# Chapter 1 Introduction to Visual Analytics by an example:

The main purpose of the data analyst is to create prediction models.

Data visualisation plays a crucial role each part of this process.

This is indeed the most efficient way to convey non trivial information to human cognition

## 1 What is visual analytics? p3

Computational results might not enable to fully grasp a data related topic.

Visualisation supports computation methods by allowing the analyst to take decisions at each step of it’s workflow.

Eg:

Interactive visual interfaces enable to see and interpret prediction models.

It offers the possibility to how parameters tweaking or different use of data subsets alters it’s outputs

When a computational method has been chosen to identify groups, those data subsets can be prompt by a visualisation technique to see If the split is relevant or if the differentiation criteria should be reconsidered (eg: days in a week).

Finally, visual analytics is before being the science of careful and effective use of computational technics, the science of human **analytical reasoning**.

## 1.2 A motivating example: p5

Investigating an epidemic outbreak

This example outlines how visual analytics supports data science workflow.

Data come from the **IEEE VAST Challenge 2011** and is synthetic.

### 1.2.1 Data task and description

Vastopolis: city, population ~ 2 milion, fear of a possible epidemic

Two data set:

1 = messages on a blog with gps localisation

2 = city map informations

3 questions:

.Identify approximately where outbreak started

.Present a hypothesis on how the infection is being transmitted

.Is the outbreak contained

### 1.3 Discussion: How visual analytics has helped us

Well-conceived visualisations are necessary to understand large volume of data: “A picture worth a thousand words” .

However, a bad conceived might be misleading.

Tp answer our questions we followed the following steps:

.Data preparation, in which the potentially relevant records has been selected on a basis of key word relevance.

.analysis of the temporal distribution of the records, in which we identified the start-time of the epidemic.

.analysis of the spatio-temporal distribution, in which we identified how the outbreak evolved, discovered differences between the temporal patterns in two most affected areas and came to the hypothesis that they could be affected by two different diseases.

.comparision between the texts of the messages posted into two most affected areas, in which we confirmed our hypothesis of the existence of two different diseases

.reasoning about diseases transmission mechanisms, in which we observed patterns to the context information concerning the weather (wind) and geographic features(river

.Hypothesis about a common source of the two diseases based on the observation of the special patterns

.finding relevant information for explaining the reasons for the epidemic outbreak

.putting our findings together into a story that gives answers to the questions of the challenge

Visual Analytics definition:

“The science of analytical reasoning facilitated by interactive visual interfaces.”

While visual analytics aims to help humans to extract knowledge in general from and with models whether they are only mental models or also computer models, it is perfectly fitted to help the data scientist within all of its process.

Without visual analytics data science isn’t good data science due to the incompleteness of computational methods to convey crucial information to the human analyst.

### 1.4 General definition of visual analytics: p22

Let’s dive in the ‘interactive part’ of visualisation.

What is exactly interactivity in this realm?

.zooming

.filtering

.multiple views from different perspective

.linking multiple views

Static images are just not enough to fully grasp an analytical process.

That’s why interactivity is necessary.

Human and computer must work together in synergy nowadays to plainly benefit each one another advantages.

Humans

Flexible and inventive

See the big picture

Solve problems that are hard to formalise

Can cope with incomplete/ inconsistent information

Computers

Can handle hudge amounts of data

Fast search

Fast data processing

Interlink to extend their capacity

Can render high quality graphics

Visual analytics technology combines methods of visual representation of the information,

Techniques for human-computer interaction, and algorithmic methods for computational processing.

# Chapter 2 p27

# General Concepts

The data science process involves analysis of three different subjects:

Data,

Real information portrayed by data,

Computer models derived from data

## 2.1 Subjects of analysis p27

A subject is made of multiple components related between each other.

Understanding a subject is understanding those relationships.

Relationships are studied by identifying patterns in distributions of elements of some components over elements of other components.

Depending of the nature of the components the nature of the distribution changes as well:

We might investigate a frequency distribution as well as a temporal and spatial distribution,

And joint distribution of several components.

Each type of distribution might display different patterns specific to it.

Distribution of visual marks positioned in a display space is perceived as a spatial distribution.

Indeed inclination of human perception to seeing spatial patterns draws visual analytics in a discipline displaying data in artificial spaces where distances represent the strengths of dome non-spatial relationships.

