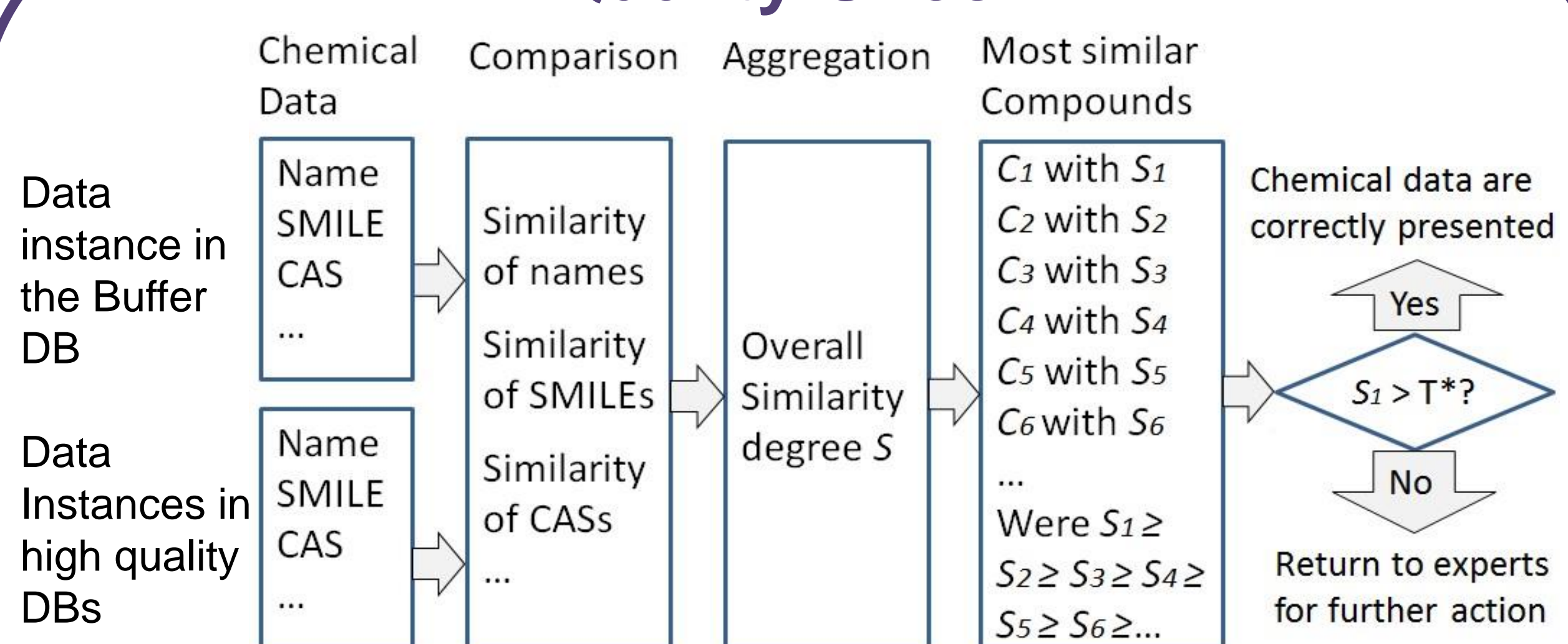


## Introduction

Due to the advance of database technologies, more and more toxicity data can be accommodated with increasingly complex data structures. Prior to the enjoyment of the convenience brought along with the abundance of data, we are facing a great new challenge: that the chemistry and toxicity data, including the metadata which describe data, need effective and efficient governance in order to be truly beneficial to final users<sup>1</sup>. This poster proposes a general framework for data reliability management, to assist in producing high quality databases, which follow gold standards practically applied and/or systematically specified by regulatory bodies, such as FDA, EPA and OECD. In particular, the multiple representations of each chemical compound in question are compared with those of each chemical compound in the so-called gold standard databases (or inventories), and the consistency result is utilised to examine the quality of chemical information; the processes of assay design, experiment execution, and experimental result documenting, are analysed based on their consistency with internationally accepted guidelines, to help to examine the quality of chemical and toxicological information. The assessed quality values can be utilised to support data fusion in the case that multiple duplicate but not identical data instances present. Two such data fusion mechanisms are also introduced in this poster.

### Quality Check



- To find out any obvious mistakes involved in a piece of collected data;
- The values of different dimensions of the piece of data in question are compared with those of data instances already stored in the existing database or some external high quality databases;
- The similarity degrees on different dimensions are combined into one;
- The piece of data in question are error-free if the highest overall similarity degree is higher than a predefined threshold; otherwise, the most similar ones are returned to the human quality controller for further action.

