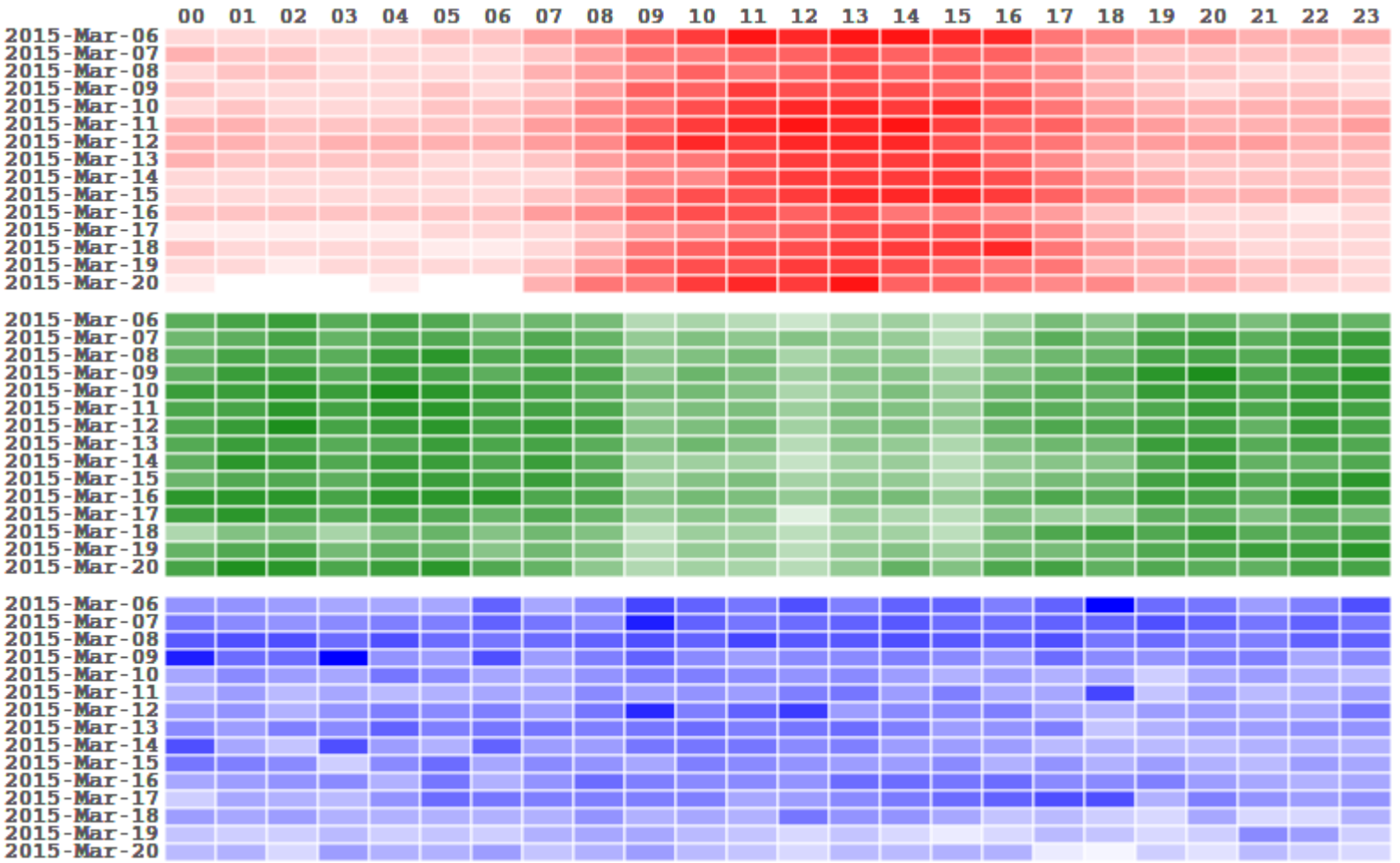


Heatmaps

https://datavizcatalogue.com/methods/heatmap.html

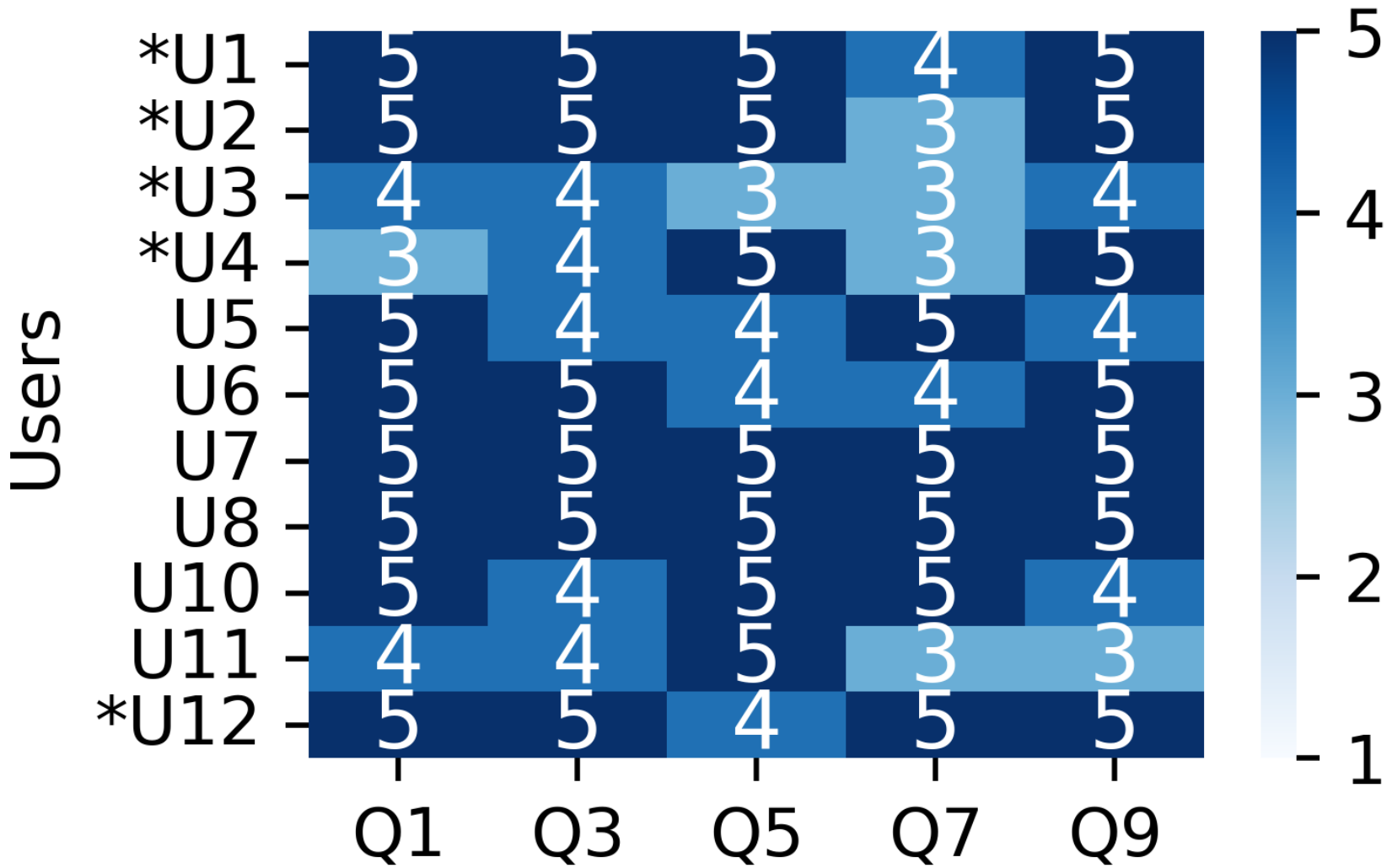
https://www.data-to-viz.com/graph/heatmap.html

Heatmaps



Can also be used for displaying time series where there is a regular pattern in time

Heatmaps



Source: Silva et al. Interacting with Computers, 2023. Evaluating visual analytics for relevant information retrieval in document collections.

Heatmaps: example

- https://academic.oup.com/bioinformatics/article/32/18/2847/1743594

Heatmaps

- Flexibility: numerical or categorical variables, time series...