Interesting block diagram displaying analyticalworkflow p29

Data science workflow has a lot in common with the cross-industry standard process for data mining knowed as **CRISP-DM** even though it is specific to business use. Visualytics can be very helpful at each of those steps.

See:

Shearer, C.: The CRISP-DM Model: The new blueprint for data mining. Journal of Data

Warehousing 5(4), 13–22 (2000)

## 2.2 Structure of an analysis subject p30

Plot of relationships between Vastopolis dataset’s variables p31

We see different type of components (variables):

.**set of discrete** entities, such as people, messages, keywords, and disease symptoms

.**space**, which is a set of locations

.**time,** whichis a set of instants(moments)

.**set of values of an attribute**, which can be numeric measure, or non-measurable(categorical, qualitative) characteristics

Cycle based relationship might occure in components relatives to time

## 2.3 Using data to understand a subject p32

Understanding data is understandind relations between components of a subject.

Some relations are higher level relationships they are relations between groups of related components.

Understand those higher level relationship require abstraction from elements to group elements metaphorically called seeing “the forest for the tree”.

### 2.3.1 Distribution p33

A distribution can be sayed as the “way things are shared among a group of individuals” Oxford.

Or “the position, arrangement or frequency of occurrence (as member of a group) over an area or throughout space and time.” Merriam-Webster4s Dictionary.

To understand the concept, we may think of a component as a base for elements of an other component the overaly

### 2.3.2 Patterns and outliers p35

Computing methods to detect patterns exist but the pattern we are searching for must be specified. Therefore visualizations allowing an overview of a distribution are once again relevant.

Eg: to perform computer models, outliers must be removed. And there detection is easily done with a visualisation.

### 2.3.3 Patterns in different kind of distributions p36

When exploring a distribution, two aspects are considered.

The first one is correspondence of the overlay to it’s base, the second is the variation between those correspondences.

All distributions have in common the frequency distribution characteristic which is frequency of distinct elements of one element over the elements of the other.

Variations (increases or decreases = trend patterns) in any distribution are called variation patterns, in a temporal distribution (temporal variations).

Change in a temporal distribution is the appearance or disappearance of elements in a moment relatively to another.

No variation not necessarily mean that there is no change, but simply that some enitities has been replaced by others.

Note that temporal distributions often display periodicity patterns.

Spatial density is the equivalent of temporal frequency but for spatial distributions.

Element of a component might be called neighbours due to small distances between their values.

However when it comes to a spatial distribution geographical barriers must be considered

(eg: slums in front of sumptuous buildings in Brazil)

### 2.3.4 p41 Co-distributions p41

Understanding relationship between two components is possible through study of their co-distributions.

It allows seeing occurrences of one component relatively to occurrences of the other and therefore to study their interrelation (correlation for numerical components).

To study it visualisations are supported in an artificial spatial environment (repère orthonormé)

Hence outliers are called spatial outliers and correspond to exceptional value combinations

Spatial clusters correspond to frequent co-occurrences of values from some intervals

Variation of spatial density tells us about the joint distribution frequency

If we want to plot more than two component joint distribution, it is possible to use such techniques as embedding, projection or dimension reduction.

However they create a distortion of distances.

Dimension reduction is a class of tools of spatialisation (translate distances between an element component by distances).

# Chapter 3 p51

Visual variables are the aspects of a plot we can change to better describe a distribution

According to component’s attributes.

## 3.1 Preliminary notes p51

Introduction to visualisation and interaction.

## 3.2 Visulisation

## 3.2.1 A motivating example p52

As with the 4 Anscom’s Quartet data sets, which display exactly equal summary statistics although they plot totally different distribution; statistic summaries can be misleading.

## 3.2.2 Visualisation theory in a nutshell p55

Graphical representation is encoding of component of data by mean of visual variables.

Position is the most important visual variable.

Data items are often plotted on vertical and horizontal axe scalled

Plannar dimensions = repère orthonormé.

Data items are represented by visual marks called in accordance to their form:

Points, line and area marks.

Visual variables which aren’t position are called retical by Bertin.

Those are size, value(lightness), colour(hue), texture, orientation and shape.

Table illustrating visual variables depending on whether they are describing variables or separating those.