### Data Collection

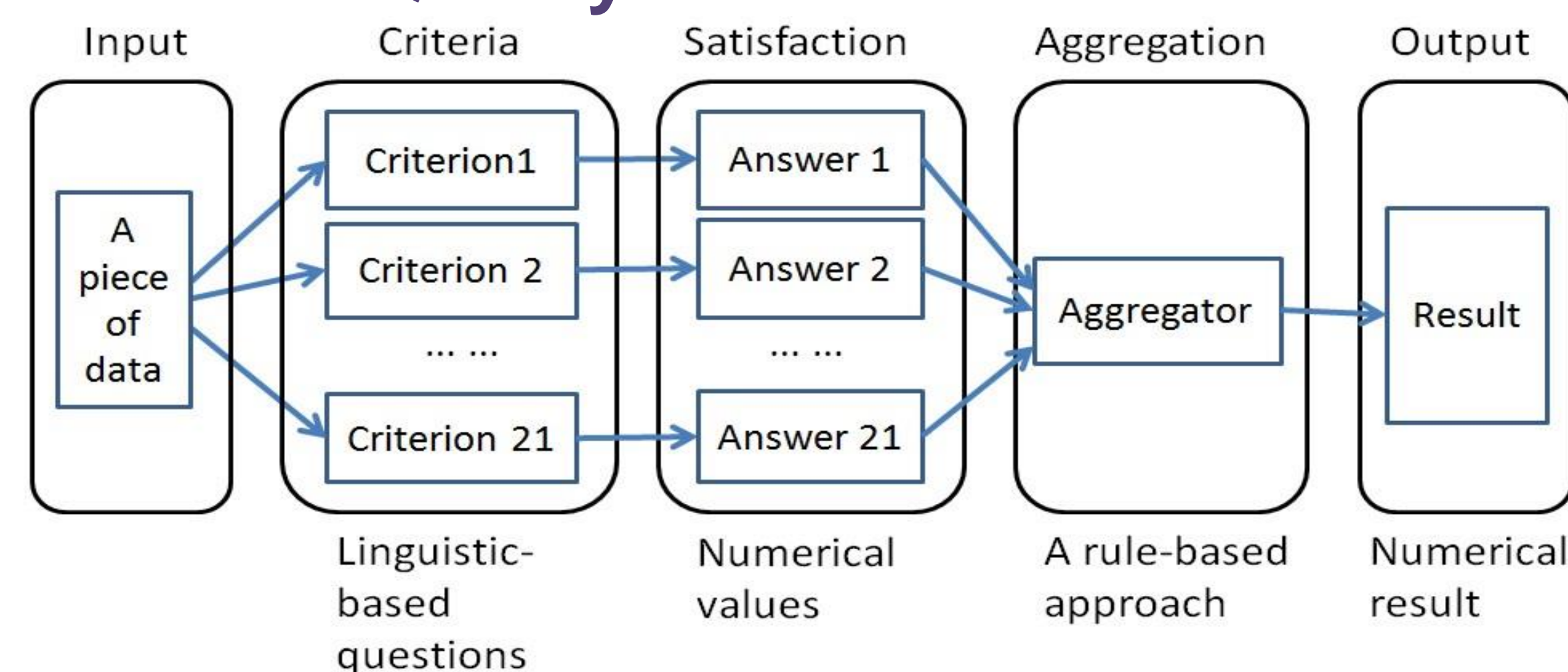
### Quality Check

### Data Deposition

### Quality Assessment

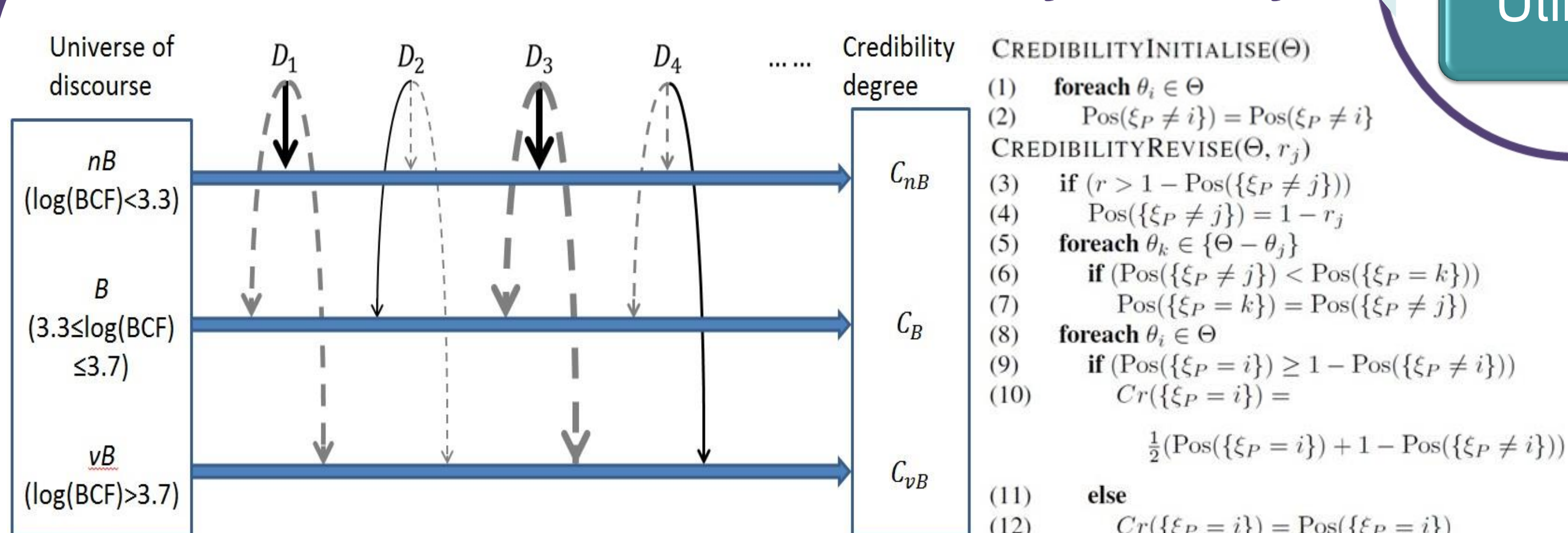
### Data Utilisation

### Quality Assessment<sup>1</sup>



- To assess data reliability to support decision-making for the cases where multiple inconsistent data instances are present;
- An extension of ToxRTool which evaluates toxicity data by 21 linguistic-based criteria;
- Extending ToxRTool by allowing partial satisfaction of criteria and considering the uncertainty of assigned partial satisfaction degrees;
- Satisfaction degrees assigned by human experts;
- A rule-based system to aggregate the assigned values to a numerical quality result.

### Data Fusion Based on Credibility Theory<sup>2</sup>



- To support decision-making with multiple data instances assisted by quality values;
- Credibility degree of a category representing the extent to which an experiment result is possibly in the category;
- Each data instance supporting one category to the degree of its quality value, and against all other categories to the same degree.
- Calculating the credibility degree of each uncertain events based on all the currently available information by applying the credibility theory;
- For any piece of new coming data, the credibility value of every category being updated by the credibility revision algorithm such that the piece of data taking into consideration.

### Data Fusion with Possibility-Probability Distribution Model<sup>3</sup>

$$U = \{u_1, u_2, \dots, u_m\} \quad P = \{p_0, p_1, p_2, \dots, p_n\}$$

A controlled point  $u_2$  is shown in the matrix. The matrix represents the possibility values  $\pi_{u_j}(p_i)$  for each controlled point  $u_j$  and probability value  $p_i$ .

- To support decision-making based on multiple data instances assisted by quality values;
- Fuzzy random variable to model the coexistence of fuzziness and randomness;
