Approaching Big Data in Statistics: Parallelization of Classic PLS Algorithm:

Cristina Montañola-Sales [(a)] [(b)], Alba Martínez-Ruiz [(c)]

[(a)] Universitat Politècnica de Catalunya [(b)] Barcelona Supercomputing Center [(c)] Universidad Católica de la Santísima Concepción, Chile

Challenges of Statistics for the Big Data era

Statistics
Computer Science
Big data analysis
Mathematics
Application

Classical statistical methodologies have long demonstrated the powerful capabilities of statistics to analyse data for current policy or business analytics. However, in the era of Big Data how can we approach datasets of Teras or Petas with those techniques? We need new solutions to deal with increasing size in datasets and complex operations.

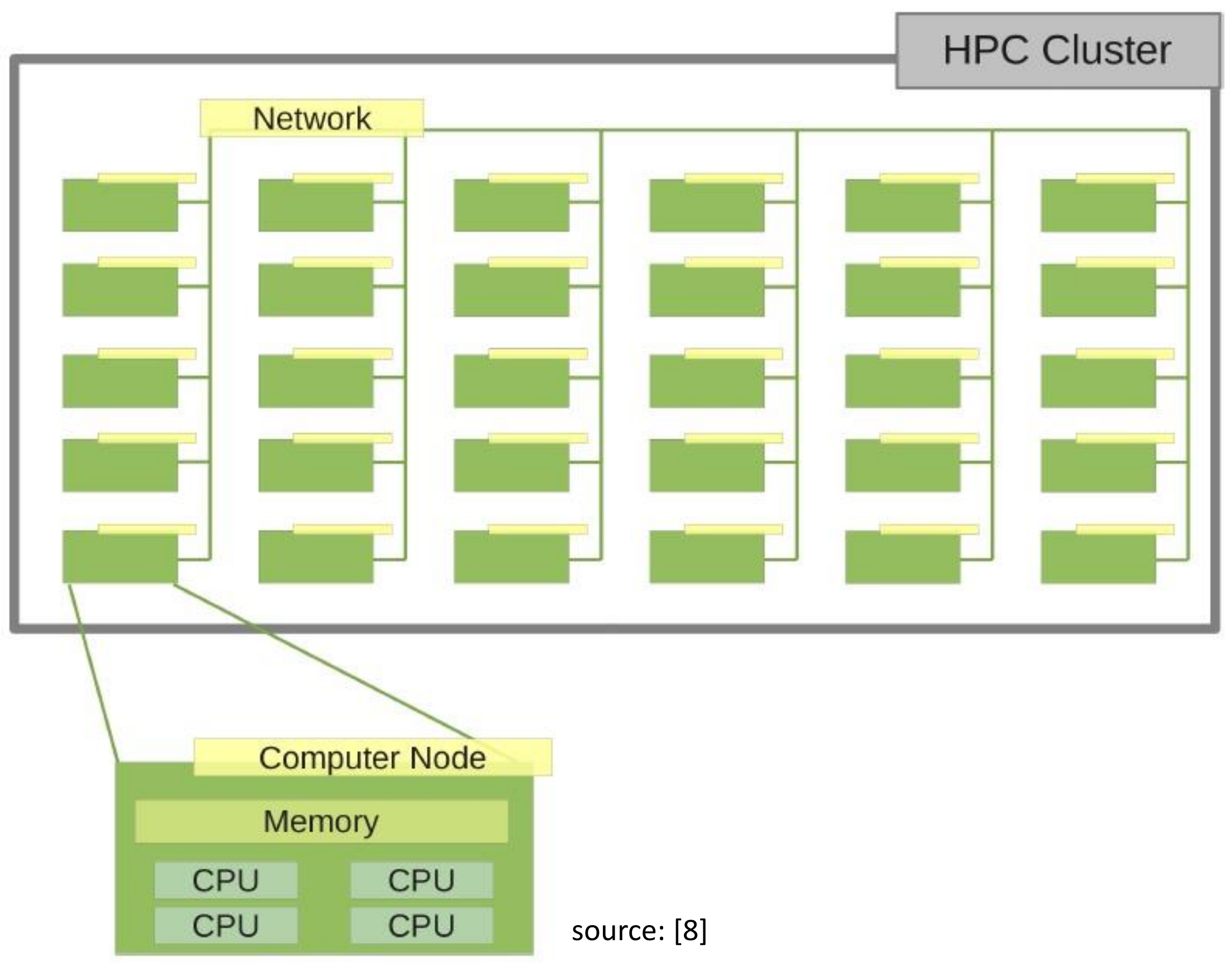
Distributed and parallel computing can provide a solution to perform complex mathematical operations by dividing the problem in small sizes to apply classical statistical methods.