Selectivity is the effectiveness of distinction allowed by a visual variable to perceive different values (the most effective one being space).

Associativity is the effectiveness of distinction allowed by an aesthetic to perceive different group of values from others (the most effective one being space).

“The visual variables must have a level of organisation at least equal to that of the component which they represent” Jacques Bertin

To do proper visualisations Bertin created levels of organisation(a priority order of the principles which must be respected when building a visualisation).

Those level of organisation also known as scale of a data component may be nominal, ordinal or numeric.

1 quantitative (numeric scale)

2 ordered (ordinal scale)

3 selective (nominal scale)

4 associative (nominal scale)

Strongest associative power are position and colour

To check out which visualisation variables are the best fitted for which variable chec out diagrams of the chapter.

Length of variable is the number of categories or steps that an aesthetic allows to distinct.

Length should be equal or greater to the number of distinct values present in component to be visually efficient. If length isn’t large enough, the observer will perceive some of the different data as being identical

Charts of most well fitted aesthetic depending on the scale class to depict is present p60

First is from Bertin

Second is from a psychologic perspective

Third is from a cartographer perspective check out there:

MacEachren, A.M.: How maps work: representation, visualization, and design. Guilford

Press (1995)

Further detail in there:

Andrienko, N., Andrienko, G.: Exploratory analysis of spatial and temporal data: a system-

atic approach. Springer Science & Business Media (2006)

Associativity plays an important role in abstraction (capacity of perceiving only relevant elements by ignoring unimportant details).

Colour has a strong associativity power.

“display time” used in animated display has a high level of organisation is suited for supporting an overall view of how something develop over the time but is less supportive for tasks requiring attention to details and detection of small changes.

### 3.2.3 The use of display space p61

Position in the display or the dimensions of the display are the most important visual variables having the greatest expressive power.

.Space partitioning: often call juxtaposition (facet\_wrap() in R)

.Space embedding: histogram of distribution of interventions by categorie by commune on a map for example

. Space sharing a.k.a superposition or overlay: representation of two our more data component in a common display space

. Fusing data components: use a two or three display dimension for representing a larger number of data components.

However when there is no intrinsic relationship between two components a plot using position visual variable can be misleading.

Eg: keywords in set of texts, in this case world cloud is a relevant plot to use cause position has no meaning.

### 3.2.4 Commonly used visualisations p63

#### 3.2.4.1 Representing distributions of numeric values

Bar plot:

May be ordered by ordinal categories depending on the inspected component.

Pie Chart:

Are better suited to display fraction of a whole than bar plots. However, bar plot are better to compare values.

Dot Plot:

Depicts a distribution of numeric values along an axis

Line graph, line plot, or line chart: points linked by a line on planar dimensions.

Heat map or heatmap: hue graduation depending on a component value on any spatial basis.

Histogram (frequency histogram): displays frequency of a component, altering bin size (width of histogram’s bars) is useful to fully grasp a distribution. Hue can be used to identify classes or categories.

Violin plot: kind of bow plot showing shape of distribution

Box plt a.k.a box -and-whisker-plot: two sides of the plot are depicted by two parallel lines representing q1 and q3 however it can be connected to several different values.

Dots are outliers.

#### 3.2.4.2 Representing temporal information

Time graph or time plot: line chart but points represent time

2D time chart: time-referenced values of an attribute regarding one time cycle(eg hours in a days) and the other to the linear component consisting of the repetition of this cycle (eg multiple consecutive daysd) or to another time cycle (eg days in a week)

Timeline view: one dimension for time the other for event occurrence

Gantt chart: portraying activities in task management. Blocks of activities that may be related to each other by arrows.

Time histogram: histogram of frequencies but with time and events

2D time histogram: like the time graph displayin histograms or other kind of symbols in the matrix

#### 3.2.4.2 Representing spatial information

Dot map: geographic distribution of discrete entities which are represented by dots.

Choropleth map: depicts the distribution of discrete geographical areas (polygones)

Colour(hue) is used for categorical attribute (qualitative) and value (lightness) for numerical

(quantitative) attributes.

Diagram map or chart map: Represents spatial distribution on a map through symbols, diagrams or even pie charts drawn on the precise location of the distribution.

