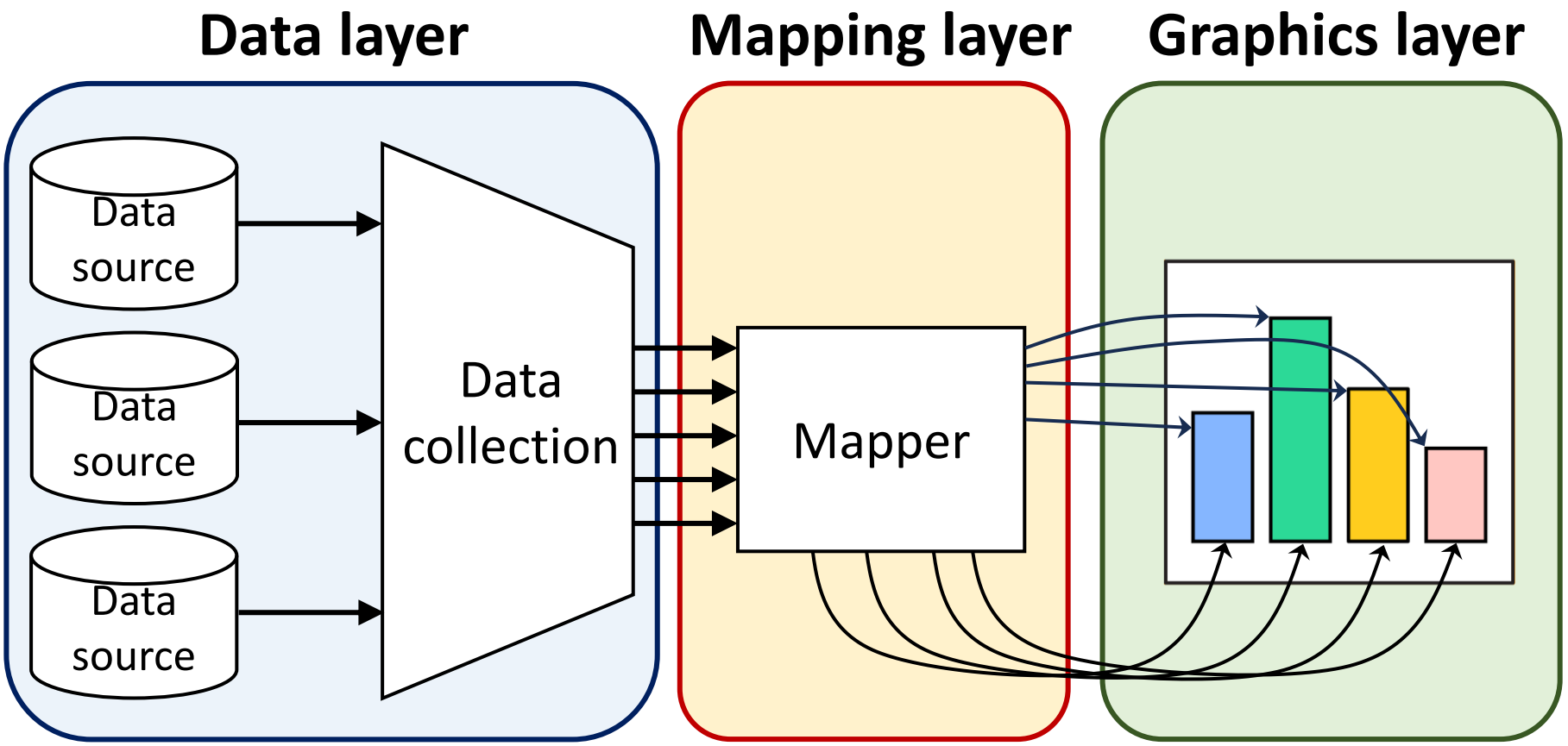
Data Visualization

Nguyen Ngoc Thao
nnthao@fit.hcmus.edu.vn

Content outline

- Basic charts
- Advanced representations

Data visualization framework



Data visualization framework

- Locate and obtain data
- Import data in proper format
- Relate data for proper correspondence
- Data analysis and aggregation

Data layer

- Associate appropriate geometry with corresponding data channels
- Data analysis and algorithms (e.g. contouring)

Mapping layer

- Conversion of geometry into displayable image
- Decorations
- Managing interaction

Graphics layer

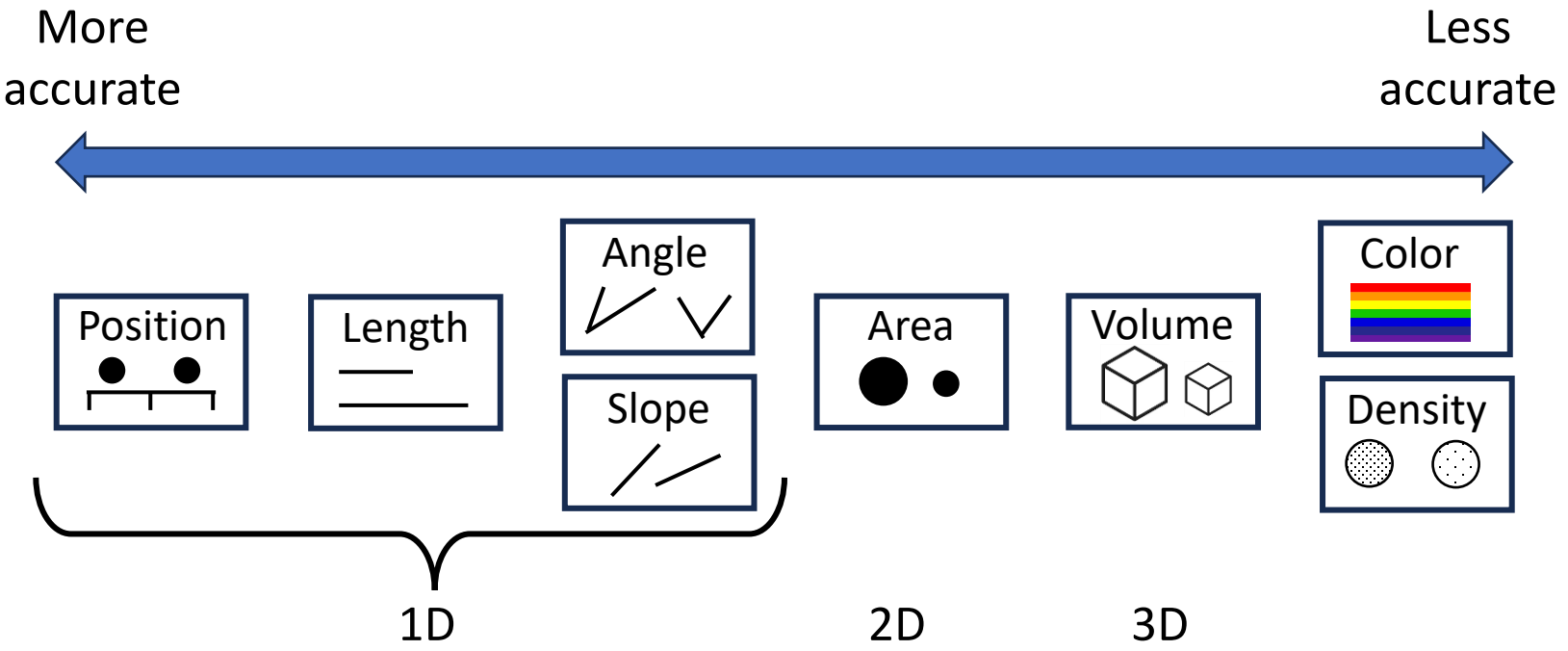
Data types

	Discrete (no between values)	**Discrete** (values between)
Ordered (values are comparable)	**Ordinal** e.g., sizes: S, M, L, XL **Quantitative** e.g., counts: 1, 2, 3, …	**Fields** e.g., altitude, temperature
Unordered (values not comparable)	**Normal** e.g., shapes: □○△ **Categories** e.g., nationality	**Cyclic values** e.g., directions, hues

Data as variables

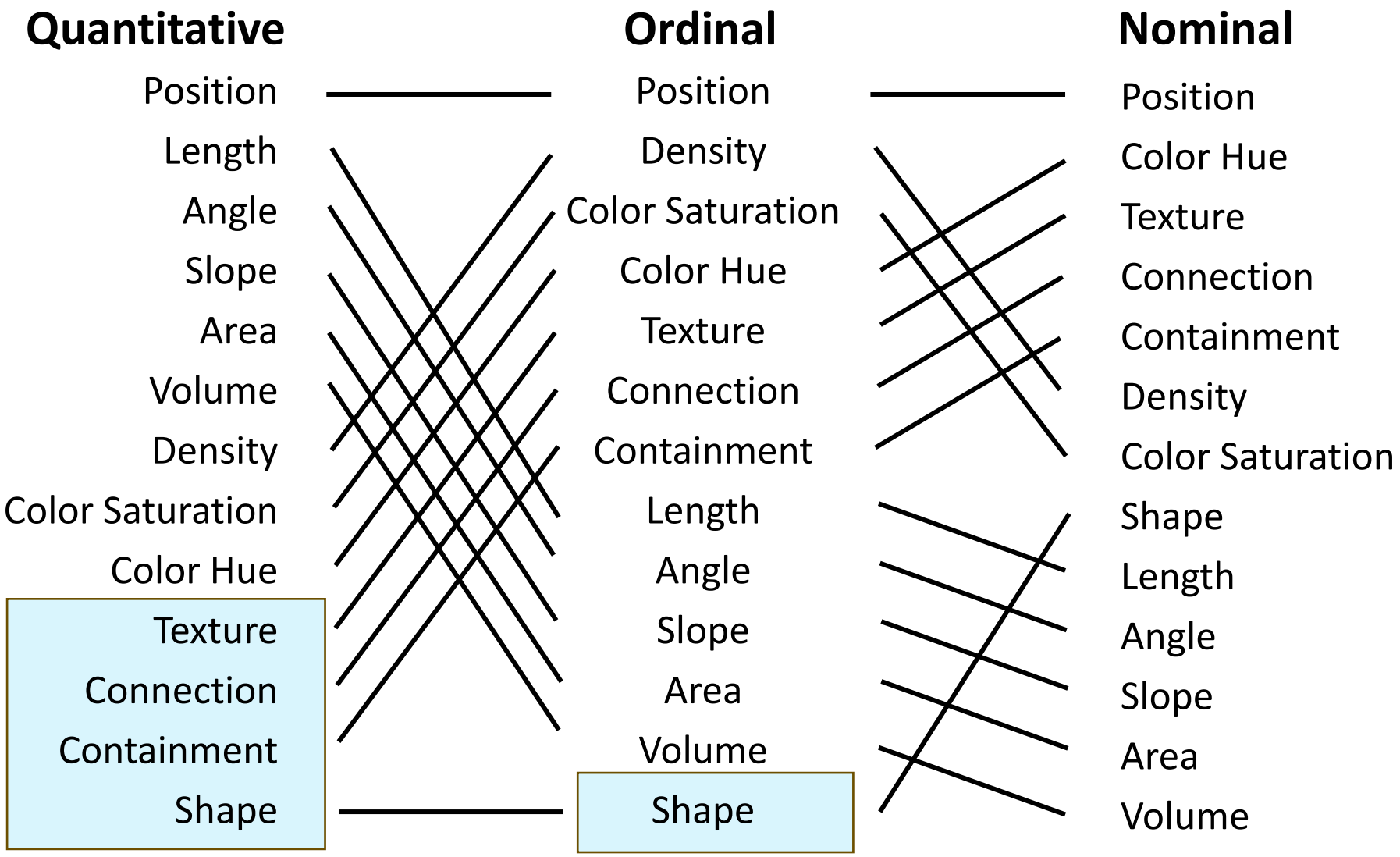
Science	Databases	Data warehouses
Independent variable	Key	Dimension
Dependent variable	Value	Measure

Ranking of perceptual tasks



Accuracy ranking of quantitative perceptual tasks.
Higher tasks are accomplished more accurately than lower tasks.
Cleveland and McGill empirically verified the basic properties of this ranking.

Cleveland, W. S., and McGill, R. Graphical perception: Theory, experimentation and application to the development of graphical methods. Journal of the American Statistical Association, 79(387) 1984



Ranking of perceptual tasks.

The tasks shown in the gray boxes are not relevant to these types of data.

J. Mackinlay, Automating the Design of Graphical Presentations of Relational Information, ACM Transactions on Graphics 5(2), 1986