The current statistical community relies in R as a preferred software. The reason is that R provides a good environment to run statistical methods widely used by the community. Moreover, it allows to publish and make available new packages with new developed techniques. Therefore, there is a need to include solutions in R to handle big data problems. Large problems can be overcome by distributing mathematical calculations across processors, so distributed datasets could be processed with current statistical instruments.

Currently, there are some R packages who can help on distributing R algorithms in High Performance Computing systems. The most important ones are Snow [1], Parallel [2], Rmpi [3] and pbdR [4].

In this poster, we present our current work-in-progress to use pbdR for large multi-block data analysis. Specifically we propose an approach to parallelize the classic PLS algorithm [5][6][7]. This algorithm is useful in many applications in areas as diverse as sensometrics, information science and systems, marketing, strategic management and technological change.

High Performance Computing Systems for Big Data



source: [8]

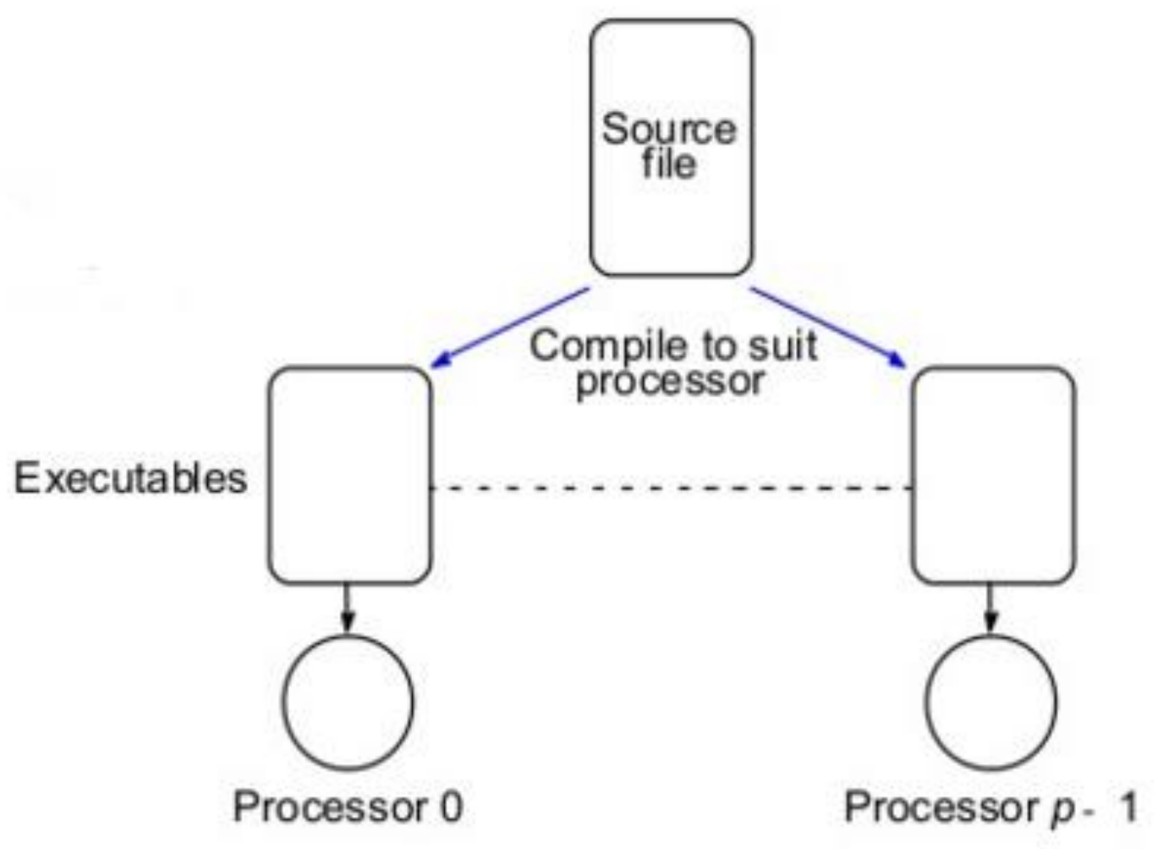
To distribute and processing,

- **Clusters of computers**, a set of computer nodes interconnected with high-speed networks offering HPC at a relatively low-cost.
- **Clouds**, computers connected through a real-time communication network, typically the Internet, which provides dynamic and scalable resources as services to the end-user.
- **Grids**, computers connected through a real-time communication networks as clouds, which require more control by the end user.