- Because of their reliance on color to communicate values, heatmap charts are better suited to displaying an overview of numerical data
  - it’s hard to accurately tell the differences between color shades
  - If informing precise values is important one may display the data values inside the cells

Effect of data range and outliers in color mapping

- Typically, a continuous colorscale uses a linear mapping to associate a range of values with the available color shades
  - e.g., if we have N = 10 shades and scalar values in a range [min, max]

- (max – min) values mapped to N shades: each shade will correspond to *k* distinct values, $k = \frac{(\max - \min)}{N}$

Effect of data range and outliers in color mapping

- $C = \{c_i\}_{i = 1, \ldots N}$ shades, given a scalar value *v*, its color is given by c_i
- Highest value mapped to c_{10}, lowest in c_1, others following a linear correspondence
- The presence of outliers compresses the lower values to a few color shades, and renders the linear mapping ineffective

Effect of data range and outliers in color mapping

- $C = \{c_i\}_{i = 1, \ldots N}$ shades, given a scalar value *v*, its color is given by c_i
- Highest value mapped to c_{10}, lowest in c_1, others following a linear correspondence
- A similar scenario occurs if a single colormap is employed to visualize variables/attributes given in different ranges of magnitude: that is why the choice of scaling is so important

Heatmaps

- See this (very informative) StatQuest video:

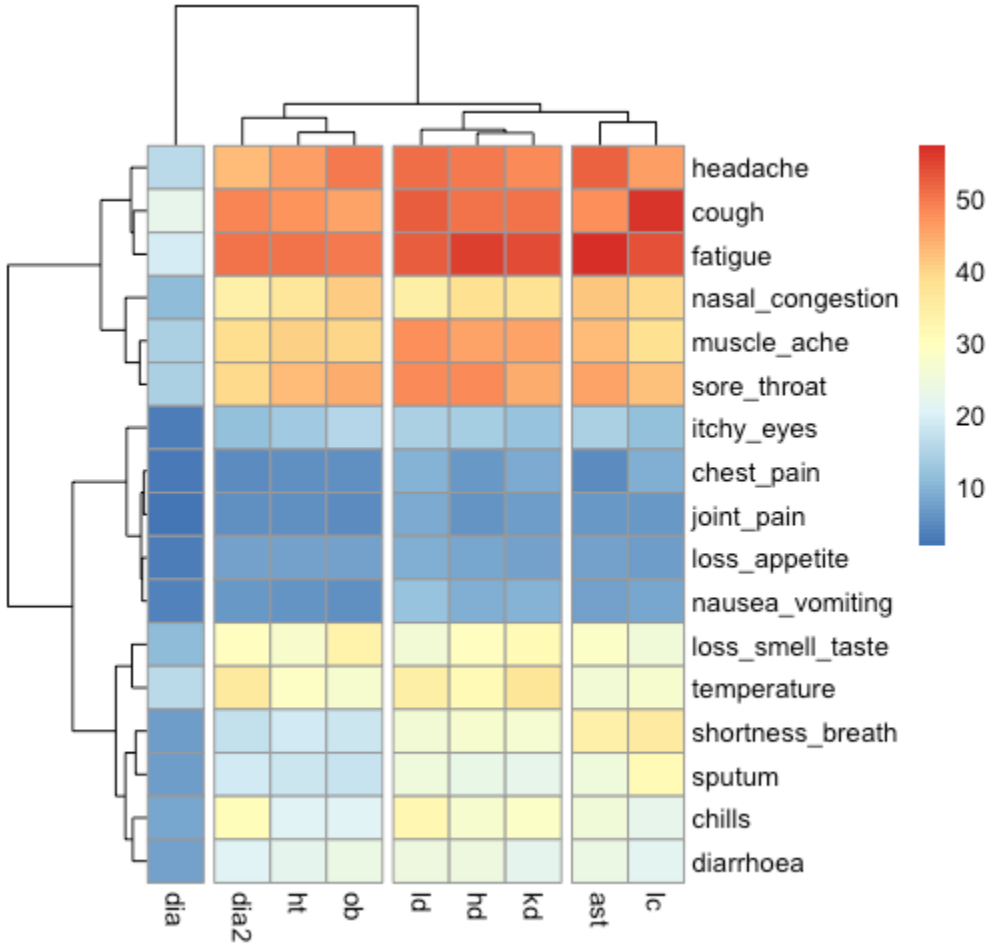
  https://www.youtube.com/watch?v=oMtDyOn2TCc&ab_channel=StatQuestwithJoshStarmer (~17 min)