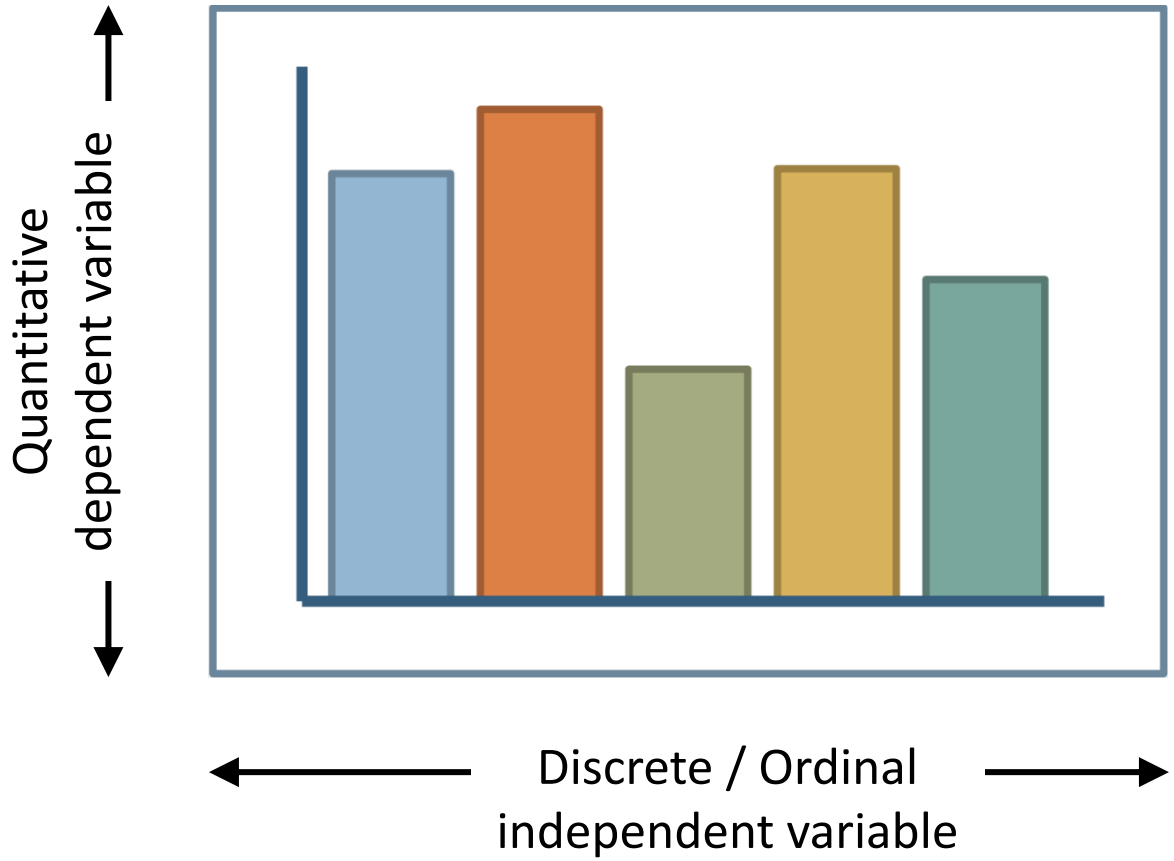
Basic charts

Bar charts

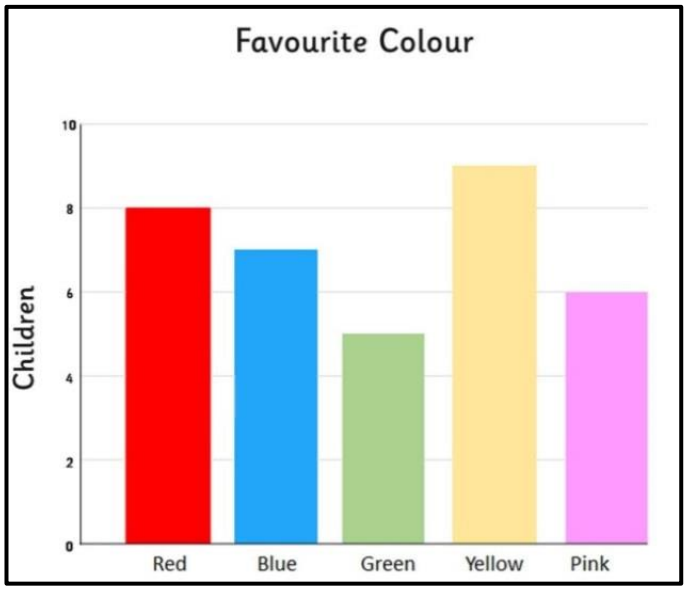
Position

Length

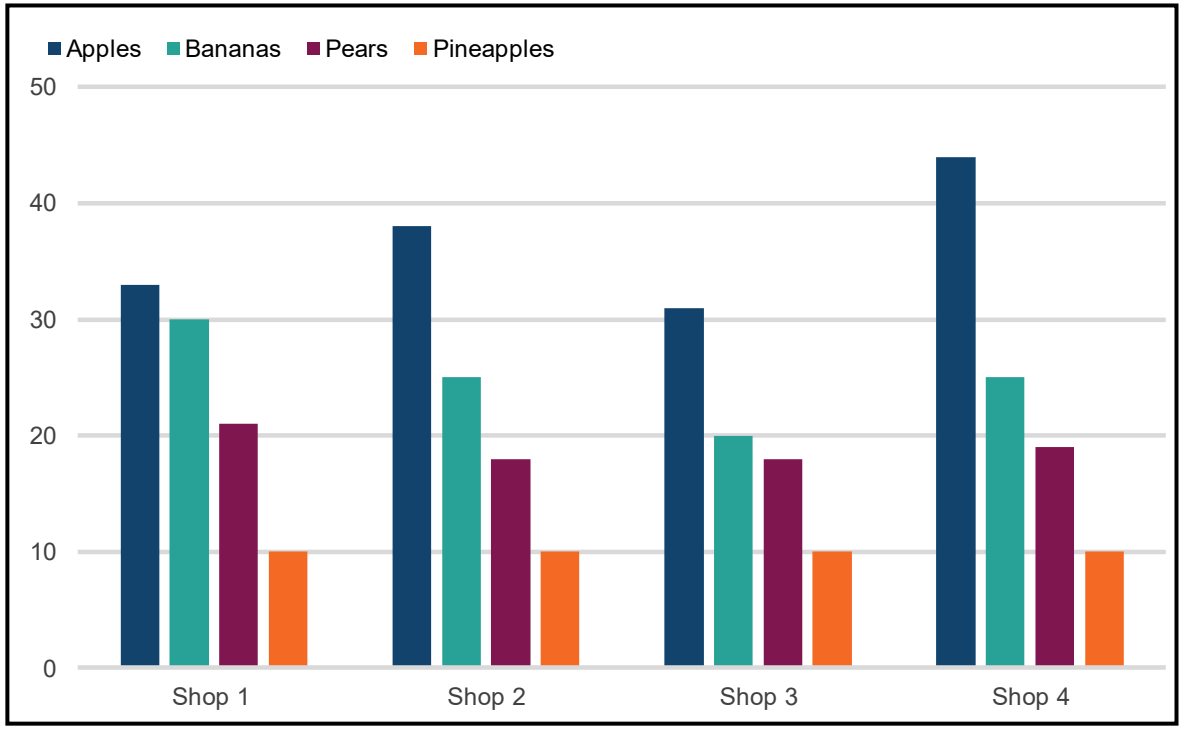
- Bar charts benefit from both the position (bar top) and length (bar size).



Bar charts: Examples



Comparisons can be made by considering bars within each group.



Bar chart for a single attribute (left) and bar chart for contrasting four attributes (right).

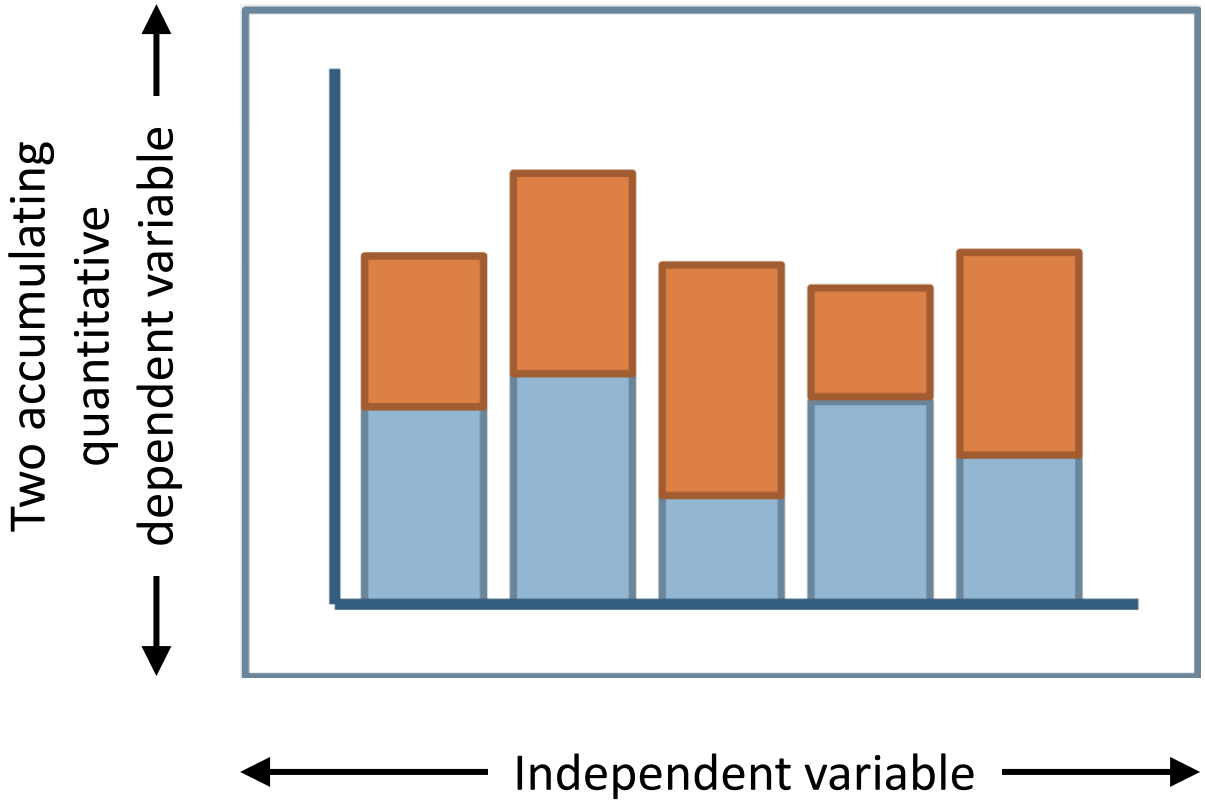
Image credit: Engnovate, Analysis Function

Stacked bar charts

Position

Length

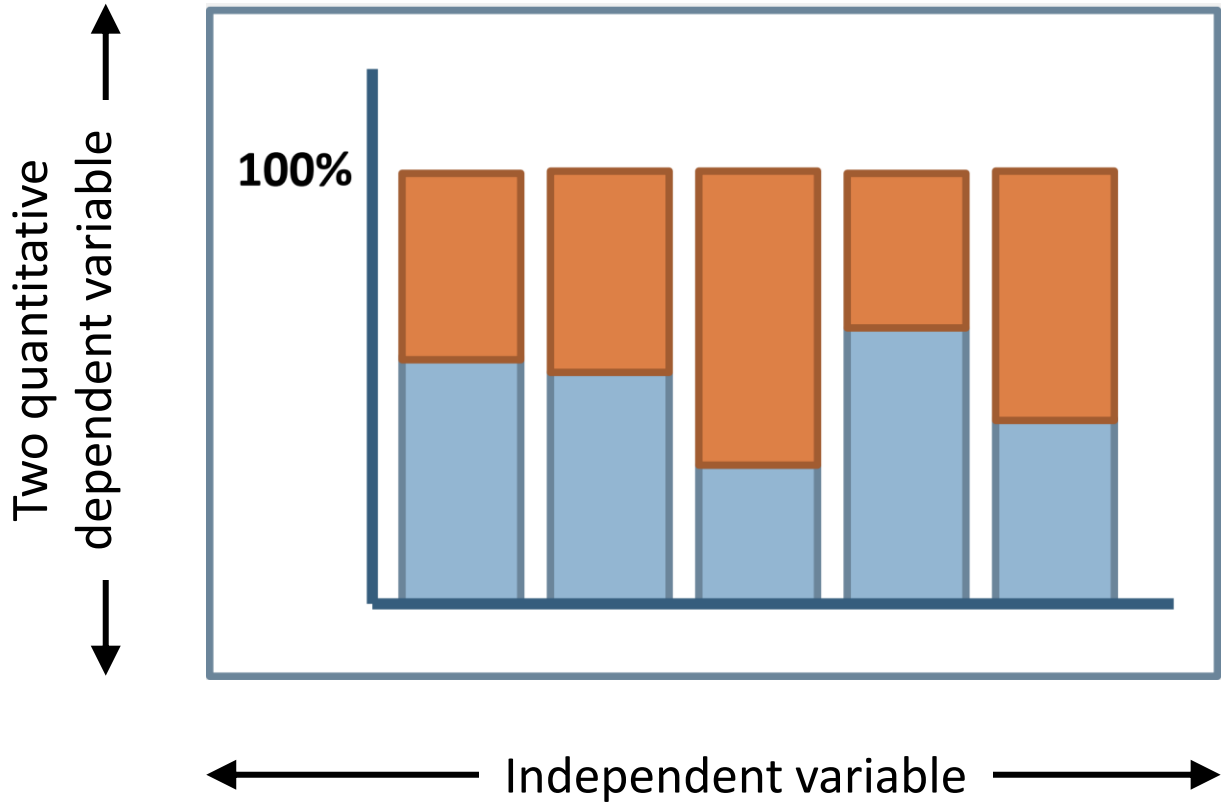
- Central limit theorem → as more bars are added, sums will vary less.



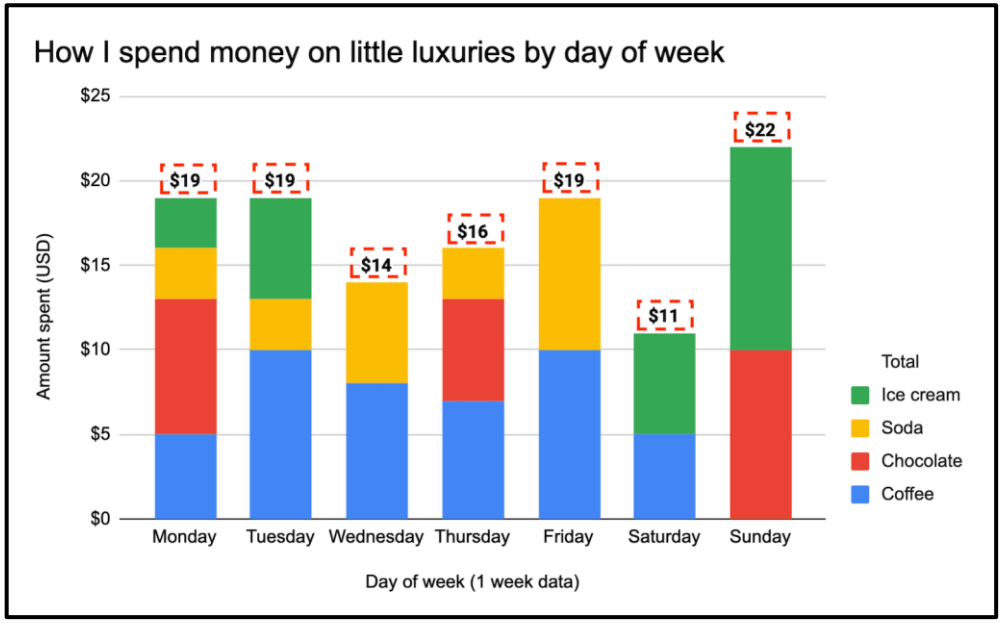
Relative stacked bar charts

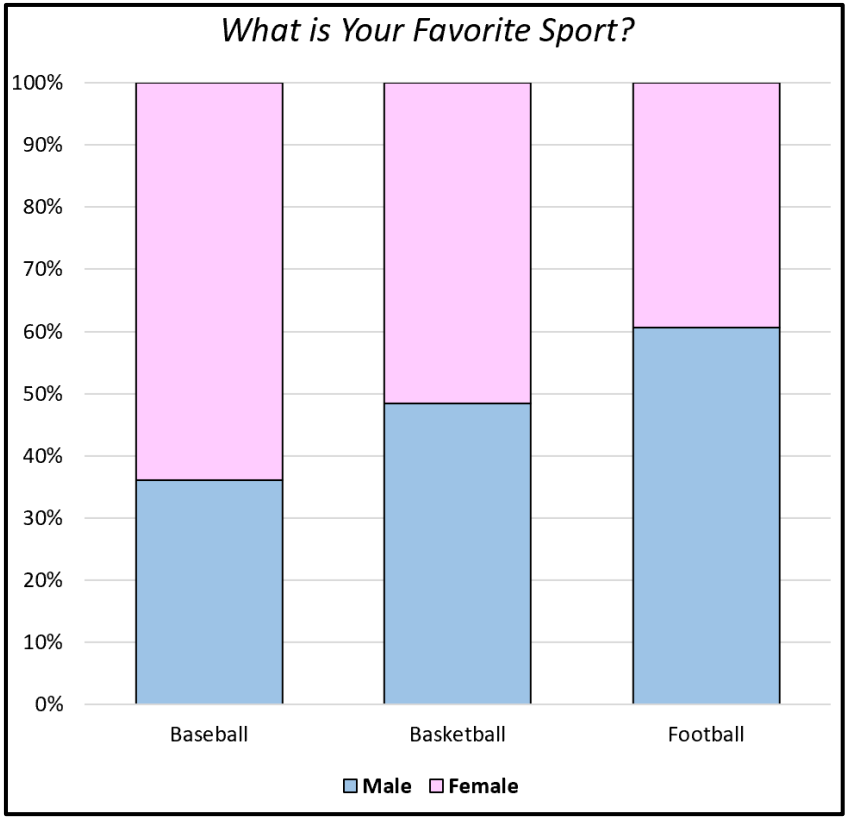
Position

Length



Stacked bar charts: Examples



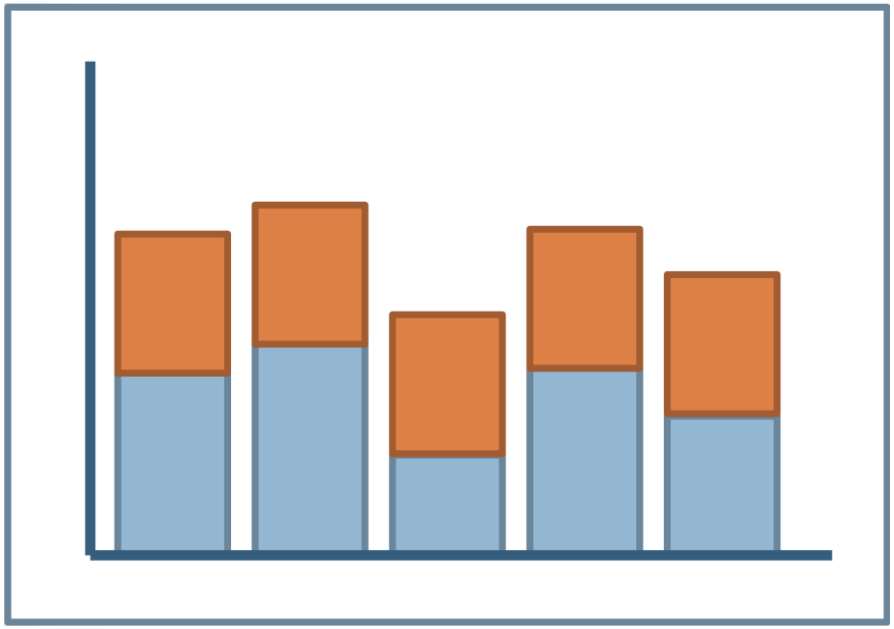


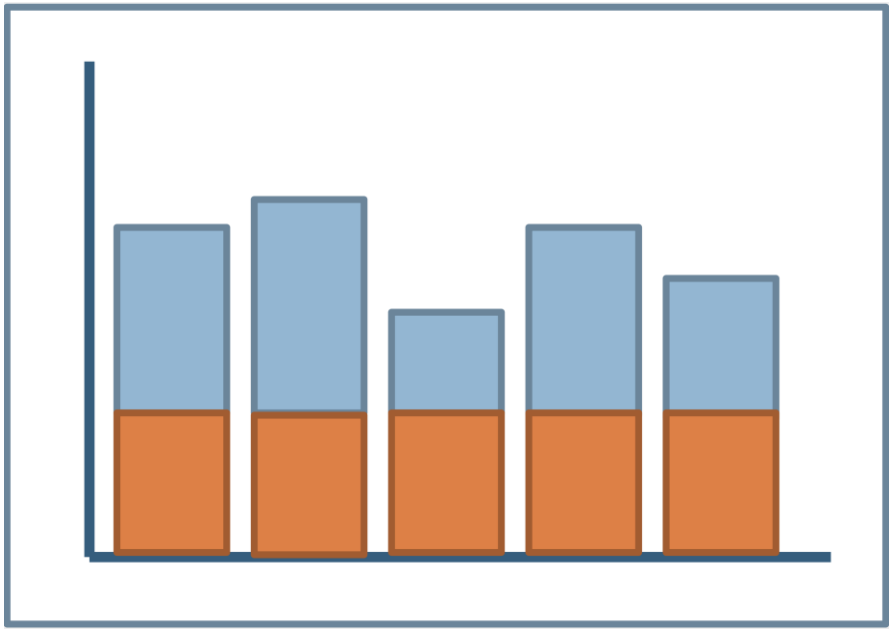
Stacked bar chart (left) and relative stacked bar chart (right)

