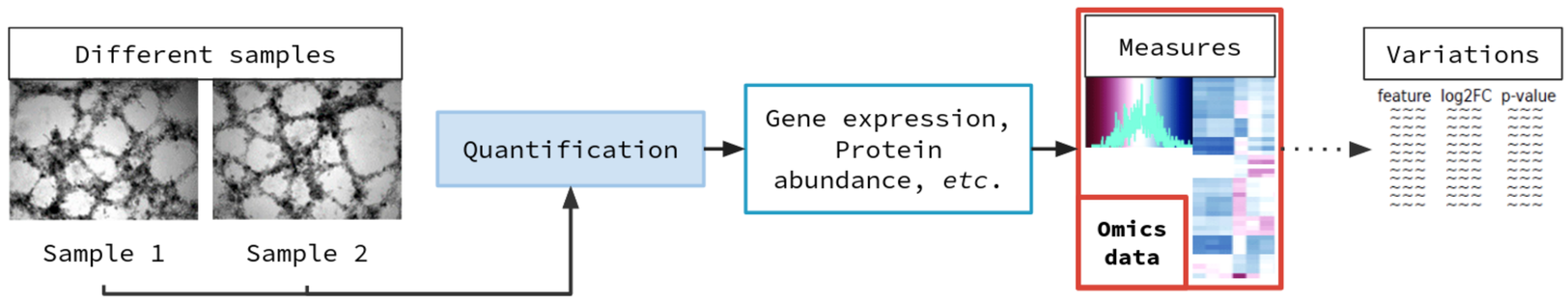
# Evolving communities detection in node-attributed networks

Nina Singlan* and Claude Pasquier

* Corresponding author : nsinglan@i3s.unice.fr

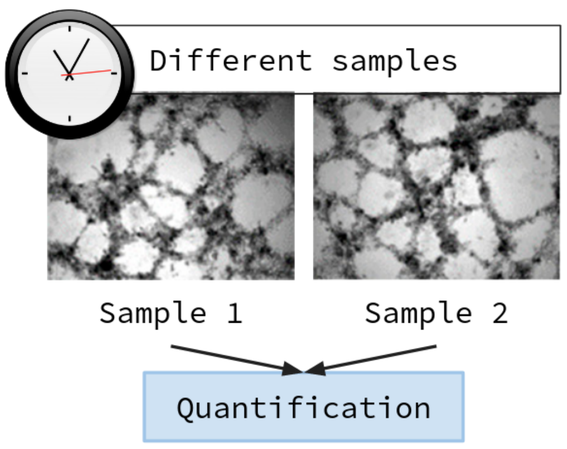
## 1. Origin of data

Now, with the new molecular biology tools, it is possible to measure, at the scale of a whole genome, the changes that occur at the cell level under different experimental conditions. This represents a huge amount of data and Artificial Intelligence (AI) can help us to identify, in this mass of data, the key elements that are important in the changes that occur.




## 2. Temporal aspect of data

The values obtained by quantification are time-dependent. For example, some genes may be essential at the onset of the disease and become inactive during its evolution. And vice versa...

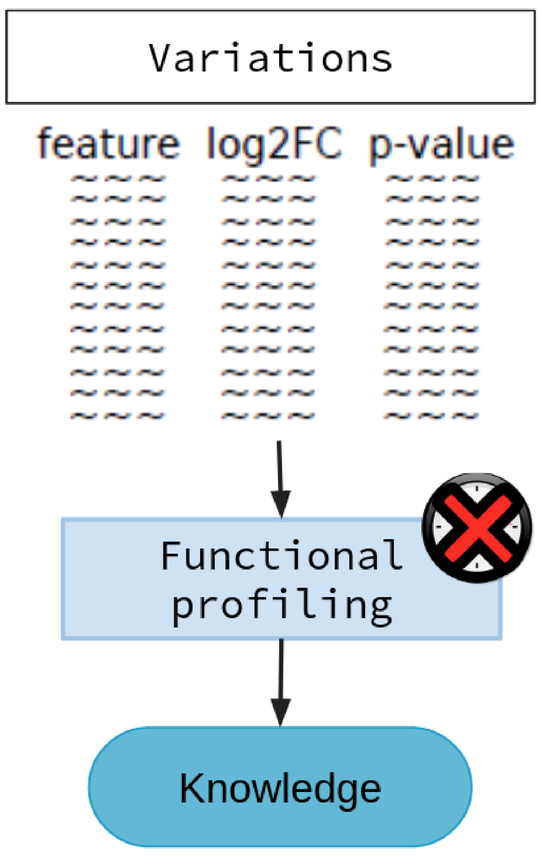



## 3. Limit of standard data analysis

Standard methods of data analysis use the genes with the highest scores, that is, those that vary the most. These methods may miss groups of genes that have low individual scores but are together involved in the process being studied.