Area cartogram:

Distortion of geographic areas sizes to represent a numerical attribute (eg population)

#### 3.2.4.4 Representing co-distributions and interrelations.

Scatterplot: two numeric attributes on two distinct axis. Pairs of attribute values are represented by dots.

Scatterplot matrix: Matrix of scatterplots representing different pairs of attributes (attribute = component). Often diagonal of the matrix is used to depict the distributions of the individual attributes (by frequency histograms).

2D histogram:

Represents co-distributions by bins which can take any form such as bars, circles or whatever

Parallel coordinates: see fig 2.10 p 44 (I don’t get it)

#### 3.2.4.5 Representing relationships between entities

Node-link diagram:

Linked nodes, nodes representing entities and lines relations between them.

Can be Fastly cluttered, visibility highly dependent on a suitable layout.

Connectivity matrix:

Matrices de contingence. Links between variables can be represented by any fited visual variable.

Treemap: Hierachical sized components often in the shape of rectangles.

Many of the visualisations that are listed in this section can be part of composite graphics depicting complementary portion of information multiple simultaneous perspectives of the data.

### 3.2.5 general principles of visualization p68

1: Use space at first: the intrinsic relationship exists represent those relations with positions

2: Respect properties: consistency of visual variable for representing same value domain

3: Respect semantics: Domain knowledge and common sense should be used when creating visualisations, some intrinsic relationships have no need to be explained

4: Enable seeing the whole:

Enable overview: avoid overlapping and sear for possibilities to associate multiple items in group.

5. Include titles, labels, and legends:

Interpretation of labels strongly depends on appropriate labelling of data components and explanation of how they are encoded by visual means.

6. Avoid excessive display ink: avoid visual components that do not display information

7. Consider employing redundancy: presenting same data under two different plots can be usefull to fully understand it

8. Enable looking at data from multiple perspectives:

When the number of components is higher than the number of visual variables interactive juxtaposition of plots depicting different components can be useful to see the data from different perspectives.

9. Rely on interactivity:

Interactivity is pretty useful.

For eg instead of trying to depict exact values, interactively changing symbols can be pretty helpful to find them

### 3.2.6 Benefits of visualisation

A good visual representation helps to confirm expected and discover unexpected patterns.

Understanding the distributions: visualisations are often much more informative and understandable than numerical summaries of data.

Spatial and temporal distributions cannot be adequately expressed by any numerical summary

Comparison:

Juxtaposition, superposition and explicit encoding(comparison metric calculated then visualized)

Discovering the unexpected:

Summary statistics can be unhelpful or misleading to discover patterns. For example, mean and standard deviation are only meaningful for a normal distribution.

### 3.2.7 Limitation of visualisation p73

Implicitly defined and unchecked assumptions:

Since many statistical summaries implicitly assume certain property of the data, visual representations involving these summaries inherit these assumptions.

An example is boxplots using the quartiles of the data.

The quartiles can provide an adequate statistical summary only when the data distribution is unimodale. The boxplot is therefore non suited for bimodales distributions.

However this limitation is only due to lack of knowledge in statistics.

Fortunately a distribution can be found unimodal or bimodal thanks to an histogram.

Lack of visual resolution:

Attempt to display all items in the display space often results of the majority of marks being clustered. Alleviating this problem can be done by zoom and pan interactions in an implementation, and computationally, by performing pixel-based aggregations or using data reduction techniques to select a smaller number of representative points.

Visual clutter:

Zoom and filter interactions may help, giving the user the ability to remove irrelevant items or details irrelevant for the task however this can lead to losing context.

Therefore it is appropriate to combine a zoomed-in view of a part of the data with a more aggregated view of the whole dataindicating what part is currently explored.

A computational solution is to employ data reduction/sampling methods; in particular, context-aware filtering.

Obvious patterns obliterate more interesting patterns:

Adding zoom and pan interactions and filtering is an obvious pattern to manage this however it might be hard to guess interesting patterns.

A more sophisticated approach is to use a model for removing the expected patterns and revealing deviations from this.

Eg: normalising geographical data by population or area may result in that the data are not dominated by uninteresting structures.