Image credit: DATAtab, Statology

Stacking order matters

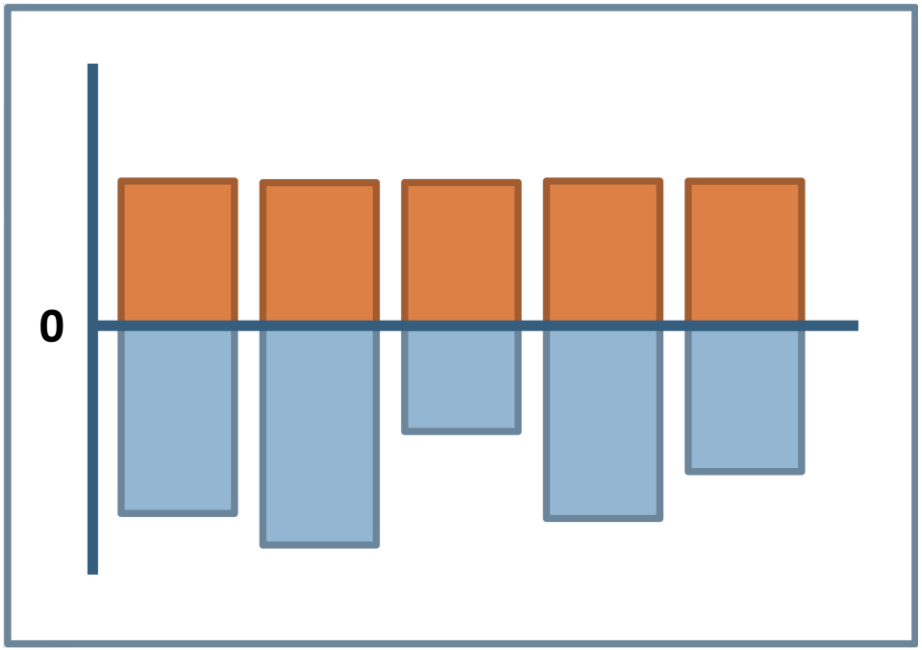
- Variance of lower stack elements influences perception of upper elements.

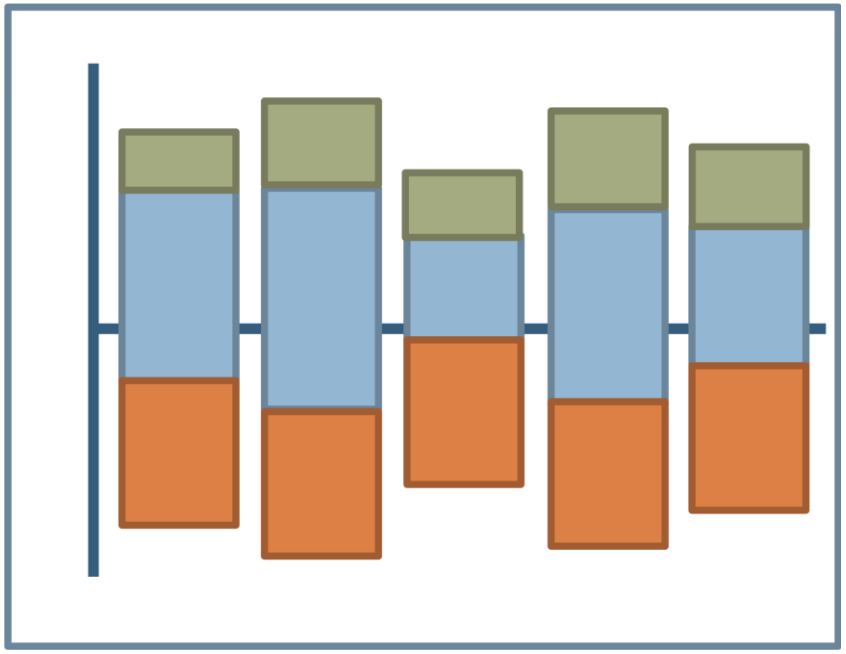




Position > Length

Diverging stacked bar charts





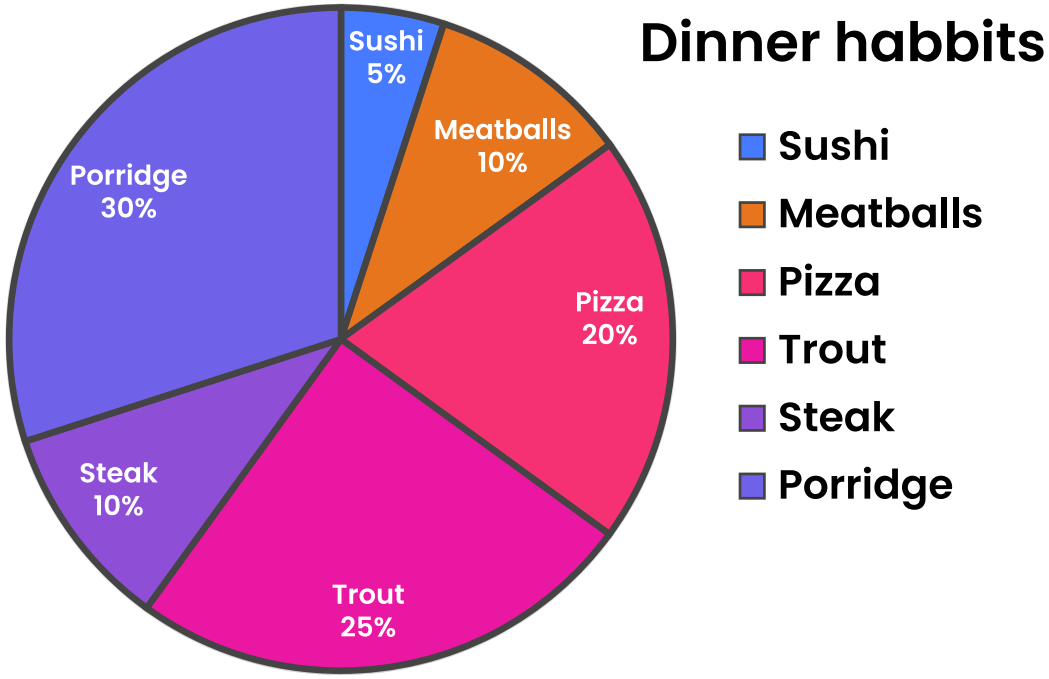
- Benefits from position and length
- Only work for two variables
- Negative connotation for lower bars

- Only indicates length
- Work for many variables
- Bar trends can still be obscured by neighboring bar variance

Pie charts

Angle

- Pie charts indicate relative portions of a quantitative dependent variable of a single dimension via the angles of slices, which sum to 100%.



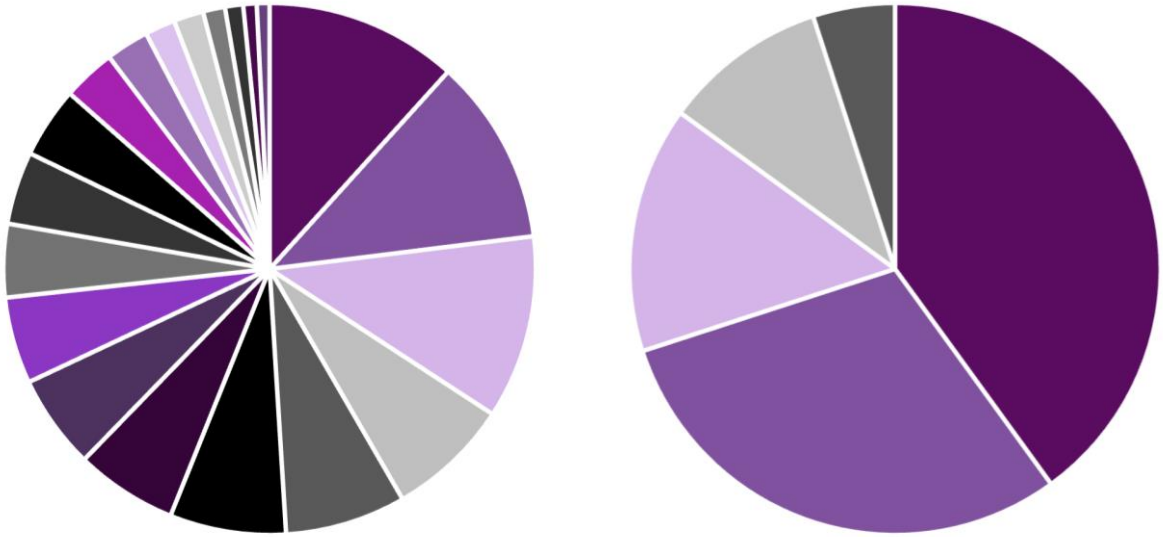
A pie chart showing several options for dinner habbits.

Image credit: HOUSEofMATH

Pie chart: Drawbacks and Fixes

- **Drawback:** When there are too many categories, the pie chart becomes less intuitive.

Many categories



Less categories

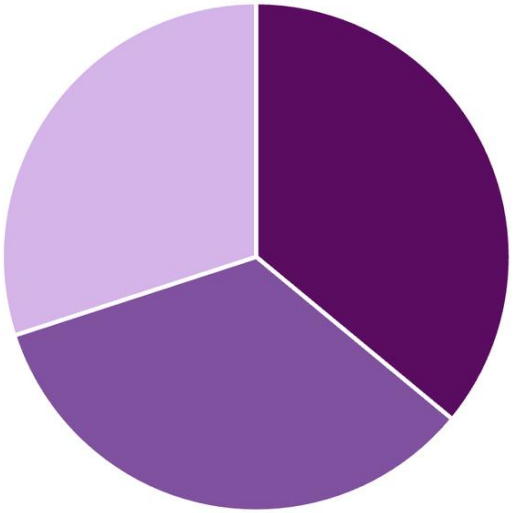
- **Fix:** group smaller or alike categories to reduce the number of slices or consider bar and column charts as better alternatives

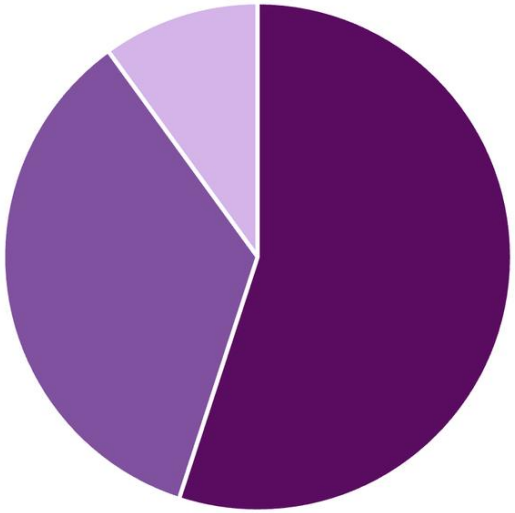
Image credit: EvalAcademy

Pie chart: Drawbacks and Fixes

- **Drawback:** It is difficult to distinguish similar-sized slices as angles are harder to interpret than lengths.

Many categories





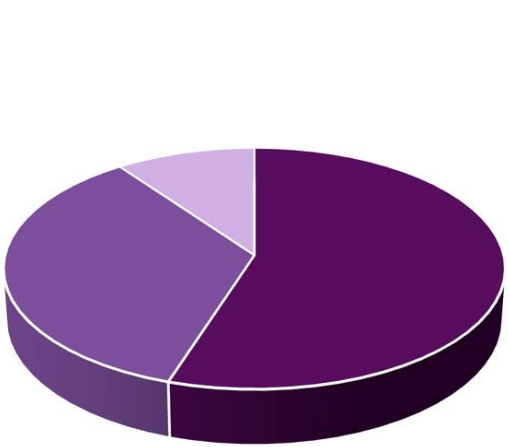
Less categories

- **Fix:** Bar charts are better for close comparisons between categories.

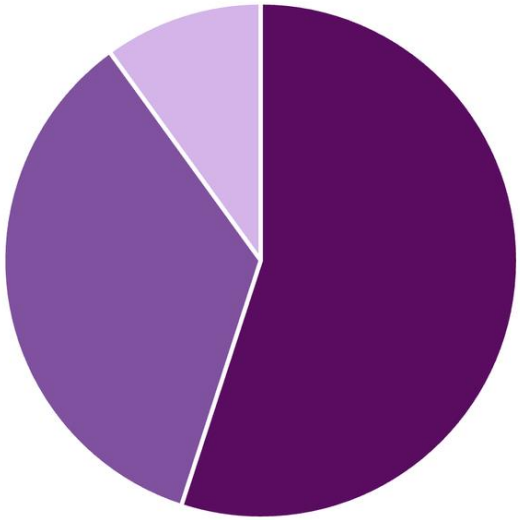
Image credit: EvalAcademy

Pie chart: Drawbacks and Fixes

- **Drawback:** The 3D perspective can distort slice sizes, making some appear larger based on their position.



3D representation



2D representation

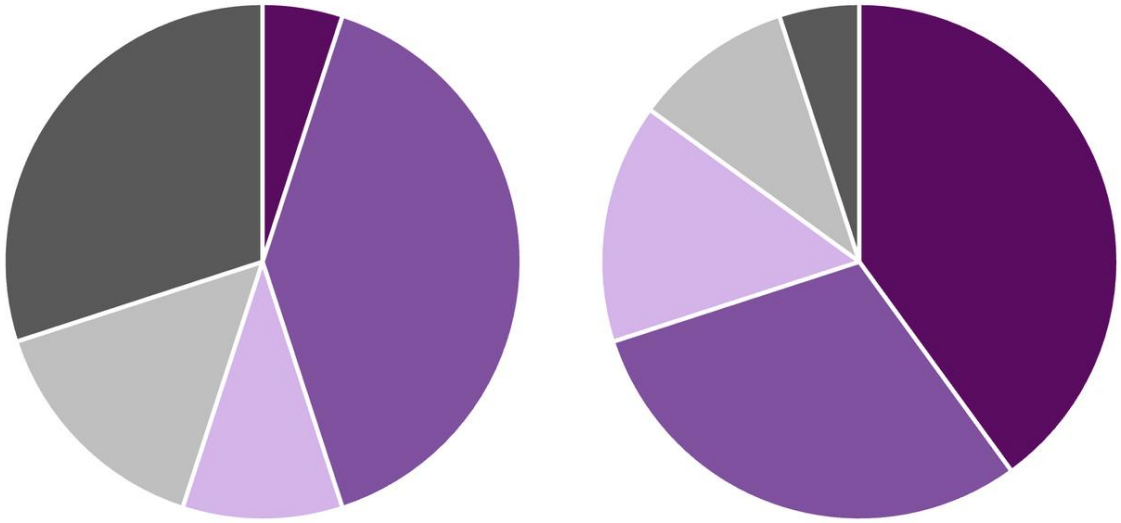
- **Fix:** Use 2D pie charts for easier interpretation.

Image credit: HOUSEofMATH

Pie chart: Drawbacks and Fixes

- **Drawback:** Without the logical ordering of categories (e.g., largest to smallest) it becomes difficult to extract meaningful insights from data.

