



Algorithms. Matching

Part II. Preference model.

B9 - Algorithms Matching

M-ALG-102

Overview of day 2

Classic visualization methods

- Classification of data types
- Classic methods

Intermediate methods

- Parallel coordinate plot
- Miscellaneous
- Hierarchies

Visualization platforms

challenges

...

- └ Classic visualization methods
- └ Classification of data types

Classification of data

Let us summarize the different types of data that can be represented:

...

└ Classic visualization methods

 └ Classification of data types

Classification of data

Let us summarize the different types of data that can be represented:

- ▶ univariate data (time series)

...

- └ Classic visualization methods
- └ Classification of data types

Classification of data

Let us summarize the different types of data that can be represented:

- ▶ univariate data (time series)
- ▶ two-dimensional data (geographical coordinates)

...

- └ Classic visualization methods
- └ Classification of data types

Classification of data

Let us summarize the different types of data that can be represented:

- ▶ univariate data (time series)
- ▶ two-dimensional data (geographical coordinates)
- ▶ multidimensional data (experiments, climatic data)

...

- └ Classic visualization methods
- └ Classification of data types

Classification of data

Let us summarize the different types of data that can be represented:

- ▶ univariate data (time series)
- ▶ two-dimensional data (geographical coordinates)
- ▶ multidimensional data (experiments, climatic data)
- ▶ text (newspaper, web pages)

Classification of data

Let us summarize the different types of data that can be represented:

- ▶ univariate data (time series)
- ▶ two-dimensional data (geographical coordinates)
- ▶ multidimensional data (experiments, climatic data)
- ▶ text (newspaper, web pages)
- ▶ hierarchical and relational data (web links, hierarchical relationships in an organization)

...

- └ Classic visualization methods
 - └ Classic methods

Classic visualization methods

We will now go through some visualization methods for data and discuss their assets and drawbacks, and when they are relevant.

...

- └ Classic visualization methods
 - └ Classic methods

Pie chart

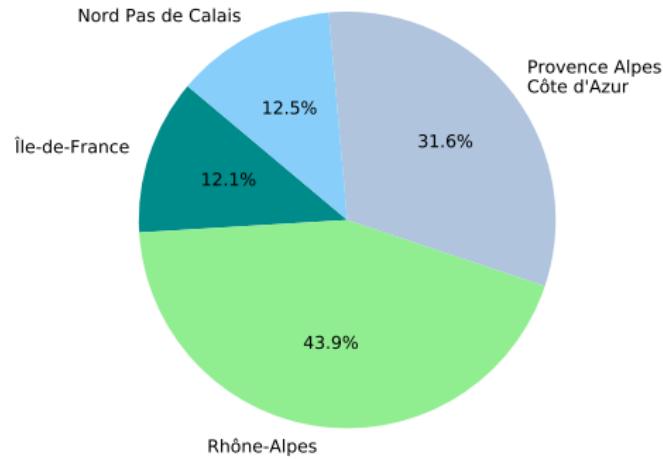


Figure: Areas of four region in France

...

- Classic visualization methods

- Classic methods

Pie chart

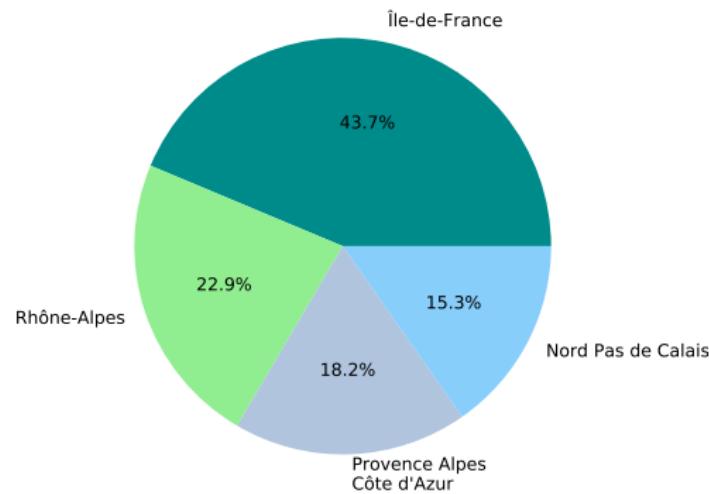


Figure: Population of four region in France

...