Another example is time series ; removing expected trends and seasonal components ighlights deviations and is able t uncover surprising patterns

False patterns appear when a user tries to attach meaning:

For instance, smoothing techniques might make occur nonsense patterns depending on the processing degree. To avoid this, an interactive control of the processing degree can be used as proposed in

Wang, Y., Xue, M., Wang, Y., Yan, X., Chen, B., Fu, C., Hurter, C.: Interactive structure-

aware blending of diverse edge bundling visualizations. IEEE Transactions on Visualization

and Computer Graphics 26(1), 687–696 (2020). DOI 10.1109/TVCG.2019.2934805

for edge bundling

Slow rendering for large datasets:

Big data amounts can slow down processes of both computational pattern detection and visual exploration.

To alleviate this problem the generally used technical approach is data reduction including aggregation, sampling, dimensionality reduction feature selection methods.

Important to check if important properties of the data set still are represented in the sample.

Another approach is design of so called anytime method:

An algorithm wich will always output a correct answer and win in precision over time until giving the perfect one. It allows to interrupt it anytime to get an answer.

Some applications need to be applied to streaming data.

To make them efficient to human perspective, change blindness must be alleviated by automatically highlighting important changes while data is streamed.

## 3.3 Interaction p75

Static plot is often not enough, interaction can be used multiple ways to help analyst doing its job.

Those ways are :

Changing data representation

Focusing and getting details

Data transformation

Data selection and filtering

Finding corresponding information pieces in multiple views

### 3.3.1 Interaction for changing data representation

Time can be represented either in a linear way either in a cyclic way.

Bins of an histogram can be switch with interaction, it changes shape of distribution.

Cumulative frequency curve displays a constant distribution shape when changing its steps though.

### 3.3.2 Interaction for focusing and getting details P79

It is now a standard that graphical displays provide access to exact data values upon pointing (mouse hovering) at visual marks.

Other operations that are used for seeing more details are zooming and panning.

An example of zooming would be to select an interval of values to visualize rather than the all picture

Color rescaling can be efficient for this purpose see p91

### 3.3.3 Interaction for data transformation p80

In the process of exploring data with the use of visual displays, motives for transforming data may arise:

.Simplify the display or the data

.Reduce the amount of the data

.Disregard excessive details an facilitate abstraction

.Others

Transforming continuous data into classes is an example of transformation and is called discretisation.

Logarithmic transformation can be applied to values of a numeric attribute when the most of the values are small and there is a few high values.

It allows to enhance visualization of differences between low values.

However the representation will be harder to interpret than a display of original values.

Aggregating also is a useful data transformation technic.

Eg: Aggregate all communes by its department

Data smoothing: The simplest way to integrate outliers is by replacing them by the mean.

A more sophisticated way is to attribute values with values of its neighbourhood

Time series specific smoothing methods exist as exponential smoothing and double exponential smoothing.

Data changing operations for time series:

Calculation of changes with respect to previous time step or a selected time step.

Changes can be express in the form of differences or ratios

### 3.3.4 Interaction for data selection and filtering p83

Data selection ~ querying ~ filtering

Selection is taking a portion of the data to visualize this portion only.

Filtering is temporarily removing the data that are currently out of interest from already existing visual display.

Those can be operated on different criteria such as:

.attribute value( attribute base filter)

.position in time, either along the time line or in a temporal cycle (temporal filter)

.position in space (spatial filter)

.references to particular entities (entity-based filter).

.relationship to other data (filter of related data)

All filtering operation should be assisted by a representation of the distribution

Brushing with a mouse may also allow highlighting as a form of filtering.

A good practice in a composed visualisation is that selecting an element on a plot would allow to select it simultaneously on other related plots

It is also possible to select data by point neighbourhood.

It can be done with the so called Mahalonobis distance or other distance functions

### 3.3.5 Relating multiple graphical views p85

Brushing concept that we just saw in the previous section is part of coordinates multiple views(CMV ) approach to visual analysis.

The different techniques of this approcha are:

Brushing: Highlight part of the data

Filtering: Select only a part of the dat

Common symbolisation: Use same visual variable attribute for a similar component attribute across the visualisations

Conditioning: Create several instances of a display showing different data portions

It is also possible to let the user chose to allow or not interaction between displays

### 3.3.6 Limitations and disadvantages of interaction p87

Interactive exploration lacks of a systematic approach.

For a more systematic interactive exploration, it is useful to keep track of the inspected data components and performed operations and document the visual data analysis process which helps to identify data components and portions that have not yet been considered.