Unsorted categories



Sorted categories

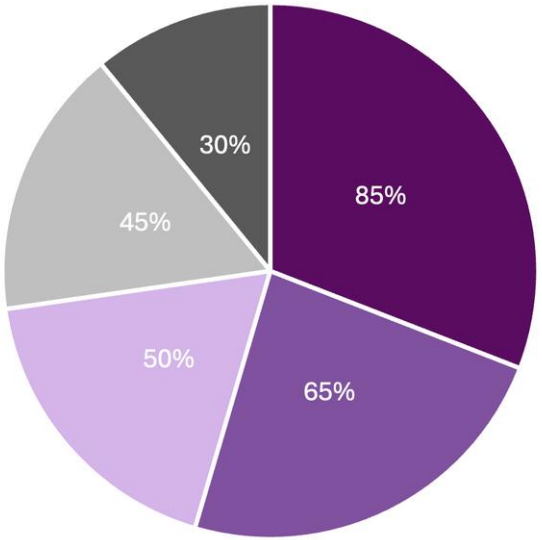
- **Fix:** Order categories from largest to smallest improves the readability of pie charts.

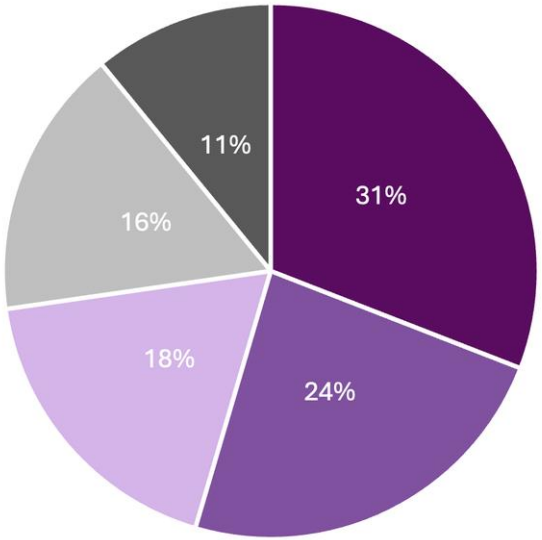
Image credit: HOUSEofMATH

Pie chart: Drawbacks and Fixes

- **Drawbacks:** Using pie charts to visualize non-proportional data (i.e., proportions exceeding 100%) often leads to confusion.

Non-proportional





Proportional

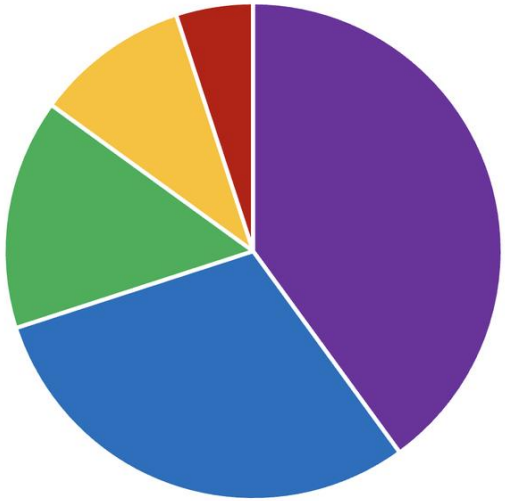
- **Fix:** Use alternative data visualizations, such as bar or column charts.

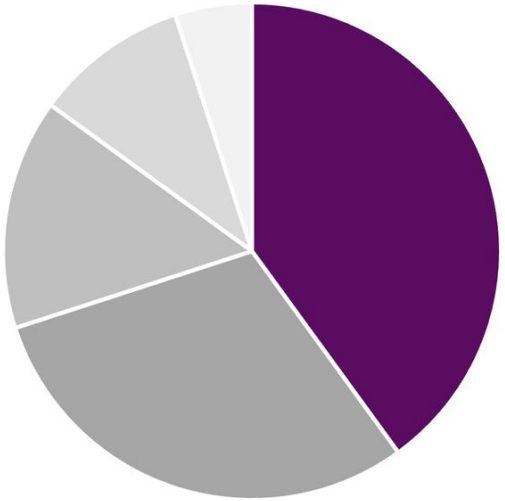
Image credit: HOUSEofMATH

Pie chart: Drawbacks and Fixes

- Too much colour used can detract from the message of the pie chart.
- Additionally, certain colour combinations can be difficult to distinguish, especially for those with colour vision deficiencies.

Too much colors





Selective colors

- Use colour strategically to highlight the most important point in the chart.
- Applying muted tones, such as greyscale, to less relevant data allows the primary colour and key message to stand out.

Image credit: HOUSEofMATH

Line charts

Position

- Line charts benefit from position but not length.

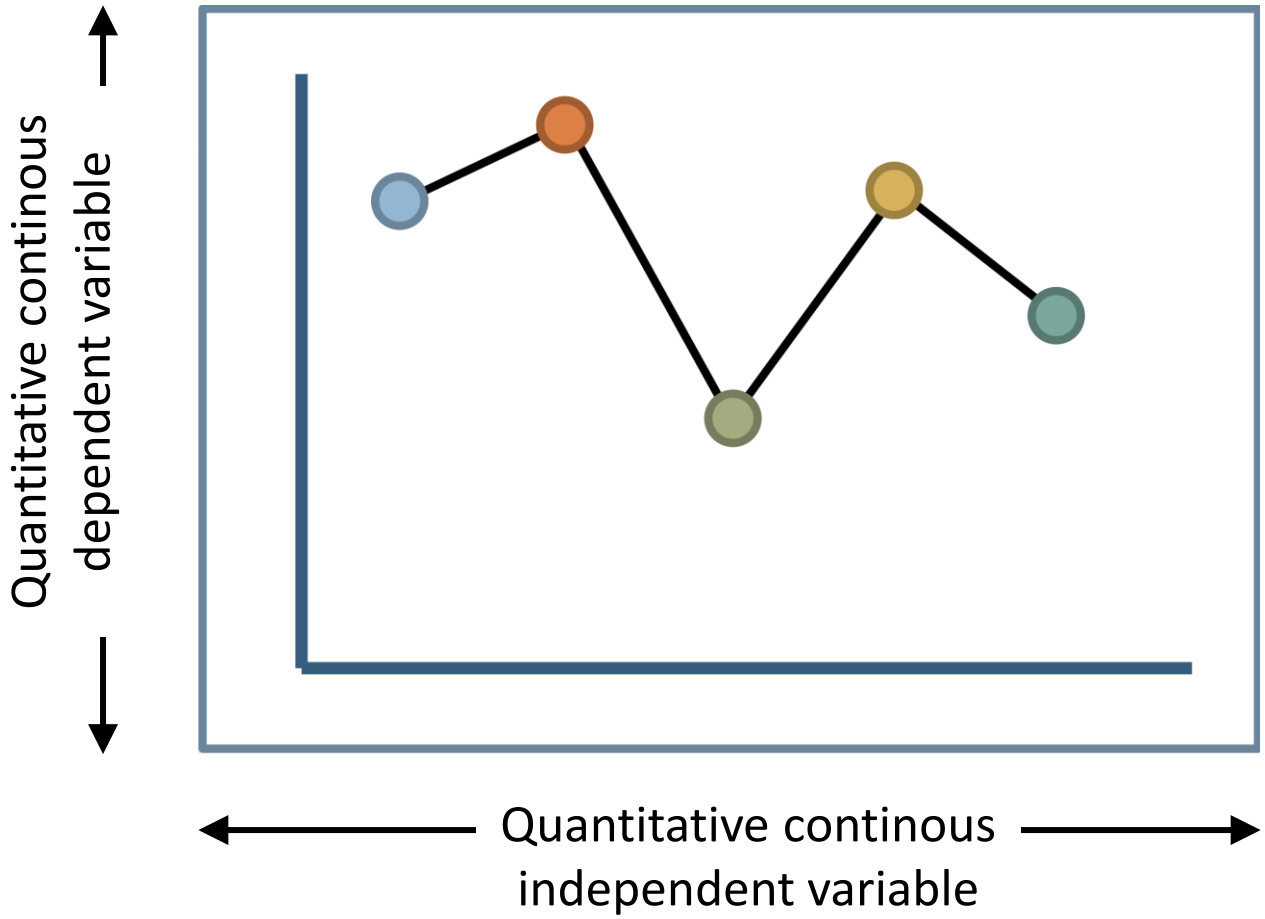
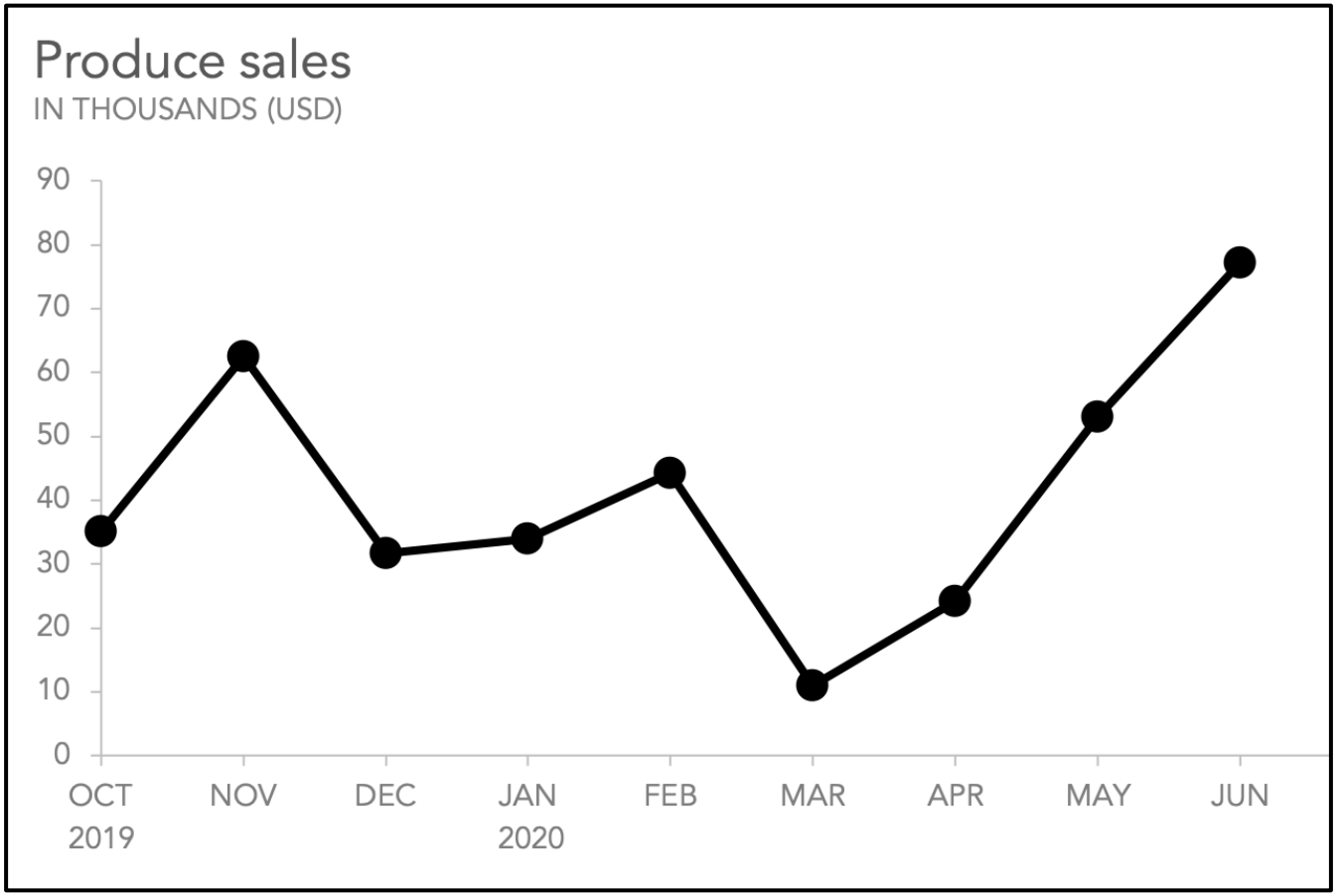


Image credit: Engnovate

Line charts: Examples



A line chart that show product sales (in thousands USD) for each month from Oct 2019 to Jun 2020.