To parallel programming models,

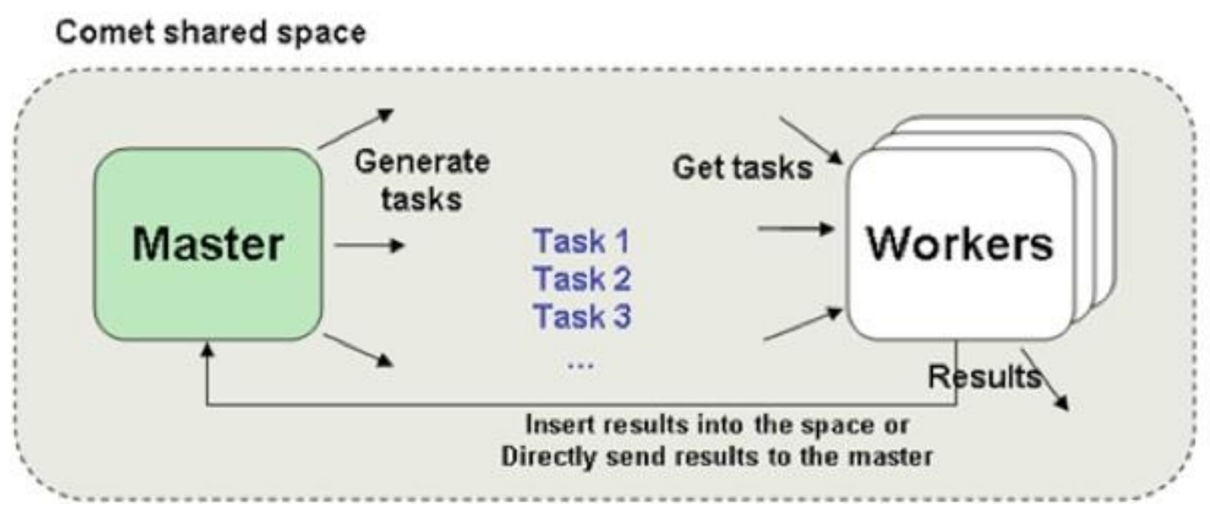
- **Data parallelism**, a set of computer nodes interconnected with high-speed networks offering HPC at a relatively low-cost.
- **Task parallelism**, computers connected through real-time communication network, typically the Internet, which provides dynamic and scalable resources as services to the end-user.

Data Parallelism vs. Task parallelism



Single Program Multiple Data Approach

source: B. Wilkinson
http://webpages.uncc.edu/abw/



Master – Worker Approach

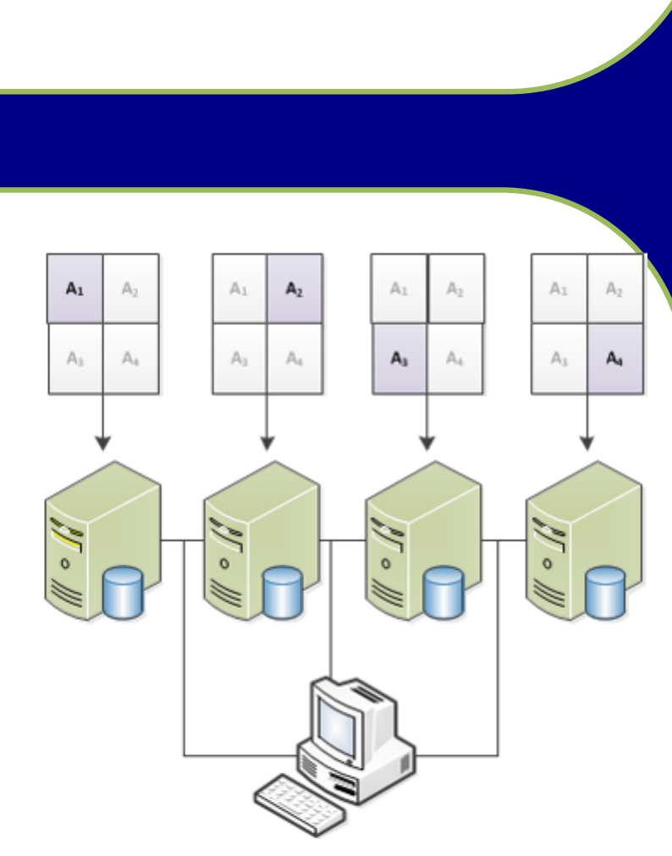
source: Rutgers Discovery Informatics Institute
http://nsfcac.rutgers.edu/

