Artificial Immune Systems and Data Mining: Bridging the Gap with Scalability and Improved Learning

Olfa Nasraoui, Fabio González
Cesar Cardona, Dipankar Dasgupta
The University of Memphis

A Demo/Poster at the National Science Foundation Workshop on Next Generation Data Mining, Nov. 2002

Inspired by Nature…

- living organisms exhibit extremely sophisticated learning and processing abilities that allow them to survive and proliferate
- nature has always served as inspiration for several scientific and technological developments, exp: Neural Networks, Evolutionary Computation
- immune system: parallel and distributed adaptive system w/ tremendous potential in many intelligent computing applications.

What is the Immune System?

- **Protects** our bodies from foreign pathogens (viruses/bacteria)
- **Innate** Immune System (initial, limited, ex: skin, tears, …etc)
- **Acquired** Immune System (**Learns** how to respond to NEW threats adaptively)
- **Primary** immune response
  - First response to invading pathogens
- **Secondary** immune response
  - Encountering similar pathogen a second time
  - **Remember** past encounters
  - Faster and stronger response than primary response

Points of Strength of The Immune System

- **Recognition (**Anomaly detection, Noise tolerance)
- **Robustness** (Noise tolerance)
- **Feature extraction**
- **Diversity** (can face an entire repertoire of foreign invaders)
- **Reinforcement learning**
- **Memory** (remembers past encounters: basis for vaccine)
- **Distributed** Detection (no single central system)
- **Multi-layered** (defense mechanisms at multiple levels)
- **Adaptive** (Self-regulated)

Major Players: B-Cells

- Through a process of ***recognition*** and ***stimulation***, B-Cells will **clone and mutate** to produce a ***diverse*** set of antibodies adapted to different antigens

- **B-Cells** secrete **antibodies w/ paratopes** that can bind to specific antigens (epitopes) and destroy their host invading agent through a *KILL, SUICIDE, or INGEST* signal.

❖ **B-Cells antibody paratopes** also can bind to antibody **idiotopes** on **other** B-Cells, hence sending a STIMULATE or SUPPRESS signal → hence the ***Network*** → ***Memory***

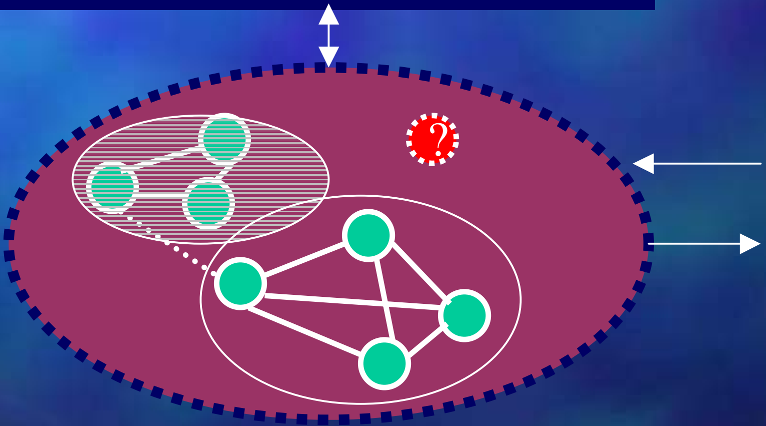
Requirements for Clustering Data Streams (Barbara, 02)

- **Compactness of representation**
  - Network of B-cells: each cell can recognize several antigens
  - B-cells compressed into clusters/sub-networks
- **Fast incremental processing of new data points**
  - New antigen influences only activated sub-network
  - Activated cells updated incrementally
  - Proposed approach learns in **1 pass**.
- **Clear and fast identification of "outliers"**
  - New antigen that does not activate any subnetwork is a potential outlier ➔ create new B-cell to recognize it
  - This new B-cell could grow into a subnetwork (if it is stimulated by a new trend) or die/move to disk (if outlier)

General Architecture

Evolving data →

1-Pass Adaptive Immune Learning



Evolving Immune Network (compressed into subnetworks)

