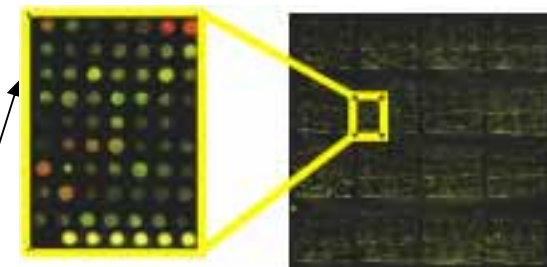
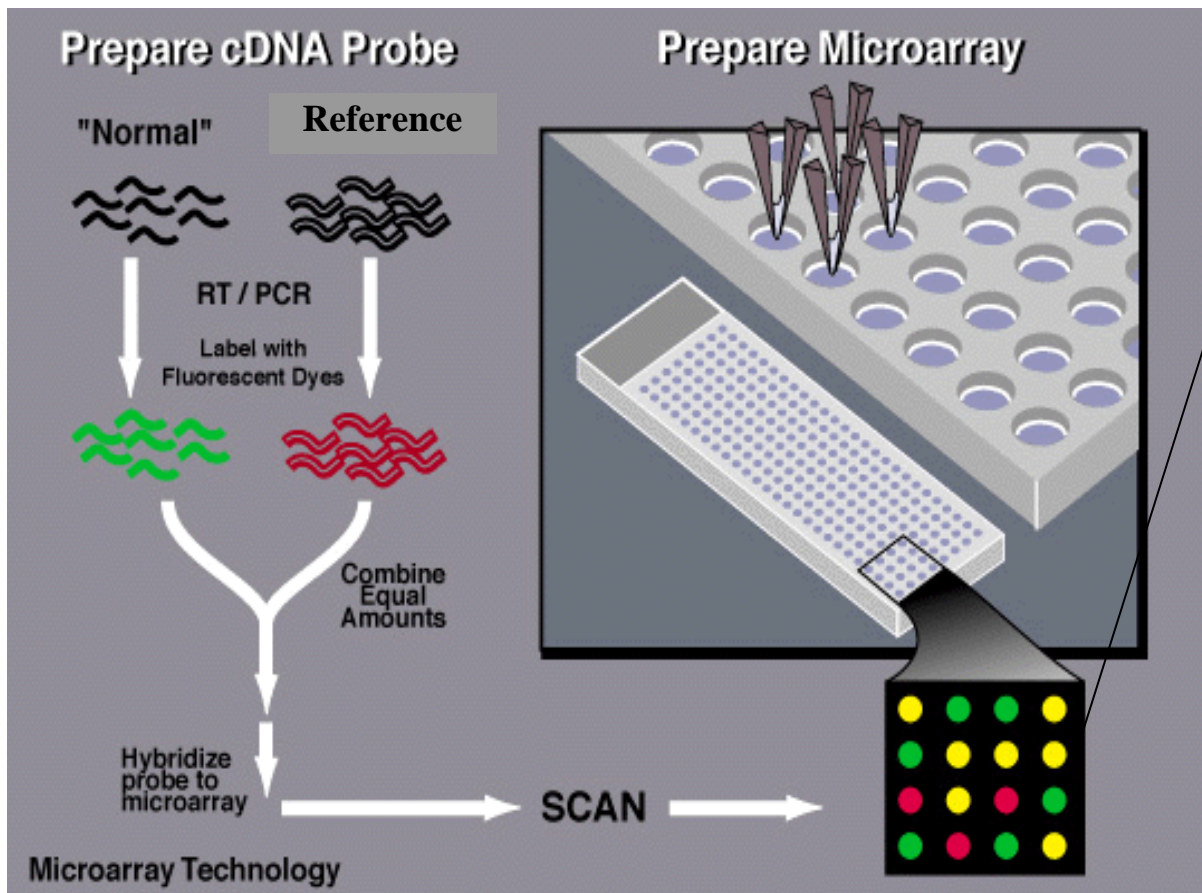


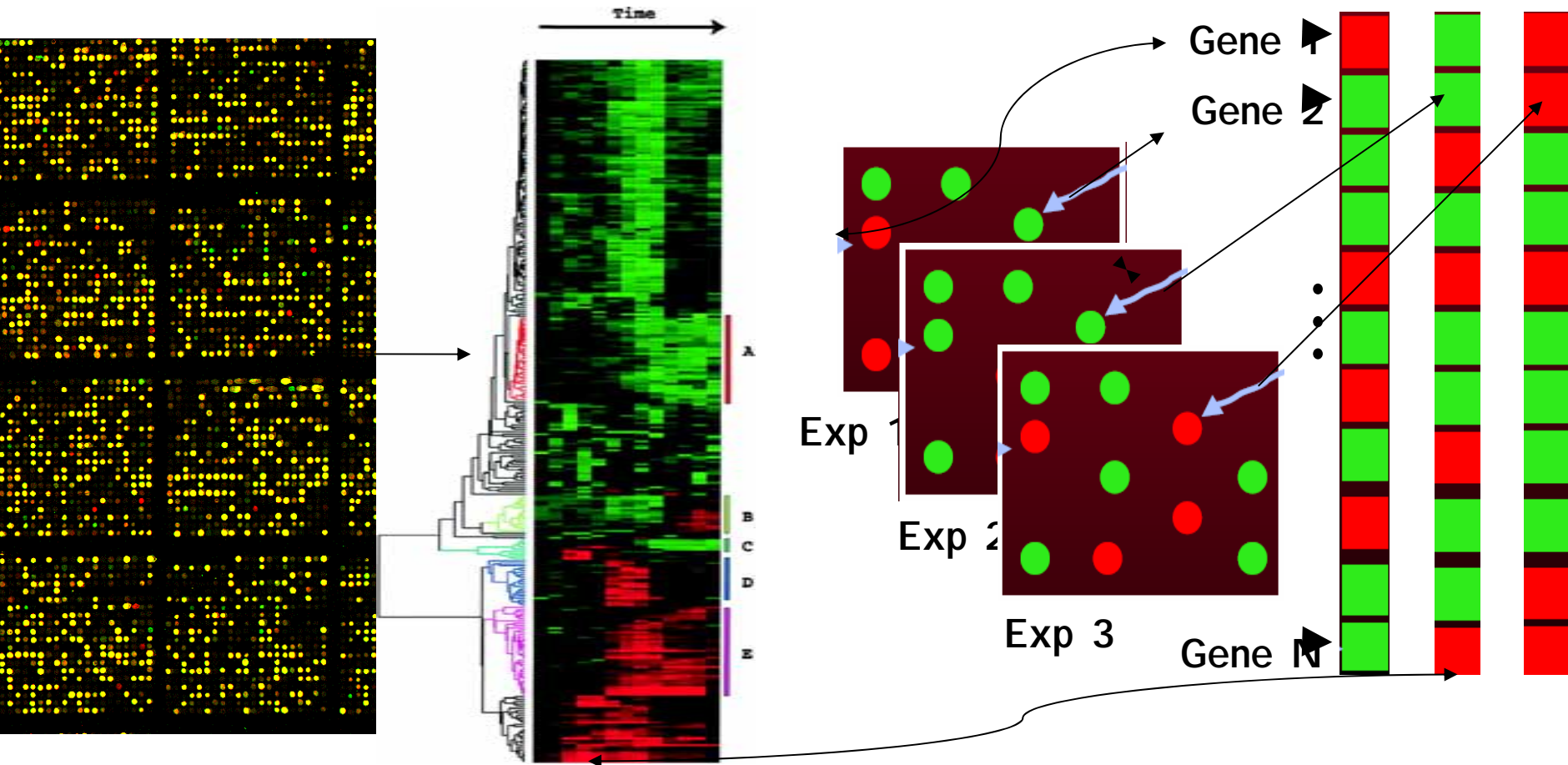
Transcriptomics
The genome-wide study of gene expression levels, measured in the variation and presence of mRNA

