Visualisation on a closed surface

Walter A. Kosters and Jeroen F. J. Laros
Leiden University

NWO

ToKeN2000

Universiteit Leiden

Outline.

Given a set of points in high-dimensional space, we give a visualisation projected on a 2-dimensional torus.

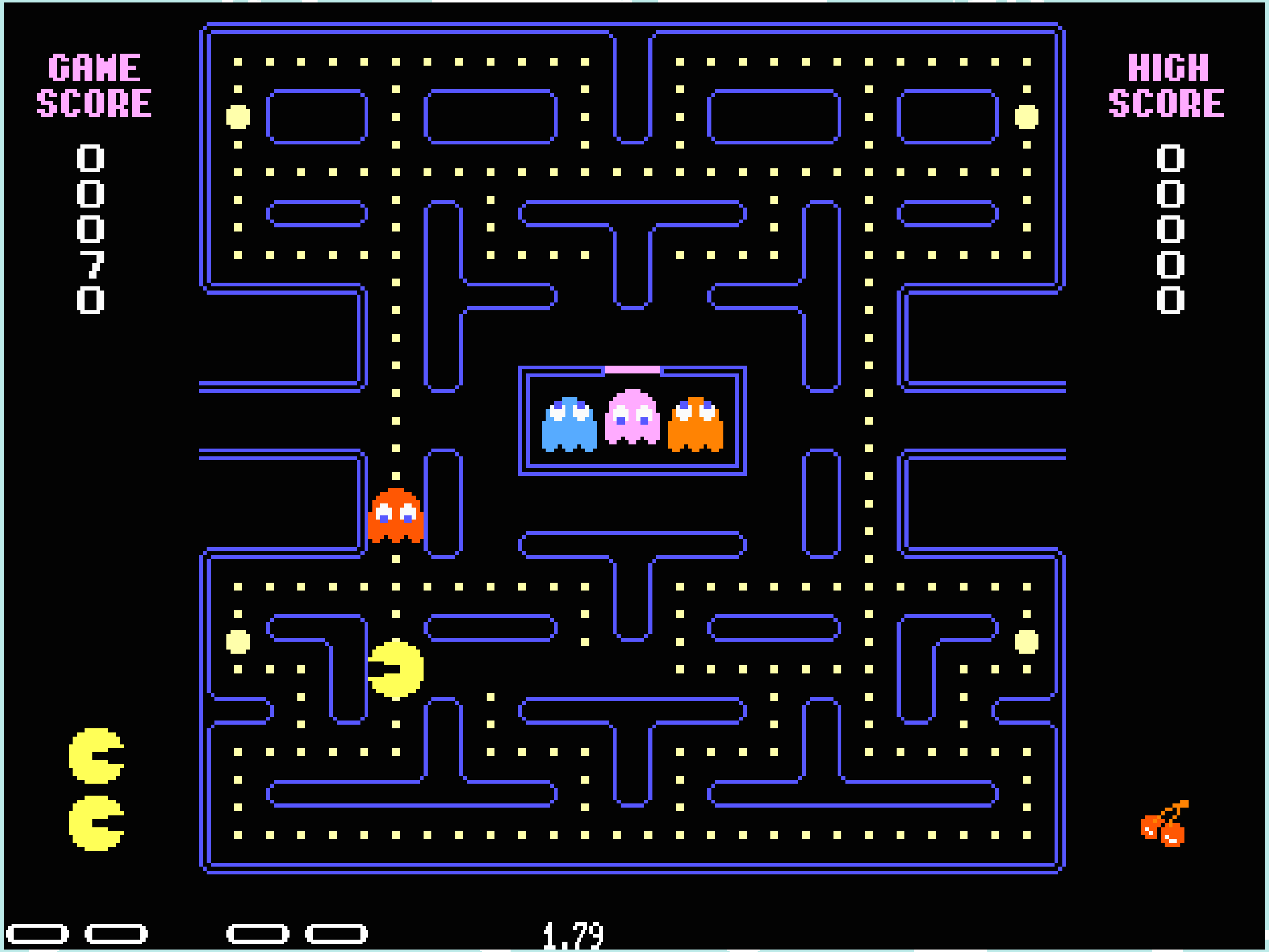
The shape of the torus can be exploited and is a generalisation of the flat unit square.

