# Collection of medical original data with search engine for decision support

**Wolfgang Orthuber**
University Kiel, UKSH, Department of Orthodontics
Email: orthuber@kfo-zmk.uni-kiel.de

## Problem

**Medicine is becoming more and more complex and humans can capture total medical knowledge only partially. Therefore merging of medical experience is important, so that therapeutic decisions less depend on the chosen doctor.**

## Current approaches

Helpful can be enhancement of the own knowledge and in certain cases special decision support systems. Already patients use text search engines like Google, similarly the doctor can do this. Also the IBM Watson system uses natural language. A universal and genuine approach would be a search engine in original data. Published are e.g. results from text search in original data
(e.g. University of Michigan, Electronic Medical Record Search Engine EMERSE, http://www.ncbi.nlm.nih.gov/pubmed/25979153 ). The approach is interesting. Problems: The transfer of individual (pseudonymous) patient data to external persons is in many countries prohibited due to data protection laws. Text search is unsharp. No "individual study".

## New precise data base

Appropriate **quantitative description** is well defined and (in case of precise data basis) much more accurate than the description by natural language. The cardinality of available quantitative spaces is much greater than that of language vocabulary. To make quantitative data searchable and suitable for decision support, they are represented by sequences of numbers as vectors. Every vector is called "**Domain Vector**" (**DV**). It is element of a metric space which is called "**Domain Space**" (**DS**). DSs can be defined by all owners of web space. The URL of the DS definition also serves in the data (in every DV) as identifier of the DS.

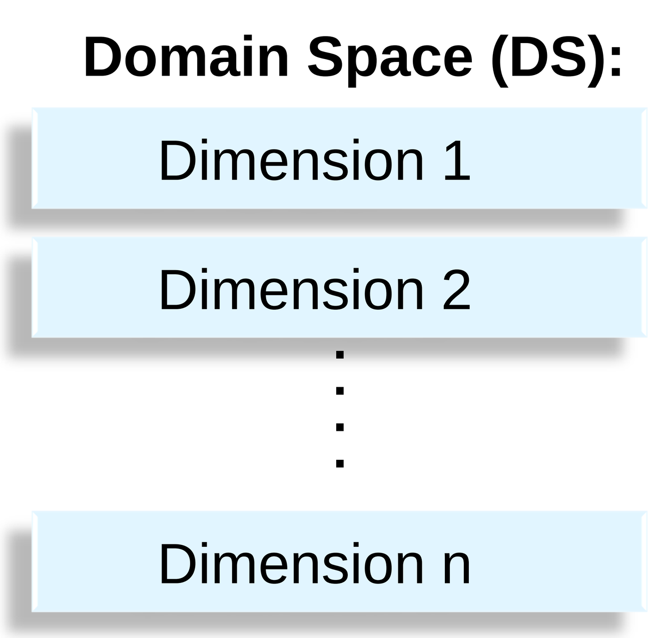


**Fig. 1:** A n-dimensional Domain Space (DS). The DS and every dimension have a URL.

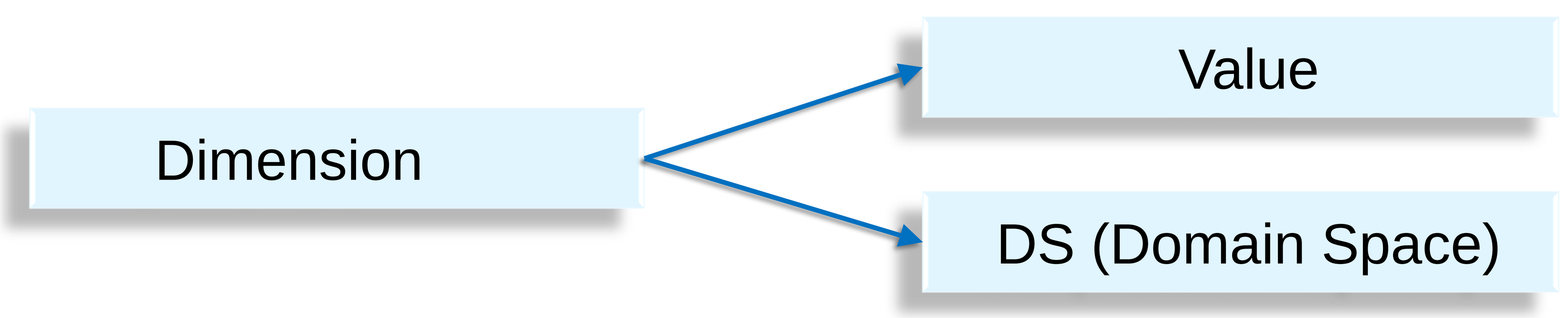


**Fig. 2:** A dimension of a DS can represent a value or again a DS, therefore DS definitions can be nested and reused within new DS definitions. Values are usually numbers (quantitative). This definition can be extended later.

