



POLITÉCNICA

Seminario de Investigación
Master en Ciencias y
Tecnologías

Modelos de Computación no Convencionales y Aplicaciones Big Data

Sandra María Gómez Canaval
Grupo de Investigación en Modelos Matemáticos
y Algoritmos para Biocomputación

Introducción

Modelos de Computación No convencionales

- Características
- Dominios de Problemas Complejos
- Estado del Conocimiento

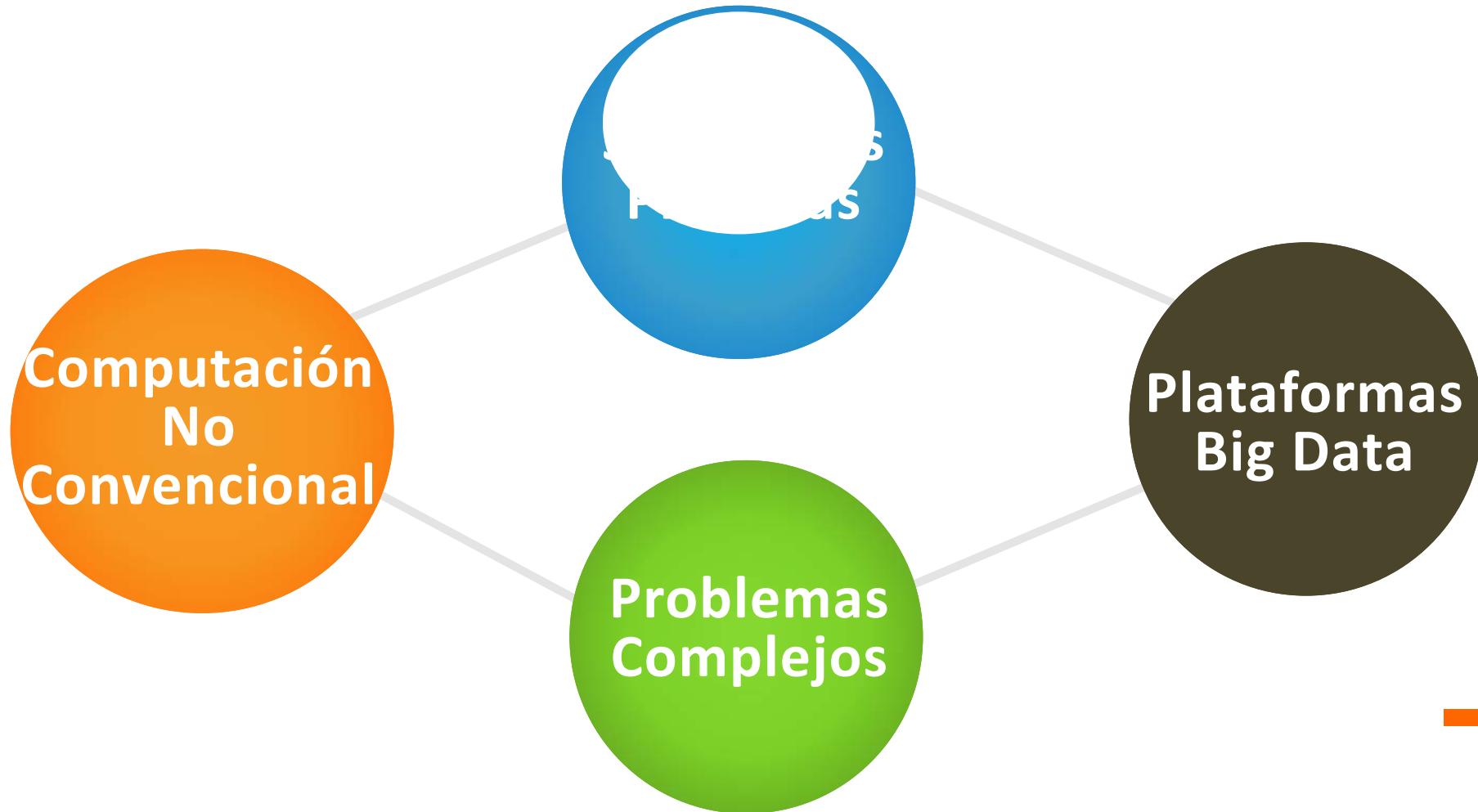
Plataformas de Computación Big Data

- Tipos
- Aplicaciones

Simulación de Modelos Bio-inspirados en Plataformas Big Data

Problemas abiertos y líneas futuras

Introducción: Contexto Científico



Marco Conceptual

Modelos de Computación No Convencionales

- Modelos, métodos, técnicas, conceptos e ideas disruptivas.
- Soluciones eficientes a un conjunto de problemas complejos o intratables.
- Generalmente se inspiran en fenómenos de la naturaleza:
 - Químicos, Físicos, Biológicos
- Perspectivas: ciencias de la computación y ciencias biológicas

Modelos de Computación No convencionales

Ciencias Naturales

Ciencias de la Computación

- Computación Evolutiva
- Computación con Membranas
- Computación Cuántica
- Inteligencia de Enjambres
- Biología Sintética
- Biología de Sistemas
- Computación Celular

