Data mining: finding relevant info from data

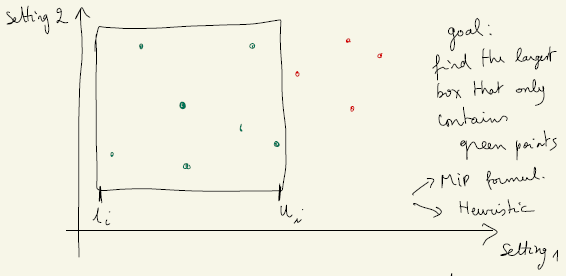
-> retrieve variables that influence an output variable

-> find set of rules (intervals on the variables) for which output is maximized

Formal Definition:

* D: input variables
* x\_i , a datapoint, a vector of input variables for item i *(all except last column)*
* c\_i, the output variable of item i *(last column of the file)*
* C, a given threshold
* X, the set of data points for which c\_i > C
* Y, the set of data points for which c\_i < C
* B, a box defining a set of rules over data points
  + intervals defined by vector l and u
  + *lower&upper bounds for each input variable*
  + 

Notes:

* C is a threshold put over the output variable to determine what is “good” (c\_i < C) and what is “bad” (c\_i > C)
* 

**[Q1] MIP Problem: Find vector u and l such that **

**Objective function:**



**Let:**

**Constraints:**