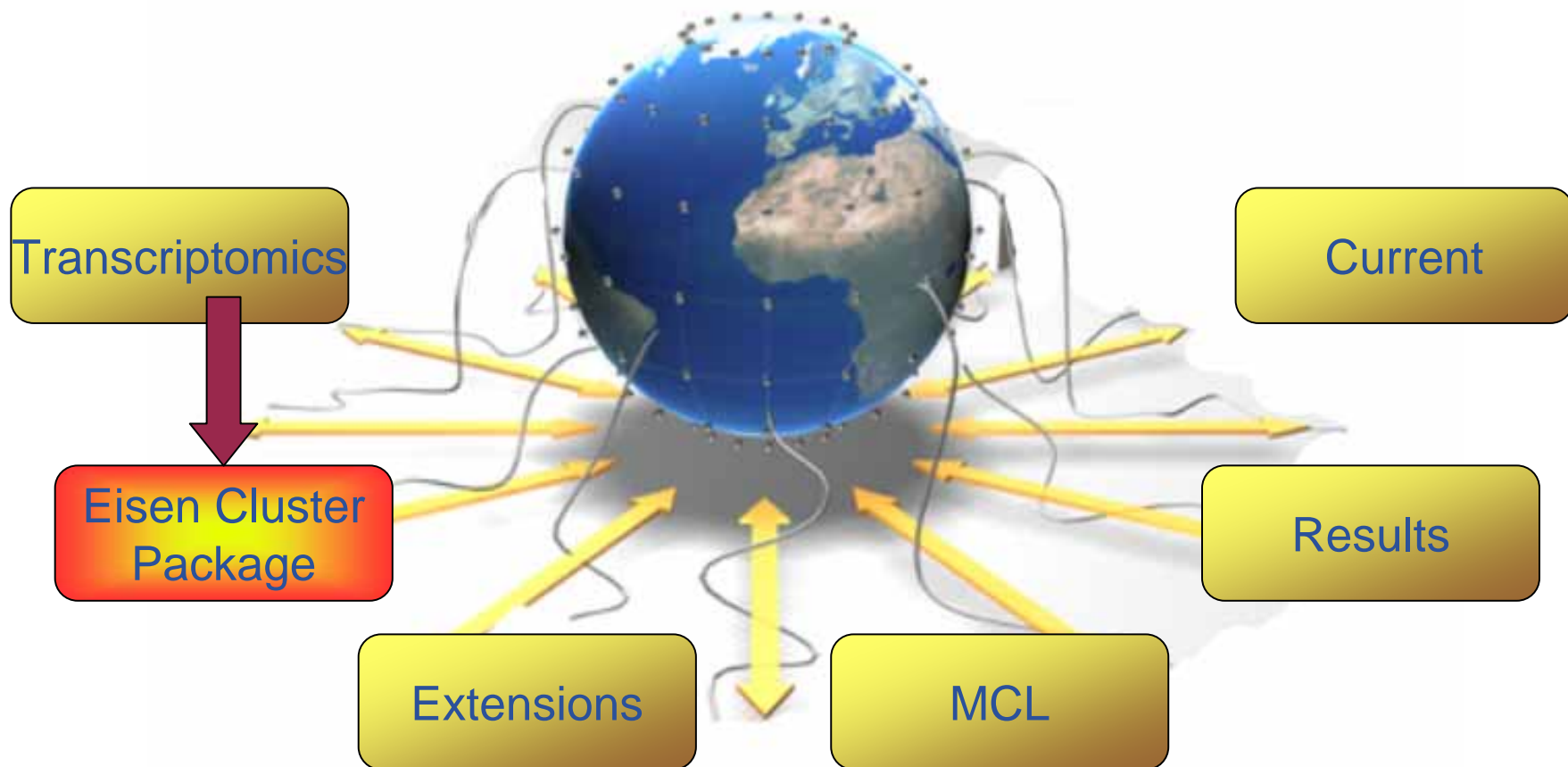


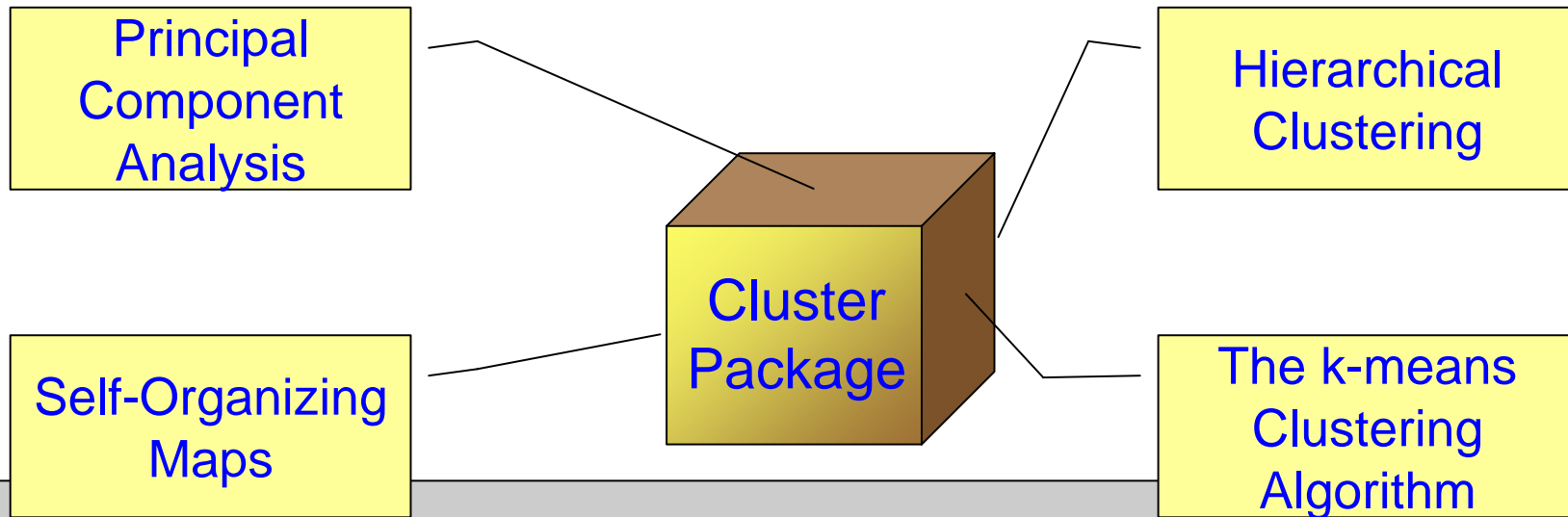


**Calculate intensities
and ratios**

YORF	0 minutes	30 minutes	1 hour
YAL001C	1	1.3	2.4
YAL002W	0.9	0.8	0.7
YAL003W	0.8	2.1	4.2
YAL005C	1.1	1.3	0.8
YAL010C	1.2	1	1.1