Modelos de Computación No convencionales

- Autómatas Celulares
- Redes Neuronales
- Algoritmos Genéticos
- Algoritmos de Enjambres de Partículas, Colonias de Hormigas, Abejas, etc.
- Sistemas de Membranas
- Redes de Procesadores Bio-inspirados

Problemas de Optimización

Problemas NP completos

Modelos de Computación No convencionales

Técnicas bio-inspiradas

- Precisión
- Espacios de búsqueda Grandes
- Combinatoria
- Data intensive
- Computation intensive

Modelos bio-inspirados

- De propósito general
- Combinatoria
- Data intensive
- Computation intensive
- Escalabilidad
- Soluciones eficientes
- Intrínsecamente masivamente paralelos y distribuidos

Modelos de Computación No convencionales

Técnicas bio-inspirados

- Optimización
-
- Clasificación
 - Clustering
 - Reconocimiento de Patrones

Modelos bio-inspirados

NP Complete:

- 3-Colorability Problem
- Problema de satisfacibilidad booleana
- Camino Hamiltoniano
- Problema de la Mochila
- Problema de Cobertura de Vértices
- Problema del Cliqué

I. Boussaïd, J. Lepagnot, P. Siarry. A survey on optimization metaheuristics. Informationciences, 237, 82-117. 2013

Applied Soft Computing. 26:74-89. <http://dx.doi.org/10.1016/j.asoc.2014.09.029>

G. Zhang, M. Gheorghe, L. Pan, M. Pérez-Jiménez, Evolutionary membrane computing: A comprehensive survey and new results, Information Sciences, 279:20,528-551, <http://dx.doi.org/10.1016/j.ins.2014.04.007>. 2015.

A. Ortega, M. Echandía (Eds). Networks of Bioinspired Processors. Publicacions URV. Triangle, 7, 1a ed. 2013-939X. 2012

Modelos de Computación No convencionales

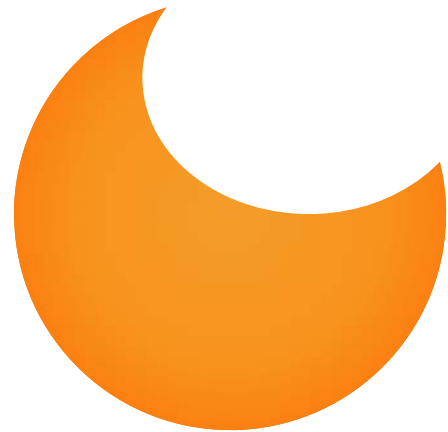
Desventajas

Técnicas bio-inspirados

- Ausencia de generalidad
- Pocas implementaciones distribuidas y paralelas
- Problemas de escalabilidad
- Diversidad de métodos

Modelos bio-inspirados

- No hay computadores naturales
- Simulaciones in silico: crecimiento exponencial
- Implementaciones hardware: problemas de escalabilidad y de reconfiguración
- Ámbito teórico



PLATAFORMAS DE COMPUTACIÓN BIG DATA



Big Data

- Data Intensive Scientific Discovery Paradigm
- Big Data: **V**olumen, **V**elocidad, **V**ariedad
 - **V**eracidad / Valor
- Crecimiento de datos a una tasa de entre el 40 y el 60% al año.
- Creación de plataformas de computación ultra-escalables, distribuidas y paralelas:
 - Computación intensiva de datos
 - Computación intensiva en cálculos

Plataformas de Computación Big Data

Problemas Big Data

- Tradicionales de data mining /Análisis de datos
 - Clasificación y clustering
 - Reconocimiento/Descubrimiento de Patrones
- Optimización
- Predicción/Evolución de flujos de datos
- Análisis de Data Streaming
- Visualización
- Análisis de Grafos Complejos

Plataformas de Computación Big Data

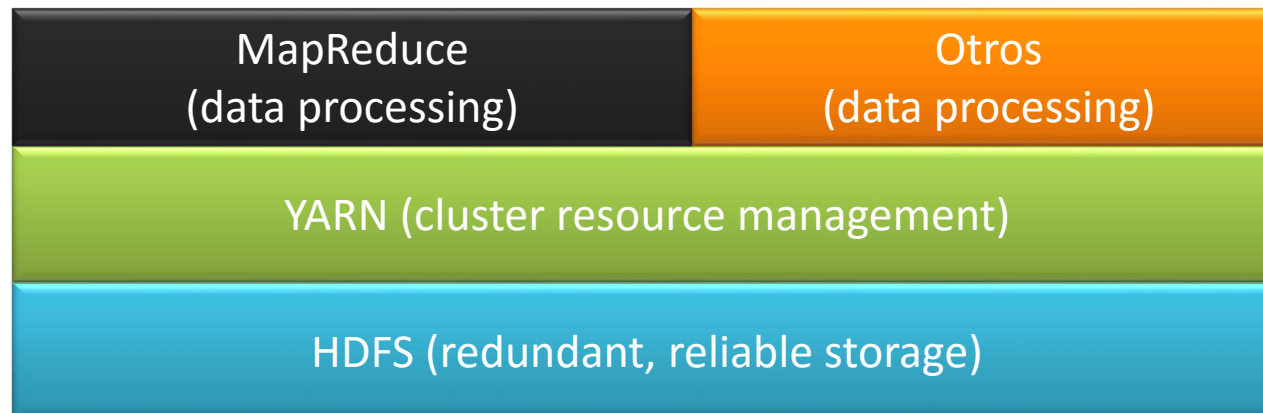
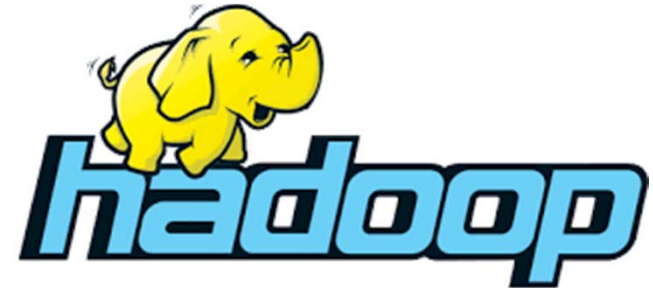
Plataformas Big Data