The dimension reduction technique is a push and pull oriented one, which uses an aggressive correction function.

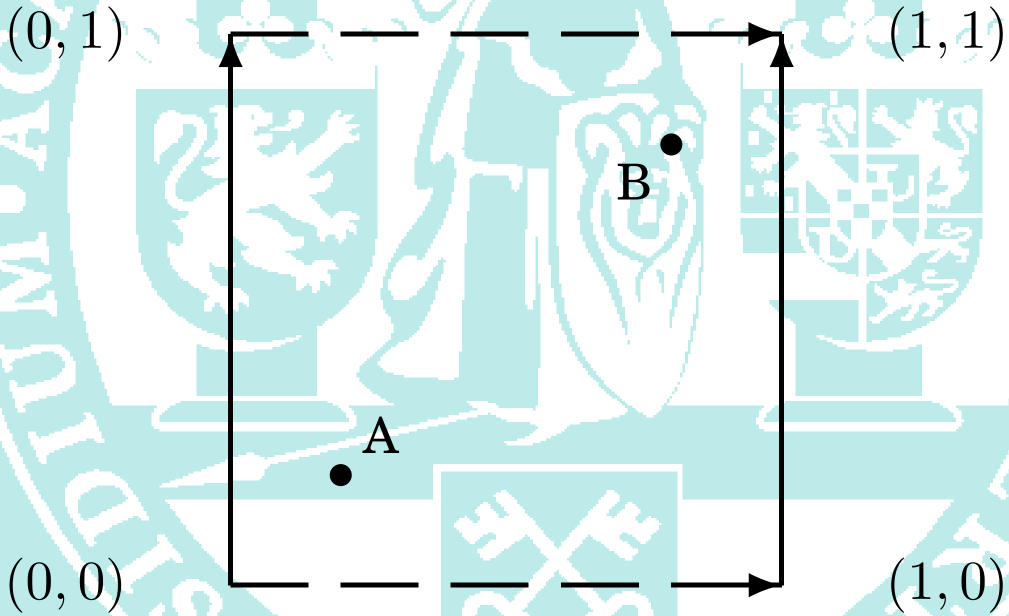
Some practical applications are given, criminal analysis being the most prominent one.

Universiteit Leiden

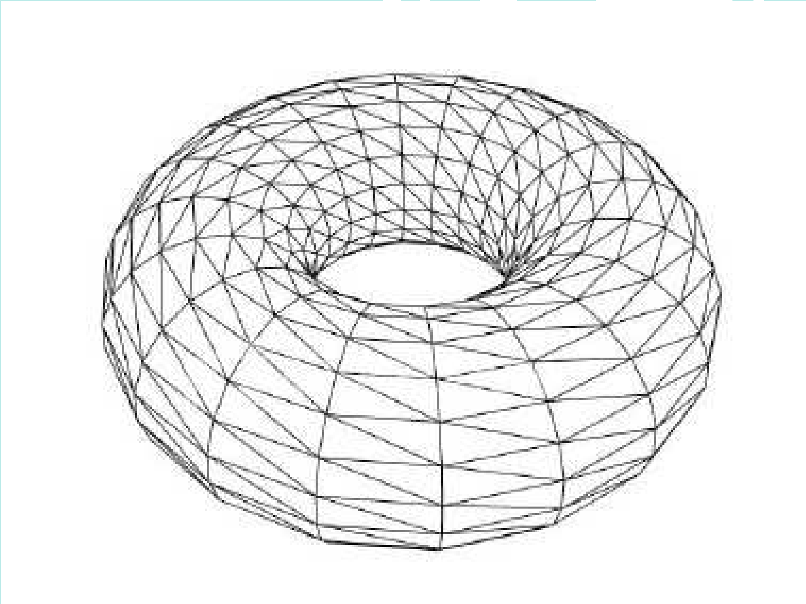
A 2d-torus.