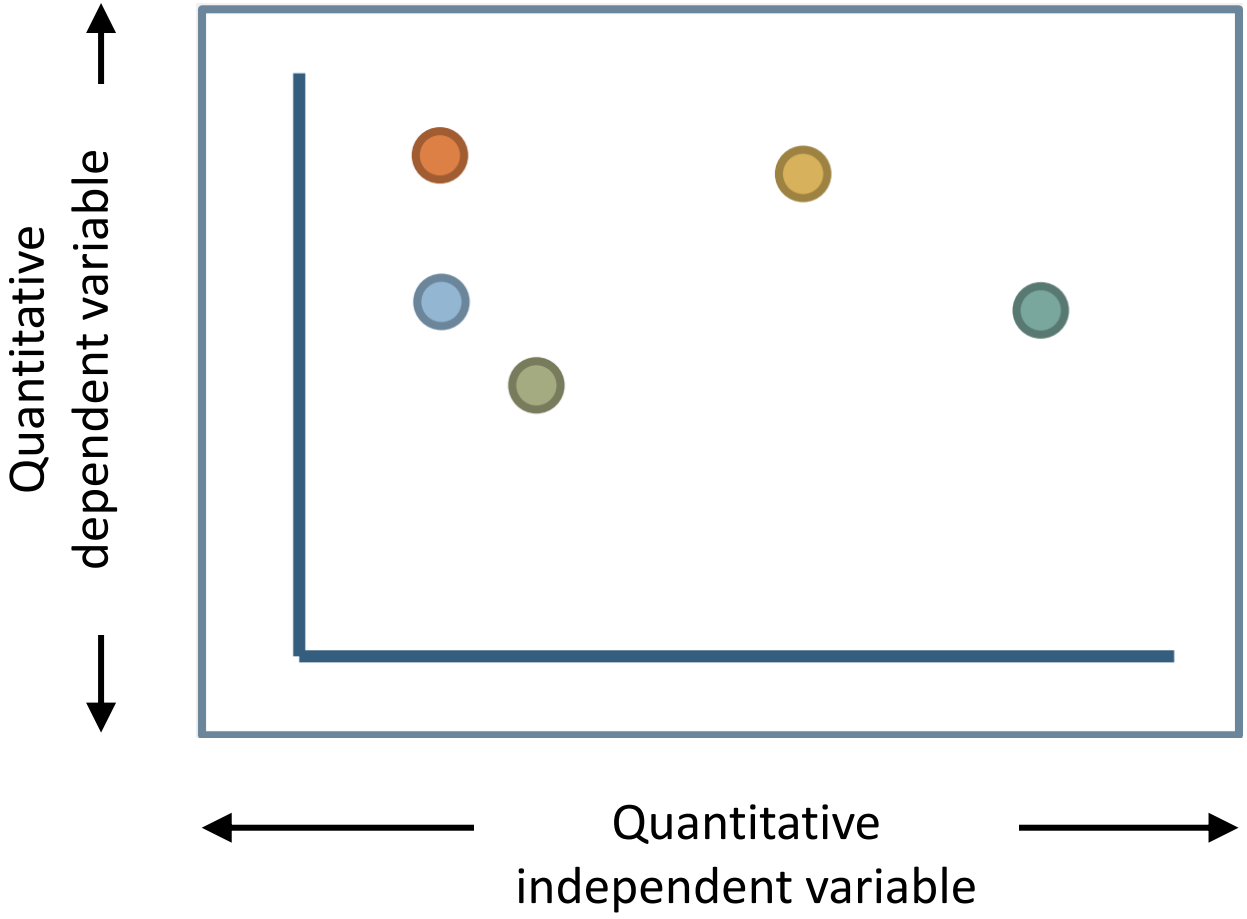
Image credit: Engnovate

Scatter plots

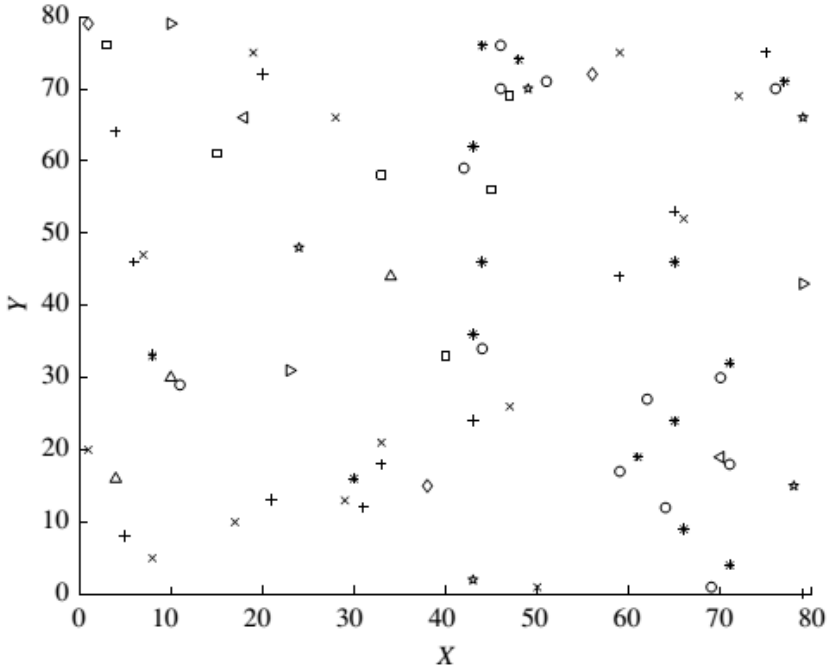
Position

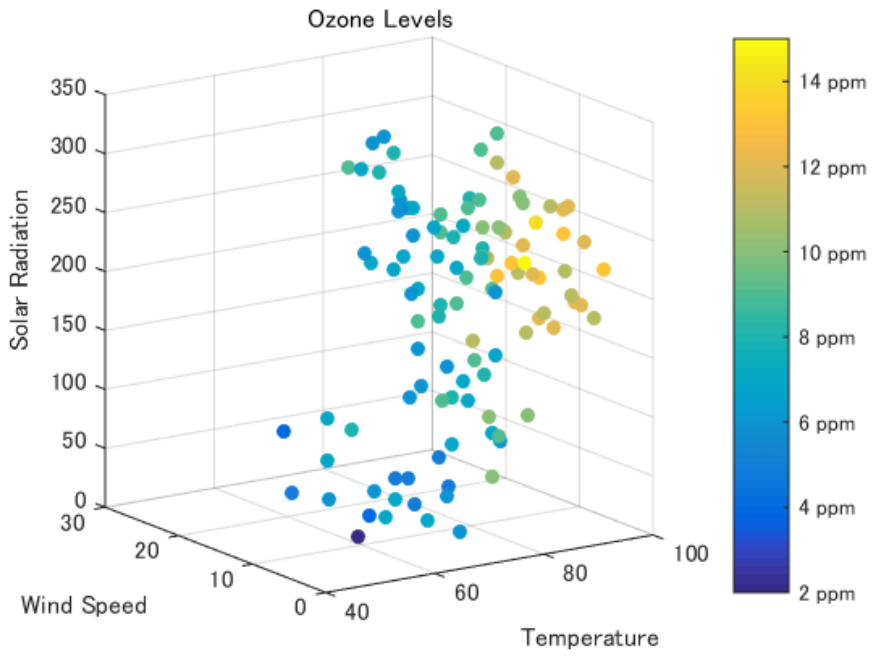
Density

- A scatter plot relies mostly on position, but clusters also yield density.



Scatter plot

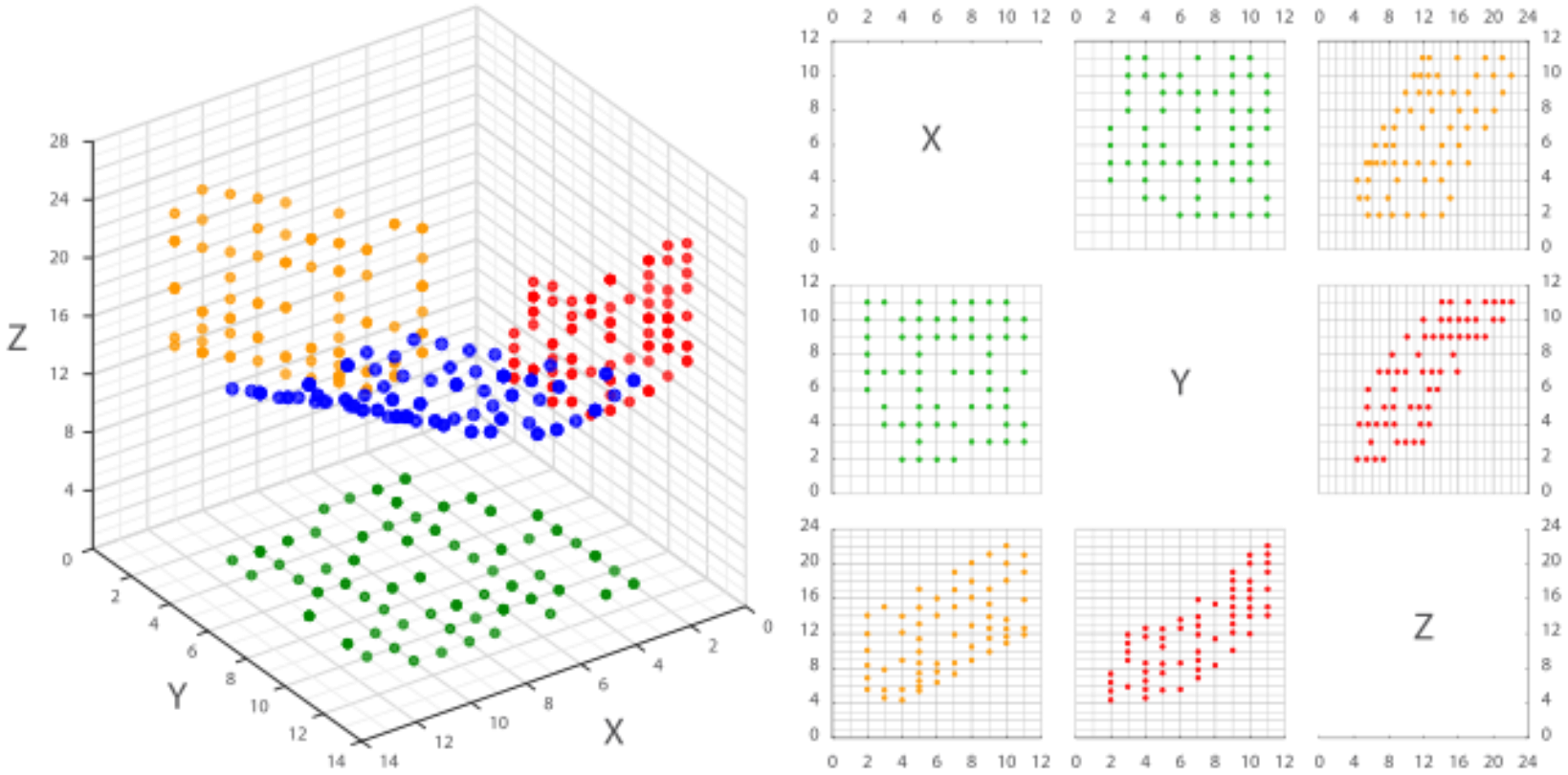




2D scatter plot (left) and 3D scatter plot (right)

Scatter plots: Scatter-plot matrix

- A matrix of 2D scatterplots for all pairs of dimensions from k-dimensions.



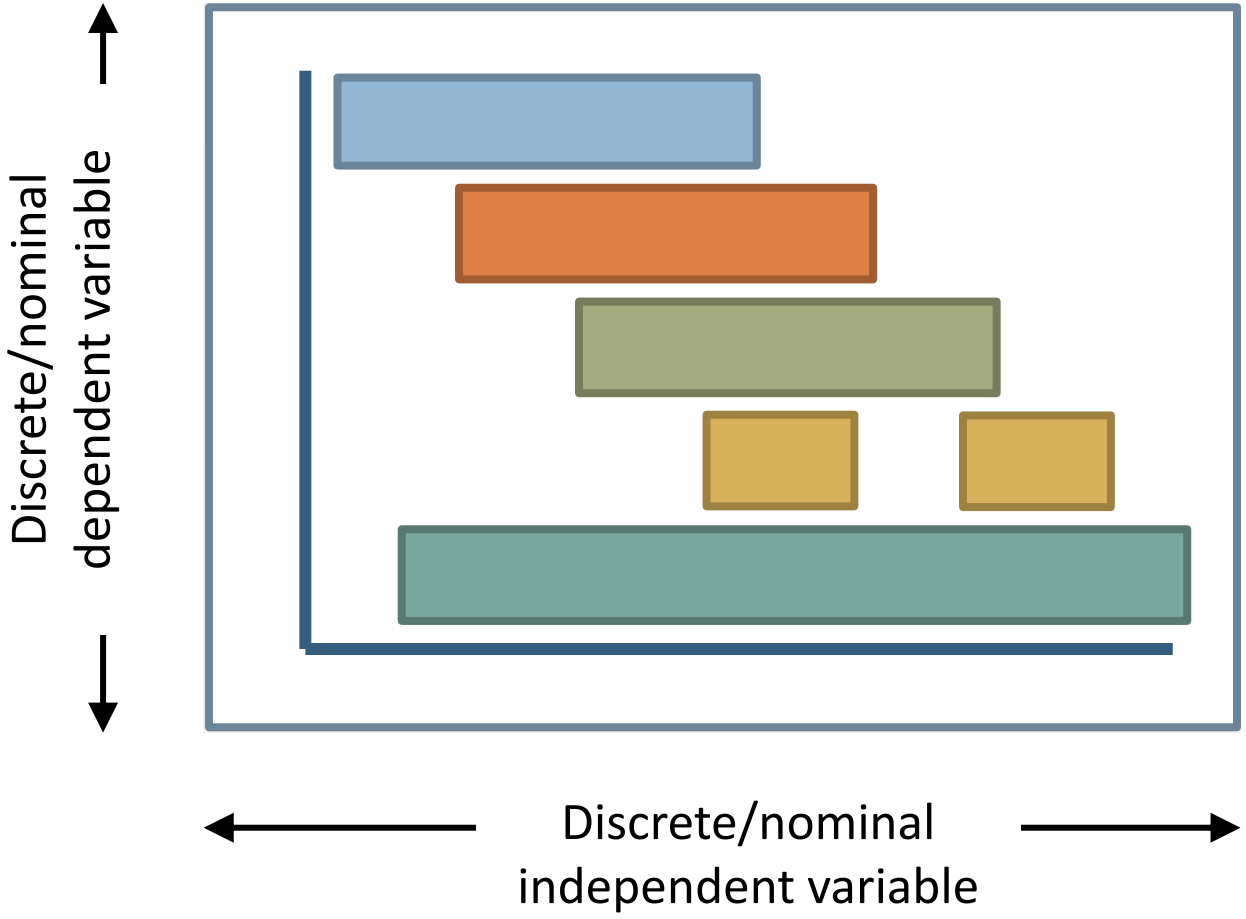
Visualization of 3D data (left) and the correspondent scatterplot matrix (right).

Image credit: Wikipedia

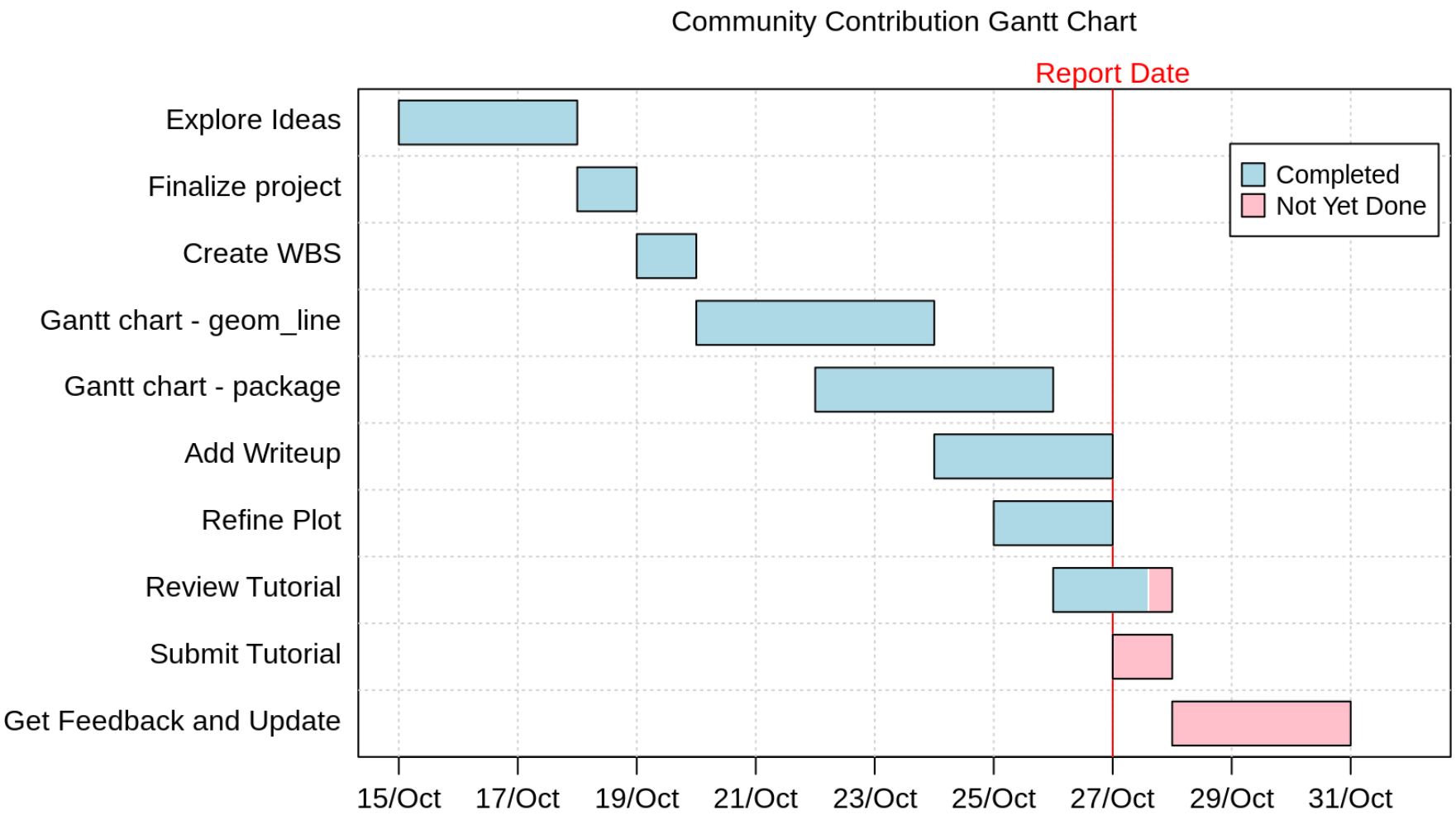
Gantt charts

Position
Length

- A Gantt chart benefits from both position and length.

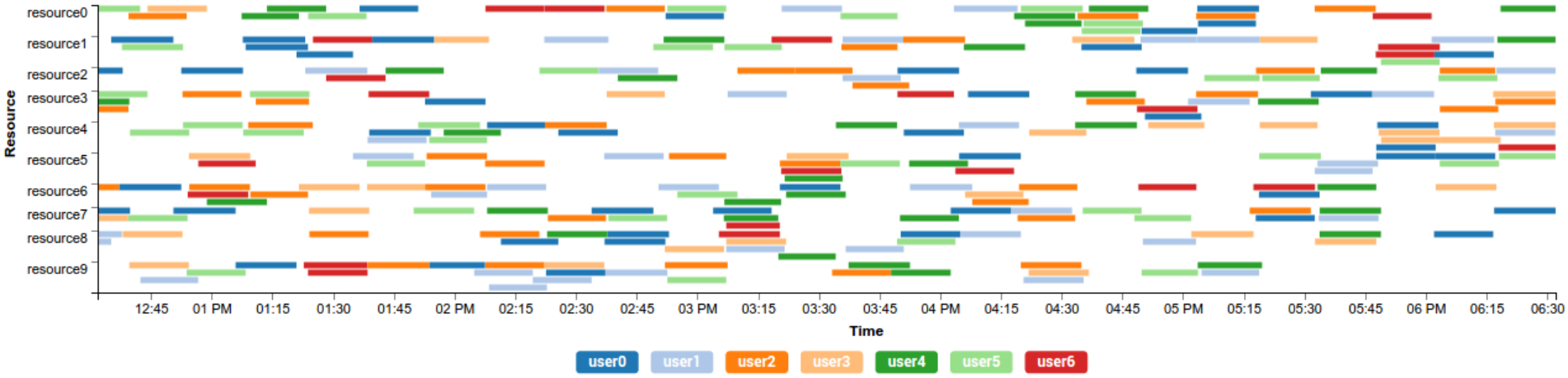


Gantt charts: Examples



A Gantt chart showing the tasks of a project (usually order in time), how long each task takes, and whether a task is fully completed.