- Ultraescalables
- Computación masivamente paralela y distribuida
- Disponibilidad de componentes, librerías y aplicaciones para diversos dominios Big Data
- Orientadas a computación intensiva en datos e intensiva en cálculos

Plataformas de Computación Big Data

Apache Hadoop

- Permite procesamiento distribuido de grandes datasets a través de un cluster de ordenadores tolerante a fallos.
- Implementa el modelo Map/Reduce
- Arquitectura:



Plataformas de Computación Big Data

Apache Hadoop

- MapReduce es ineficiente en el funcionamiento de algoritmos iterativos.
- Múltiples acceso a disco para gestionar los resultados.
 - Cuello de botella que degrada significativamente el rendimiento
- Orientada a grafos: Adecuado para algoritmos/problemas basados en grafos.

Plataformas de Computación Big Data

Pregel y Giraph

- Pregel fue propuesto por Google en 2010
- Inspirado por el Bulk Synchronous Parallel (BSP) model
- Procesamiento altamente escalable y tolerante a fallos.
- Orientado a grafos
- **Computación**: Secuencia de iteraciones llamadas Supersteps.
 - Cada superstep representa una unidad atómica de computación paralela.
- Un procedimiento de chequeo global determina si la computación ha finalizado.

Plataformas de Computación Big Data

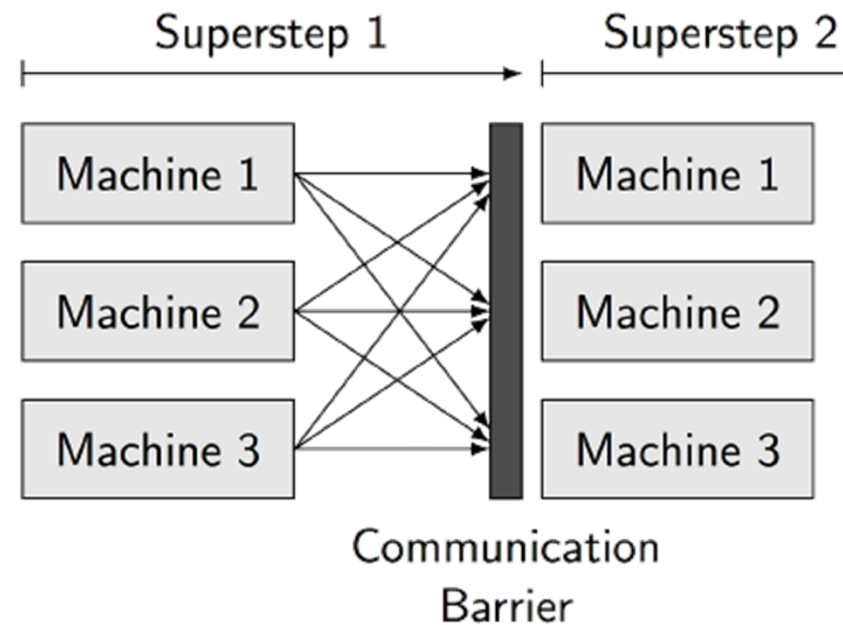
Pregel y Giraph

- Supersteps:

Computación en paralelo: cada vértice ejecuta su función de forma independiente como una computación local

Comunicación: los vértices intercambian datos vía mensajes con otros vértices

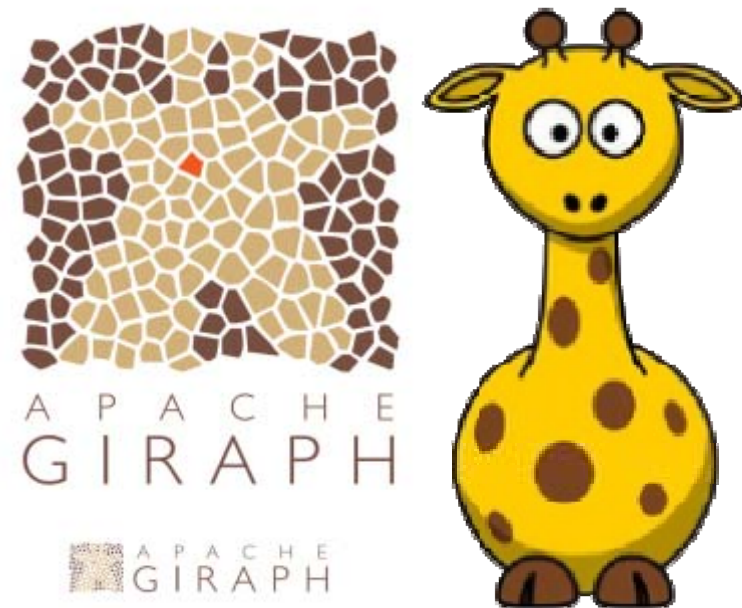
Barrera de Sincronización: todos los vértices esperan hasta que el proceso de comunicación haya finalizado