- └ Classic visualization methods
 - └ Classic methods

Pie chart

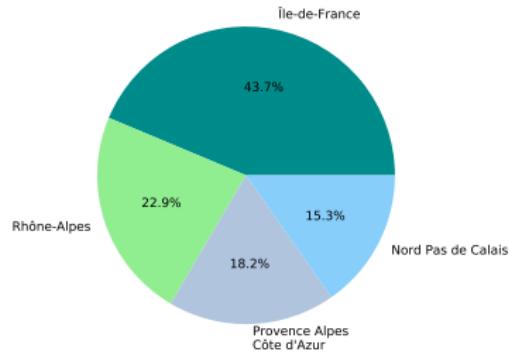


Figure: Population of four region in France

You have a pie chart template in **simple_graphs/pie_graph.py**

...

- └ Classic visualization methods
 - └ Classic methods

Line graph (graphique linaire)

We studied the line graph yesterday.

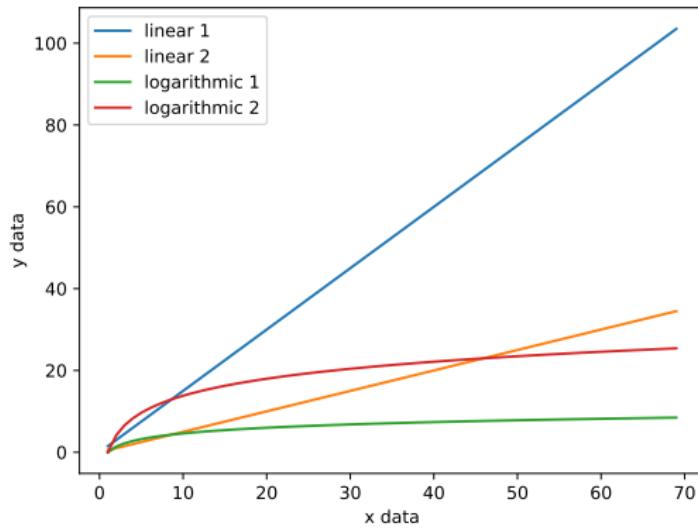


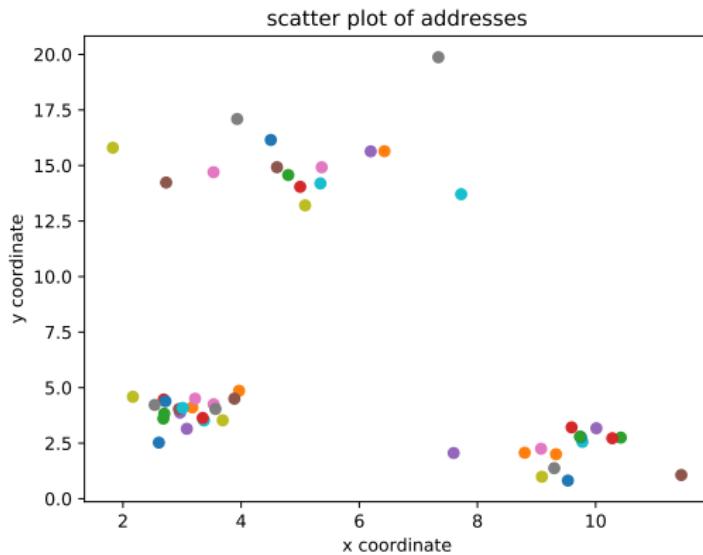
Figure: Logarithmic function

...

- └ Classic visualization methods
 - └ Classic methods

Scatter plot (nuage de points)

We also studied the scatter plot yesterday.



...

- └ Classic visualization methods
 - └ Classic methods

Scatter plot (nuage de points)

However, we need to discuss the case of a multivariate random variable.

For instance the classic iris dataset.

...

Scatter matrix

- ▶ Let us generalize the idea of a scatter plot.

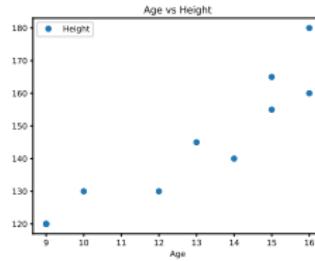


Figure: Exemple scatter plot

- ▶ Given a dataset, we can build all the possible scatter plots and store them in a matrix, called the **scatter matrix**.

Scatter matrix

- ▶ Let us generalize the idea of a scatter plot.

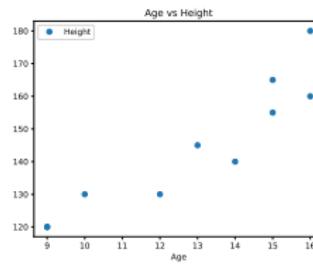


Figure: Exemple scatter plot

- ▶ Given a dataset, we can build all the possible scatter plots and store them in a matrix, called the **scatter matrix**.
- ▶ What will happen on the diagonal of the matrix ?