Because DSs are metric spaces, the contained quantitative data (DVs) are precisely searchable. Their rank in the search result is the higher, the lower the "distance" to the searched data is. For calculation of the distance a metric or distance function is used, e.g.

$$d_1(X,Y) = \sum_{j=1}^{n} \left| r_j (x_j - y_j) \right| \qquad d_2(X,Y) = \sqrt{\sum_{j=1}^{n} \left| r_j (x_j - y_j) \right|^2}$$

(1) Manhattan dist (2) Euclidean dist.

where $X = (x_1, x_2, \ldots, x_n)$ and $Y = (y_1, y_2, \ldots, y_n)$ represent n-dimensional DVs and for j = 1..n every $r_j > 0$ the weighting faktor of dimension j. The distance d is nonnegativ. In case of d = 0 all compared dimensions are identical. The summation is done only over searched dimensions. If for example only one dimension j is searched with $r_j = 1$ (default weighting faktor), then $d = |x_j - y_j|$ .

**The approach has been implemented with search engine using a local database, online in http://numericsearch.com**

## Usage of the new data base

1. The doctor makes a first principal diagnosis, e.g. ICD. The code system should be free, reproducible and clear.
2. Using the code (later also additional finer quantitative findings) the software shows frequencies of fine (quantitative) diagnostics chosen by other doctors in such a case. This reduces the risk to overlook something important. The representation should be clear and the interface of the software user friendly.
3. The doctor decides about finer diagnostics. This includes collection of relevant collateral data e.g. about daily food intake, sports etc..
4. **The multidimensional results of finer diagnostics are provided to the software (if possible, more and more automatically). The completeness, resolution and efficiency of this quantitative information it not nearly attainable by non quantitative code systems (example: GPS coordinates compared to postal addresses). Most important quantitative results are provided to the search engine.**

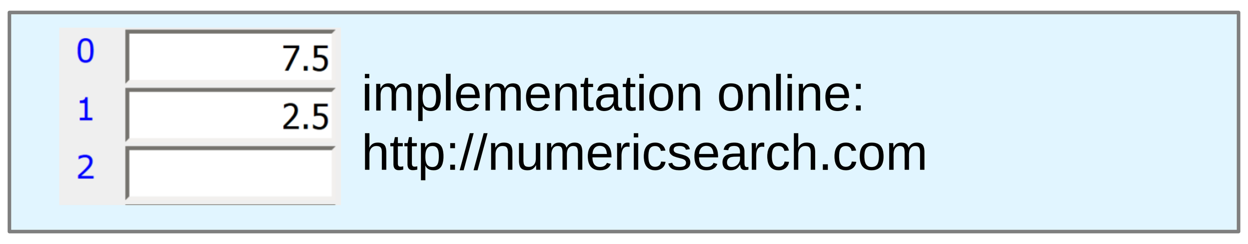


**Fig. 3:** Excerpt from the search mask of the implementation used for selection of records with similar data. The software searches in the available data collection for records (DVs) with similar data.