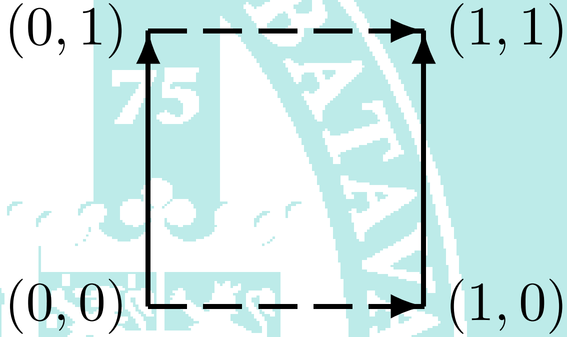


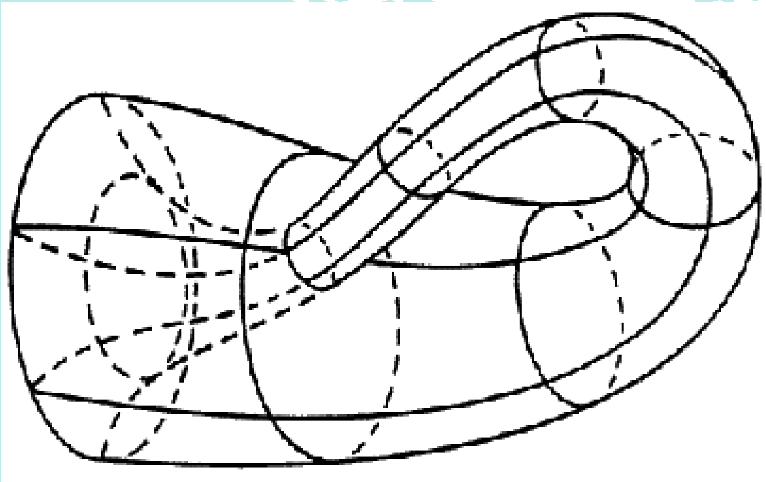
The unit square with sides identified: a torus.

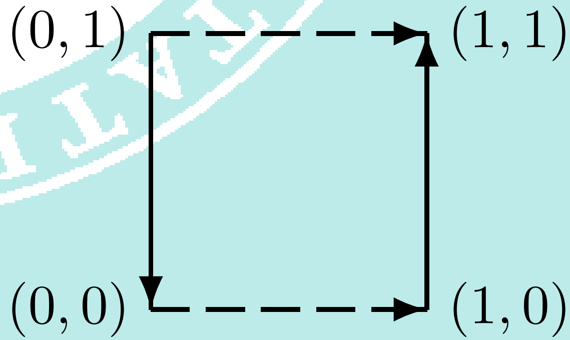


- A torus does not have a boundary.
- The metric on a torus is Euclidean.
- There is a maximum distance on a torus.



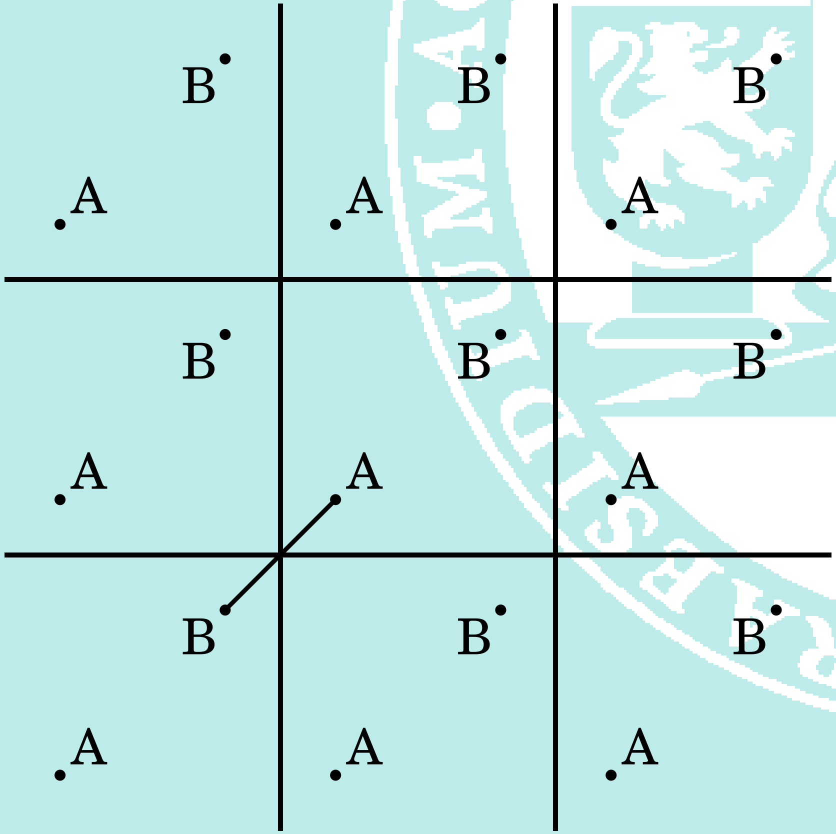






Universiteit Leiden

Finding the distance between two points.