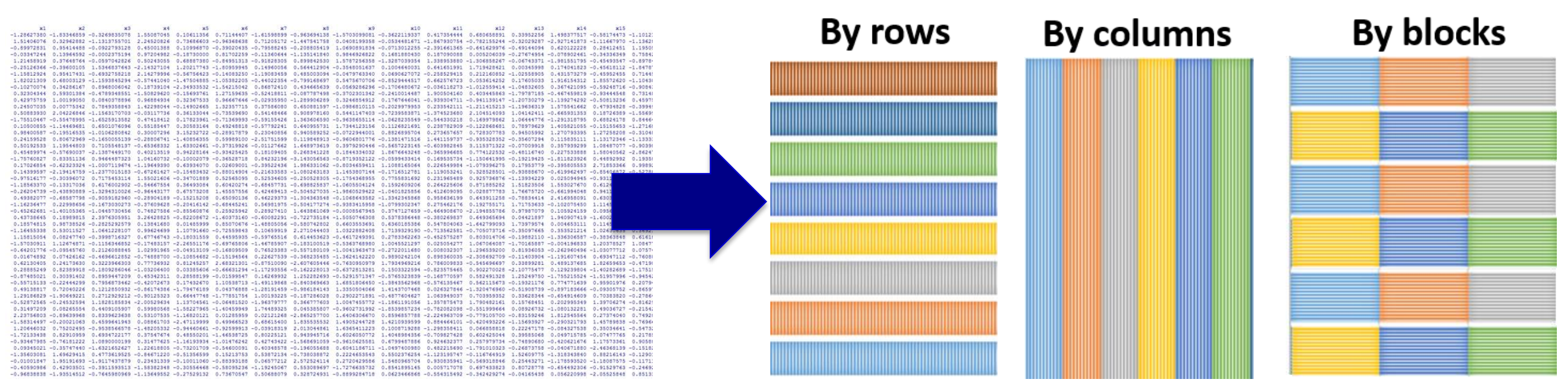
Big Data in Multi-Block Data Analysis

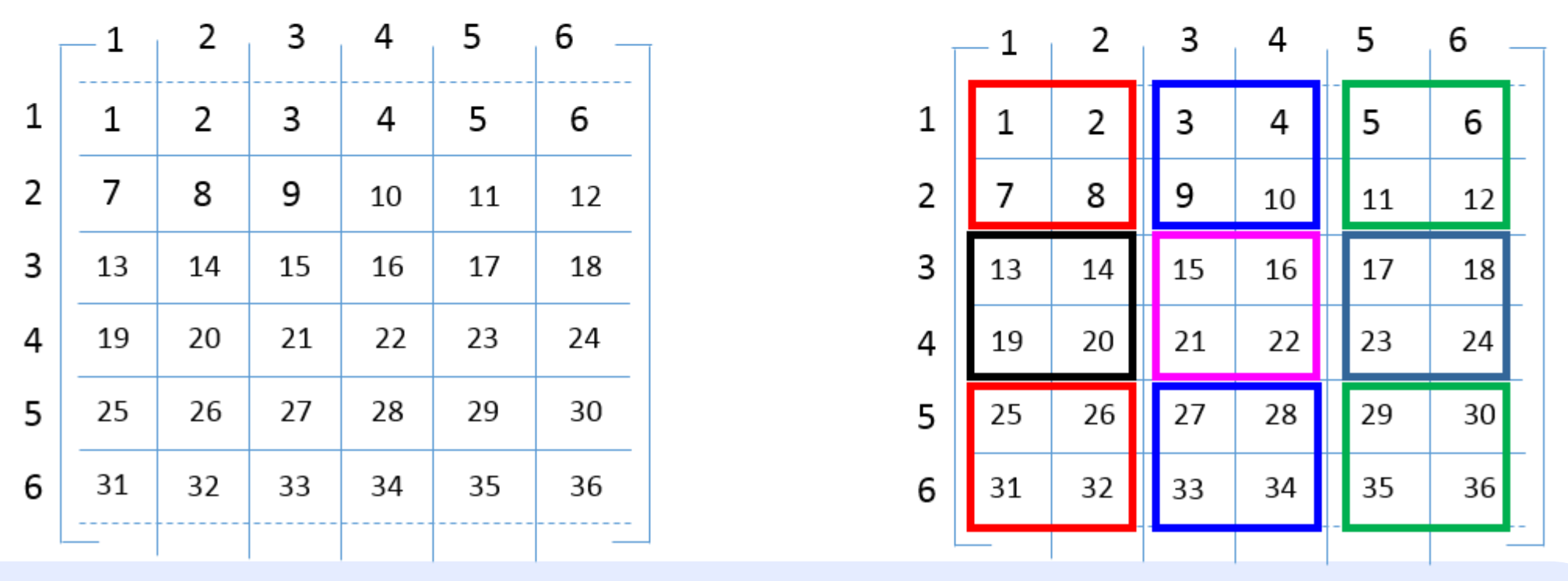
- **Multi-block data analysis** consisting of a set of well-established methods, is properly positioned to meet the big data analysis challenge, since the methods are based on the analysis of components.

MPI processor ranks processor grid $\begin{bmatrix} 0 & 1 & 2 \\ 3 & 4 & 5 \end{bmatrix}$