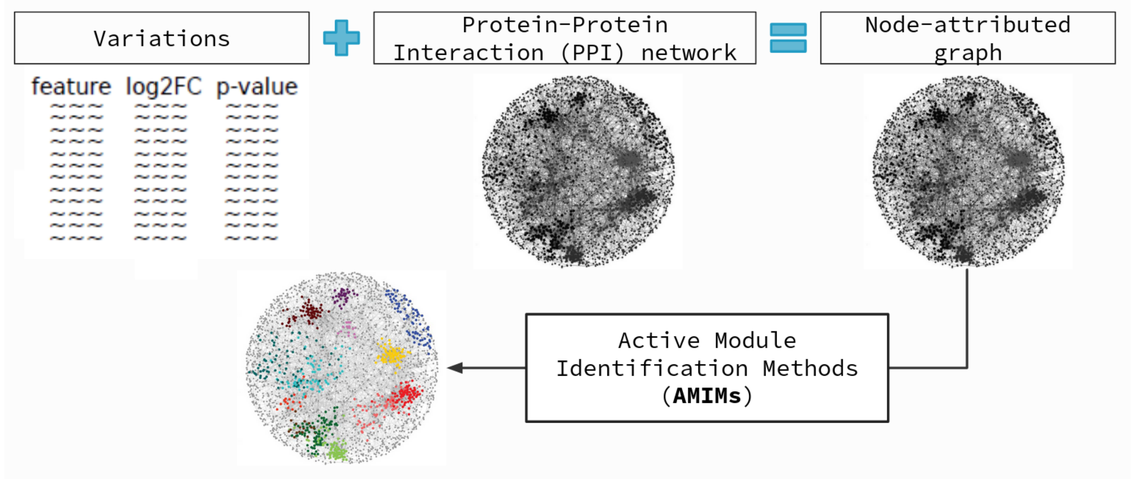
*A small but coherent difference in the expression of all the genes in a pathway should be more significant than a larger difference occurring in unrelated genes.* [1]

Moreover, these methods do not take into account the temporal dimension of the data. Important genes are identified for each time point separately.




## 4. The biological problem as a graph problem

The interactions between genes can be represented by a graph in which each vertex represents a gene and each edge represents an interaction. For this purpose, Protein-Protein Interaction (PPI) networks can be used. Each gene can be associated with its variations measured at different time points. This results in a graph whose nodes are associated with values. The problem then consists of finding modules in this graph (of several thousand nodes) and succeeding in following their evolution over time.



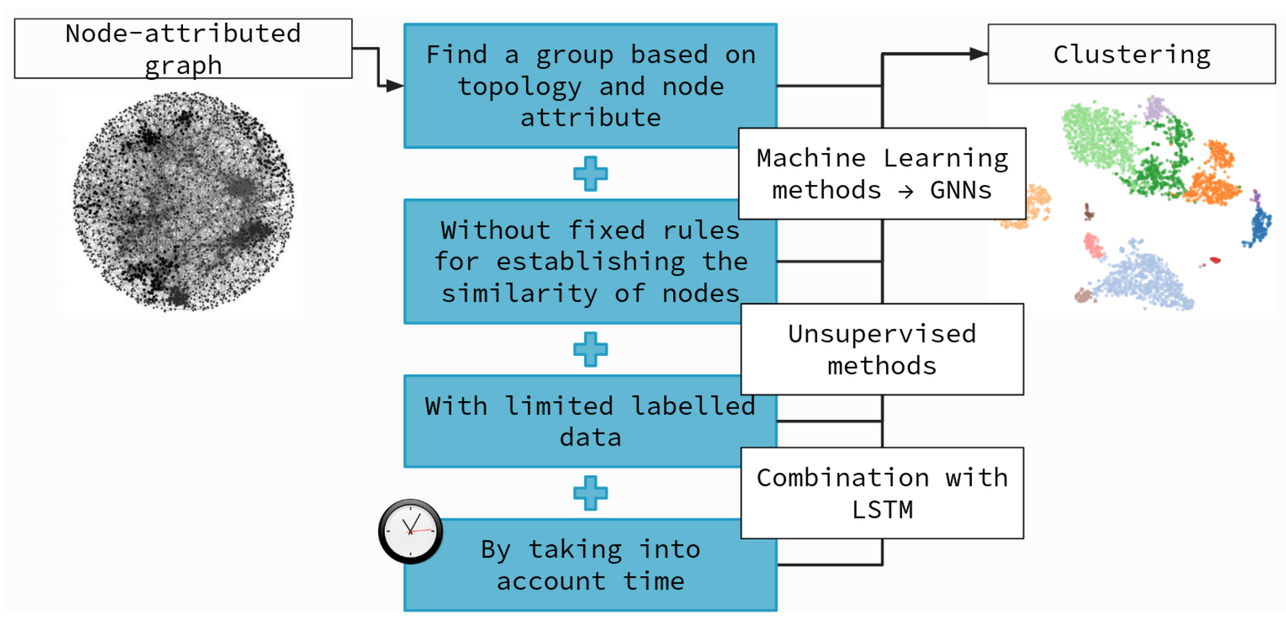

## 5. An unusual clustering problem

To identify groups based on the nodes and the topology of the graph, it seems natural to look at clustering techniques. However, in computer science, topologically speaking, a cluster has a precise definition: the nodes belonging to a cluster have more edges between them than with other nodes. Unfortunately, this is not the case in the active modules. Existing methods and metrics in the field of clustering are therefore not usable on this data. Moreover, there is no consensus on the rules governing group membership. It is therefore necessary to start by re-defining a cluster in this specific context.

**Gene communities are not clusters in the computer sense**

## 6. Proposed solution

As the membership of a module is not strictly defined, it seems interesting to use machine learning techniques. Graphs Neural Networks (GNN) have been specially designed to work on this data type. As we have little labelled data, unsupervised methods seem appropriate and some algorithms already offer unsupervised learning methods on graphs with GNNs [2]. The technique will be combined with a Long-Short Term Memory (LSTM) to track the evolution of the modules over time.




## References

[1] Franck Rapaport, Andrei Zinovyev, Marie Dutreix, Emmanuel Barillot, and Jean Philippe Vert. Classification of microarray data using gene networks. *BMC Bioinformatics*, 2007.

[2] Anton Tsitsulin, John Palowitch, Bryan Perozzi, and Emmanuel Müller. Graph clustering with graph neural networks, 2020.