Scatter matrix

- ▶ Let us generalize the idea of a scatter plot.
- ▶ Given a dataset, we can build all the possible scatter plots and store them in a matrix, called the **scatter matrix**.
- ▶ What will happen on the diagonal of the matrix ?

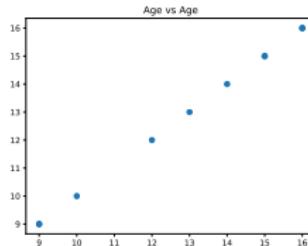


Figure: Variable plotted against itsself : All the points are on the $y = x$ line

...

Scatter matrix

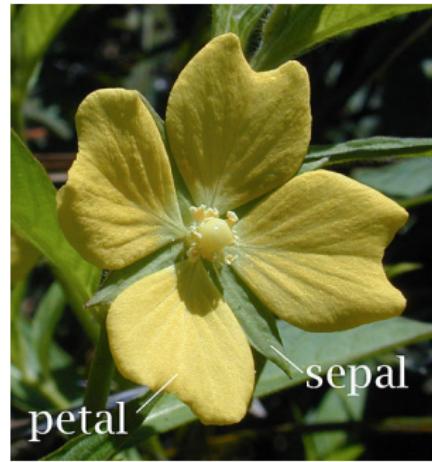
- ▶ Let us generalize the idea of a scatter plot.
- ▶ Given a dataset, we can build all the possible scatter plots and store them in a matrix, called the **scatter matrix**.
- ▶ On the diagonal, one can plot histograms or the density probability
- ▶ The scatter plot can be a good way to start analyzing a dataset when we don't know which variables could be correlated

...

- └ Classic visualization methods
 - └ Classic methods

Iris dataset

- ▶ 150 samples of iris flower
- ▶ 3 species
- ▶ 4 attributes : petal width and length, sepal width and length



...

- └ Classic visualization methods
 - └ Classic methods

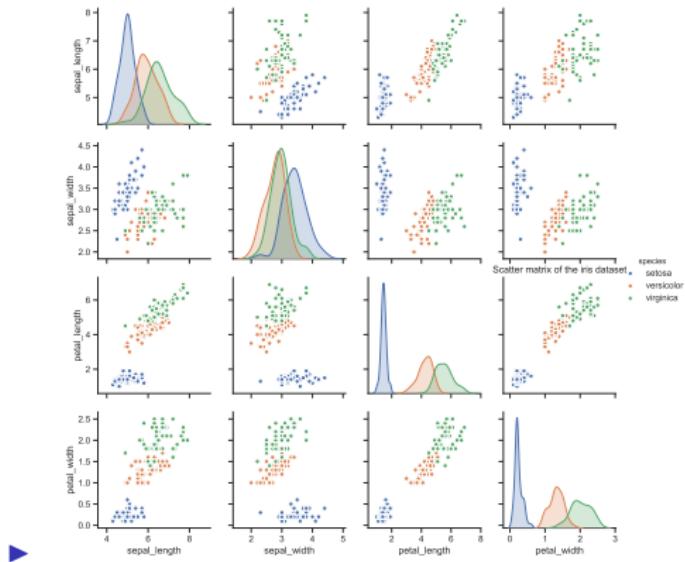


Figure: Plotted in `iris_scatter_matrix.py`. Is there a variable that can discriminate between the species ?

...

- └ Classic visualization methods
 - └ Classic methods

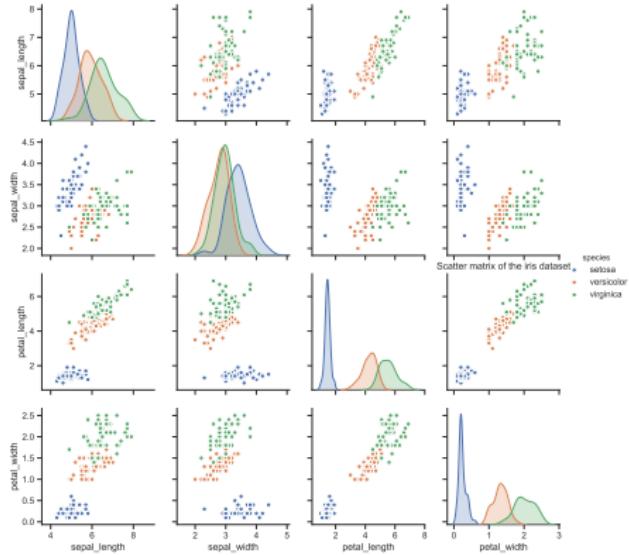


Figure: It seems that **petal width** is a parameter that separates the three species. On the contrary, **sepal width** is not able to do so.