Plataformas de Computación Paralelas y Distribuidas

Pregel y Giraph

- Giraph is la versión Open Source de Pregel.



Plataformas de Computación Big Data

Spark

- Desarrollado por la Universidad de Berkeley 2009-2010. Ahora es un Proyecto Apache
- Basado en MapReduce
- Mejora los problemas de rendimiento relacionados con los cuellos de botella del acceso a disco de Hadoop
- Provee tolerancia a fallos sin replicación.
- Soporta computación iterativa



Plataformas de Computación Big Data

Spark

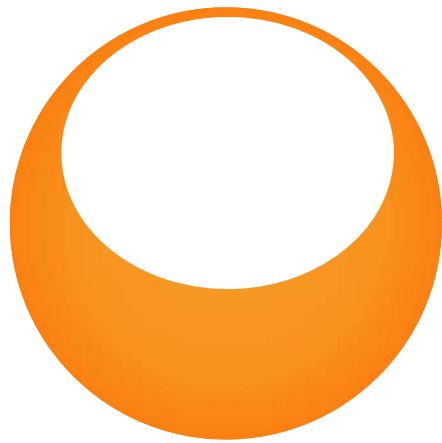
Utiliza computación en-memoria

- Resilient Distributed Datasets (RDD) (read-only distributed shared memory)
 - Almacena en memoria datos para las computaciones
 - Realiza computaciones en los datos cargados en memoria generando resultados intermedios
 - Ejecuta los cambios generales disminuyendo el numero de operaciones de lectura y escritura
- Eliminando así limitación de sobrecarga de disco para las tareas iterativas

Plataformas de Computación Big Data

Otras:

- Orientada a grafos:
 - Dato (antes GraphLab): Desarrollada por la Universidad Carnilie y Mellon en 2009
 - PowerGraph
- Para data Streamming:
 - Apache Storm
 - Flink
- Escalabilidad vertical:
 - Graphics Processing Unit (GPU's)



SIMULACIÓN DE MODELOS BIO- INSPIRADOS EN PLATAFORMAS BIG DATA



Simulaciones en Plataformas Big Data

- Estas plataformas se convierten en idóneas para:
 - Simular modelos computacionales no convencionales
 - Realizar implementaciones de métodos y técnicas bio-inspiradas que requieran paralelización en computaciones intensivas en datos y cálculos

Simulaciones en Plataformas Big Data

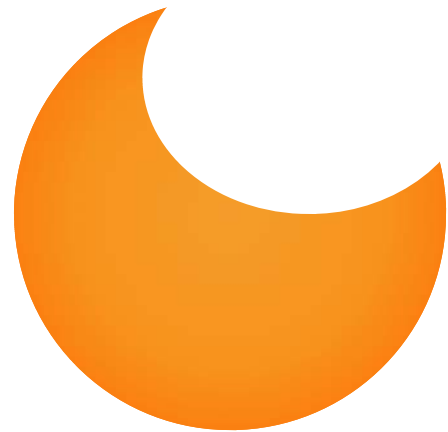
Ventajas

Técnicas bio-inspirados

- Escalabilidad
- Mayor generalidad
- Mapeo directo de paralelización de datos o de tareas
- Implementaciones paralelas y distribuidas
- Reducción de costes computacionales en los métodos existentes

Modelos bio-inspirados

- Reducción de la brecha entre la teoría y la práctica
- Aplicación a problemas reales
- Mapeo directo entre arquitectura teórica y arquitectura real
- Estas plataformas son idóneas para encapsular el paralelismo inherente de estos modelos



**UNA SOLUCIONES PRÁCTICA
ACTUAL PARA REDES DE
PROCESADORES BIO-INSPIRADOS**



Redes de Procesadores Bioinspirados (NBP)

- Masivamente paralelo y distribuido
- Procesamiento simbólico
 - Bio-operaciones (Point mutations DNA sequences, splicing, Crossover)
 - Datos representados como cadenas
- Resuelven problemas NP completos eficientemente
 - Universal y computacionalmente completos
 - Simples y Expresivos
 - Extensibles y Escalables

[Mitrana, 2001] Solving NP-complete problems with NEP. LCNS. 2041. 621.