Gantt charts: Examples

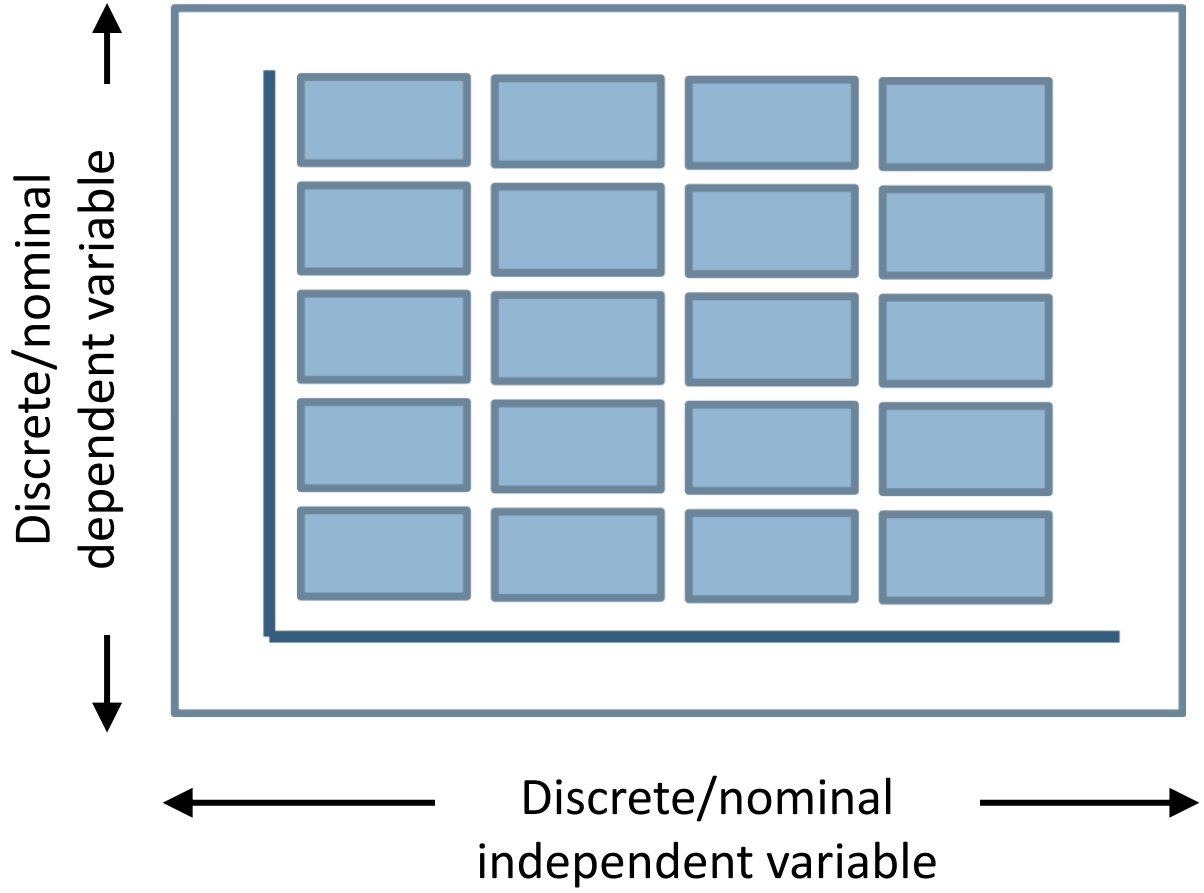


A Gantt chart can be quite complicated with several entities and detailed timeline.

Tables

Position

- A table benefits from position only.



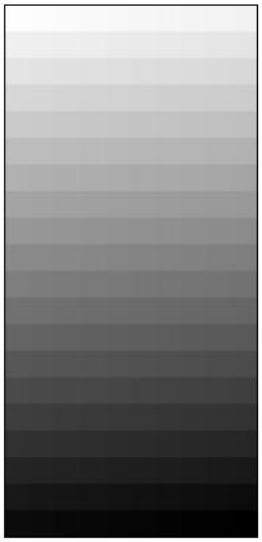
What chart to use?

Dependent	Quantitative continuous	Bar	Line
	Quantitative discrete	Bar	Bar
Independent	Quantitative continuous	Gantt	Scatter
	Quantitative discrete	Scatter	Gantt
		Nominal or Quantitative discrete	Quantitative continous
		Independent	

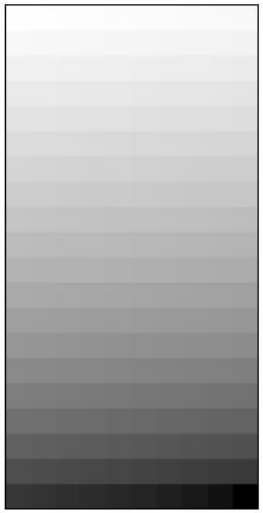
Advanced charts

Pixel-oriented visualization

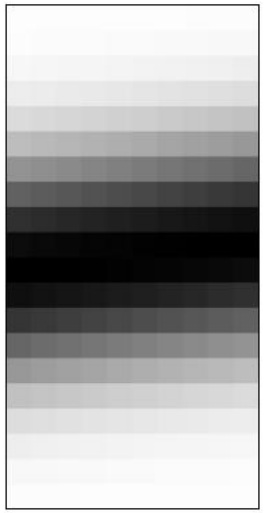
- Consider a dataset of m dimensions and n examples.
- Create a window of n pixels for each dimension
- For each example, the i^{th} value is mapped to the matching pixel in the i^{th} window.
- The colors of the pixels reflect the value range.



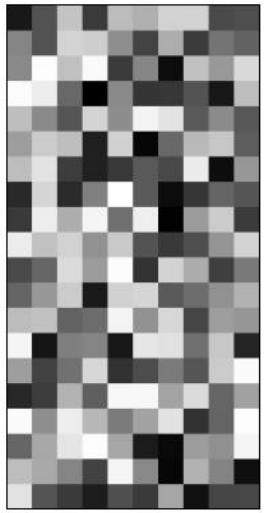
income



credit limit



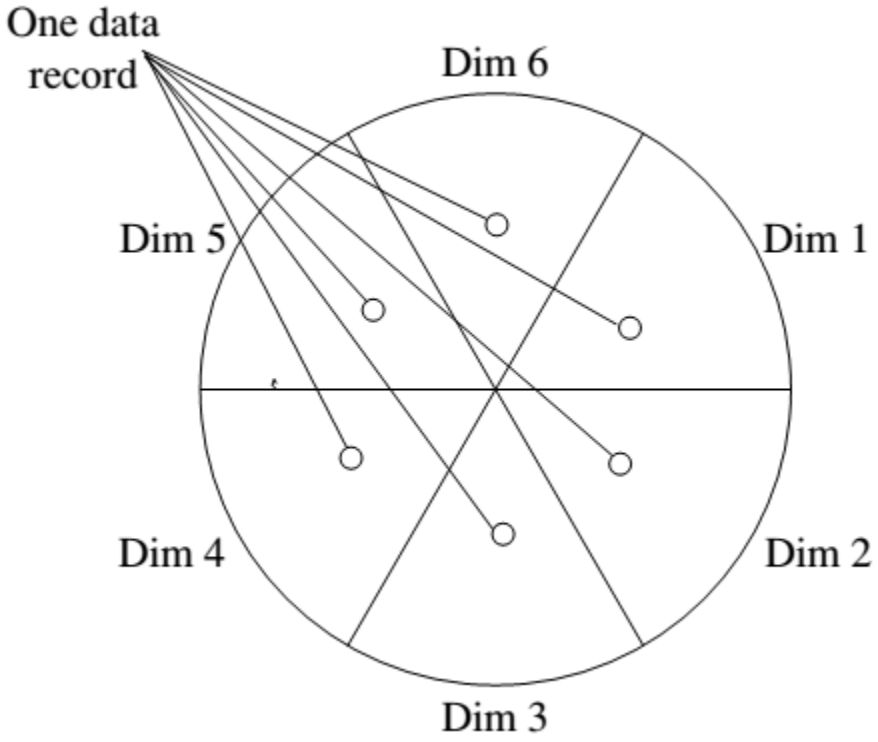
transaction volume



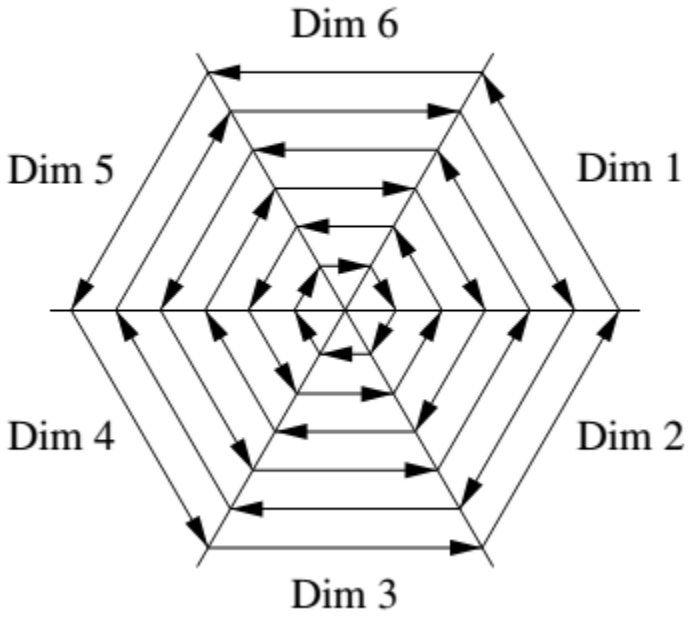
age

Laying out pixels in circle segments

- Space-filling is often done in a circle segment to save space and show the connections among multiple dimensions.

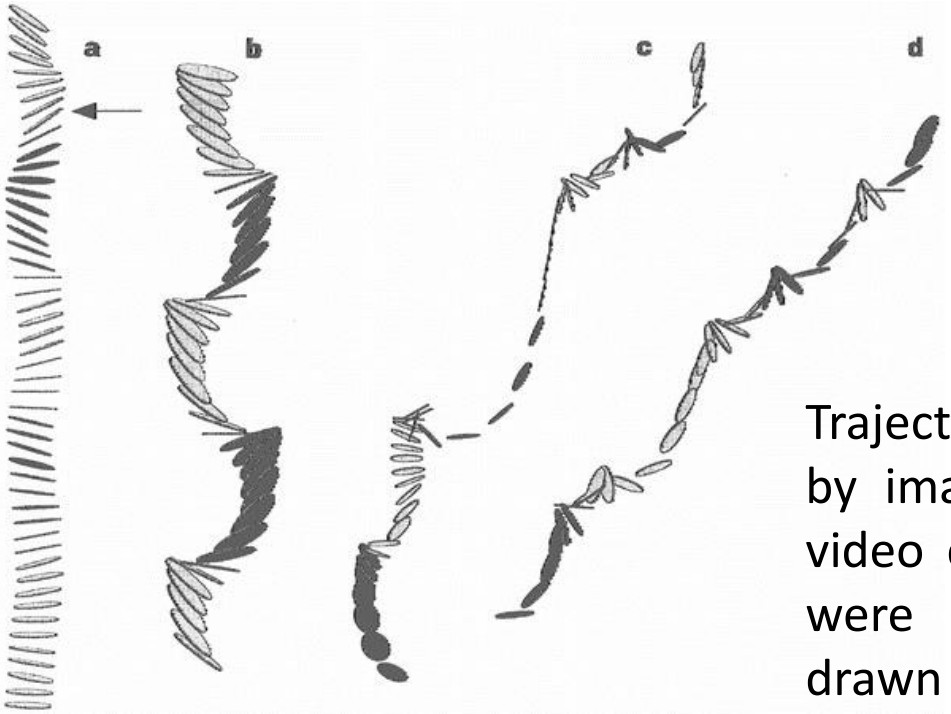


Representing a data record in circle segment



Laying out pixels in circle segment

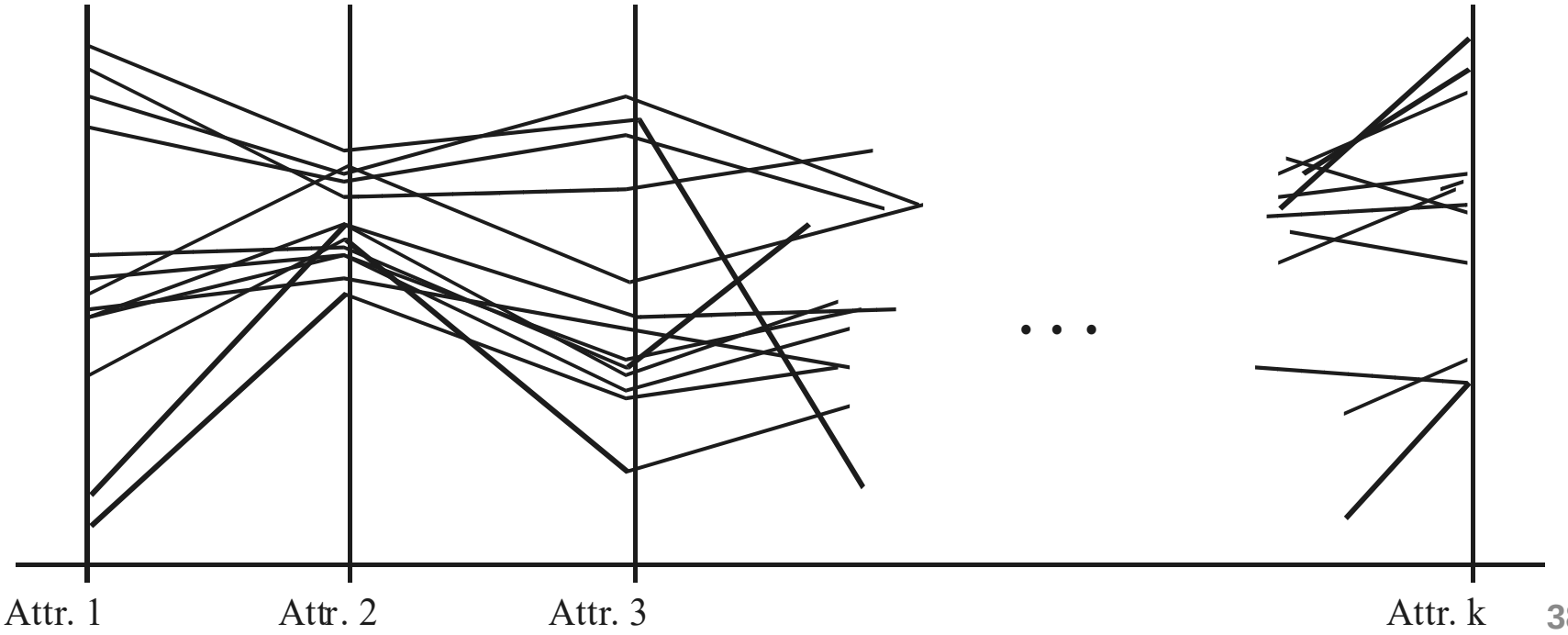
Direct visualization



Trajectories of falling disks, obtained by imaging from the side using a video camera; the digitized images were measured, and computer-drawn images are shown.