...

- └ Classic visualization methods
 - └ Classic methods

Scatter plot

The scatter plot is considered to be relevant to represent around 10000 points at most.

...

- └ Classic visualization methods
 - └ Classic methods

Scatter plot

The scatter plot is considered to be relevant to represent around 10000 points at most.

If we have a large dataset, we must select a fraction of the points to represent.

...

- └ Classic visualization methods
 - └ Classic methods

Scatter plot

The scatter plot is considered to be relevant to represent around 10000 points at most.

If we have a large dataset, we must select a fraction of the points to represent.

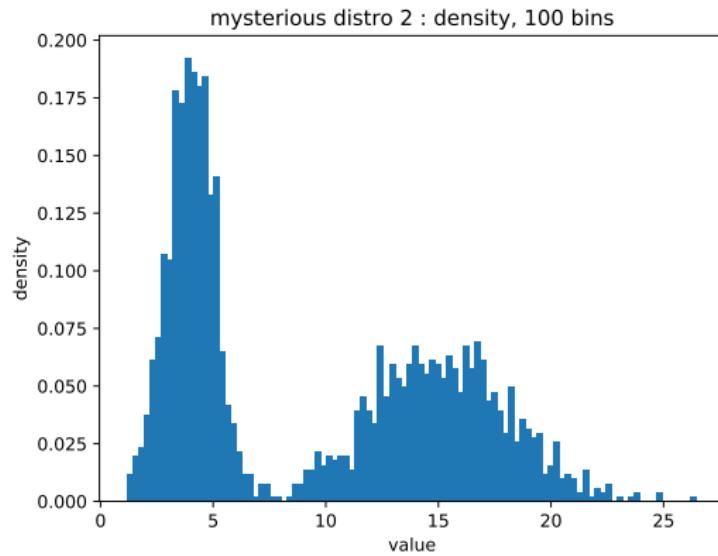
Exercice 1 : Number of possible plots

How does the number of possible scatter plots depend on the number of datapoints ? and on the number of dimensions ?

...

- └ Classic visualization methods
 - └ Classic methods

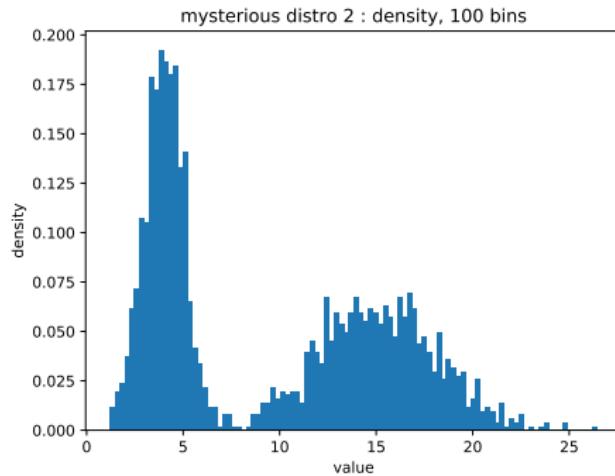
Histogram



...

- └ Classic visualization methods
 - └ Classic methods

Histogram



As in the case of the pie chart, this histogram can only represent **one dimension**.

...

└ Intermediate methods

 └ Parallel coordinate plot

Parallel coordinate plot

The parallel coordinate plot is a tool used to show multidimensional data.

...

└ Intermediate methods

 └ Parallel coordinate plot

Parallel coordinate plot

The parallel coordinate plot is a tool used to show multidimensional data.

Each dimension will correspond to one column.

Each datapoint is represented by a line between columns.

...

└ Intermediate methods

 └ Parallel coordinate plot

Parallel coordinate plot

Exercice 2 : Showing the plot

cd parallel_coordinate/ and use **parallel_coordinate.py** in order to to a parallel coordinate plot of the data from the file **data.npy**.

...