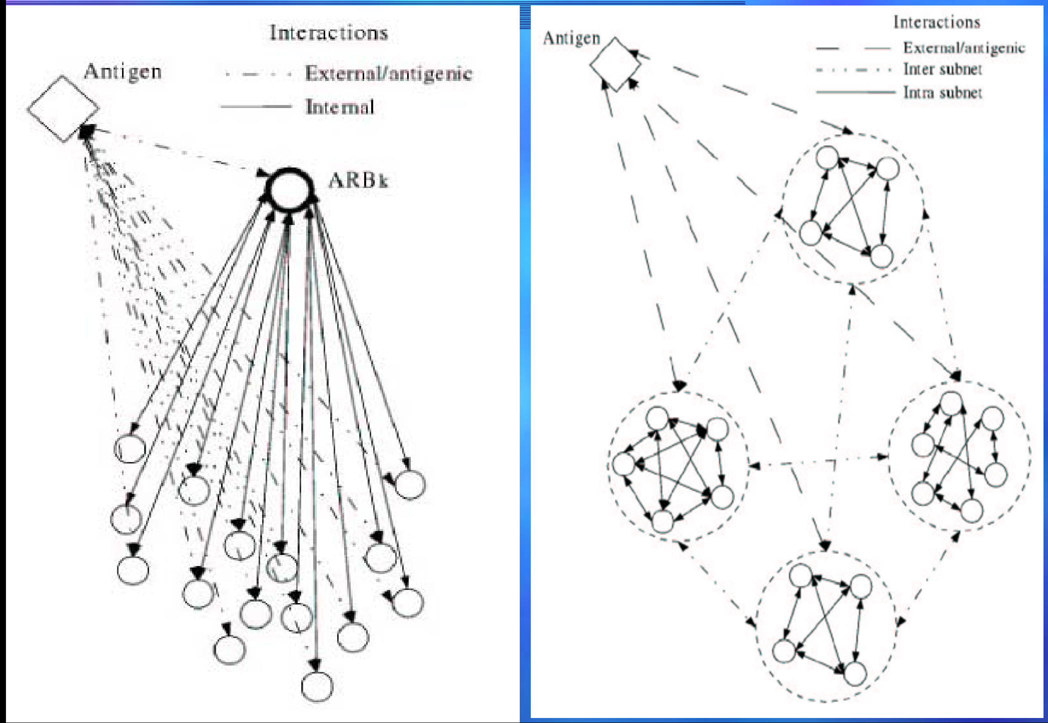
Immune network information system

Stimulation (competition & memory)

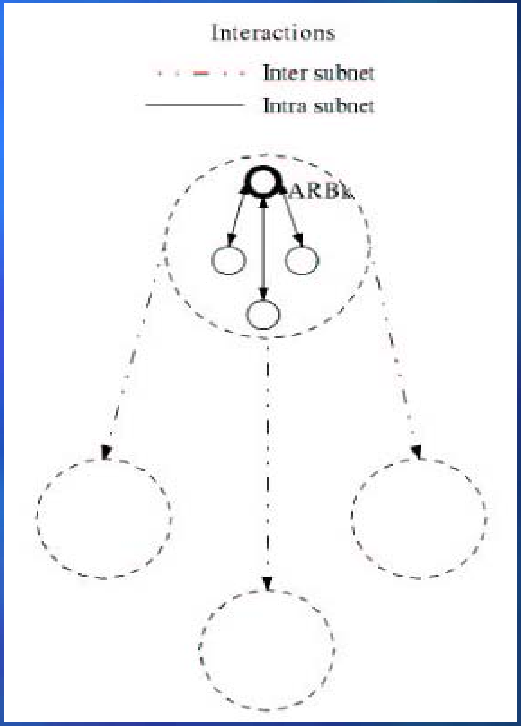
Age (old vs. new)

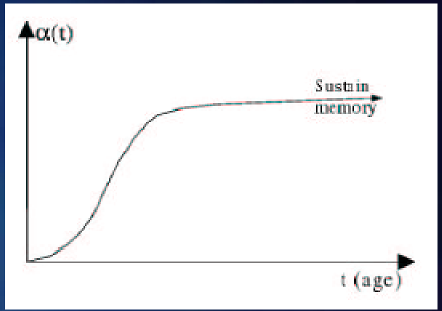
Outliers (based on activation)