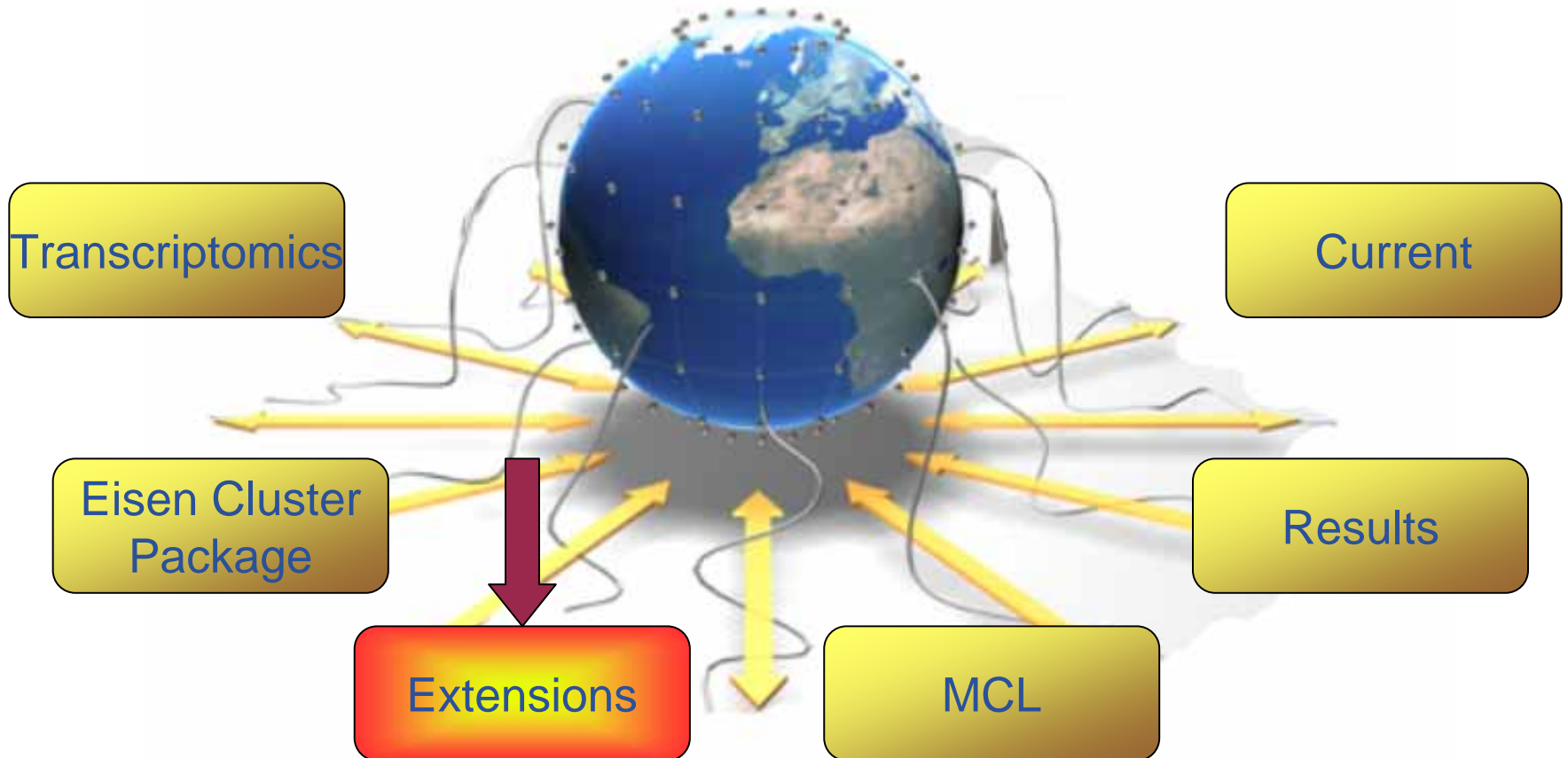


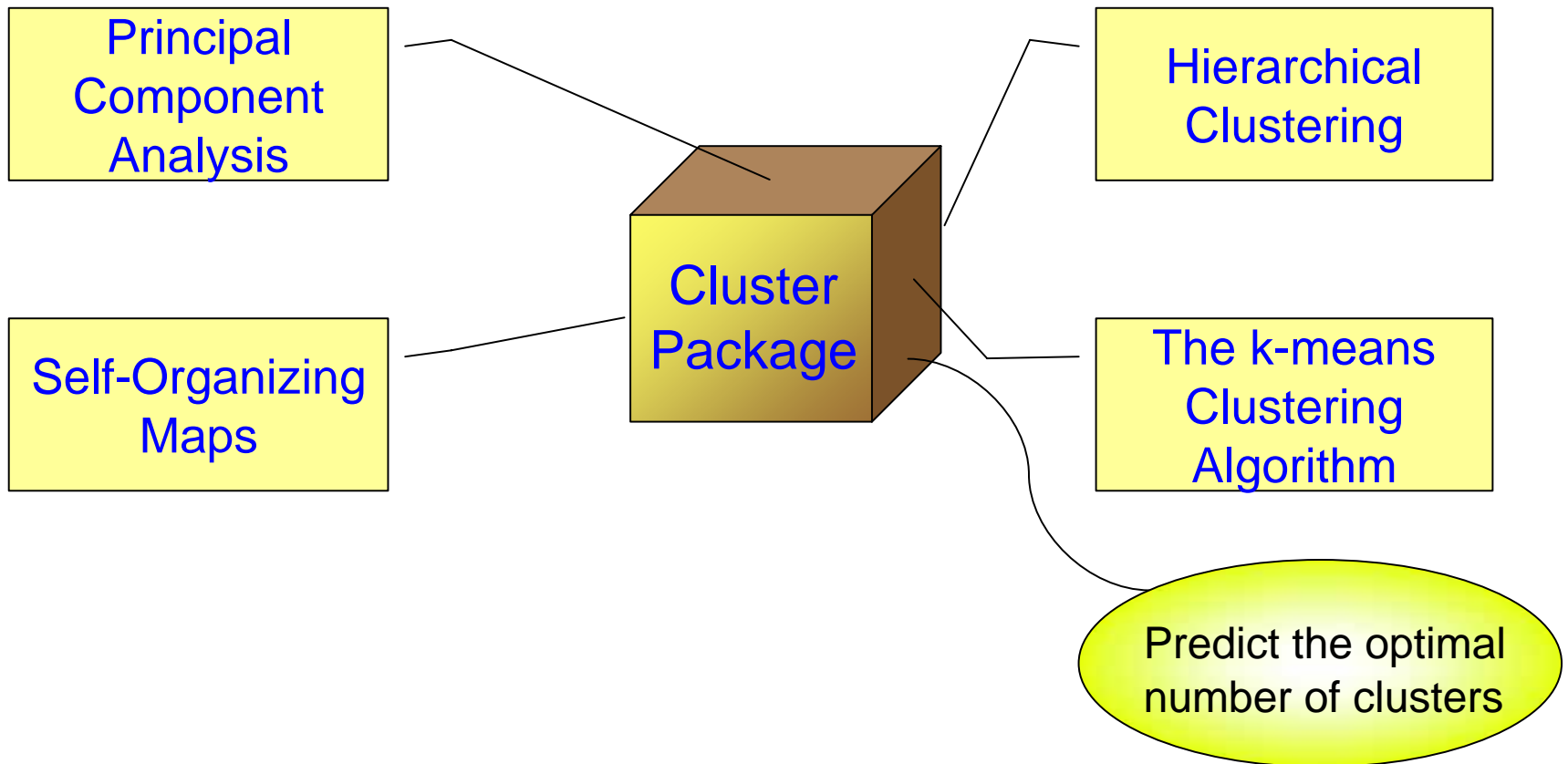
Distance measures

- Pearson's Correlation Coefficient
- Absolute PCC
- Uncentered PCC
- Absolute Uncentered PCC

