

# A Theoretical View on Reverse Engineering Problems for Database Query Languages

Pablo Barceló

University of Chile (CL)

**Abstract.** A typical reverse engineering problem for a query language  $L$  is, given a database  $D$  and a sets  $P$  and  $N$  of tuples over  $D$  labeled as positive and negative examples, respectively, is there a query  $q$  in  $L$  that explains  $P$  and  $N$ , i.e., the evaluation of  $q$  on  $D$  contains all positive examples in  $P$  and none of the negative examples in  $N$ ? Applications of reverse engineering problems include query-by-example, classifier engineering, and the study of the expressive power of query languages.

In this talk I will present a family of tests that solve the reverse engineering problem described above for several query languages of interest, e.g., FO, CQ, UCQs, RPQs, CRPQs, etc. We will see that in many cases such tests directly provide optimal bounds for the problem, as well as for the size of the smallest query that explains the given labeled examples. I will also present restrictions that alleviate the complexity of the problem when it is too high. Finally, I will develop the relationship between reverse engineering and a separability problem recently introduced to assist the task of feature engineering with data management tools.