Blocking factor 2x2





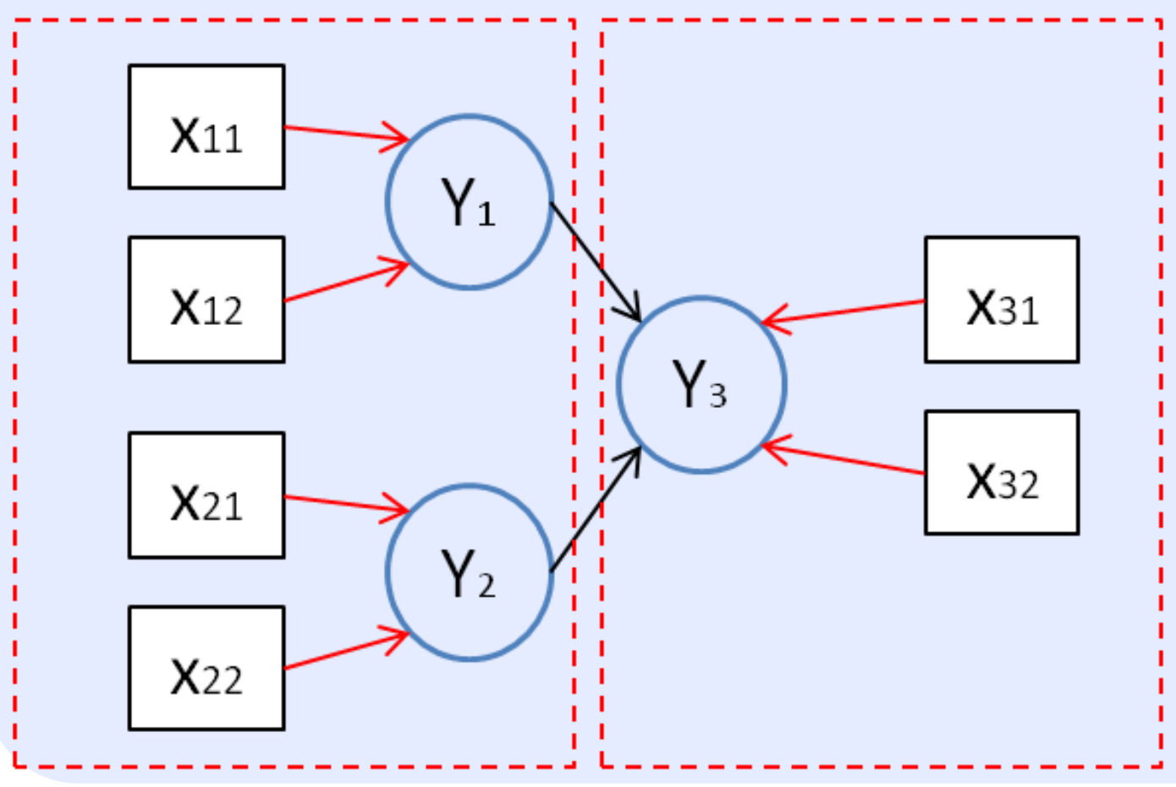


	1	2	3	4	5	6
1	1	2	3	4	5	6
2	7	8	9	10	11	12
3	13	14	15	16	17	18
4	19	20	21	22	23	24
5	25	26	27	28	29	30
