1. **Qué ofrezco y en qué contexto**

Highly frequent co-occurrences of data items have been a target for several types of data mining frameworks for decades, in all types of data, including medical data. Most commonly studied notions for this sort of analysis are frequent sets and variants thereof, such as frequent closed sets or association rules. In most cases, frequent set mining actually returns a large textual list; then, a visualization of this result can allow us to achieve a better understanding of the data, as well to discover implicit and potentially useful information. But these notions are not reaching out to end users, mainly due to the difficulty of finding explanatory descriptions. In fact, the results of these notions of data analysis are found in spaces of huge dimensionality, and their reductions almost never offer enough interpretability.

For a useful interpretation and better understanding of data, I propose to employ decomposition techniques on Gaifman graphs as a graphical tool for exploratory data analysis on medical data. The Gaifman graphs record the co-occurrences of data items in datasets and, then, graph decompositions provide valuable information that is not directly observable on the data, since the Gaifman graph decompositions displays a hierarchical visualization of the co-occurrences.

Gaifman graphs are a notion originated in mathematical studies of logical structures that, actually, can support a relational database model. Gaifman graphs can be advantageously employed for exploratory data analysis via their decomposition in terms of so-called clans, either in their original form or in one of its variants.

The decompositions of Gaifman graphs have the potential to reveal “co-occurrence” patterns or, alternatively, “incompatibility” patterns, since they relate together data items that, pairwise, appear together somewhere in the whole dataset. Generalizations of the notion allows to adjust the co-occurrence thresholds so as to account for particularly frequent joint occurrences, or for different intervals of co-occurrence counts.

Specifically, in the presentation I will focus on a variant of Gaifman graph, the exponential variant, that shows how this combinatorial technic is applicable to medical data corresponding to the joint diagnostics of patients. I believe that the resultant visualization, like others that can be created with this technic, can act in a useful way complementing existing statistical approaches; for example, by pointing out specific pairs of elements, possibly conditioned to other elements, whose correlation studies could be candidates for priority analysis.

1. **Porqué es importante para mi interlocutor/para otros**

The construction of data sets like the one I have worked on is not simple. Initially, many diagnoses are expressed as natural language expressions, and the information with ICD coding require a separate and subsequent process with often specialized people or even companies. I think that a graphical tool trained in frequent concurrent diagnoses can help accelerate this type of process by offering common options for automatic completion and/or by checking for double verification of rare cases, which could be either correct or the result of coding errors (as extreme example, a co-occurrence between the prostate surgery with a normal delivery).

1. **Como lo he hecho/hago/haré**

The graphical tool developed is based on the Gaifman graphs, a mathematical structure introduced in Logic, that can be used in an advantageous way for exploratory data analysis. The method used is a modular decomposition method, in terms of the so-called strong clans.

The construction of a Gaifman graph is based on co-occurrence, or lack of it, of items in the medical data. From it can be obtained a complete graph with an equivalence relation among its edges (so-called 2-structure). The strong clans allow to decompose a Gaifman graph into a tree-like form.

The dataset that I used for the development and tests of the graphical tool, was provided by the Hospital de la Santa Creu i Sant Pau. The dataset contains information of all hospitalizations for the years 2015-2016. Specifically, it was analysed a part of this dataset, I used the information about the diagnostics and procedures. The global dataset consists of 7741 values (patients). Therefore, I has been necessary to consider a minimum frequency threshold, since I couldn't display all values graphically. For example, in the graph that is showed in the presentation, I was needed to consider 100 as a threshold. This means that diagnostics and procedures that appear less often than 100 times were not taken into account for the visualization.

With the graphical tool I can demonstrate a novel application of Gaifman graphs and their decomposition, providing a general visualization of the medical data behaviour that could be used as a tool to complement statistical approaches. Indeed, in order to obtain the results, I had to carefully observe the behaviour of the data. In general, at the moment, the human brain us essential during the exploration.

That said, I believe that my graphical tool can act in a useful way complementing the statistical approaches, as an example among many others, it could point to the medical staff specific pairs of elements possibly conditioned to other elements, whose correlation studies could be candidates for priority analysis.

In future contributions, we would like to offer self-descriptive, more informative, perhaps even animated, visualizations that trained medical staff can immediately capture.

1. **Qué pido y a quién**

I am a recent PhD who has just created a Spin-off with my thesis co-directors, and I am looking for a collaboration project for conducting research studies with the Hospital de la Vall d'Hebron in Barcelona. I'm asking to the director of this hospital, who is represent by my teacher.