When large data amounts are being visualized interactions can be slow down.

To alleviate this problem different approach can be applied such as restricting the possible interactions.

It must be considered though that those techniques takes time to apply and that time is the most precious resource of the analyst.

Therefore, any use of interaction should be justify and computations should be used whenever it’s possible.

## 3.4 Concluding remarks p88

References to go further on spoken subjects

# Chapter 4 Computational Techniques in Visual Analytics p89

Combining interactive visualisation with computational techniques has to sides.

One side is computational support to visual analysis.

The other side is visual support to application od computational methods.

This chapter discusses of some techniques of the first side:

Spatialisation done with data embedding techniques and grouping with clustering algorithms.

It also summarizes the second side

## 4.1.1Visualisation for supporting computations p90

Select right computational methods to apply depends on the nature and properties of the data under analysis and the current analysis task.

This task can and must be supported by visualisations.

Then data must be prepared to the application of the chosen method.

Here visualisation has once again a purpose; the one of uncovering existence of outliers and dissimilar subsets. And if exist, how do they impact a model.

Models to be understood by humans must be visualized.

If there is no reason to choose a model over another, both must be created then compared.

A single model results on different runs must be analysed as well. Consistency between the results validates the model while inconsistencies require investigation, explanation, and making a justified choice or drawing a justified overall conclusion.

The measures aren’t sufficient for supporting understanding and substantiated judgements, visualisations are once again the key to alleviate this.

Bullet point on performed task by visualisation in data science workflow:

.Before applying computational methods: enable investigation of data propreties in order to:

.Chose suitable computational methods and make appropriate parameter setings

.Detect outliers, which may need to be removd

.Detect disparities between part of data, which may require different approaches to analysis

.During applying computational methods:

.inspect how the methods uses and processes the input data

.investigate the intermediate structures constructed by the method

.examine the current state of the model being built by the method and understand whether is develops in the right direction.

.After applying computational methods:

.enable evaluation of method results for understanding their meaningfulness

And usefulness and seeing the variation of the quality across the input data

.enable comparison of different results obtained by varying the input data or values of method parameters, or by choosing alternative methods

### 4.1.2 Computations for supporting visual analysis p92

.

When it’s impossible to create a visualisation out of a huge amount of data either for technical or for readability reasons, two approaches can used to alleviate the problem.

One approach would be to only visualize small pieces of data extracted through explicit condition queries or with data mining techniques that search for combinations of data items having particular properties or frequently occurring in the data set.

This approach is taken when only such data pieces are of interest and there is no need in getting a big picture of the whole dataset.

Techniques that are used for extracting data pieces support data analysis but do not play a special role of supporting visualisation.

Visualisation can be supported by computational techniques that summarize and/or organize information so that it become better suited for visual representation and human perception and comprehension. The main goal is to enable overall view of the while bulk of information.

This is possible thanks to a high level of abstraction, details can be obtain by means of interaction.

There are two major approaches to creating an overall view of a dataset:

.Spatialisation:

Position the data items in a two or three dimensions artificial space according to a certain principle, usually according to their similarity or relatedness.

.Grouping:

Organise the data in groups according to their similarity.

The resulting groups are treated as units, each group is a single object.

The groups are characterised by statistical summaries of characteristics of their element.

The statistical summaries can be represented visually.

Spatialisation can be achieved by mean of the class of computational techniques called data embedding or, more frequently, dimensionality reduction or dimension reduction.

These methods, generally, represent data items by points in an abstrzct space with a chosen number of dimensions (2 or 3).

One categorie of data embedding methods apply some transformations to the original components of the data called “dimensions”, in order to derive a smaller number of components by which the data items can be described so that substantial differences between them are preserved.

Another category of the data embedding methods use numeric measures of the similarities or relatedness between any two data items.

According to chosen criteria of similarity, less similarity between two data elements exist, more the value of the measurement will be high.

A measure fulfilling this principle is called **distance.**

A mathematical formula or an algorithm for determining the distance between given two items is called a **distance function.**

Data embedding methods using distances aim at reproducing the distances between the data items as accurately as possible through the distances representing the data items.

Some of those methods ignore the original components (dimensions) of the data and use only the distances irrespective of how they have been determined.