- Spearman's rank
- Kendall's r
- Euclidean distance
- Manhattan Distance







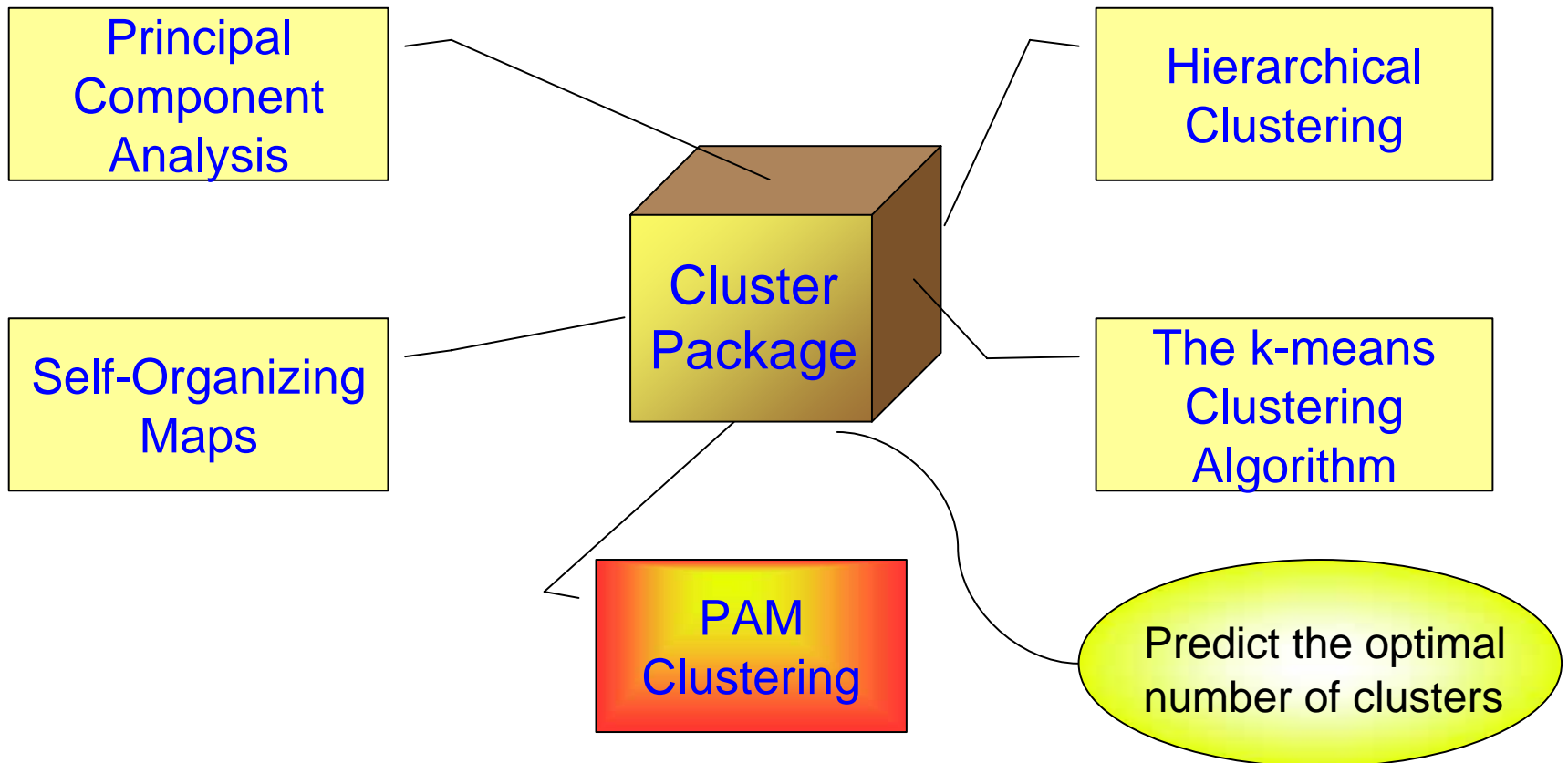
Many different methods have been proposed to serve this goal but none can be characterised as flawless the method we chose to implement for the purpose of this project was introduced by Kaufman and Rousseeuw.

The Silhouette widths algorithm shows whether an observation lies well within the cluster it is placed in, whether it falls between two different clusters or whether it should actually belong to a different one.

By doing so for all observations or samples it calculates the average correctness of a clustering result. If run for different number of clusters, it gives an indication of which number gives the most successful clustering on average.

Predict the optimal number of clusters

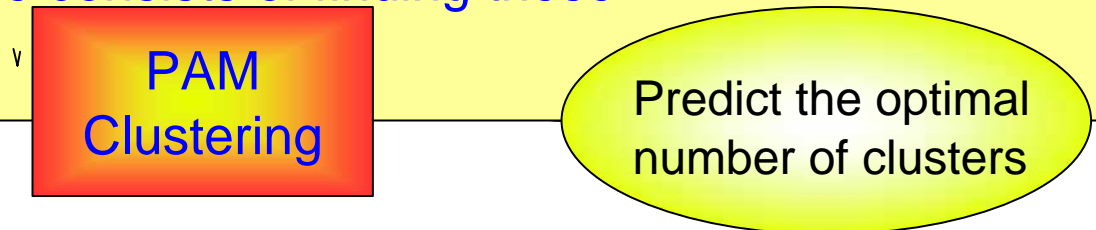


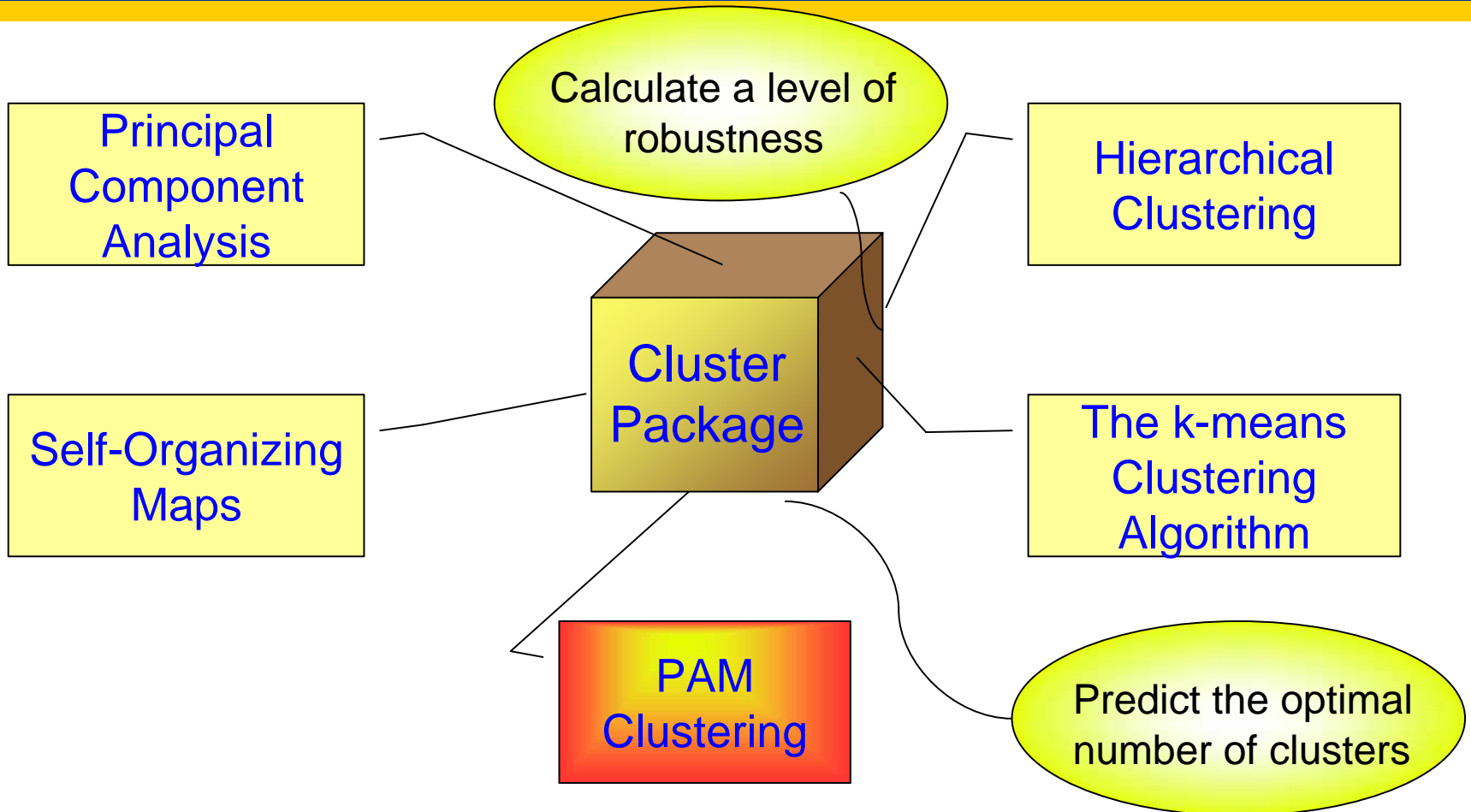


The partitioning around medoids algorithm (PAM) was implemented in order to serve as the clustering algorithm for finding the optimal number of clusters in a dataset.