$$d_{\text{current}}\,((x_1, y_1), (x_2, y_2)) ::$$

$$\quad \textbf{var } x_3 \leftarrow x_2;$$

$$\quad \textbf{var } y_3 \leftarrow y_2;$$

$$\quad \textbf{if } \; x_1 - x_2 > 0.5 \; \textbf{ then } \; x_3 \leftarrow x_3 + 1.0;$$

$$\quad \textbf{if } \; x_1 - x_2 < -0.5 \; \textbf{ then } \; x_3 \leftarrow x_3 - 1.0;$$

$$\quad \textbf{if } \; y_1 - y_2 > 0.5 \; \textbf{ then } \; y_3 \leftarrow y_3 + 1.0;$$

$$\quad \textbf{if } \; y_1 - y_2 < -0.5 \; \textbf{ then } \; y_3 \leftarrow y_3 - 1.0;$$

$$\quad \textbf{return } \; (x_1 - x_3)^2 + (y_1 - y_3)^2 \; ;$$

Dimension reduction.

We use a push and pull oriented technique, where the correction factor depends on a number of variables.

- The ideal distance (d_{desired}).
- The real distance (d_{current}).
- The maximum distance in the space ($\frac{1}{2}\sqrt{2}$).
- The distribution of the desired distances.

If the distribution is not between 0 and 1, a simple correction can be used to correct this.

Universiteit Leiden

Correction functions in general.

Each correction function must adhere to the following restrictions:

- $f(0) = \rho$
- $f(0.5) = -\rho$
- $f(d_{\text{desired}}) = 0$

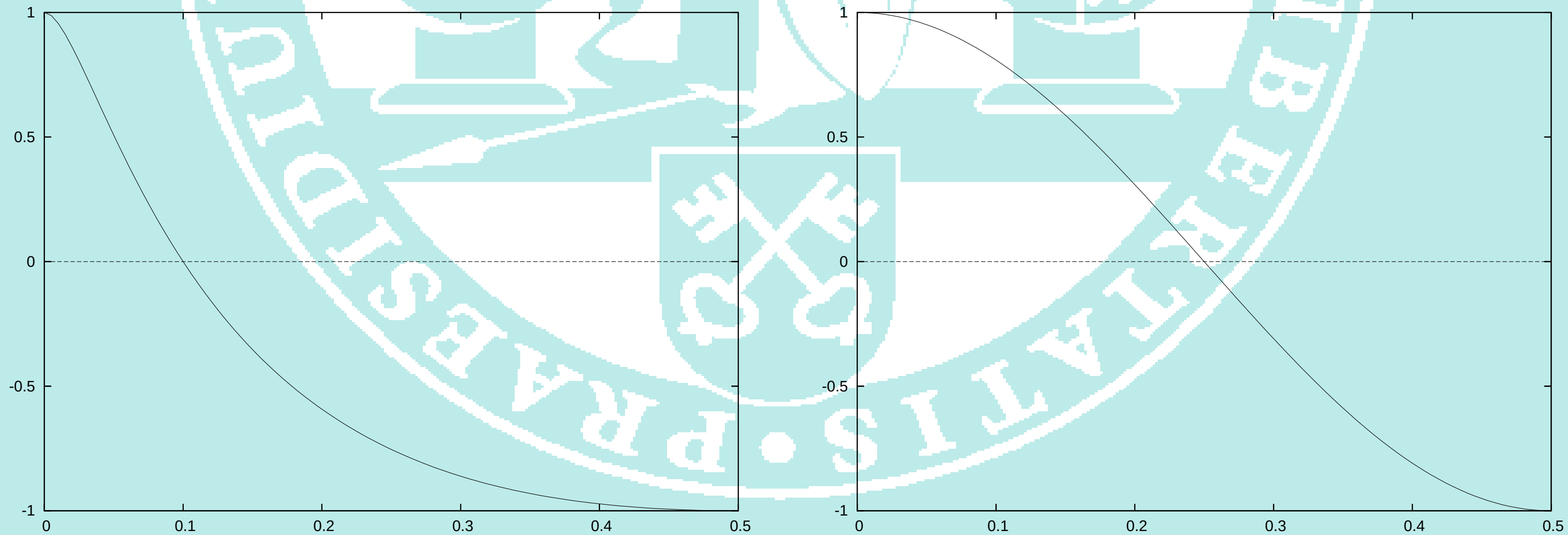
Here $\rho \in (0, 1]$ is the so-called *correction multiplier* (usually 1).