Heatmaps

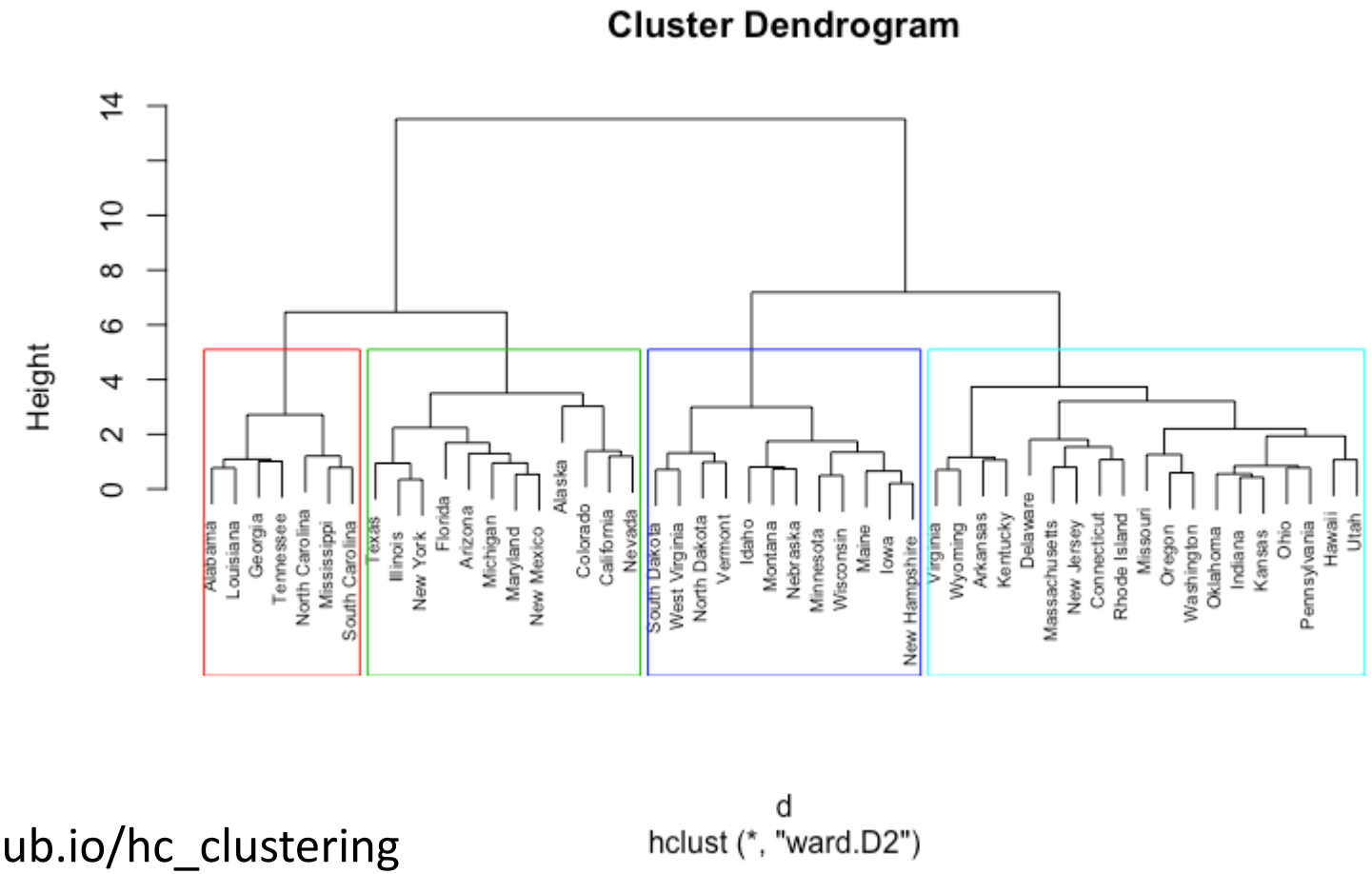
- The way you scale the data is very important (affects what you can see) (local vs global scaling, scale rows, scale columns, or scale both…)
- Outliers can also introduce distortions when using global scaling
- Often shown in association with results from a (hierarchical) clustering