6	31	32	33	34	35	36

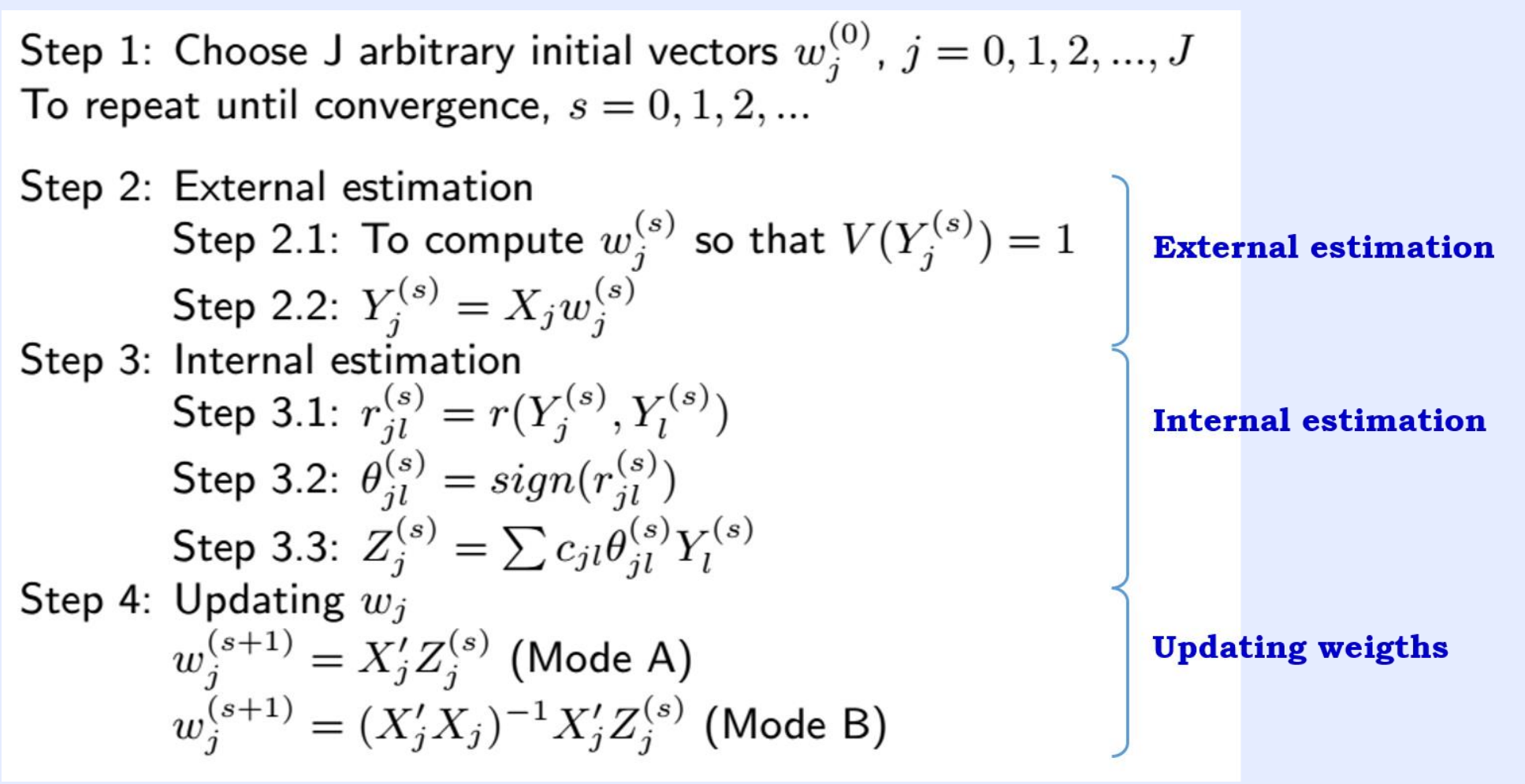
	1	2	3	4	5	6
1	1	2	3	4	5	6
2	7	8	9	10	11	12
3	13	14	15	16	17	18
4	19	20	21	22	23	24
5	25	26	27	28	29	30
6	31	32	33	34	35	36