Notice that we use 0.5 as maximum distance, since we work with quadratic distances.

Example correction function.

$$f_{d_{\text{desired}}}(x) = \begin{cases} \rho \ \cos\left(\pi \log_t(2x(t-1)+1)\right) & \text{if} \ \ d_{\text{desired}} \neq 0.25 \\ \rho \ \cos\left(\pi 2x\right) & \text{if} \ \ d_{\text{desired}} = 0.25 \end{cases}$$

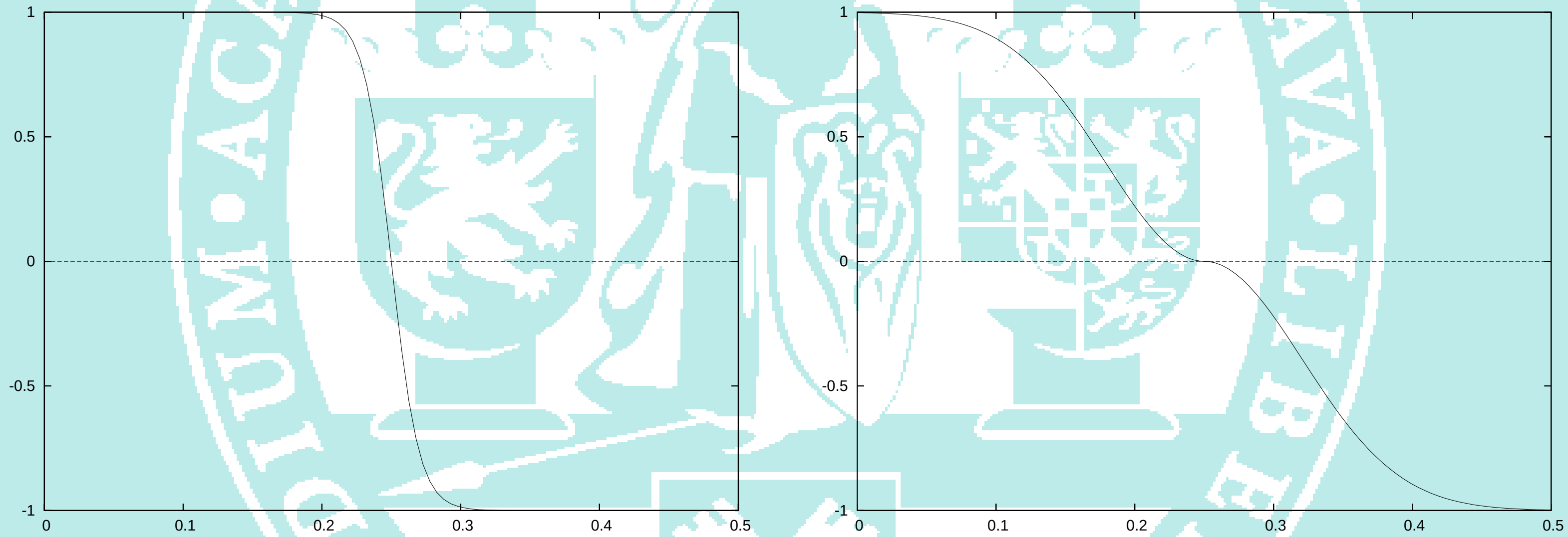
where $t = (1 - 1/(2d_{\text{desired}}))^2$.



Correction function for $d_{\text{desired}} = 0.1$ resp. 0.25.