└ Intermediate methods

└ Parallel coordinate plot

Parallel coordinate plot

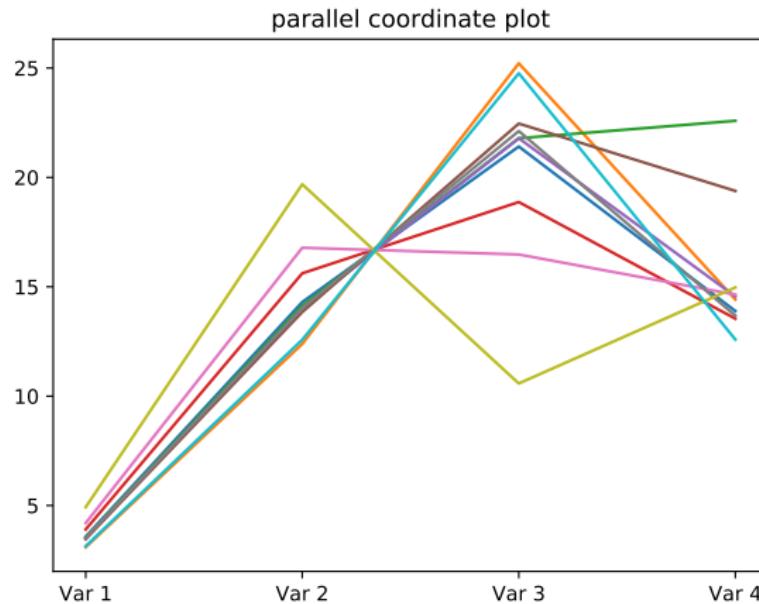


Figure: What are the relationships between the variables ?

...

└ Intermediate methods

 └ Parallel coordinate plot

Parallel coordinate plot

Assets:

- ▶ Can help identify patterns or correlations in the data.

Drawbacks:

- ▶ Can be overcluttered

...

- └ Intermediate methods
 - └ Parallel coordinate plot

Parallel coordinate plot

We will use `plotly` to do "brushing". (`pip install plotly`)

<https://plot.ly/>

Use the file **parallel_coordinate_plotly.py** in order to show an interactive plot of the iris dataset.

...

- Intermediate methods

- Parallel coordinate plot

Practical application

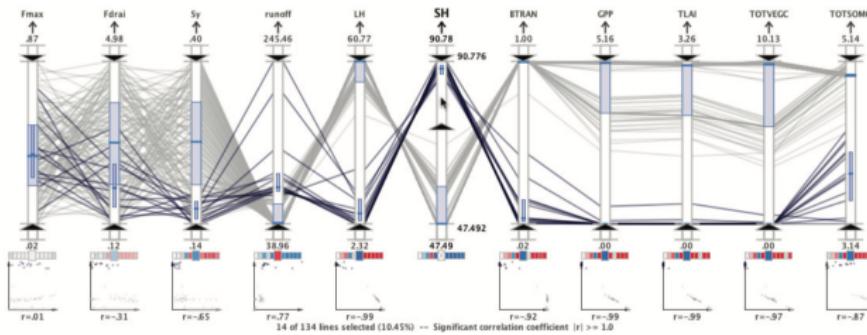


Figure: [Steed et al., 2014]

Radar plot

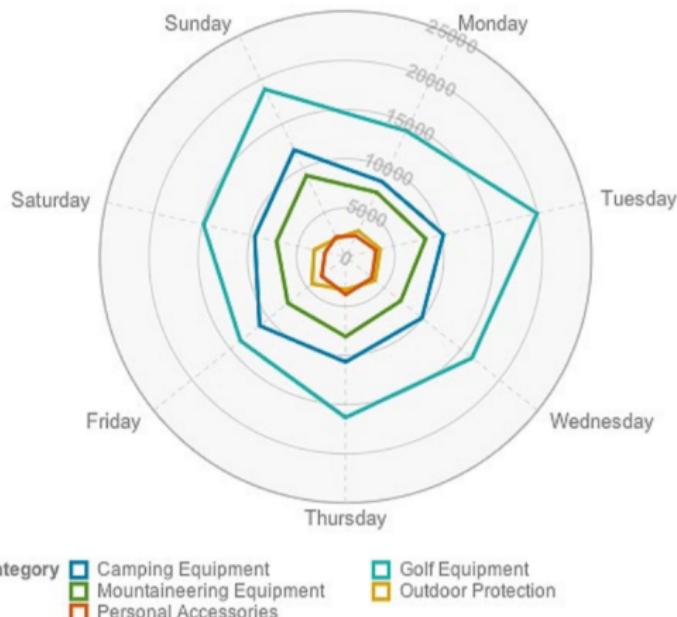


Figure: [IBM, 2013]

...

- └ Intermediate methods
- └ Miscellaneous

Heat map

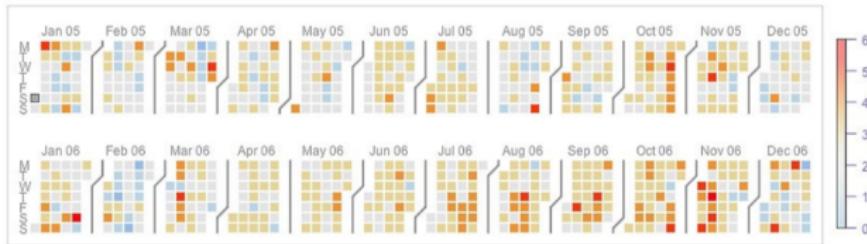


Figure 3: Calendar heat map example that shows two years of changes (in percentages) in customer web orders by year (row), month (column), day of week (sub row), week (sub column) and day.