PLS multi-block analysis

1. Estimation of standardized constructs using an iterative algorithm
2. Endogenous constructs are regressed on the exogenous constructs to estimate effects



To parallelize an algorithm



```
Step 1: Choose J arbitrary initial vectors $w_j^{(0)}, j = 0,1,2,...,J$
To repeat until convergence, $s = 0,1,2,...$

Step 2: External estimation                                          External estimation
        Step 2.1: To compute $w_j^{(s)}$ so that $V(Y_j^{(s)}) = 1$
        Step 2.2: $Y_j^{(s)} = X_j w_j^{(s)}$
Step 3: Internal estimation                                          Internal estimation
        Step 3.1: $r_{jl}^{(s)} = r(Y_j^{(s)}, Y_l^{(s)})$
        Step 3.2: $\theta_{jl}^{(s)} = sign(r_{jl}^{(s)})$
        Step 3.3: $Z_j^{(s)} = \sum c_{jl}\theta_{jl}^{(s)} Y_l^{(s)}$
Step 4: Updating $w_j$                                               Updating weigths
        $w_j^{(s+1)} = X_j' Z_j^{(s)}$ (Mode A)
        $w_j^{(s+1)} = (X_j' X_j)^{-1} X_j' Z_j^{(s)}$ (Mode B)
```

- The efficiency of a distributed processing system will depend on a proper mathematical design, factoring procedure and computer system.

- Block-partitioned algorithm, “recasting algorithms in forms that involve operations on submatrices, rather than individual matrix elements" [Choi et al. 1994].
- Careful design of the parallelization process.
- To examine the performance of the algorithm under different conditions.
- It is necessary to understand how modern computer architecture operates to find the best way to distribute either data and tasks.

References

[1] L. Tierney, A. J. Rossini, N. Li, and H. Sevcikova, “snow package for R,” 2015. [Online]. Available: https://cran.r-project.org/web/packages/snow/.

[2] R. Calaway, Revolution Analytics, S. Weston, and D. enenbaum, “Parallel package for R,” 2015. [Online]. Available: https://cran.r-project.org/web/packages/doParallel/index.html.

[3] H. Yu, “Rmpi package for R,” 2004. [Online]. Available: https://cran.r-project.org/web/packages/Rmpi/index.html.

[4] A. M. Raim, “Introduction to Distributed Computing with pbdR at the UMBC High Performance Computing Facility,” 2013.

[5] H. Wold, “Partial least squares”. In S. Kotz and N.L. Johnson, eds, Encyclopedia of statistical sciences (pp. 1–54), vol 6. Wiley, 1985.

[6] A. Martinez-Ruiz, T. Aluja. Two-step PLS path modelling mode B: Nonlinear and interaction effects between formative constructs. In Chin, W., Esposito Vinzi, V., Russolillo, G., Abdi, H., Trinchera, L. (eds), New Perspectives in Partial Least Squares and Related Methods (pp. 187-199), vol. 56, Springer Proceedings in Mathematics and Statistics, 2013.

[7] M. Tenenhaus, V. E. Vinzi, Y.-M. Chatelin, and C. Lauro, “PLS path modeling,” Comput. Stat. Data Anal., vol. 48, no. 1, pp. 159–205, 2005.

[8] C. Montañola-Sales, X. Rubio-Campillo, J. Casanovas-Garcia, J. M. Cela-Espín, and A. Kaplan-Marcusán, “Large-scale social simulation, dealing with complexity challenges in high performance environments,” in Interdisciplinary Applications of Agent-Based Social Simulation and Modeling, IGI Global, 2014.

Further information:
cristina.montanola@upc.edu
amartine@ucsc.cl