Heatmaps

- With agglomerative hierarchical clustering
- Dendogram + heatmap



Clustering/Dendogram



Source: https://uc-r.github.io/hc_clustering

Heatmaps

- See for an example – heatmap & hierarchical clustering

  https://www.youtube.com/watch?app=desktop&v=oGDvD3wwXsU

  Here rows are the variables and columns are the items (individuals)

RadViz

- A radial visualization based on the spring paradigm
  - Each attribute in the dataset is represented by a `dimensional anchor´
  - Dimensional anchors are evenly distributed on a unit circle

- Each data item is represented as a point that is `linked´ to every dimensional anchor by a `spring´
  - Each spring pulls the item with a stiffness proportional to the attribute value for that item
  - The point is positioned in the 2D space where the spring’s tension is minimum

(parenthesis) Spring embedder

- Force directed graph layout

https://observablehq.com/@d3/force-directed-graph-component

RadViz

- Mapping a data item $\boldsymbol{x}_i = (v_{i1}, v_{i2}, \ldots, v_{im})$, with $\boldsymbol{S}_j$ giving the position of the j^{th} dimensional anchor point
- Attributes v_{ij} normalized in range [0,1]

At equilibrium:

$$\sum_{j=1}^{m}\left(\vec{S}_j - \vec{x}_i\right)v_{ij} = 0$$

Solving for $\boldsymbol{x}_i$:

$$\vec{x}_i = \frac{\sum_{j=1}^{m}\vec{S}_j v_{ij}}{\sum_{j=1}^{m} v_{ij}}$$

RadViz

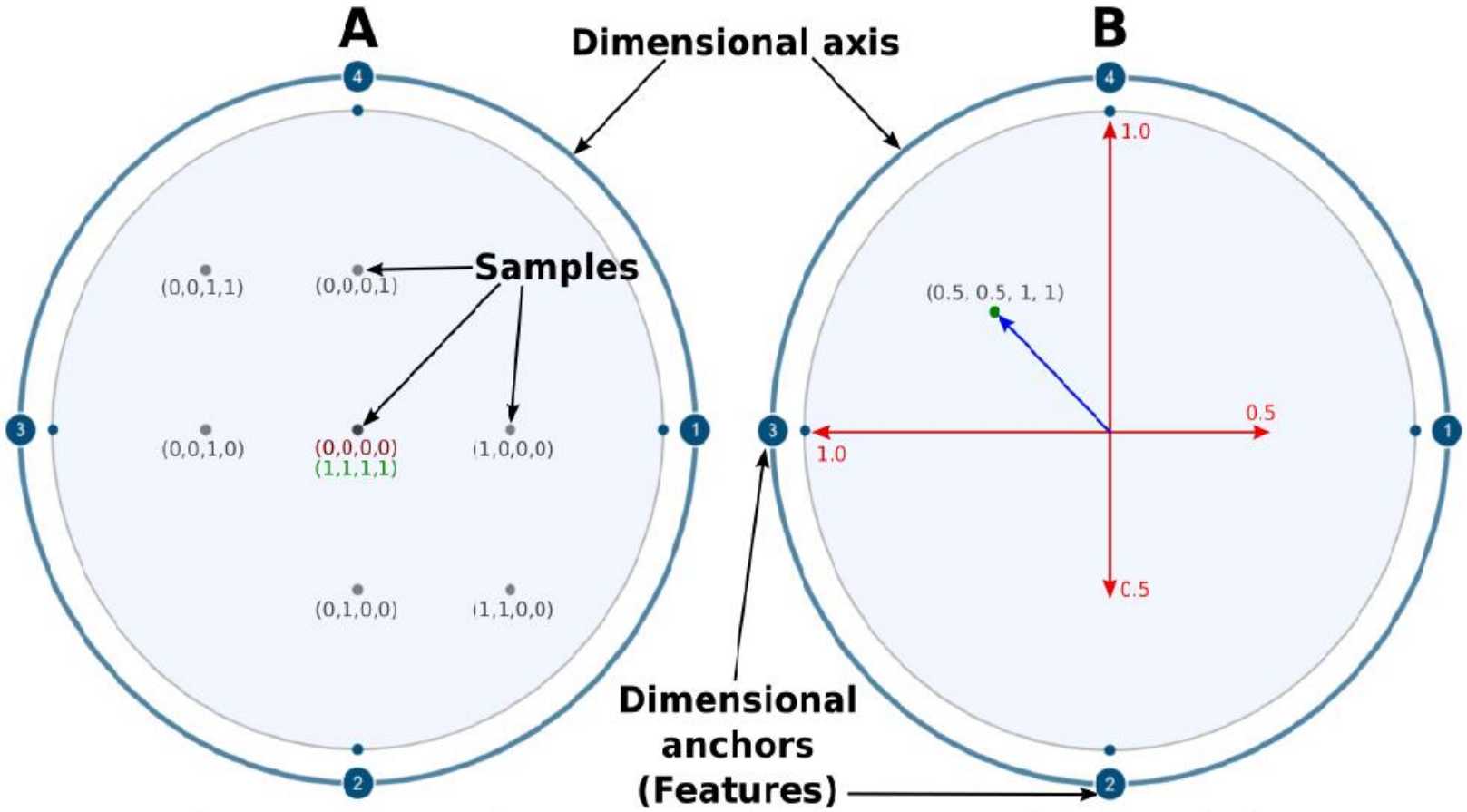


Figure 4 – Radviz visualization with 4 features placed as dimensional anchors along the dimensional axis. (A) Examples of samples with corresponding features array values are positioned inside the visualization. (B) A sample position calculated by the sum of vectors (Red lines), resulting in the vector that gives the sample position in the visualization (Blue line). The final sample position is represented by a green dot.

Source: C.D.G. Reis, Seecology: Data Visualization Framework for Soundscape Ecology Applications. DSc Thesis, ICMC-USP 2020.

RadViz

- https://observablehq.com/@saehrimnir/radviz

- https://orange.readthedocs.io/en/latest/widgets/rst/visualize/radviz.html

RadViz: discussion

- Points with approximately equal attribute values will lie close to the center
- Points with similar values whose DAs are opposite each other in the circle will lie near the center
- Points which have one or two dominant attribute values will lie close to the corresponding DAs

RadViz: discussion

- All atributes should be in the same range
  - The simplest way to achieve this is normalize each numerical atribute to its range, so that all vary in the interval [0,1]
- The position of dimensional anchors on the circle is critical
  - The best projection (as judged by the separation of points in the display) is achieved when dimensional anchors that correspond to highly correlated attributes in the data are placed closer on the unit circle