Figure: [IBM, 2013]

...

- └ Intermediate methods
- └ Miscellaneous

Networks visualization

Graphs are useful in order to represent relationships between datapoints.

...

- Intermediate methods
- Miscellaneous

Networks

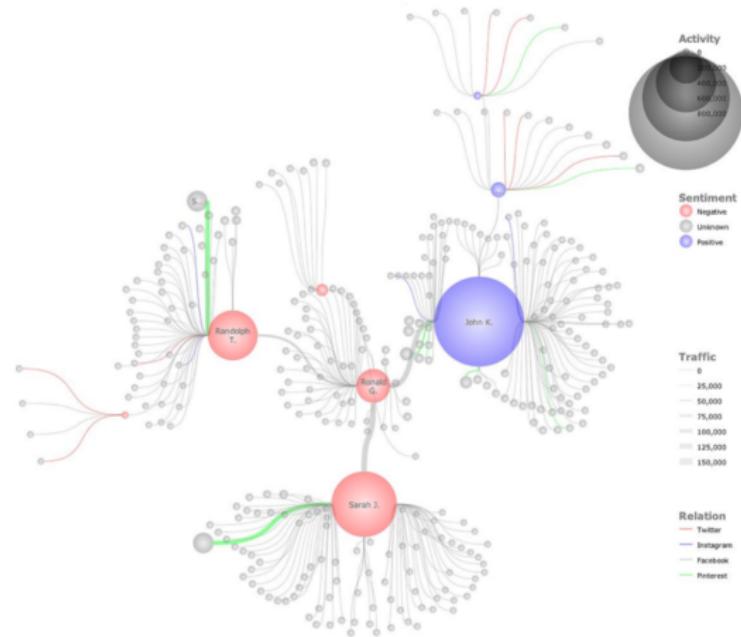


Figure: [IBM, 2013]

Hierarchies

The hierarchical clustering we performed yesterday can also be performed with libraries such as **sklearn** and **scipy**.

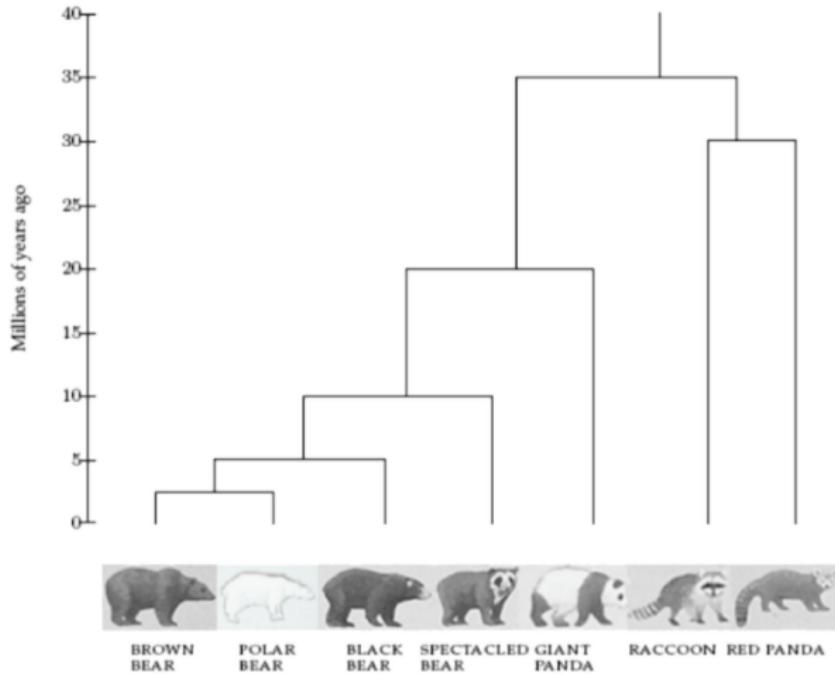
https://scikit-learn.org/stable/auto_examples/cluster/plot_agglomerative_dendrogram.html

Hierarchies

<https://docs.scipy.org/doc/scipy/reference/cluster.hierarchy.html>

...

Example application of hierarchical clustering



...

Treemaps

A **Treemap** is another representation of hierarchical data in the two-dimensional space.