PAM clusters the data around a number of representative objects called centroids or medoids.

A medoid is the object of the cluster for which the average dissimilarity to all other objects in it becomes minimal. The algorithm's main phase consists of finding these representative objects.





Calculate a level of robustness

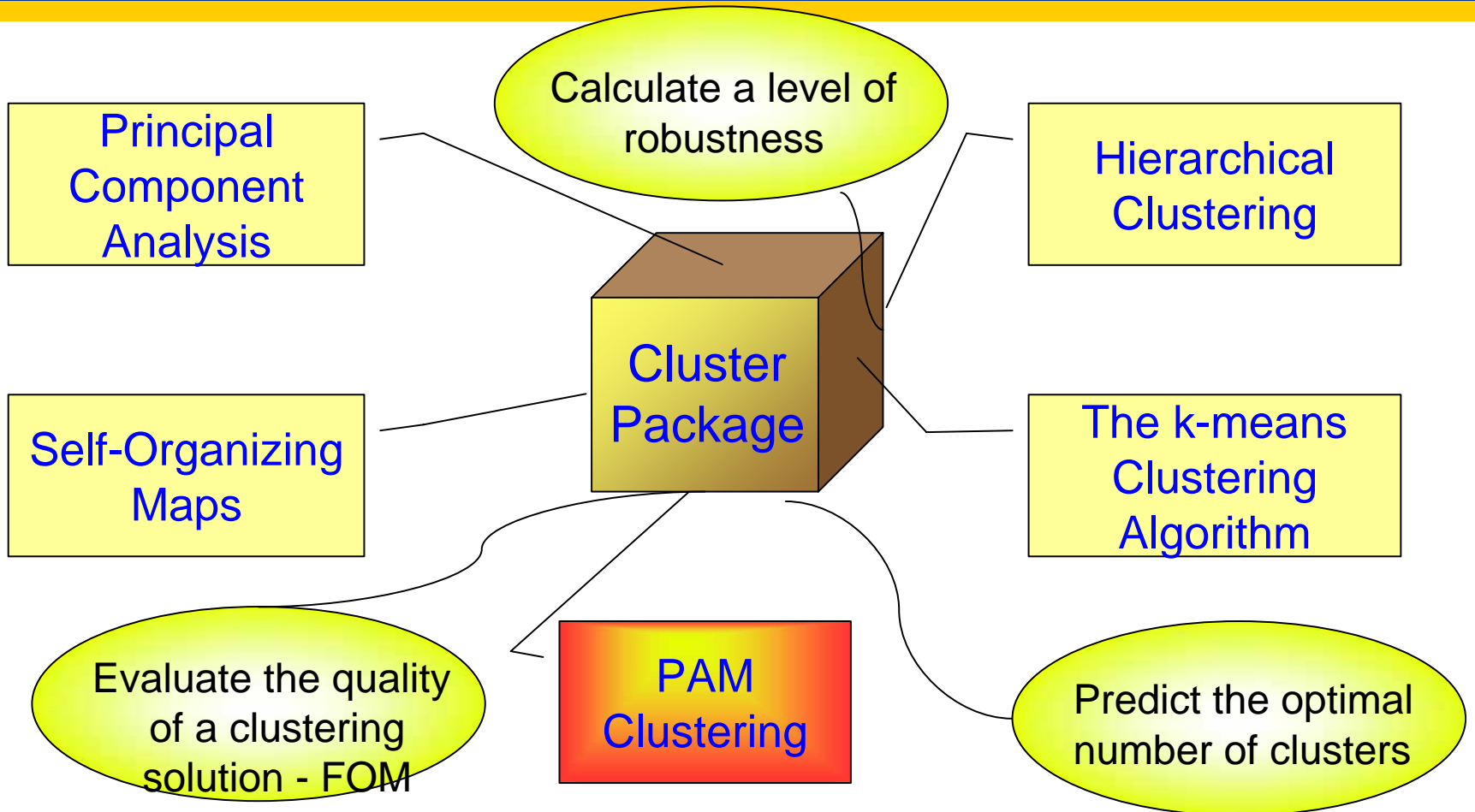
The leave-one-out cross-validation technique was implemented to depict a measure of robustness.

The clustering process is repeated n times, where n is the number of different microarrays (samples or observations).

Each time a column of the dataset is left out in turn, and the clustering algorithm takes as an input the remaining $n-1$ columns. Then a level of robustness can be defined, and then only the genes clustered together a user defined amount of times will be found in the final clustering.

This is considered to be especially useful when the dispersion of the distribution is wide or extreme values are expected to be present in the dataset.





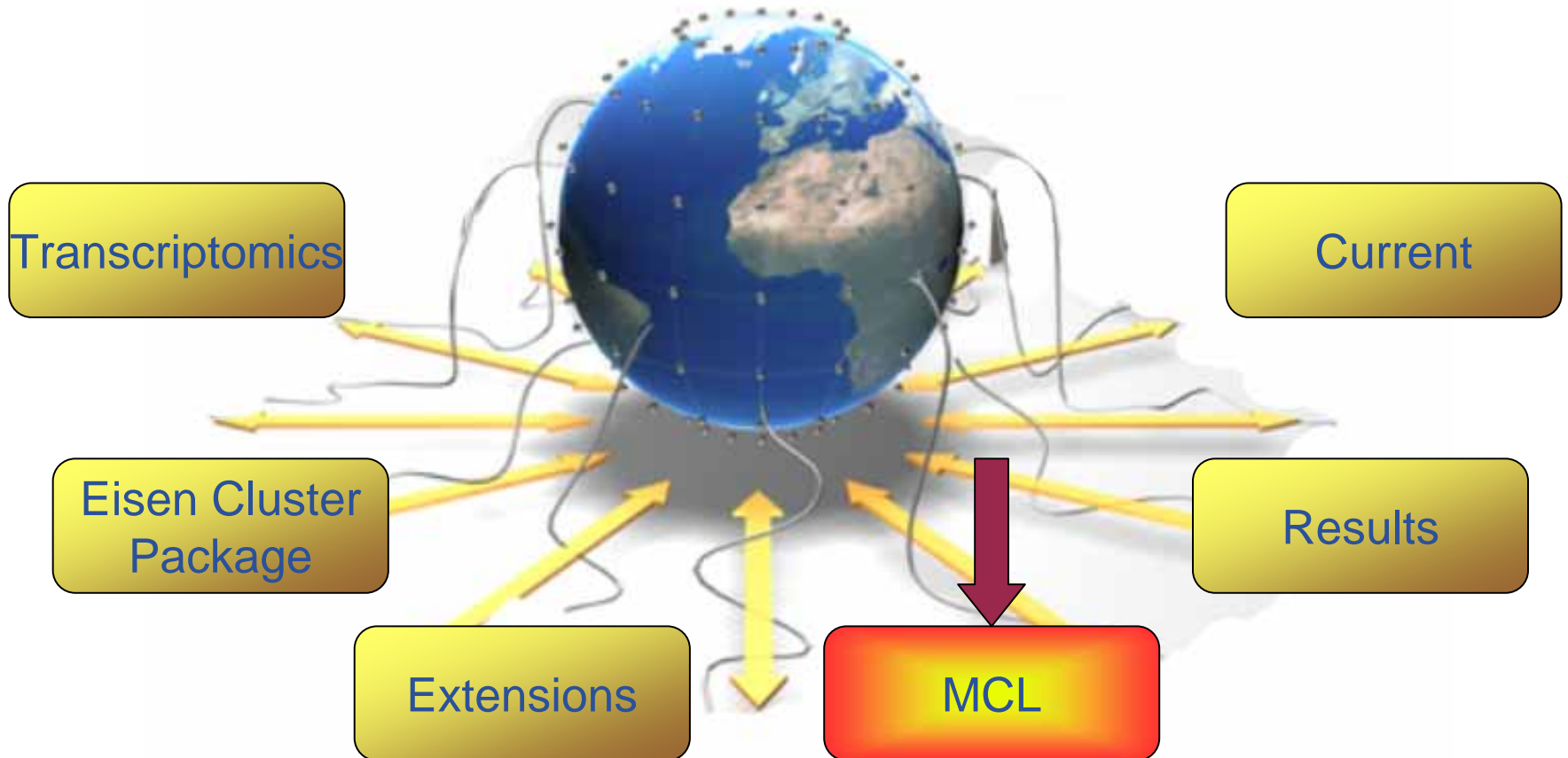
A number of different approaches have been followed to provide a data-driven framework of comparing clustering algorithms, the one implemented here is presented in Yeung et al [15]

They apply a clustering algorithm to all but one column (condition) of a dataset, iteratively, until each column has been left out in turn.