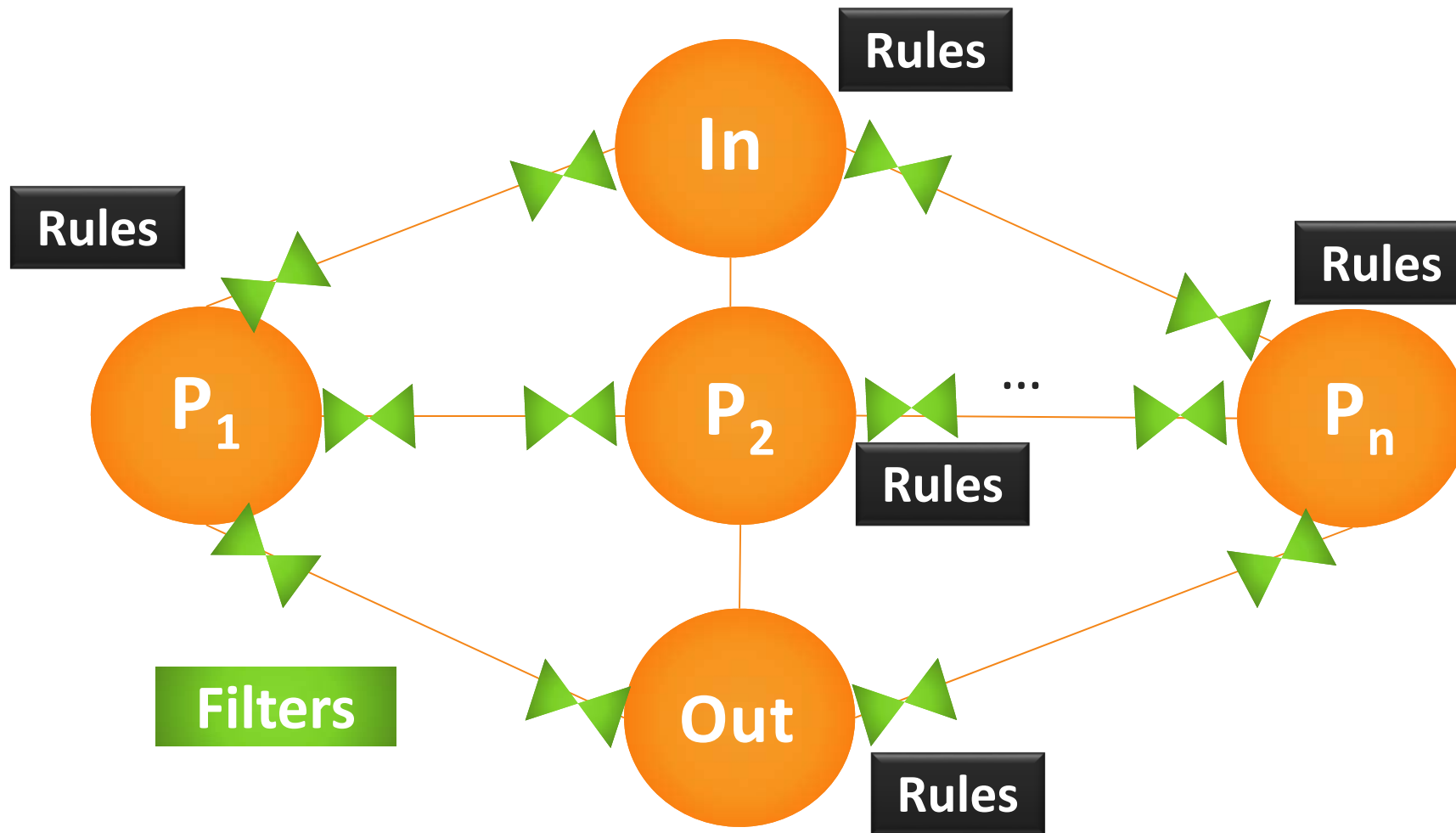
[Mitrana, 2003] Networks of Evolutionary Processors. Acta Informática 39. p.517.

[Campos, 2012] Accepting NGP are computationally complete. Theor. Comp. Sc. 456, p. 18

[Mitrana, 2014] Networks of Polarized Evolutionary Processors. Information Sci. 265. p. 189