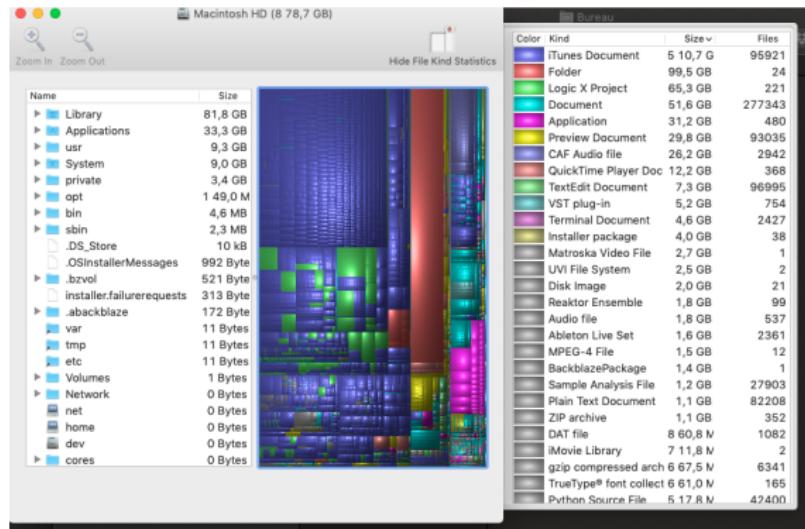


Figure: Disk Inventory X <http://www.derlien.com/>

...

Treemaps

A **Treemap** is another representation of hierarchical data in the two-dimensional space.

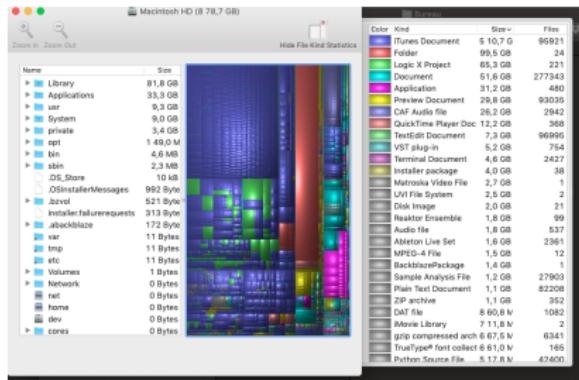


Figure: Disk Inventory X <http://www.derlien.com/>

The size of a rectangle corresponds to its size.

Building a tree map

`cd treemap/` and use **build_treemap.py** in order to build a treemap of your Desktop folder.

...

Building a tree map

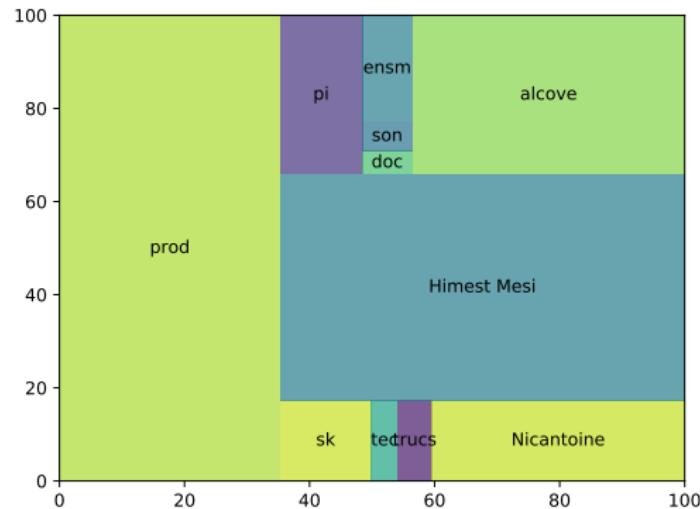


Figure: Treemap of desktop computer (desktop folder)

...