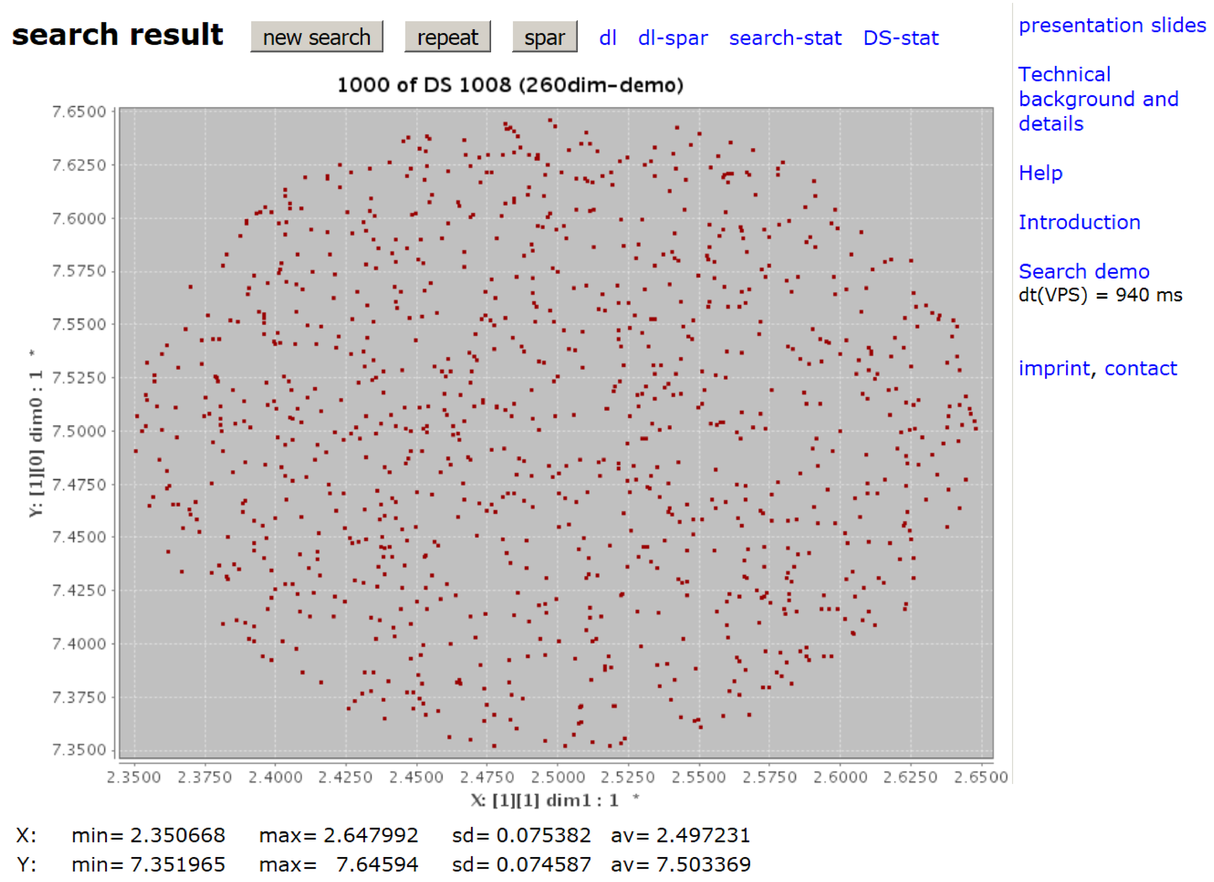


**Fig. 4: Excerpt of the search result.**
Shown are the nearest 1000 records from a 260 dimensional Domain Space with 1,500,001 records with pseudorandom numbers between 0 and 10 in the 2 searched dimensions.

If there are enough patients with similar data in the database, anonymously frequencies of further diagnoses with treatment decisions and associated results can be shown in this group of patients, like in a scientific study ("individual study").

5. Decision about further diagnostics or treatment is done and provided to software which can prepare, if wished, the draft of a medical report.
6. If necessary, later (with new data) continuation at 2. or even 1.
If the patient (or parents) agree, the own codes, decisions and results remain pseudonymously in the data collection to enlarge it. The patient can delete the own data at any time in life. The mapping of patient name to patient number is deleted at the time of death. The unerased pseudonymous data remain in the data collection and become international property. All (quantitative) data are defined online in standardized format (and so internationally uniform to ensure interoperability). Definitions are referred by their URLs. Depending on the patient's wishes, data are utilized statistically (anonymously). Data are not deleted because never "all" possibilities for medical evaluation are finished. So data collection and its medical value are growing more and more. The value of the data collection will become visible after delay.

## Conclusion

It is technically feasible to collect worldwide precise data from medical findings, therapeutic decisions and treatment results in a way, that they are anonymously available at situations where these decisions again have to be done. New medical experience can be added to the database again and again.