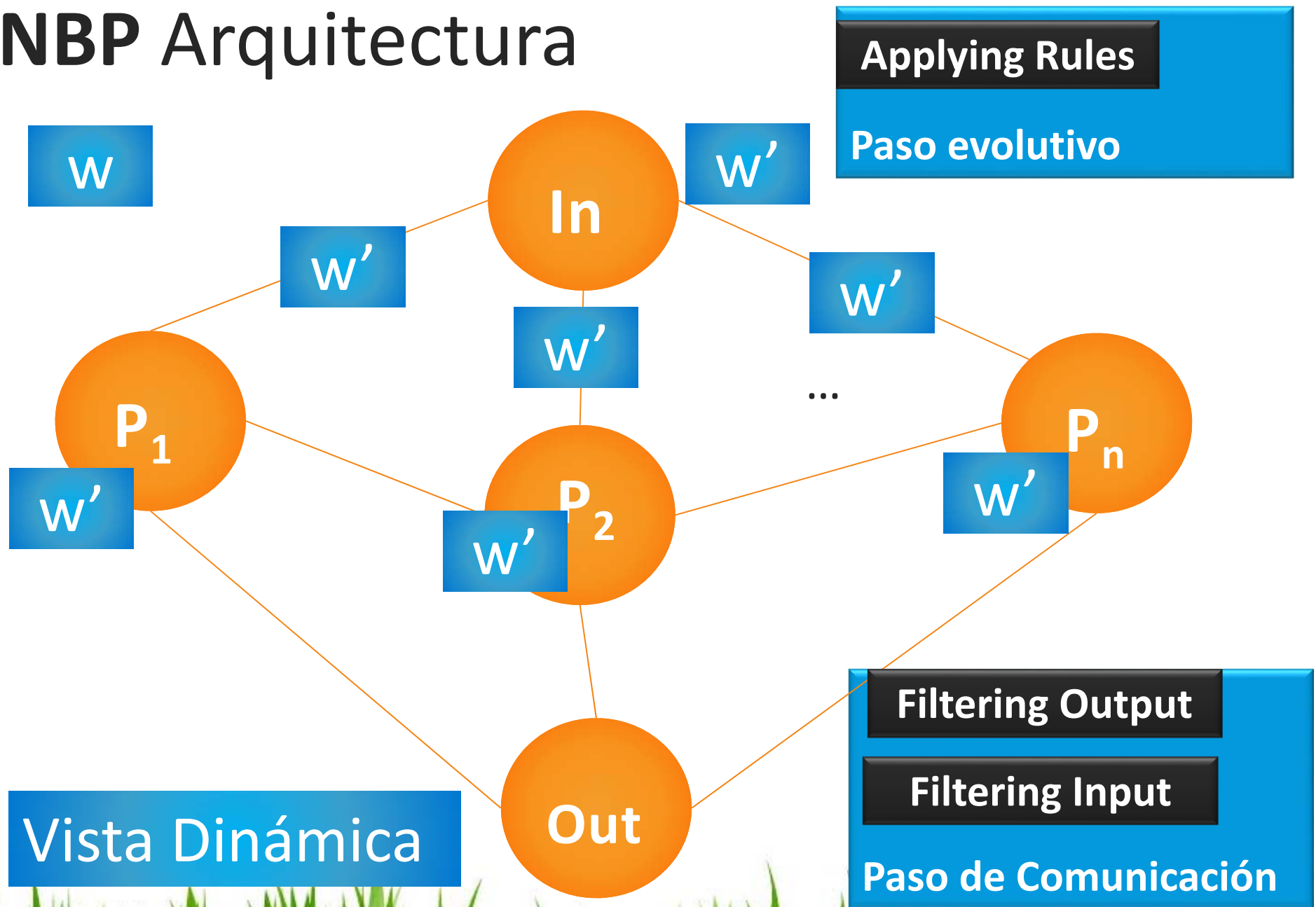
NBP Arquitectura

2 alphabets: input and network



Vista Estática

NBP Arquitectura



Dinámica NBP

- **Configuración:** conjunto de palabras presentes en un nodo en un momento dado.
- **Computación:** conjunto de configuraciones alternando pasos evolutivos y de comunicación
- **Condición de Finalización:**
 - EL nodo de salida no está vacío
 - Existen dos o más configuraciones idénticas
 - Ningún otro paso es posible