- The imprecision of probability is represented as possibility;
- The imprecision is caused by the statistics from a small data sample and the uncertainty of the data pieces themselves;
- Representing the Universe  $U$  into a set of controlled points, and representing the probability space  $P$  as a set of representative probability values;
- Applying information distribution theory to generate a possibility-probability matrix;
- Conducting mathematical analysis, such as expected values, based on the matrix.

## Conclusions

- A consistency check module based on fuzzy object comparison, to assist the human experts to quickly check if there are any obvious mistakes in the collected data before they are stored in the database;
- A quality assessment module assessing the reliability of each piece of data in the database and assigning a quality value to each of them to support decision making;
- A decision-making system based on credibility theory, to make decision based on all the currently available data instances with the help of their quality values;
- A decision-making system based on a possibility-probability distribution model, to make decision based on all the currently available data instances with the help of quality values.

## Key References

- L. Yang, D. Neagu, M.T.D. Cronin, M. Hewitt, S.J. Enoch, J.C. Madden and K.R. Przybylak. "Towards a fuzzy expert system on toxicological data quality assessment." *Molecular Informatics*. Vol. 32, no. 1, pp. 65-78, **2013**. DOI: 10.1002/minf.201200082
- L. Yang and D. Neagu. "Towards the Integration of Heterogeneous Uncertain Data". *Proceedings of the 13th IEEE International Conference on Information Reuse and Integration*, 295-302, Las Vegas, USA, **2012**. DOI: 10.1109/IRI.2012.6303023
- L. Yang and D. Neagu. "Toxicity Risk Assessment from Heterogeneous Uncertain Data with Possibility-Probability Distribution." Under review for potential conference publication.