However all data embedding methods remain called dimensional/ity reduction.

Grouping can be achieved by means of techniques for clustering.

Clustering techniques also use distance functions.

Hence distance functions are used both in spatialisation and grouping methods.

Fig 4.1 Relationship between class of computational methods supporting visualisation

Grouping and spatialisation are to classes of computational techniques used supporting data visualisation which respectively belong to the two main classes of data embedding and clustering.

Both sub classes use distance functions

For defining appropriate distance functions it is often necessary to perform **feature selection**

Which means finding a small none-redundant subset of data components that is sufficient for representing substantial differences between data items.

(Data components are called features or variables in statistics and data mining)

Topic modelling is a class of methods originally created for text analysis however it has been demonstrated as being competent as a dimension reduction technique.

However, it remains a new technique.

It must be considered that technique discussed in this chapter aren’t exclusive for supporting data visualisation, they have other applications.

Particularly, feature selection, dimensionality reduction and clustering methods can be used for building computer models and distance functions for searching by similarities.

## 4.2 Distance functions p95

Ther exist many distance functions differing in the type of data they can be applied to and in the approaches to expressing dissimilarities.

In the following subsections they will be grouped in function of their matching data type

### 4.2.1 Multiple numeric attributes

Determining the difference between values of an attribute with a cyclic value domain

Requires the following special approach:

If the result of subtracting one value from another exceeds the half of the cycle length, it needs to be subtracted from the cycle length

23 – 2 = 21

24-21 = 3

351-20 = 331

360-331 = 29

When data items differ in more than one numeric attributes, the dissimilarity can be expressed by combining in some way the differences between the values of each attribute.

Most popular method to do so is **Euclidian distance**

**Sqrt(sum(|xi-yi|²)), i<=n**

Another popular distance is **Manhattan** **distance also called Taxicab or City Block distance.**

Manhattan Distance metric is preferred over Euclidean Distance when there is a high dimensionality in the data.

**sum(|xi-yi|), i<=n**

The **Minkowski distance** generalises the possible approaches to combining multiple differences in the formula.

(sum(|xi-yi|^p)^(1/p)

When p = 1 we found Manhattan distance and when it is equal to 2 we have the Euclidian one

Go further https://www-users.cse.umn.edu/~kumar001/papers/siam\_hd\_snn\_cluster.pdf

There is however prerequisites when using any distance functions that aggregate arithmetic distances between attribute values.

.When the value ranges of the attributes significantly differ, it necessary to apply normalisation

.If some attributes are correlated their impact on the results will differ in comparison unrelated attributes. A common approach to solving this problem Is feature selection.

**Mahalonobis distance** address both problems by integrating normalisation and reducing the impact of relationships between the attributes involved.

Distance is however harder to compute and from a human perspective to interpret.

https://en.wikipedia.org/wiki/Mahalanobis\_distance

The use of arithmetic differences between attributes isn’t the only possible approach to expressing dissimilarities though.

Another approach consists in treating data items as vectors rather than points in a multidimensional space and mesure the angles between the vectors irrespective of their magnitudes.

The idea underlies the **cosinus similarity** function:

Sum(xi\*yi)/[sqrt(sum(xi²)\*sqrt(sum(yi²)]

This formula results in the cosine of the angle value between the vectors(components).

If result is -1, vectors are opposed, 0 they are unrelated,1 they are the same.

Hence this is not a distance function, values of distance functions being included between 0 and 1.

Wen components are opposed when similarities evaluated with a distance function, opposition between components is expressed by positive values when exact similarity is portrayed by 0.

To convert the angle in a distance function result, arcos of the angle must be computed, this way **angular distance** function is defined.

Then if the results is desired in degree, It must be divide by pi.

The functions of cosine similarities and angular distance are often use in analysis of texts

In general distance functions that disregard absolute value of attributes are useful when data items need to be compared in terms of their structure irrespective of quantitative differences.

Such distances can be called **scale independent.**

Apart from the angular distance, a scale independent distance function can be defined based on the correlation between two combinations of values of multiple attributes.

As for cosinus similarity function, result is given between -1 and 1 with the same significations.

The correlation based distance is given by substracting 1 from the correlation coeffitient

There are many formulas for expressing the correlation.

