# Capita Selecta AI - Entity Relationship Model - Part 2

Daan Fierens

daan.fierens@cs.kuleuven.be

## 1 Part 2 - Learning

Remember from the first part of the assignment that all probabilistic dependencies were specified in terms of some unknown probabilities $p_1$ to $p_{16}$. For the first part of the assignment (modelling) you had to choose concrete values for these probabilities yourself. For the second part of the assignment, we generated some datasets by first modelling all the dependencies (using our own values for $p_1$ to $p_{16}$) and then sampling from the resulting model.

Your task is now to find the probabilities $p_1$ to $p_{16}$ that we used to generate the datasets. To do this, take the model that you constructed in the firt part of the assignment, and apply the parameter learning algorithm in your system with our datasets as input.

### 1.1 Assignment

#### 1.1.1 Learning from Complete Data

- Learn the values of $p_1$ to $p_{16}$ by applying your parameter learning algorithm to the data in `ER_train_complete.dat`. Which values do you find?
- Using the learned parameters, calculate the likelihood of the data in `ER_train_complete.dat` (the *train data*), and of the data in `ER_test.dat` (the *test data*).[1]
- Repeat these two steps but now use only part of the training data. Concretely, instead of using all 100 examples in `ER_train_complete.dat`, use only the first 60 examples, and then only the first 30 examples. Compare the likelihood on the train and test data for 30, 60 and 100 training examples.

#### 1.1.2 Learning from Incomplete Data

Repeat all the above steps but now use the file `ER_train_incomplete.dat` for training. This file contains an incomplete dataset in which the values for some of the properties/relations are left unspecified. This dataset was constructed by sampling from the correct model, and then randomly removing $\pm 25\%$ of the sampled values.

### 1.2 Remarks

Each of the datasets that we provide is a set of examples, with each example being a set of facts. Some more information about the precise format is given in the readme file that is included in the data. Note that you might have to convert the data to another format before you can use it in your system.

The datasets that we provide do not satisfy the hard constraint given in point 10 of the modelling assignment (that each house is bought by at most one customer). If you have included this constraint in your model, please remove it before you perform learning.

[1] If your system does not calculate likelihoods, ask your supervisor what to do).