Then they define a scalar measurement to predict the significance of a clustering algorithm which is called figure of merit (FOM). FOM is calculated for each left-out column as the root mean square deviation of the gene expression levels in that column for each gene in relation to the cluster mean. The aggregate FOM is calculated by averaging the sum of all FOMs over the total number of genes in the dataset. The less the FOM is the higher the predictive power of the algorithm.

Evaluate the quality
of a clustering
solution - FOM





MCL - Markov Clustering Algorithm (VanDongen, 2000)

MCL does not require the number of expected clusters to be specified beforehand.

The basic idea underlying the algorithm is that dense clusters correspond to regions with a larger number of paths implying that a random walk has a higher probability to stay inside the cluster than to leave it soon

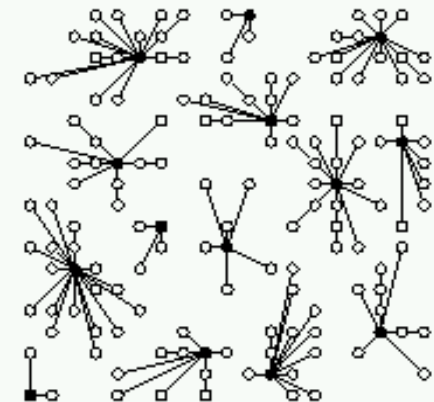
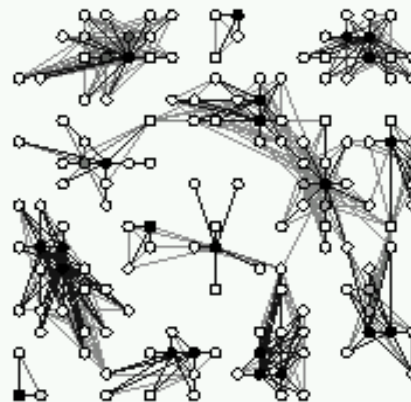
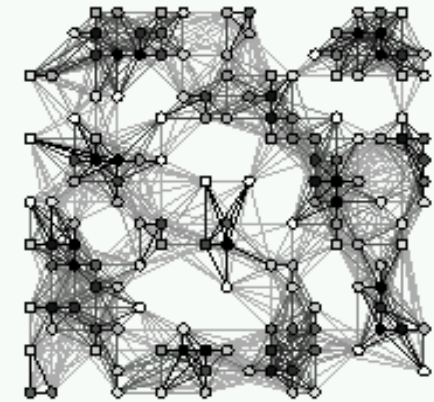
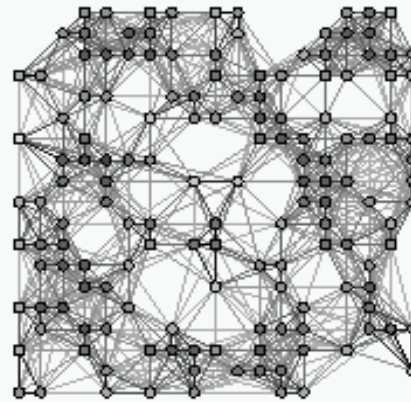
This effect is deliberately boosted by an iterative alternation of expansion and inflation steps.

