# Mineração de Dados e Descoberta de Conhecimento

*Profs. Heitor S. Lopes e Thiago H. Silva (UTFPR, 2025) - Exercise #3*

## A) Very-low-weight babies

**Objective**: The very-low-birth-weight babies dataset was constructed with data from 671 babies with birth weights below 1600 grams. There are data about the mother and the baby, including: some medications used, issues related to the childbirth, problems diagnosed in the unborn child, and whether or not the baby survived the birth. In the medical literature, some factors that **decrease** the survival chances of premature babies are already known, for example: babies with extremely-low birth-weight (<1000 grams), those born very prematurely (<27 weeks of gestational age), and the presence of comorbidities such as pneumothorax or cardiopulmonary complications resulting from lung immaturity itself. Therefore, it is desired to use association rule mining to discover **relevant associations** that are **not obvious**.

a. Orange software can be used. In this case it is necessary to reformat the dataset as the “basket” format . For this purpose, see further information here and here. Opcionally, Python can be used too.

b. Instead of generic rules, there is special interest in discovering rules that show some association between the use of **beta-methasone** but the pregnant woman and the survival, or not, of the low-weight baby. Observe the rules found by the method and, then, investigate whether the result has any scientific basis (Search Google!!).

c. Important fact: This dataset was collected at Duke University Medical Center in the state of Georgia (USA), which is one of the American states with the largest concentration of African-Americans. Keeping this fact on mind, obtain association rules (**with at least 20% support**) that involve a possible relationship between **race** and the prevalence, or not, **of low-birth-weight infant deaths**). Discuss the social implications of the findings.

# B) Análise Associativa (11) OPTIONAL CHALLENGE!

**Objetive**: The objective is to use real data about car accidents on federal highways (provided by the Federal Highway Police – PRF) that cross the Paraná State, to find relationships between the attributes in order to explore the dataset and obtain non-obvious knowledge regarding accidents with **fatal victims**.

a. Import the data into Orange, selecting **only fatal accidents**. Display the visualization of the accident points on the map of the State of Paraná, using the geographic coordinates of the accidents. This can be easily done with the **Geo Map gadget**, with the color of the points given by the variable "**Main Cause**". Using the graphical analysis of the map, report the most frequent causes of accidents with fatal victims in the three largest metropolitan regions of the State (Curitiba, Londrina, Maringá). Then, compare (**qualitatively**) with the causes of accidents on **highways**, outside the metropolitan regions. Are there significant differences?

**b.** Select only **qualitative** variables and try to generate **frequent itemsets** with the highest possible support and that contain: **Accident classification = 'With fatalities'**

c. From the itemsets, look for rules with a minimum **support of 40%** and **confidence >= 90%**. Of these rules, recall that only those with the consequent Accident Classification = 'With fatal victims' are of interest.

d. Analyze the results obtained and report the **non-obvious knowledge** obtained from the analysis, for metropolitan regions and for regions outside them.