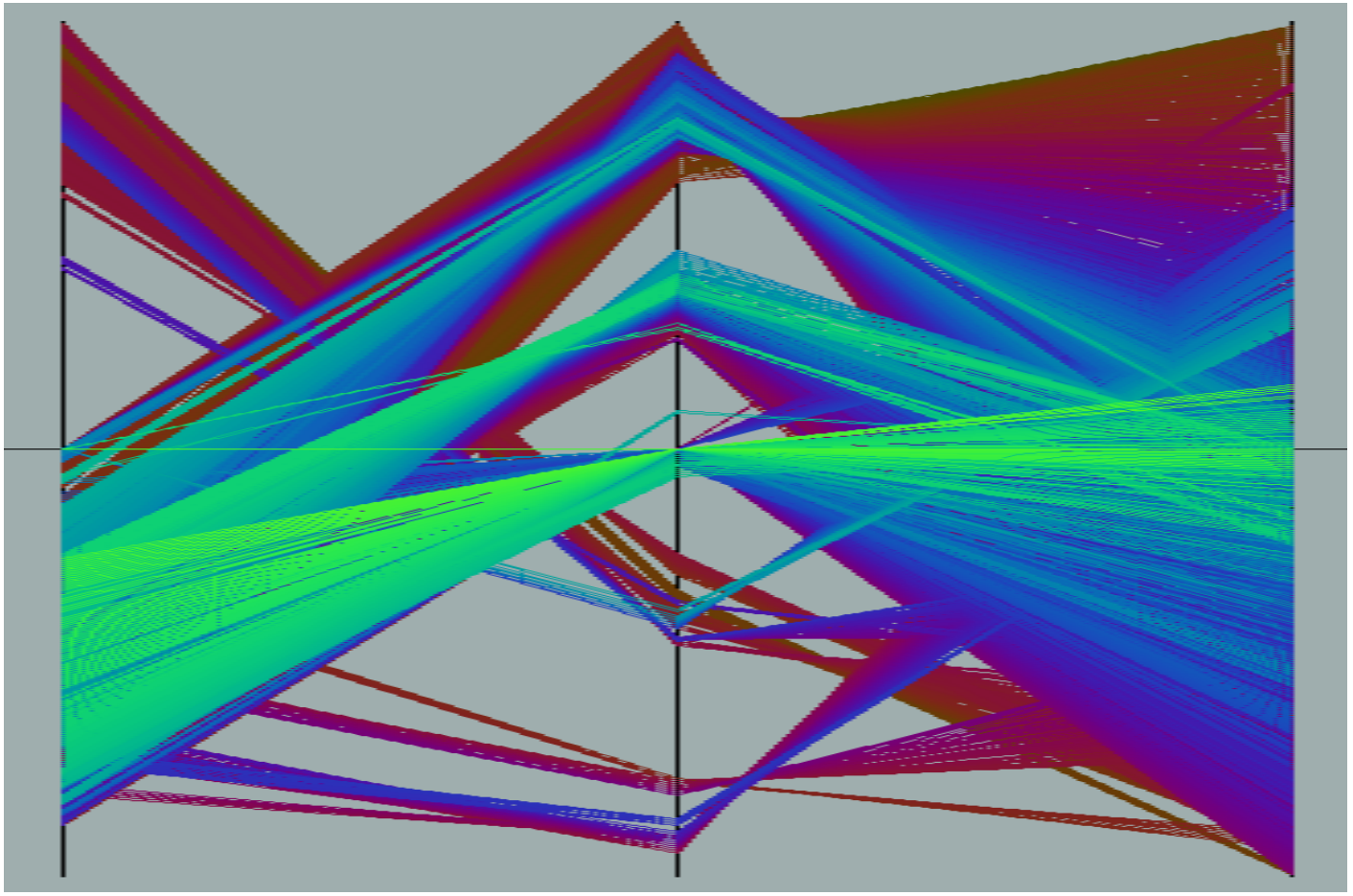
Parallel coordinates

- n equidistant axes, one for each dimension, parallel to one of the display axes
  - Each axis is scaled to the domain of the corresponding attribute
  - A data record is represented by a polygonal line that intersects each axis at the point corresponding to the associated dimension value.



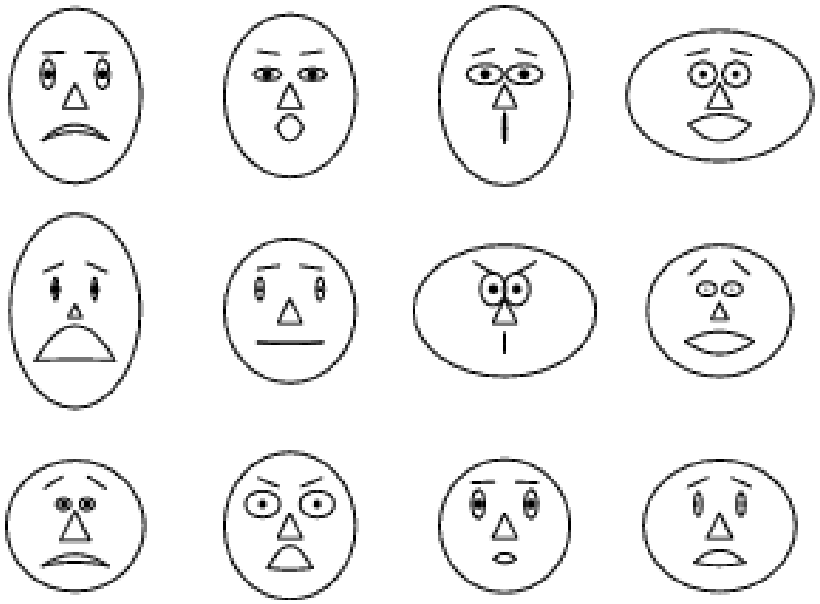
Parallel coordinates

- Parallel coordinates suffer visual clutter and overlap, often reducing the readability and making the patterns hard to find.



Icon-based visualization

- Chernoff Face: display variables on a two-dimensional surface, e.g., let x be eyebrow slant, y be eye size, z be nose length, etc.

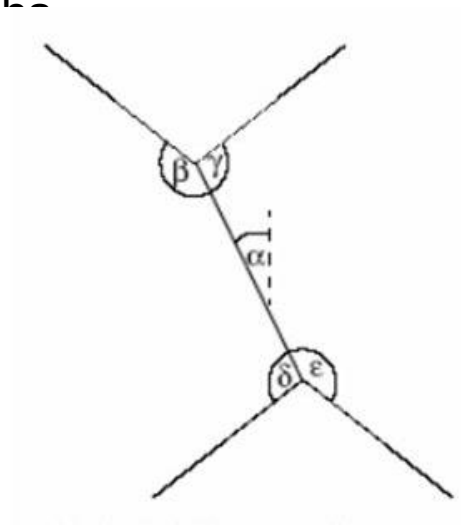


Faces produced using 10 characteristics--head eccentricity, eye size, eye spacing, eye eccentricity, pupil size, eyebrow slant, nose size, mouth shape, mouth size, and mouth opening): Each assigned one of 10 possible values, generated using Mathematica (S. Dickson)

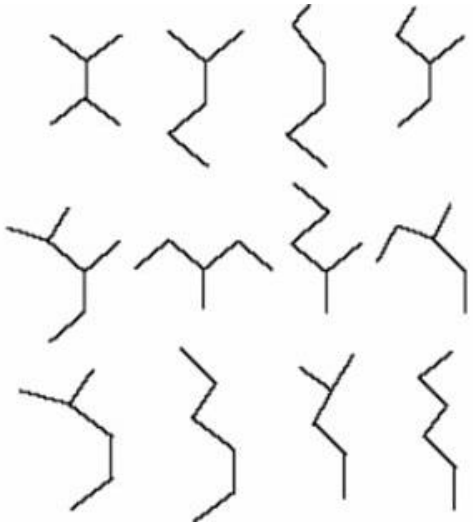
Reference: Gonick, L. and Smith, W. *The Cartoon Guide to Statistics.* New York: Harper Perennial, p. 212, 1993

Icon-based visualization

- Stick figures: Each example is a five-piece stick figure icon with one body and four limbs.
  - Two variables are mapped to X-Y axes
  - The remaining attributes are mapped to the angles and/or length of the limbs

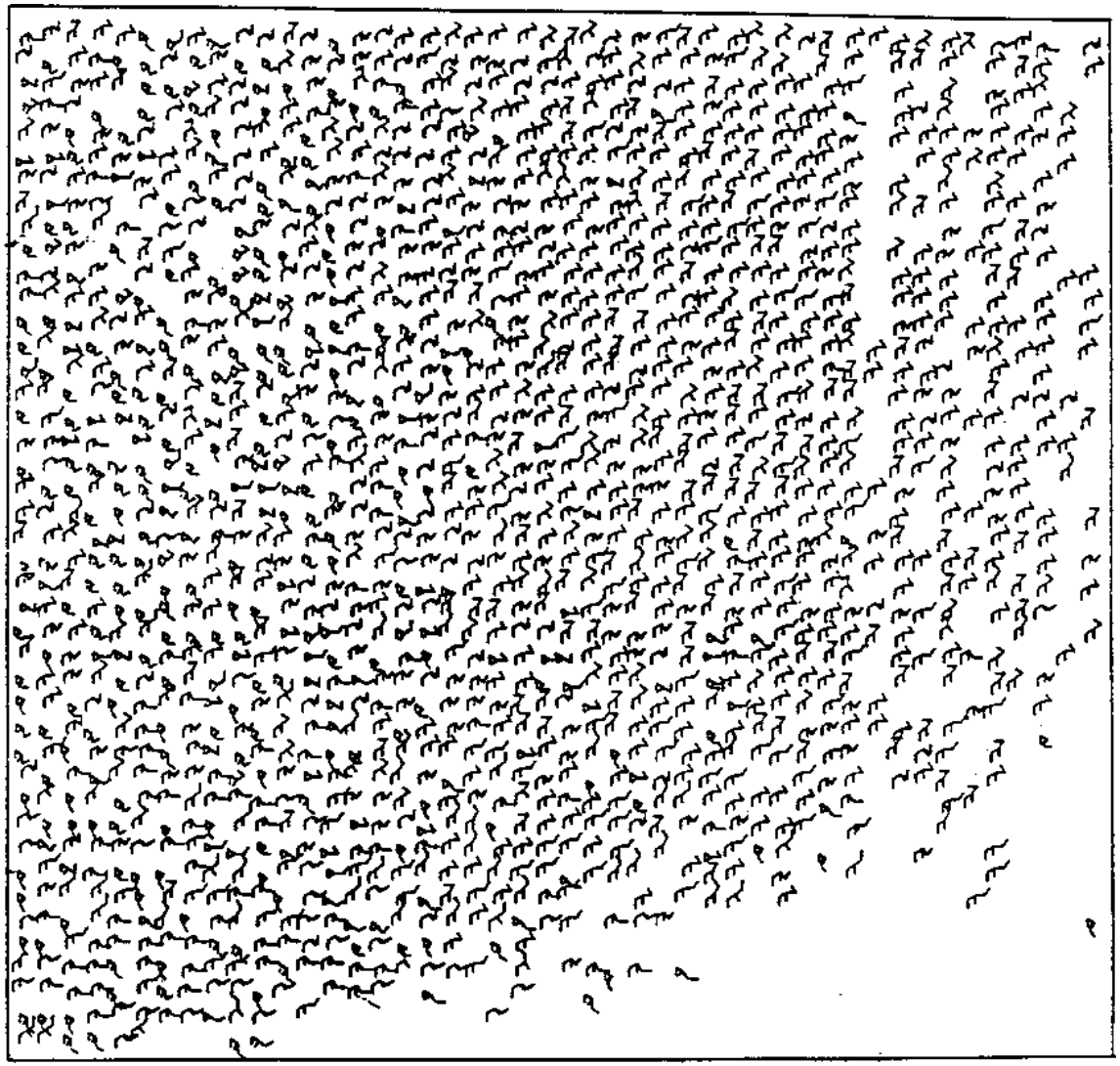


A stick figure icon



A family of stick figures

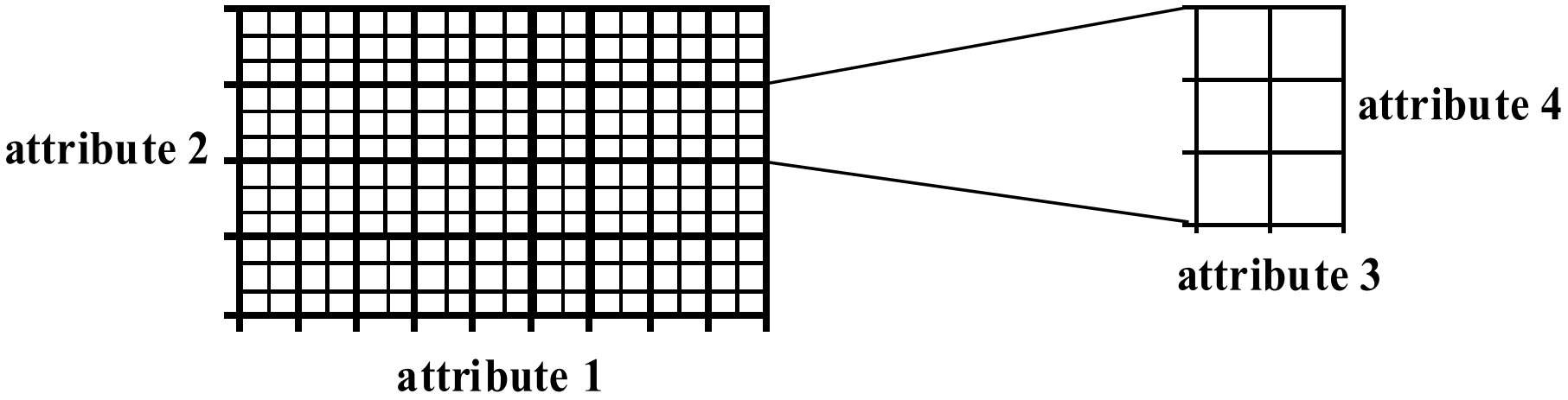
Icon-based visualization



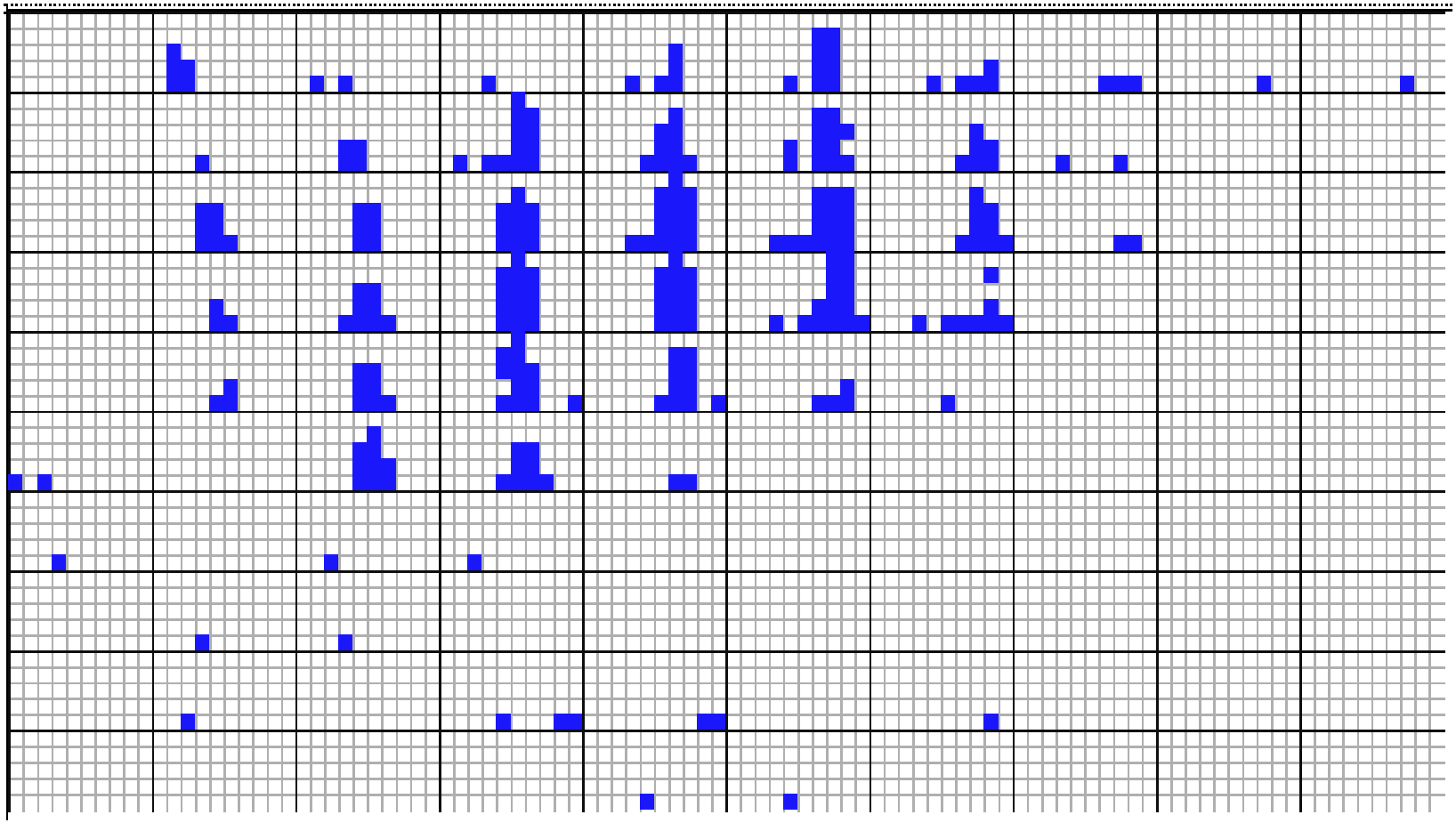
US census data showing age, income, gender, education, etc.