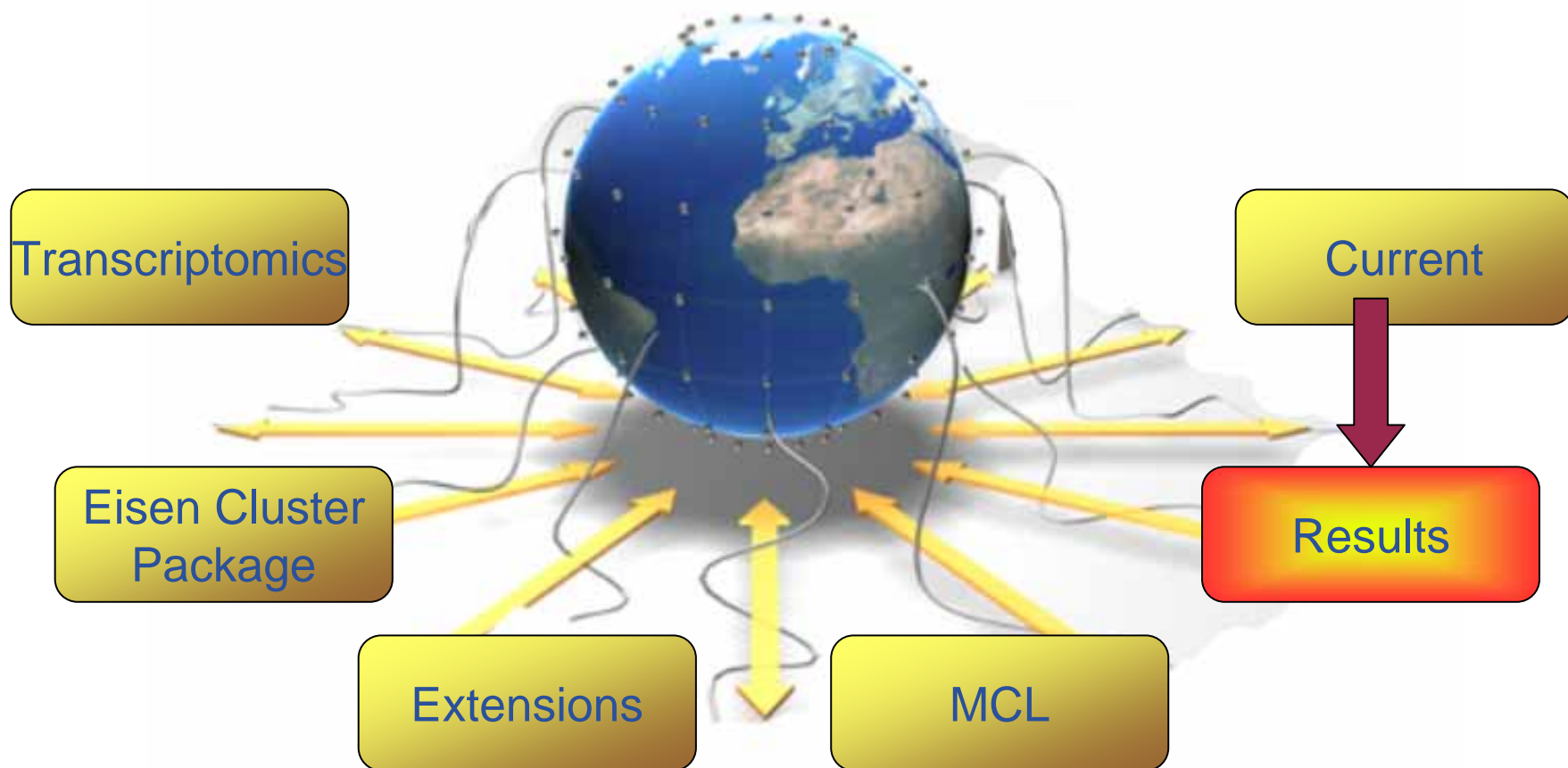


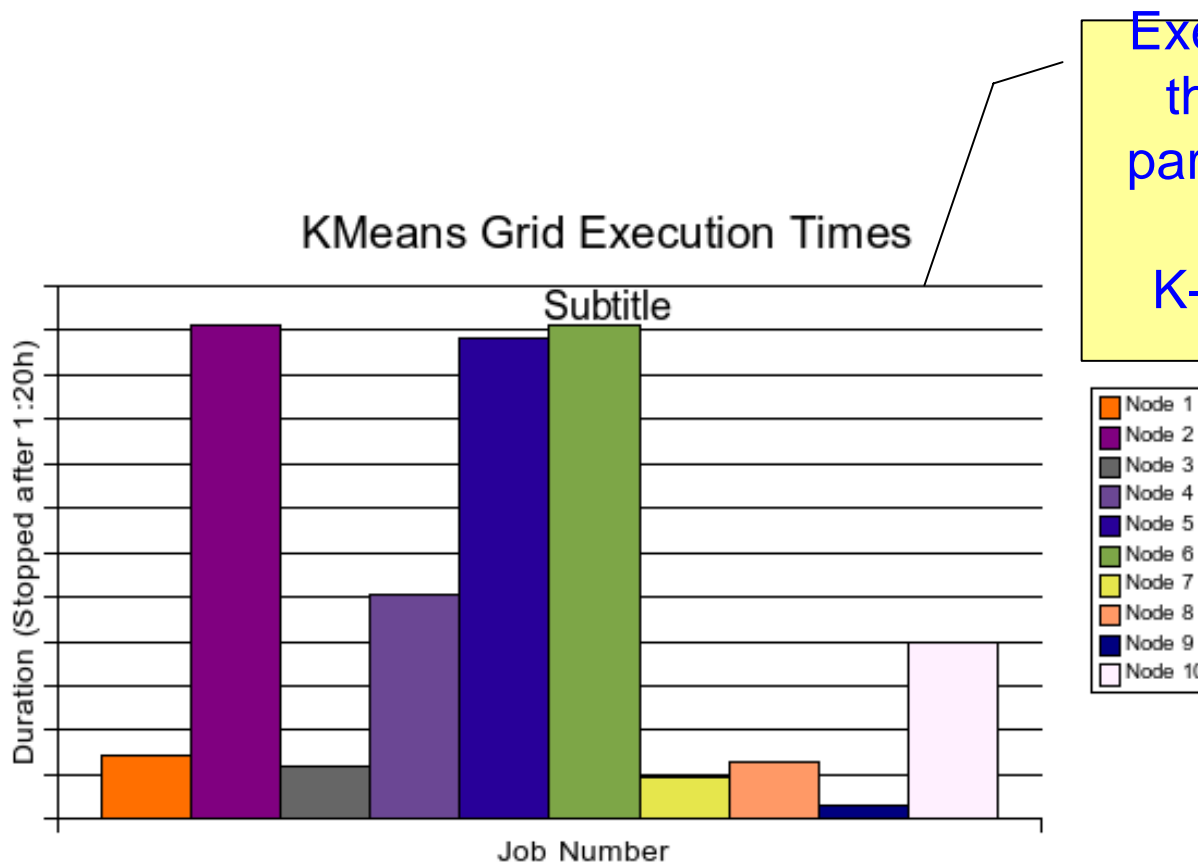
MCL Clustering:

Starts by taking a random walk on the graph described by a similarity matrix

After each step the links are weakened between distant nodes and the links between nearby nodes are strengthen





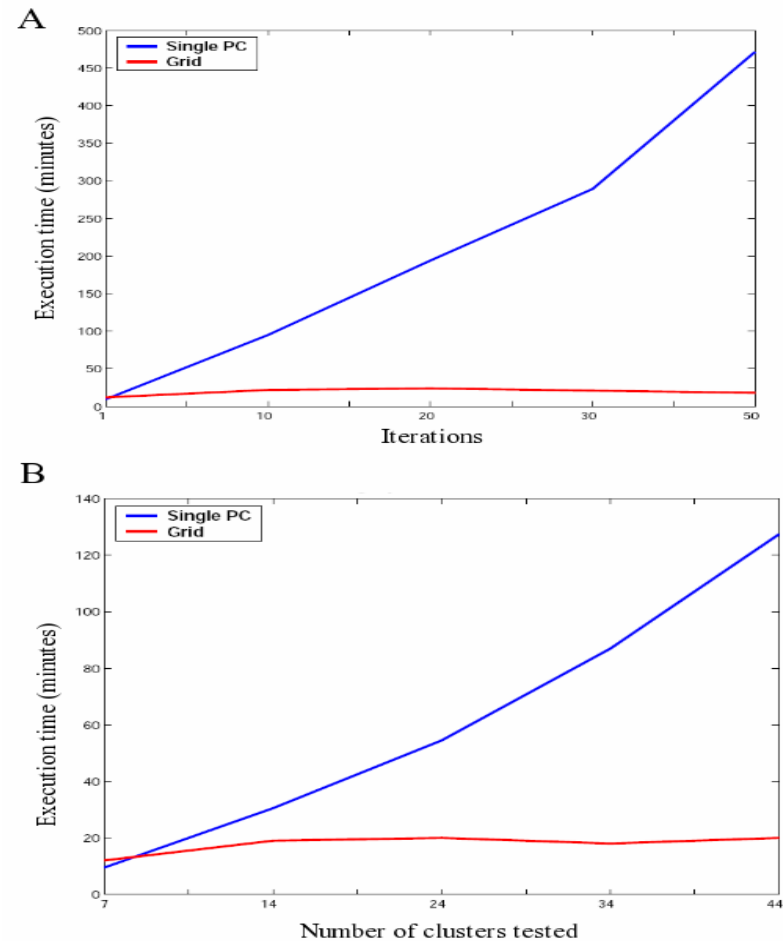


Execution times of the same job in parallel on the grid

K-Means is non-deterministic

The K-means clustering algorithm was tested on the Grid referenced against a single computer firstly with a fixed number of clusters (7) A and a varying amount of iterations and secondly for one iteration and a varying number of clusters (from 7 to 44) B.

Grid performance is on average very good regardless of the variation in execution times of similar jobs

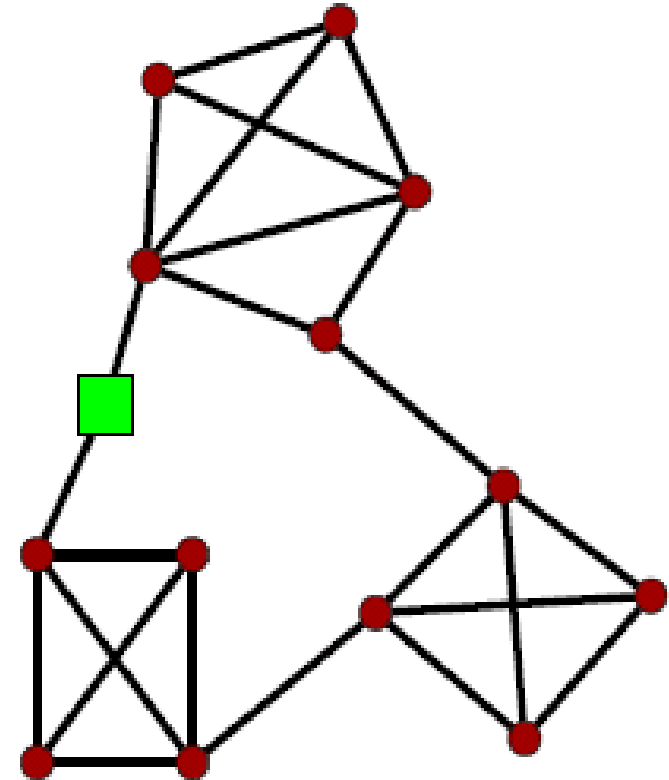


To study the robustness of the results given by MCL we considered the stability of the clustering as related to the identification of unstable nodes (Gfeller et al.,2005)

A node is unstable if it typically lies at the borders of different clusters.