Internal and External Immune Interactions: Before & After

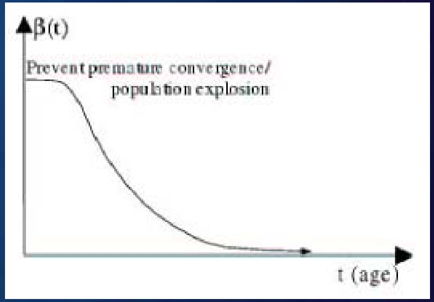


Internal Immune Interactions

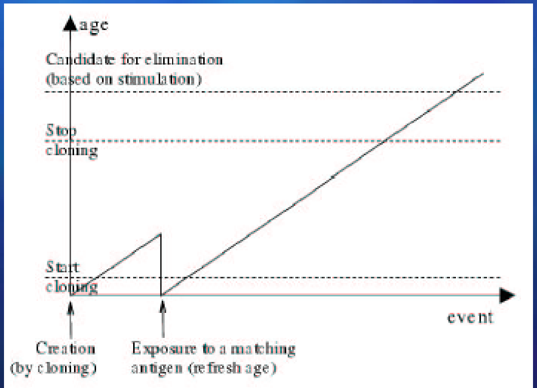




Internal Stimulation

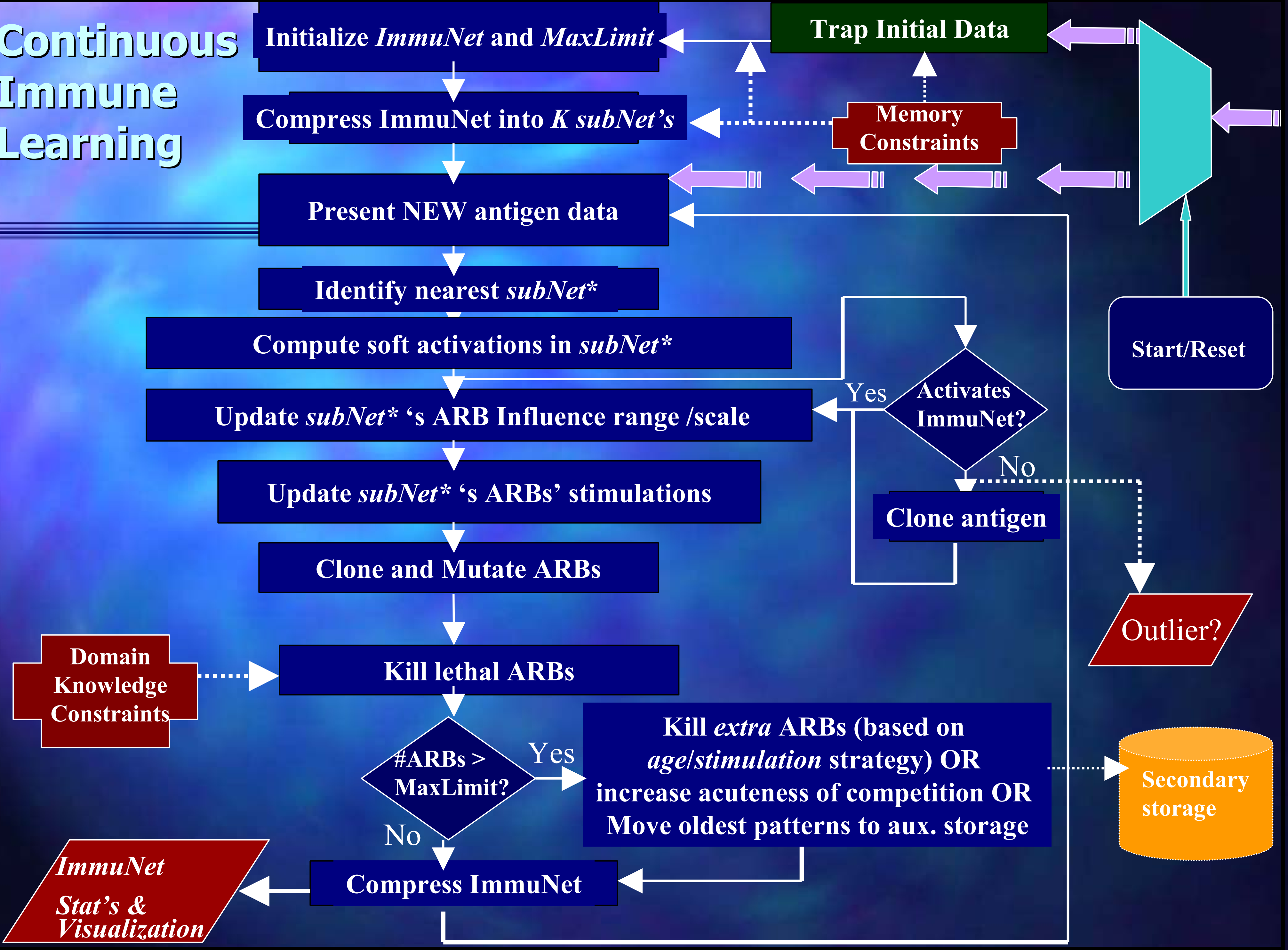


External Stimulation



Lifeline of B-cell

Continuous Immune Learning



Initialize *ImmuNet* and *MaxLimit*

Trap Initial Data

Compress ImmuNet into *K subNet's*

Memory Constraints

Present NEW antigen data

Identify nearest *subNet**

Compute soft activations in *subNet**

Update *subNet** 's ARB Influence range /scale

Activates ImmuNet?

Yes

No

Clone antigen

Start/Reset

Update *subNet** 's ARBs' stimulations

Clone and Mutate ARBs

Outlier?

Domain Knowledge Constraints

Kill lethal ARBs

#ARBs > MaxLimit?

Yes

No

Kill *extra* ARBs (based on *age/stimulation* strategy) OR increase acuteness of competition OR Move oldest patterns to aux. storage

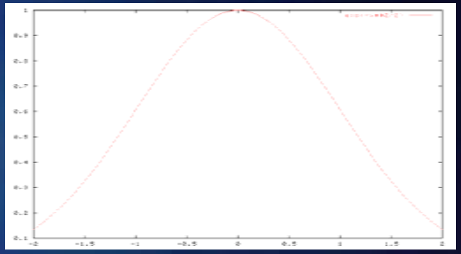
Secondary storage

ImmuNet Stat's & Visualization

Compress ImmuNet

Model for Artificial Immune Cell

- **<u>Antigens represent data</u>** and the **<u>B-Cells represent clusters or patterns to be learned/extracted</u>**
- ARB/B-cell object:
  - Represents not just a single item, but a **<u>fuzzy set</u>**
  - Better **<u>Approximate</u>** Reasoning abilities
  - Each ARB is allowed to have is own **zone of influence** with size/scale: σ_i
  - ARBs **<u>dynamically adapt their influence zones</u>**/hence stimulation level in a strife for survival.
  - Membership function dynamically **adapts** to data
  - **Outliers** are easily detected through weak activations
  - No more dependence on hard threshold-cuts to establish network
  - Can include most probabilistic and possibilistic models of uncertainty
  - Flexible for different attributes types (numerical, categorical, …etc)



Immune Based Learning of Web profiles

- The Web server plays the role of the human body, and the incoming requests play the role of antigens that need to be detected
- The input data is similar to web log data (a record of all files/URLs accessed by users on a Web site)
- The data is pre-processed to produce session lists:
  - A session list S_i for user #*i* is a list of *URLs visited by same user*
  - In discovery mode, a session is fed to the learning system as soon as it is available
- B-cell$_i$: i^{th} candidate profile:
  - List of URLs
  - Historic Evidence/Support: List of supporting cumulative conditional probabilities (URL$_k$, *prob*(URL$_k$)) with *prob*(URL$_k$) = *prob*(URL$_k$ | B-cell$_i$)
  - Each profile has its own influence zone defined by σ_i