Dimensional stacking

- Partition the n-dimensional attribute space in 2-D subspaces, which are 'stacked' into each other
  - The important attributes should be used on the outer levels.
  - Adequate for data with ordinal attributes of low cardinality, yet difficult to display more than nine dimensions



Dimensional stacking: An example

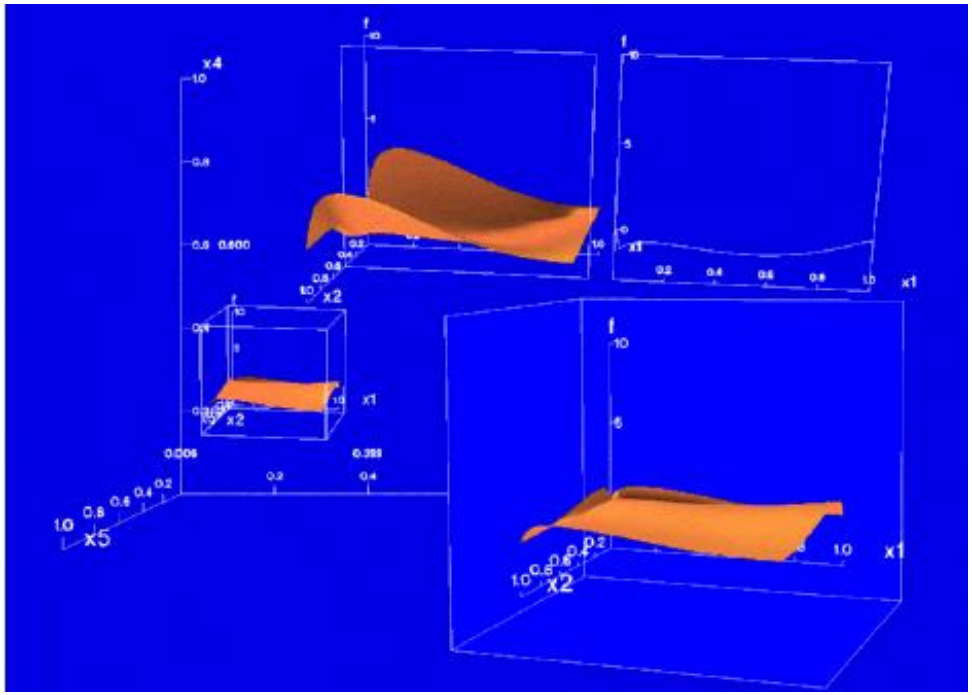


Used by permission of M. Ward, Worcester Polytechnic Institute

Worlds-within-Worlds

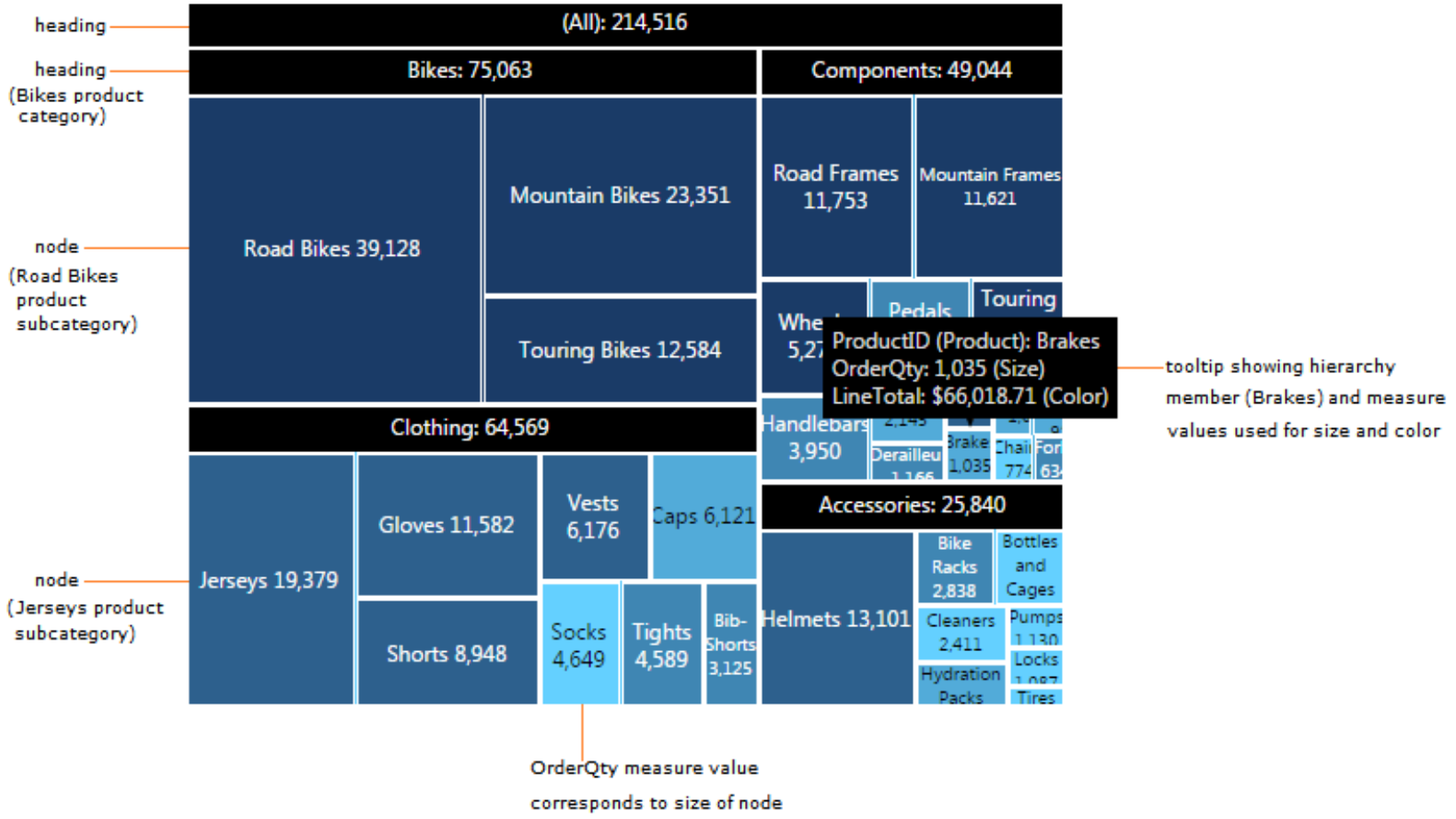
- Show an attribute F changes with respect to the other dimensions in a multi-dimensional dataset.

  - Consider a 6D dataset, where the dimensions are $F, X_1, \ldots, X_5$, observe how F changes with respect to the other dimensions
  - Fix the values of dimensions X_3, X_4, X_5 to some values, say, c_3, c_4, c_5.
  - Then visualize F, X_1, X_2 using a 3D plot whose origin locates at the point (c_3, c_4, c_5)



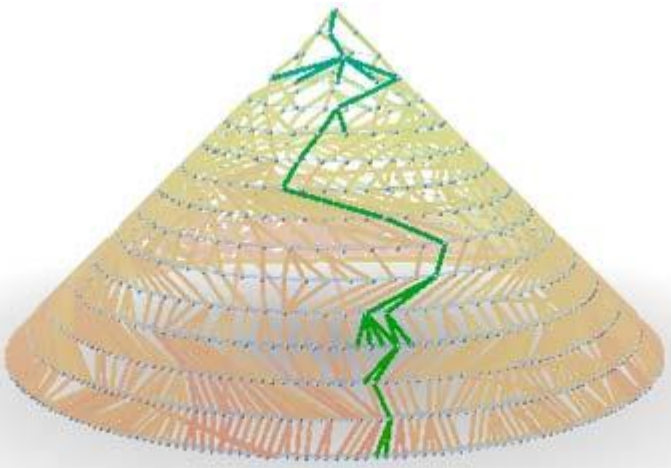
Treemap

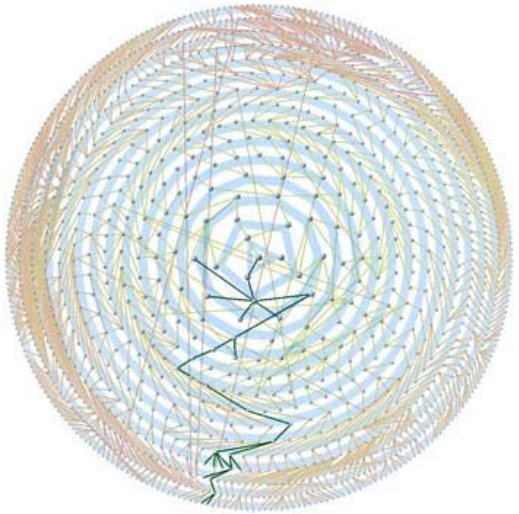
- Hierarchically partition the screen into regions depending on the attribute values



3D Cone Tree

- Typically contain more information and hierarchical data than tree diagrams due to its 3D nature
  - Work well for up to a thousand nodes or so, yet suffering overlaps when projected to 2D

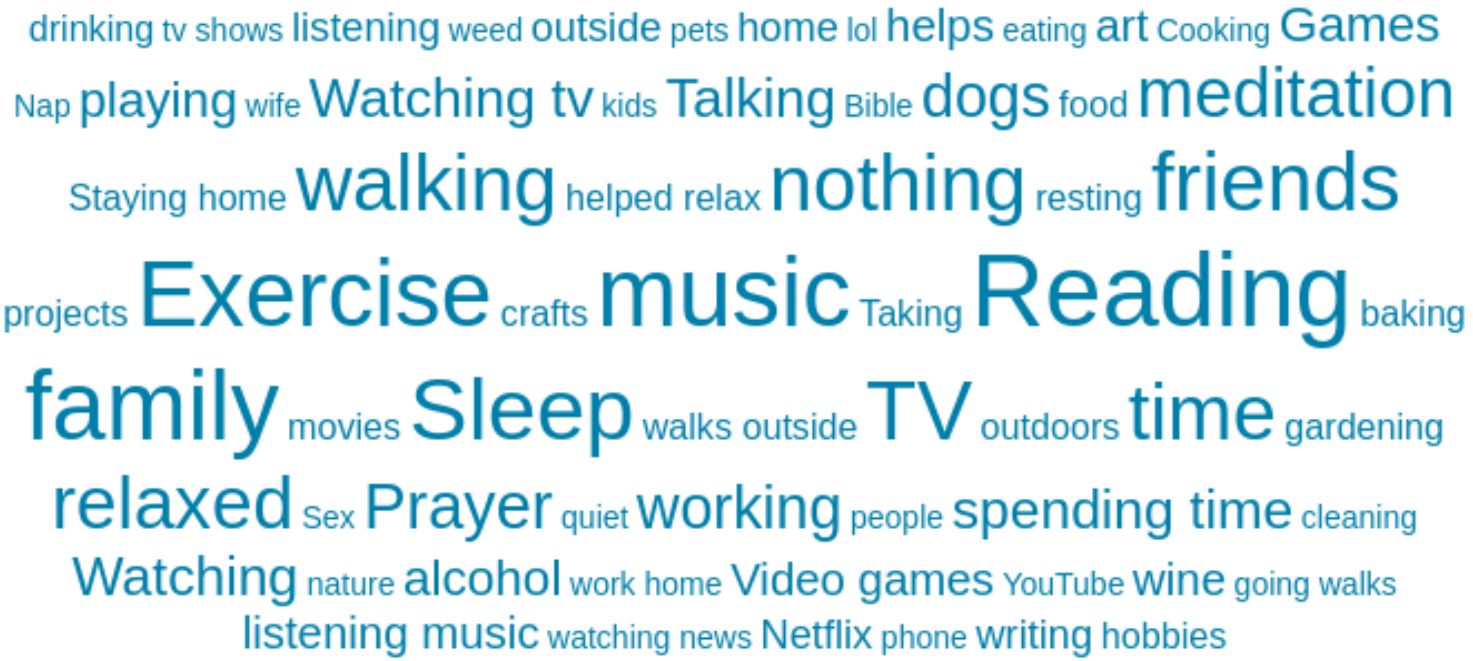




Visualize a social network dataset that models the way an infection spreads from one person to the next (Nadeau Software Consulting)

Tag cloud

- Visualize statistics of user-generated tags, whose levels of importance are indicated by font sizes or colors

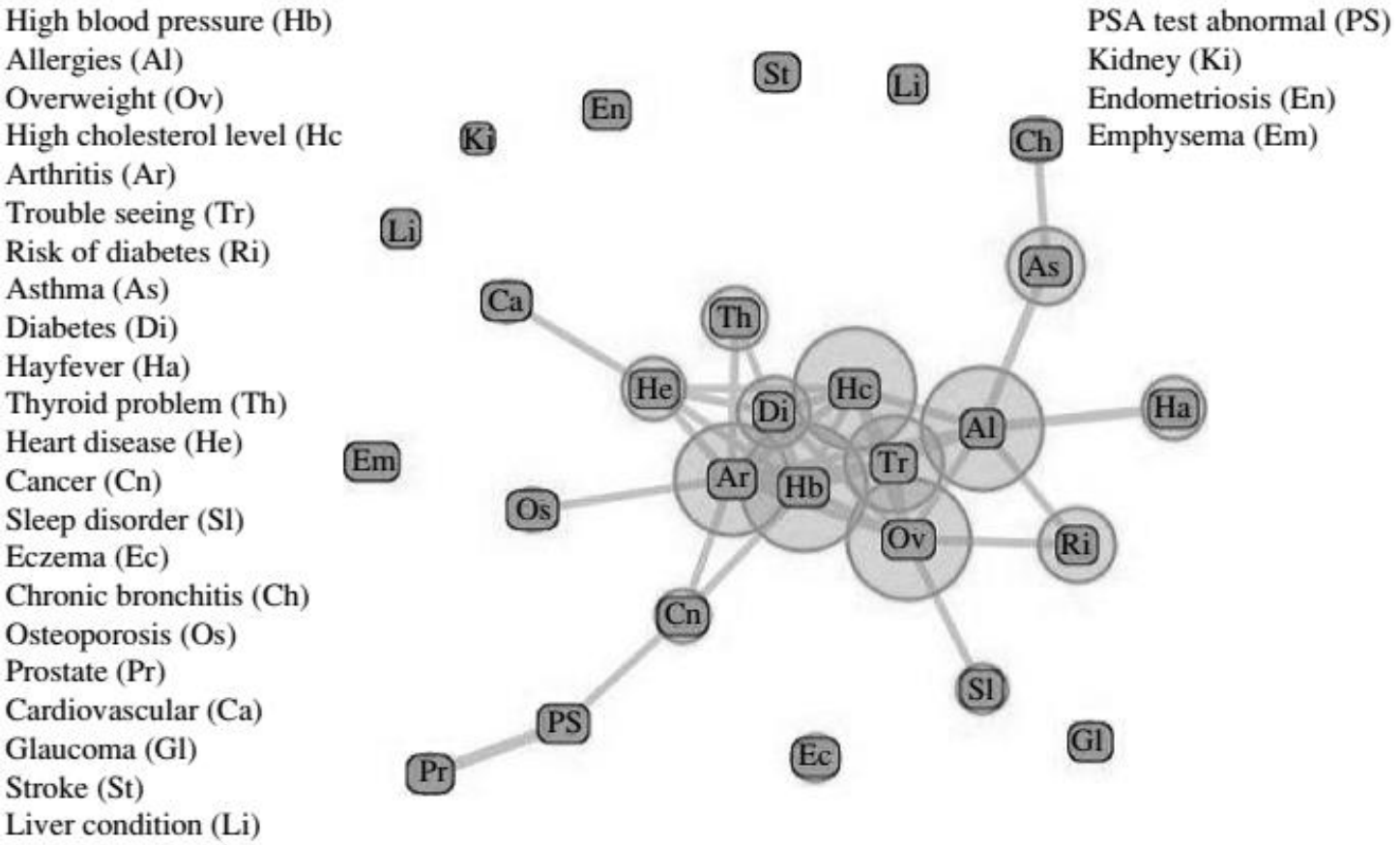


Social distancing study: How are people spending their time?

Image credit

Influence graph

- :Visualize the correlations between objects



Disease influence graph of people at least 20 years old in the NHANES dataset

References

- Jiawei Han, Micheline Kamber, and Jian Pei, 2011. Data Mining: Concepts and Techniques (3rd ed.). Morgan Kaufmann Publishers Inc. Chapter 2.
- John C. Hart. Data Visualization. University of Illinois at Urbana-Champaign
- Common Pie Chart Misuses (and How to Fix Them) (link)
- Images are obtained from the above materials and Google


...the end.