Familia de Redes de Procesadores Bioinspirados

Evolutivos

Insertion
Deletion
Substitution

Membership

Genéticos

Splicing
Crossover

Membership

Splicing

Splicing

Membership

Polarizados

Insertion
Deletion
Substitution

Polarization

Optimización

Insertion
Deletion
Substitution

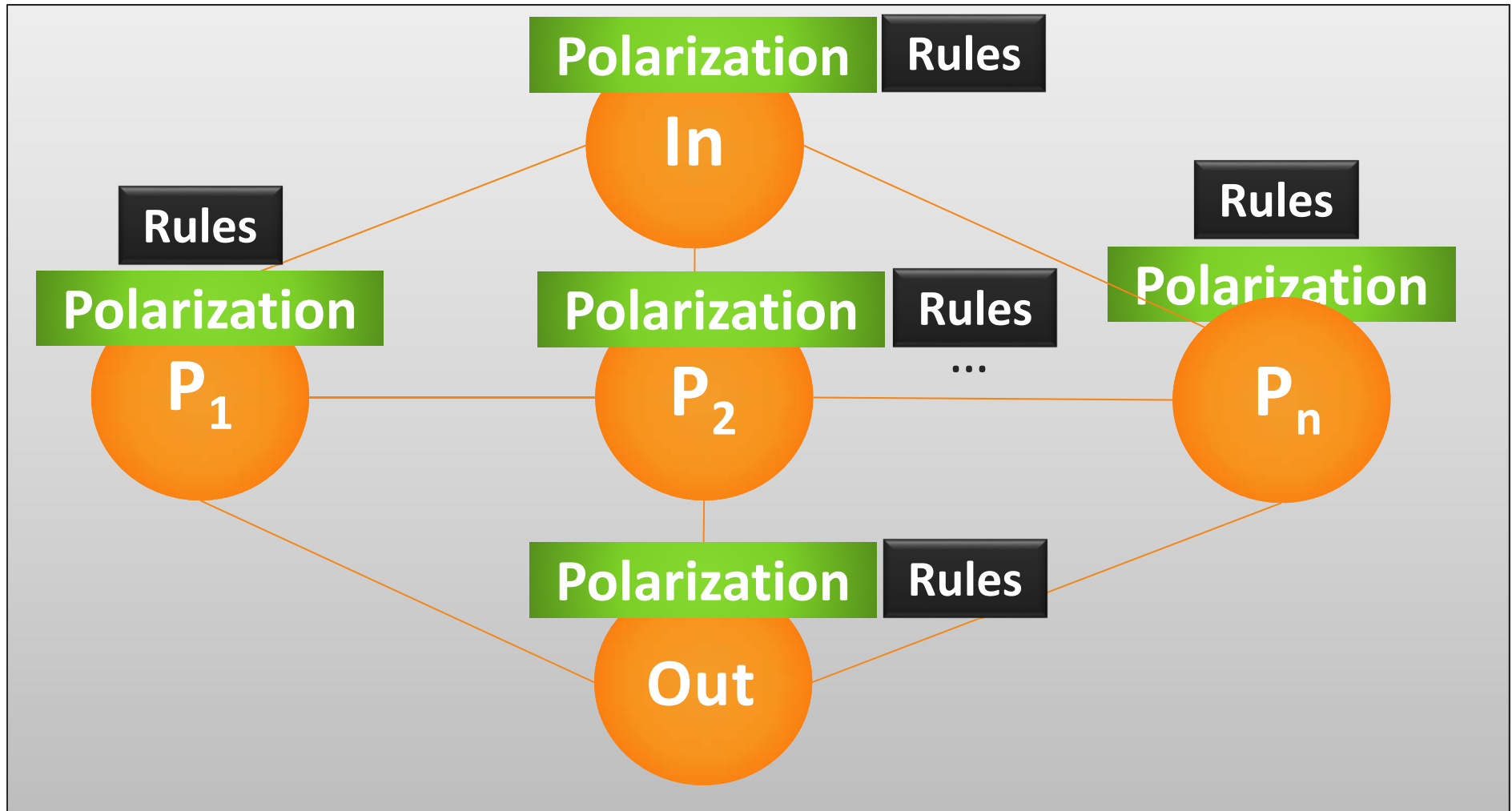
Polarization

Splicing

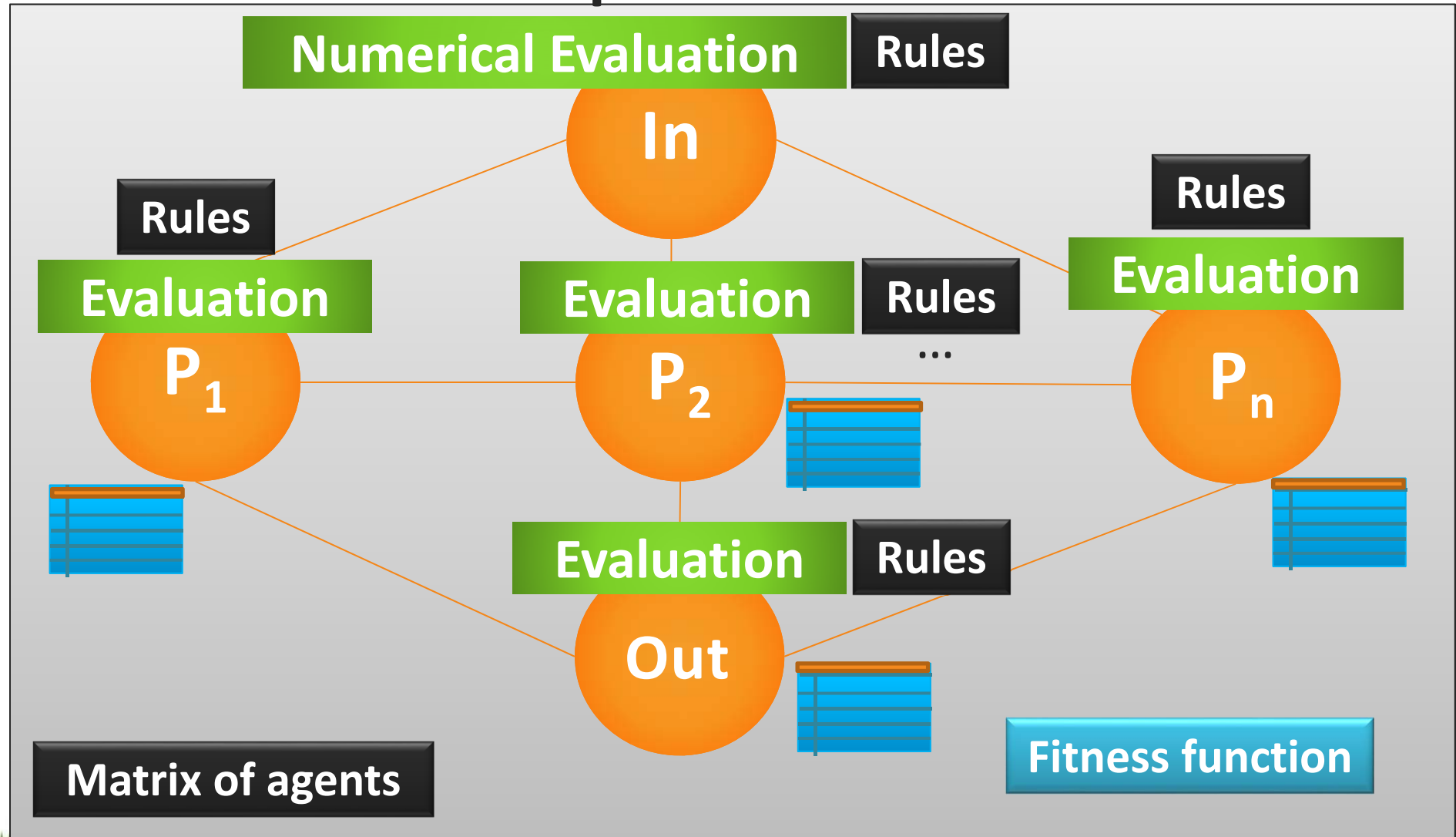
Insertion
Deletion
Substitution

Polarization

Redes de Procesadores Evolutivos Polarizados



Redes de Procesadores Evolutivos de Optimización



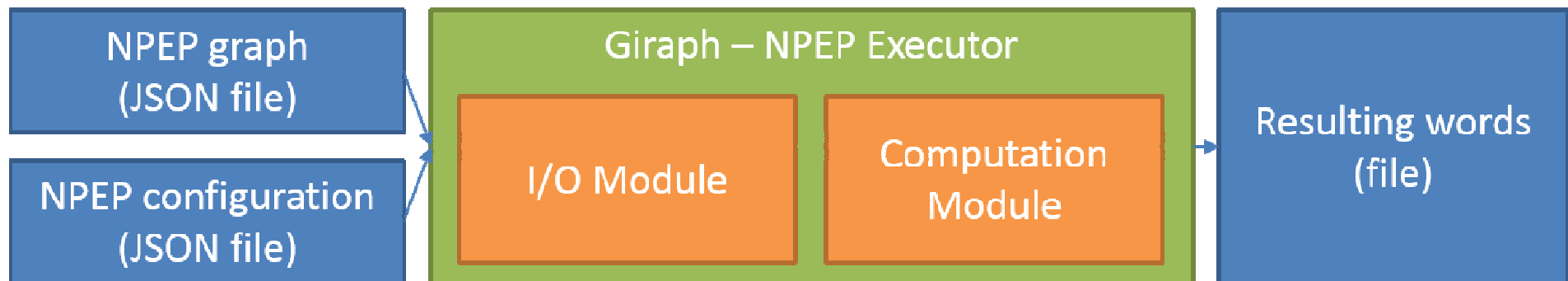
Simuladores Ultra Escalables para NBP

NPEPE

- Giraph y su implementación del modelo BSP es una arquitectura natural para implementar modelos NBP.
 - BSP y dinámica NBP
- Paralelismo masivo y computación distribuida
- Computación orientada a grafos de grafos grandes
 - Arquitectura NBP subyacente vista como un grafo virtual
- Vértice y Procesador Bio-inspirado
- Superstep incluye un pasos evolutivos y de comunicación
- El procedimiento global de chequeo es la condición de terminación de los NBP

Simuladores Ultra Escalables para NBP

Arquitectura NPEPE



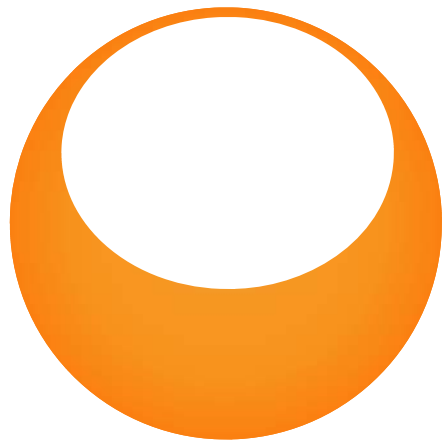
NPEPE

Se han implementado y simulado:

- El problema de los 3 colores para grafos con Procesadores Evolutivos
- El problema de la mochila con Procesadores Evolutivos de Optimización

Resultados

- Favorables en escalabilidad
- Viabilidad e idoneidad demostrables
- Adaptación de los algoritmos a la arquitectura de la plataforma Big Data de implantación



LÍNEAS DE INVESTIGACIÓN FUTURAS Y PROBLEMAS ABIERTOS

Trabajo Futuro para NBP

NPEPE

Trabajo inmediato

- Pruebas de escalabilidad con datasets con un volumen considerable de datos
- Pruebas de escalabilidad horizontal
- Probar con plataformas como Spark GraphX
- Implementación de otras familias de NBP
- Implementación de más problemas NP completos
- Optimización de componentes del modelo

Problemas abiertos

- Modelos computacionales no convencionales orientados a problemas big data (clasificación, clustering, detección de patrones, etc).
- Simulación de estos modelos a plataformas de computación big data
- Versiones paralelas de Métodos bioinspirados de optimización y metaheurísticas
- Implementación de estas versiones en plataformas big data



Thanks for your attention!