Universiteit Leiden

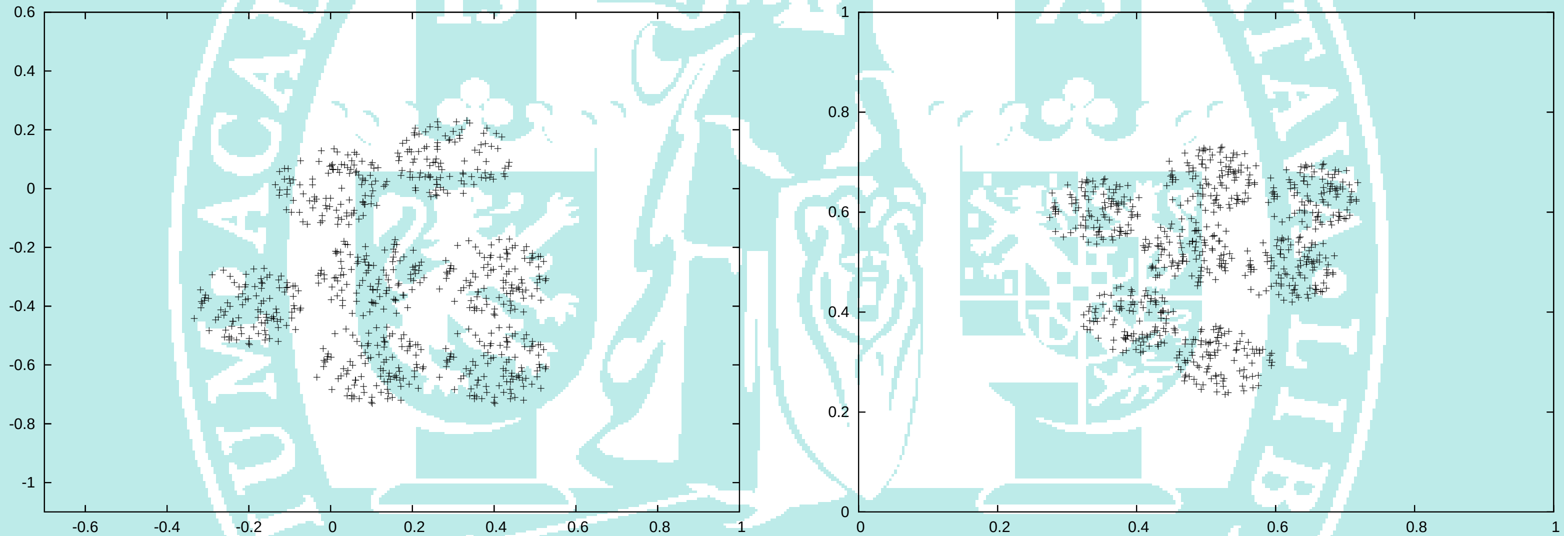
Other functions.



Other good possibilities are linear functions, sigmoid functions (left) and more exotic functions like a double sigmoid function (right).

Universiteit Leiden

Synthetic data.



The left image is the input data, from which a pairwise distance matrix is extracted. This matrix is used to produce the right image.

Notice that, although the distances are preserved, the images are not identical, they are invariant under mirroring and rotation.

Universiteit Leiden

The flat unit square as subcase.

If the data can be embedded in the normal unit square, it is also embeddable in the torus.

The trick is to normalise the distances between 0 and the maximum distance.

A flat picture will always cluster in a 0.5×0.5 subsquare.

Universiteit Leiden

Criminal data.

We have a database with more than one million criminal careers.

The crimes are divided into categories and weights are added to the categories.

The distance between two criminals is calculated with a metric designed for multisets.