The most popular is Pearson correlation:

Sum([xi-mean(x)]\*[yi-mean(yi)] / sqrt[sum(xi-mean(x))²\* sum(yi-mean(y))²]

Distance functions can however become less efficient along that dimensions increases, because greater is the artificial space, lower the distances will be in comparison, this is called the curse of dimensionality, <https://en.wikipedia.org/wiki/Curse_of_dimensionality> .

To alleviate this problem, it is necessary to use feature selection and/or dimensionality reduction.

It is admitted that if number of dimensions(columns) in a dataset exceeds number of data items (rows) data should definitely be treated as high-dimensional, however it is often the case much before.

## 4.2.2 Distributions p99

Description of different types of distributions.

Histogram mathematical sense:

“a fuction that count the number of observations that fall into each of the disjoint category (known as bins”

Sets of per-bin statistics representing two distributions can in principle, be compared in the same ways as combinations of values of multiple numeric attributes.

Distribution are typically compared using scale independent distance functions allowing to disregard quantities and compare the distribution profile such as angular distance or correlation distance.

Distributions can also be compared with other function see

<https://hal.science/hal-02299826>

In representing distributions by histograms, we must be mindful of the curse of dimensionality once again those by assuring that the number of bins is lower than the number of distributions to analyse or apply dimensionality reduction and/or feature selection.

### 4.2.3 Numeric time series p100

Numeric time series consist of values of a numeric attribute referring to a sequence of consecutive steps, which may be time moments or intervals, or just ordinal numbers.

Numeric time series does not differ from a combination of values of multiple numeric comparable attributes.

Therefore all distance functions are also applicable to time series as the time series have the same length, i.e number of steps of the other. If it is not so one time series can be re-sampled to the length of the other (see discussion on resampling in section 5.3.2).

I think that what is meant here is that it is possible to compare to different events axed on the same ordinal scale with distance functions by taking e.g first y as x and second y as y.

More information about it:

Keogh, E., Kasetty, S.: On the need for time series data mining benchmarks: A survey and

empirical demonstration. Data Min. Knowl. Discov. 7(4), 349–371 (2003). DOI 10.1023/A:

1024988512476

The distance function **Dynamic Time Warping** abbreviated as **DTW** was proposed specifically for time series. It allows to compute the distance between two time series irrespectively of their

Length(number of steps)

It has been use in speech recognition, or on human computer interaction for comparison of EEG signals. It has also been use for comparing t

rajectories of moving objects following similar routes.

In general it can be used for comparing processes unfolding with different speeds.

However it has been argued that the results of using DTW for dissimilarity assessment do not substantially differ from using the Euclidian or Manhattan distance after re-sampling distances of sequences to the same length.

It must be considered that comparison of time series may be greatly affected by noise (random fluctuation in the data). Application of a smoothing method can reduce the noisiness and thus help to compare the general patterns of the temporal variation rather than minor details.

When time series are long and may involve several variation patterns, such as an overall trend and seasonal or cyclic fluctuation it makes sense to decompose it into components and apply a chosen distance function to the components that are relevant to the analysis task.

If three or more components are involved, distance can be computed separately, and then average or weighted average of distances can be calculated.

In the case of multiple components involved when using Euclidian distance,

it also seems possible to compute sum of squared distances for each component after normalisation, then either apply square root on each distance sum then summing them or summing the distances then apply square root depending either we want to attenuate differences between events of same parameters or differences between parameters distances.

More information:

<https://en.wikipedia.org/wiki/Decomposition_of_time_series>

For periodic time series it is also possible to use the **Fourier transform**, which represents a time series as a sum of sinusoidal components i.e sine and cosine multiplied by some coefficients, which are called Fourier coefficients.

Then the distance between two time series can be expressed series can be expressed as the distance between the combination of their Fourier coefficients using one of the functions suitable for multiple numeric attribute(formulas of the Minkowski distance or angular formula)

More information:

<https://en.wikipedia.org/wiki/Fourier_series>

### 4.2.4 Categorical attributes and mixed data p102

The distance between data items described by multiple categorical attributes (eg: Gender, Socio-professional categories)

How to measure the similarity or distance between two categorical values of a single attribute (e.g attribute = Gender, values = [Male, Female])

The simplest approach is to assume that the distance between any two distinct values is the same, usually 1.