RadViz: example

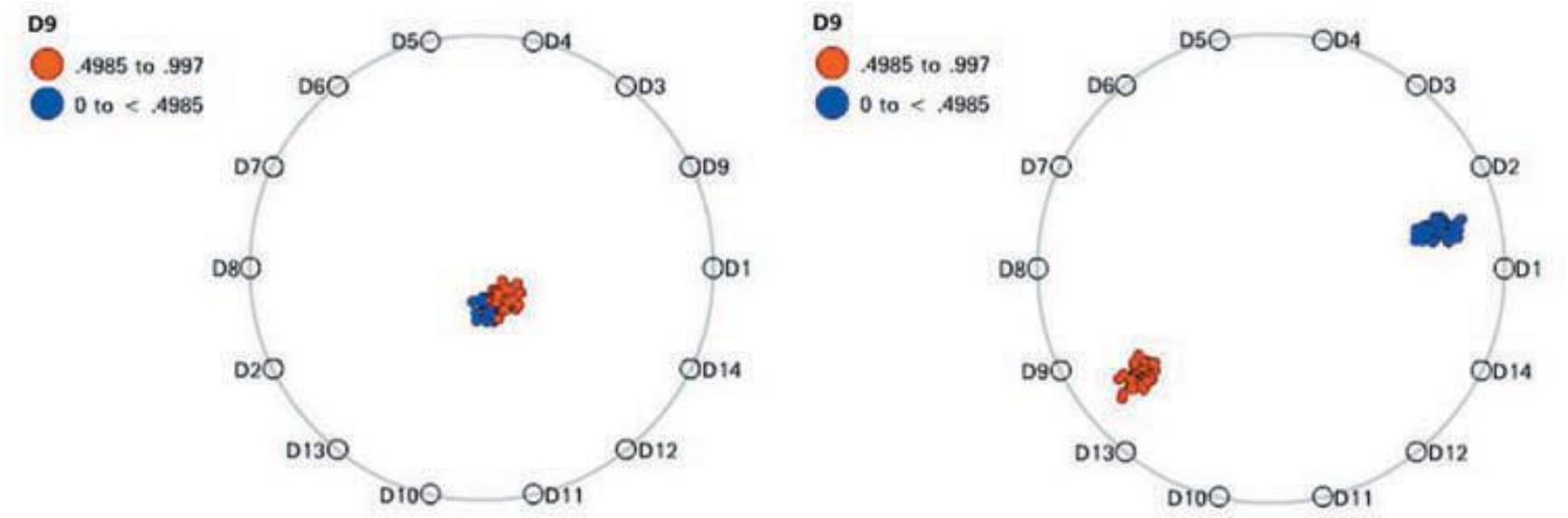


Figure 5. Reordering of 14-d data in RadViz: the arrangement of dimensional anchors on the left produces overlapping clusters for 100 data records with two inherent clusters, whereas the reordering on the right separates the two clusters. D9 is the dimension being used to color the records to provide clear cluster separation.

Source: Daniels et al. Properties of normalized radial visualizations. Information Visualization 11(4) 273–300.

RadViz: axes arrangement

- https://observablehq.com/@aware/d3-radviz
- https://aware-diag-sapienza.github.io/d3-radviz/

Effectiveness Error: Measuring and Improving RadViz Visual Effectiveness. Angelini et al. IEEE TVCG 2018.

RadViz: discussion

- Optimization of the position of dimensional anchors is also discussed in Di Caro *et al* 2012, where the authors suggest 2 metrics that can be used in an optimization process.

- Inherently interactive, e.g., see also Ono et al. Concentric RadViz

  https://www.youtube.com/watch?v=0tQX622HdAQ

- Visual clutter and point overlapping may be issues

RadViz: example

Figure 3. The battery dataset visualized with parallel coordinates.