Universiteit Leiden

The generic distance measure is

$$d_f(X,Y) = \frac{\sum_{i=1}^{n} f(x_i, y_i)}{|S(X) \cup S(Y)|}$$

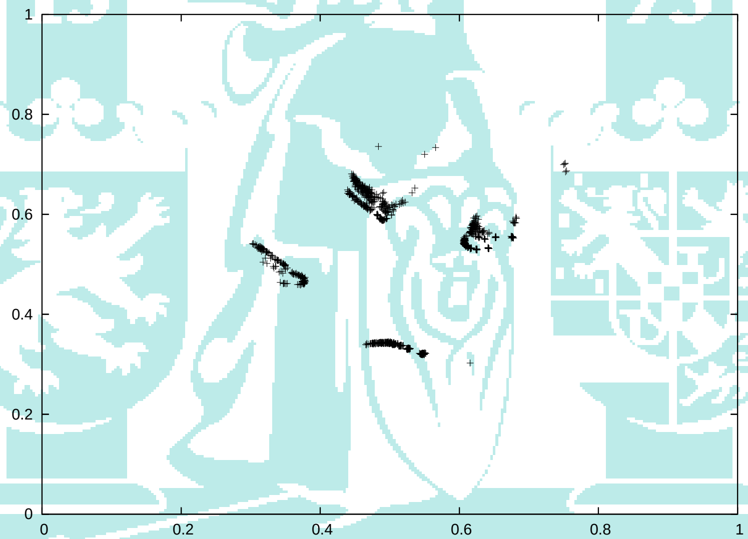
The parameter f can be tuned. It specifies the difference between the number of occurrences of a particular element in two multisets. Constructing such a function is natural and can easily be done by domain experts.

Q: What is the difference between 2 and 10 crimes of the same category?

A: $f(2, 10)$.

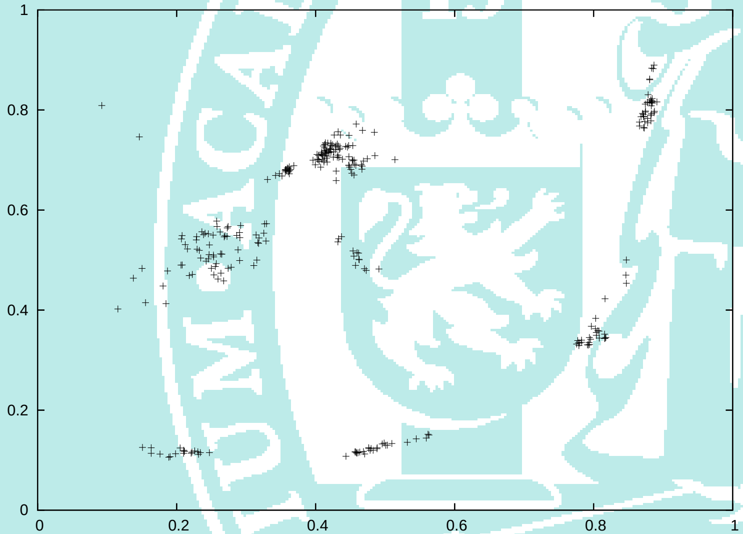
Universiteit Leiden

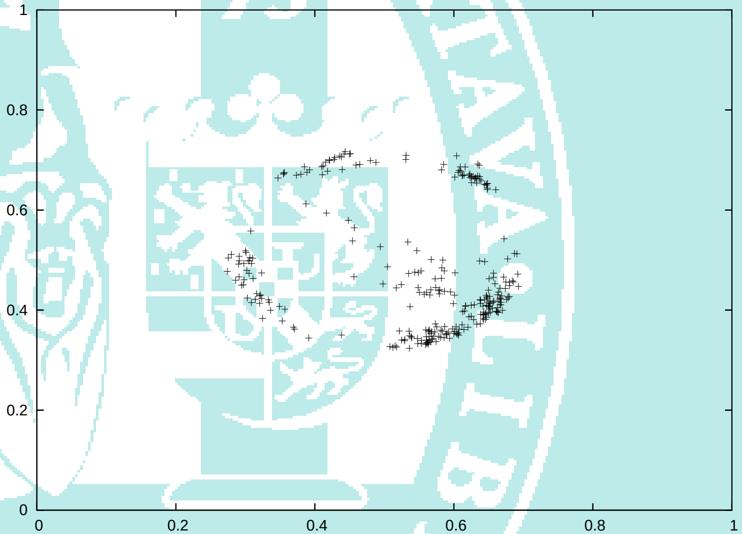
Visualisation of criminal data.



- Left: relatively light crimes.
- Top: all-rounders.
- Right: light and heavy crimes, nothing in between.
- Bottom: only light crimes.

Molecule data.





On the left we see a good embedding, on the right an inferior embedding.

Notice that the right picture is the one we would have gotten when using the normal unit square.

Universiteit Leiden

This research is part of the DALE (Data Assistance for Law Enforcement) project as financed in the ToKeN program from the Netherlands Organization for Scientific Research (NWO) under grant number 634.000.430.

NWO

ToKeN2000

Universiteit Leiden