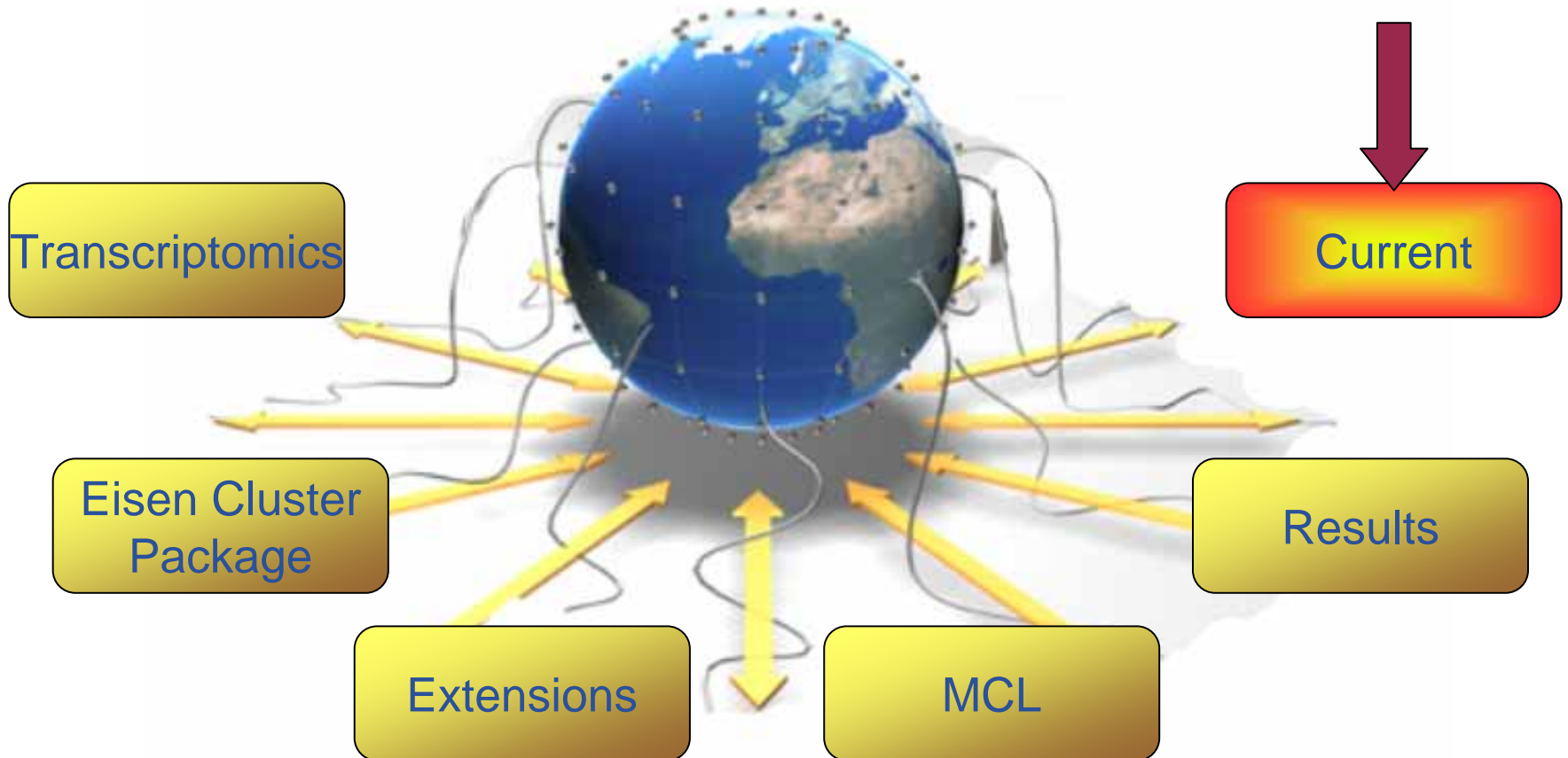
To measure the stability of the clustering patterns, we added random noise on the weights of all links in the network and studied the clustering after many realizations.

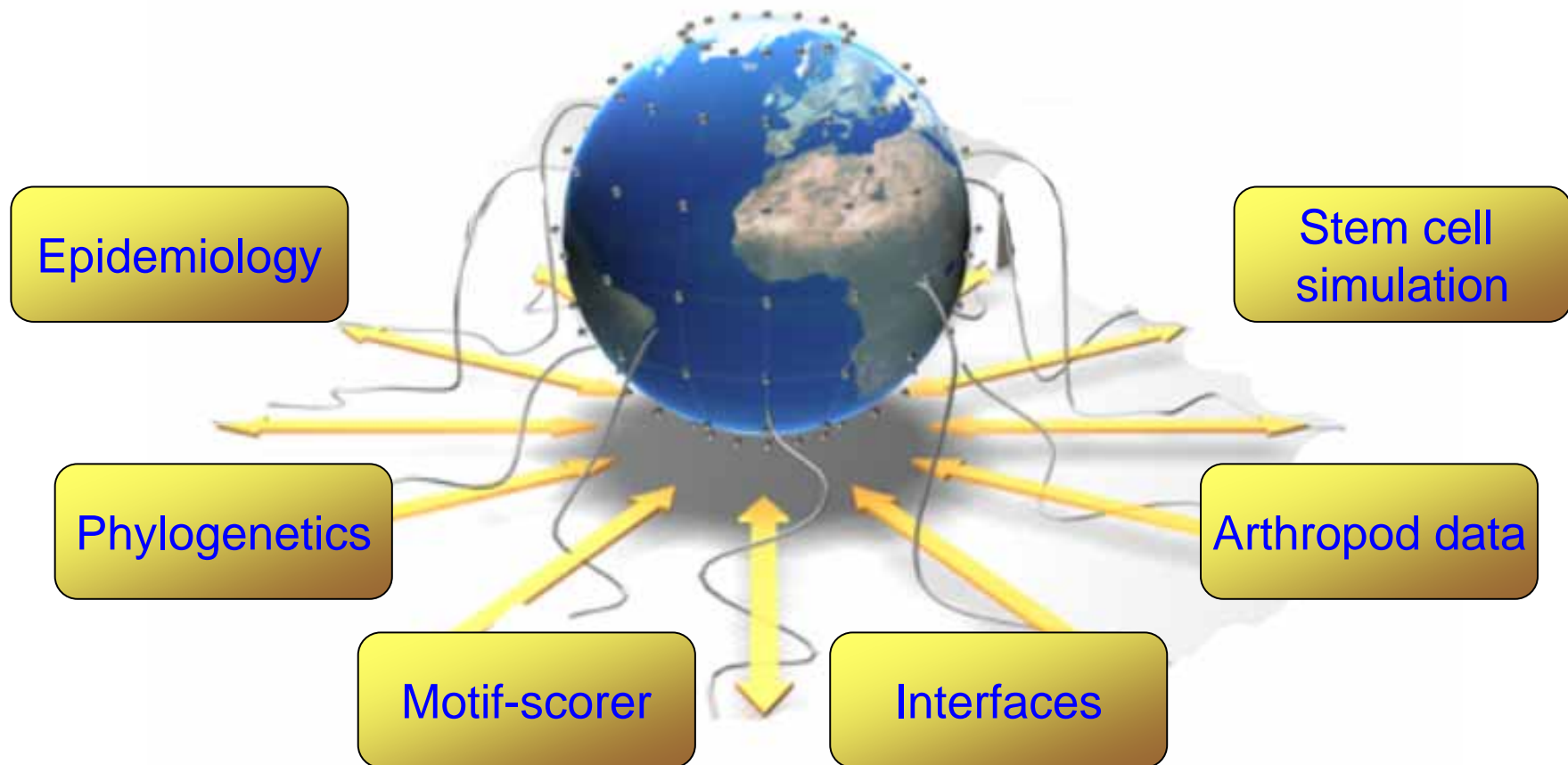
This figure shows a very simple network with a cluster structure made of three components, the green node was continuously identified as a single component that does not correspond to any cluster and thus unstable.



cilea







Constant Technical Support

Giacinto Donvito

Giuseppe La Rocca

Ivan Merelli

BioinfoGRID WP3 Management

Pietro Lio