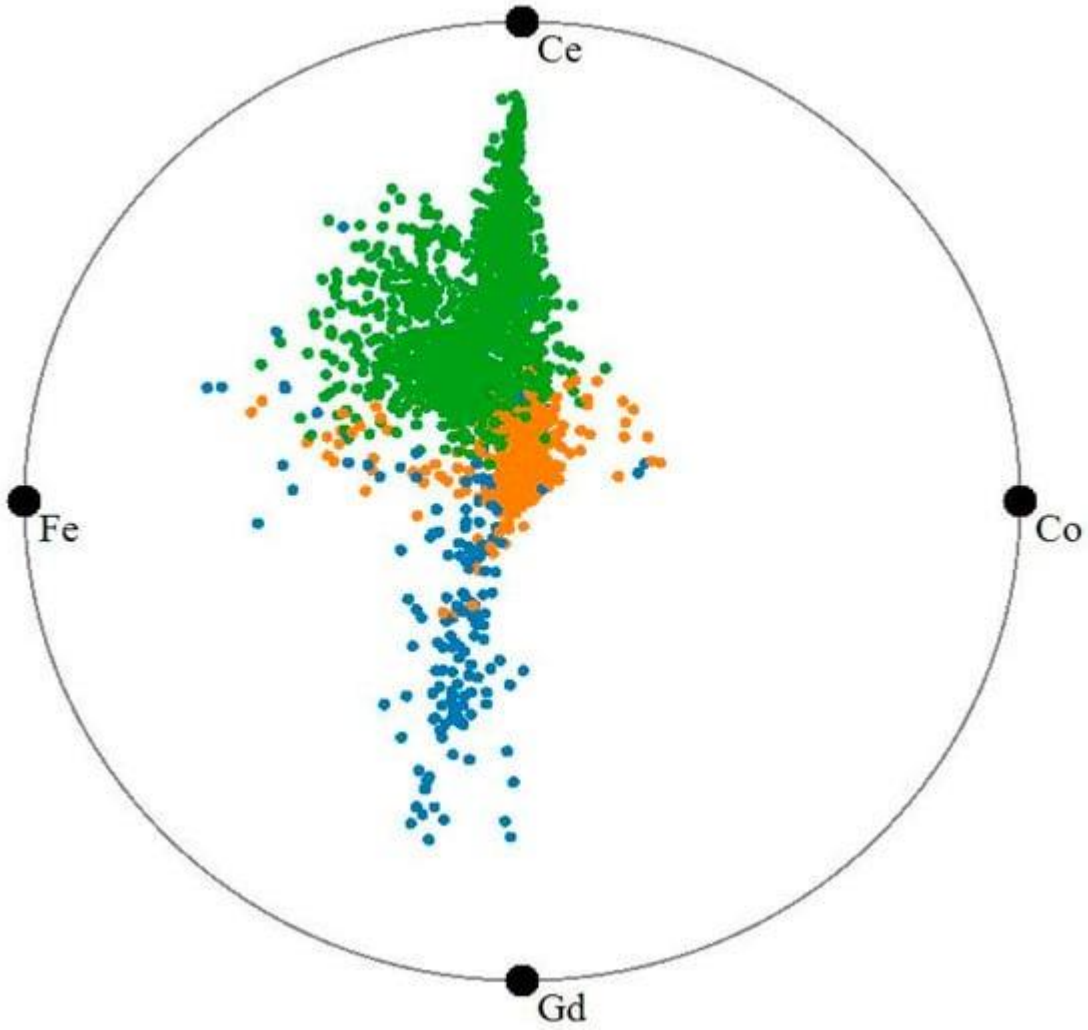


Figure 1. The battery data set visualized with RadViz.

Source: *RadViz Deluxe:* An Attribute-Aware Display for Multivariate Data· Cheng et al. *Processes*, 2027

RadViz: example

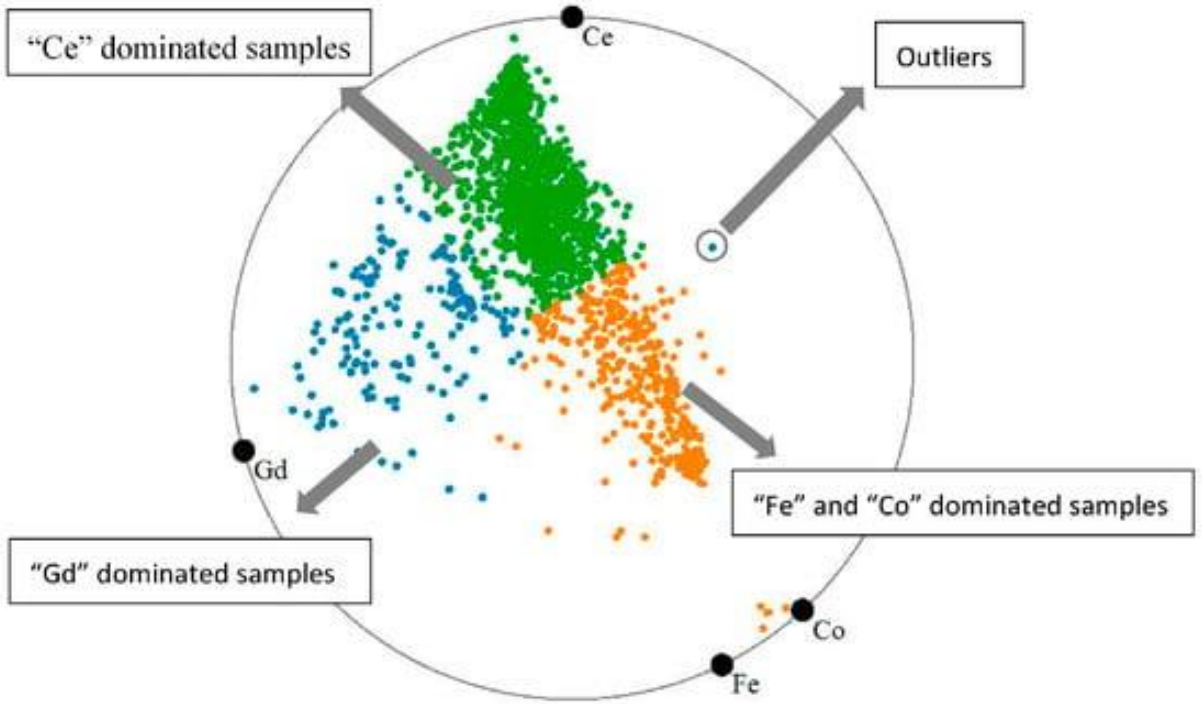


Figure 7. The battery data set visualized with RadViz Deluxe.

Source: *RadViz Deluxe: An Attribute-Aware Display for Multivariate Data*
Cheng et al. Processes, 2027

Sources/Material

- Chapter 7, Interactive data visualization, Ward et al.

- Chapter 11, Data visualization, Telea