Building a tree map

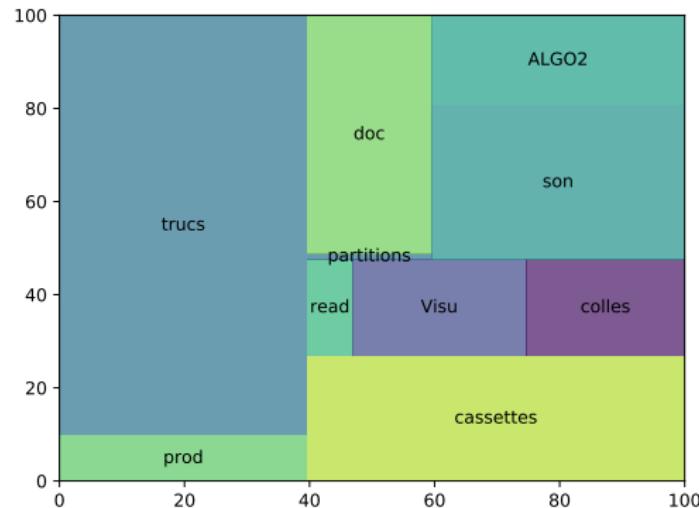
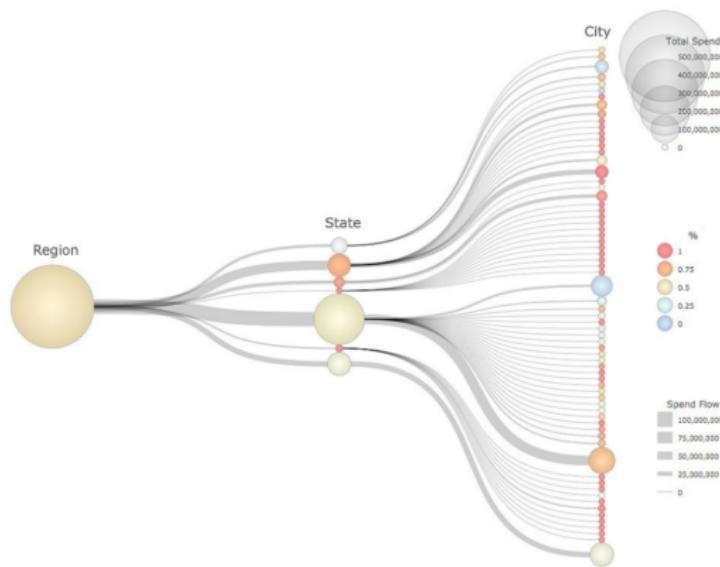


Figure: Treemap of laptop (desktop folder)

...

- Intermediate methods
 - Hierarchies

Hierarchy



Remak on visualization

The Gestalt Principles give guidelines for a good visualization.

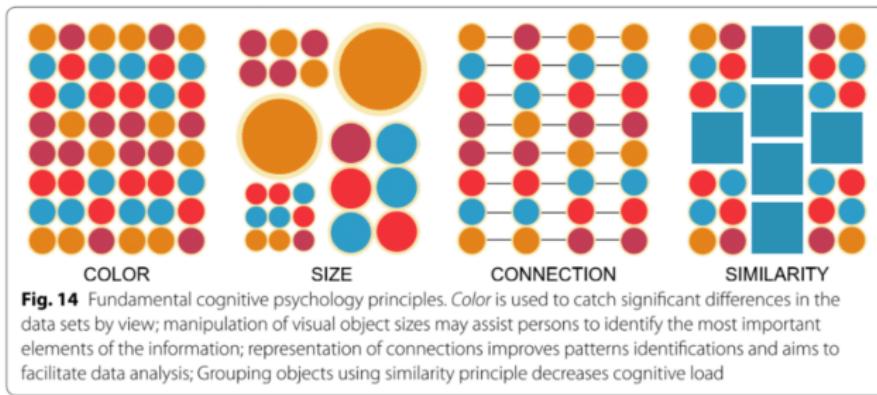


Figure: [Olshannikova et al., 2016]

...

└ Visualization platforms

Visualization platforms

Here are some useful links for data visualization and processing.

...

└ Visualization platforms

Links

<https://fcpython.com/>

...

└ Visualization platforms

Links

<https://www.dataiku.com/>

...

└ Visualization platforms

Links

<https://www.tylertech.com/products/socrata>

...

└ Visualization platforms

Links

<https://dev.socrata.com/>

...

└ Visualization platforms

Links

<https://cytoscape.org/>

...

└ Visualization platforms

Links

<https://www.tableau.com/>

...

└ Visualization platforms

Links

<http://vis.stanford.edu/wrangler/>

...

└ Visualization platforms

Links

<https://noduslabs.com/infranodus/>

Challenges

Some people argue that augmented reality or virtual reality may help us to gain more insight from data visualization [Olshannikova et al., 2016].

...

└ challenges

Project description

Description of the project.

References

-  IBM (2013).
Analytics : The real-world use of big data.
-  Olshannikova, E., Ometov, A., Koucheryavy, Y., and Olsson, T. (2016).
Chapter 4 VISUALIZING BIG DATA.
Number October.
-  Steed, C. A., Shipman, G., Thornton, P., Ricciuto, D., Erickson, D., and Branstetter, M. (2014).
Practical Application of Parallel Coordinates for Climate Model Analysis.
Procedia - Procedia Computer Science, 9(December 2012):877–886.