Semantics Analysis Monitor for the Illegal Use of the Internet

Raúl Santos de la Cámara (rsantos@hi-iberia.es) – Project Coordinator

10th December 2015

PSC Europe Forum Conference

Project outline

- European project funded by the EU Directate-General for Home Affairs within the framework of the 2012 Prevention of and Fight against Crime Programme.

- January 2014 – December 2015
- Overall goals:
  - Deliver a semi-automated tool able of processing (in near real-time) the public data gathered from OSNs in order to detect the organization and/or promotion of illegal activities.
  - Anticipate where possible physical actions such as:
    - Destruction of private property
    - Threats to individuals
    - Threats to institutions and buildings
    - Terrorist actions/ outbreaks
- Consortium
  - HI Iberia
  - Middlesex University
  - Madrid City Police

POLICÍA MUNICIPAL madrid

HI iberia

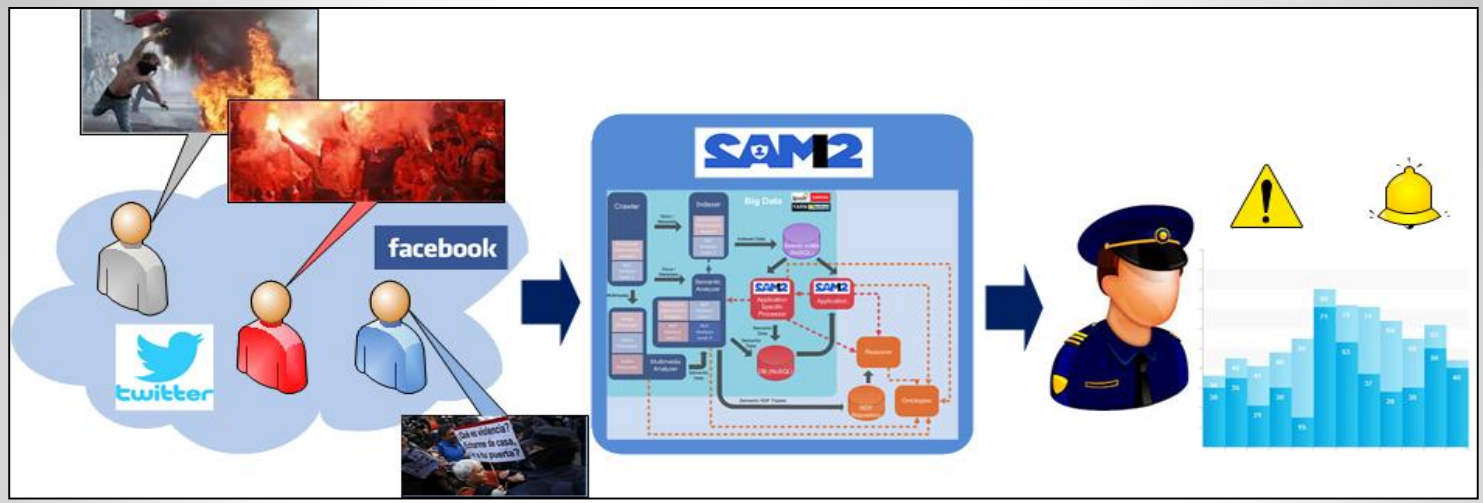
Middlesex University

SAM2

Project framework

- Social networking is one of the most popular online activities worldwide
- The deluge of information exchanged is a mirror of the sciety at large, including mostly legal but also illegal activities.
- This poses new risks that need to be balanced by the security forces:
  - The vast amount of exchanged information
  - The difficulties to detect and prevent dangerous actions among these large amounts of information with limited resources for monitoring

What is SAMi2?



- SAMi2 ensures the integrity of:
  - Social Values through the anticipation of physical actions, such as:
    - Threats to individuals
    - Threats to institutions and buildings
  - Legal and ethical Values related with Private Property and Data Protection have been taking into account about the whole project life cycle.

What is SAMi2?

HI iberia Middlesex University SAM2

Goals

- **SAMi2** is conceived as a potential tool for security forces in automizing the monitoring of OSNs. The high level goals are:
  - To provide a generic technical solution to **crawl** OSNs and extract **relevant** parts of information via the analysis of natural spontaneous text that is produced
  - To ensure that OSN user's **rights of privacy and data ownership** are respected
  - To ensure that provisions in the **law** or the OSNs **Terms of Service** are upheld
  - To enforce an **end-user centric perspective** to produce a professional tool for use by security forces

How does it work?

- Through a crawler, SAMi2 gathers information from OSNs
- The information is semantically processed using different tools for Natural Language Processing in order to make the information understandable by SAMi2
- Through the use of machine learning, SAMi2 is able to provide results based on the search profile selected by users and to learn from them

HI iberia Middlesex University SAMI2

Characteristics

- Currently, SAMi2 only processes information stored in Tweets:
  - Textual documents (previously anonymized)
  - Structured and unstructured information

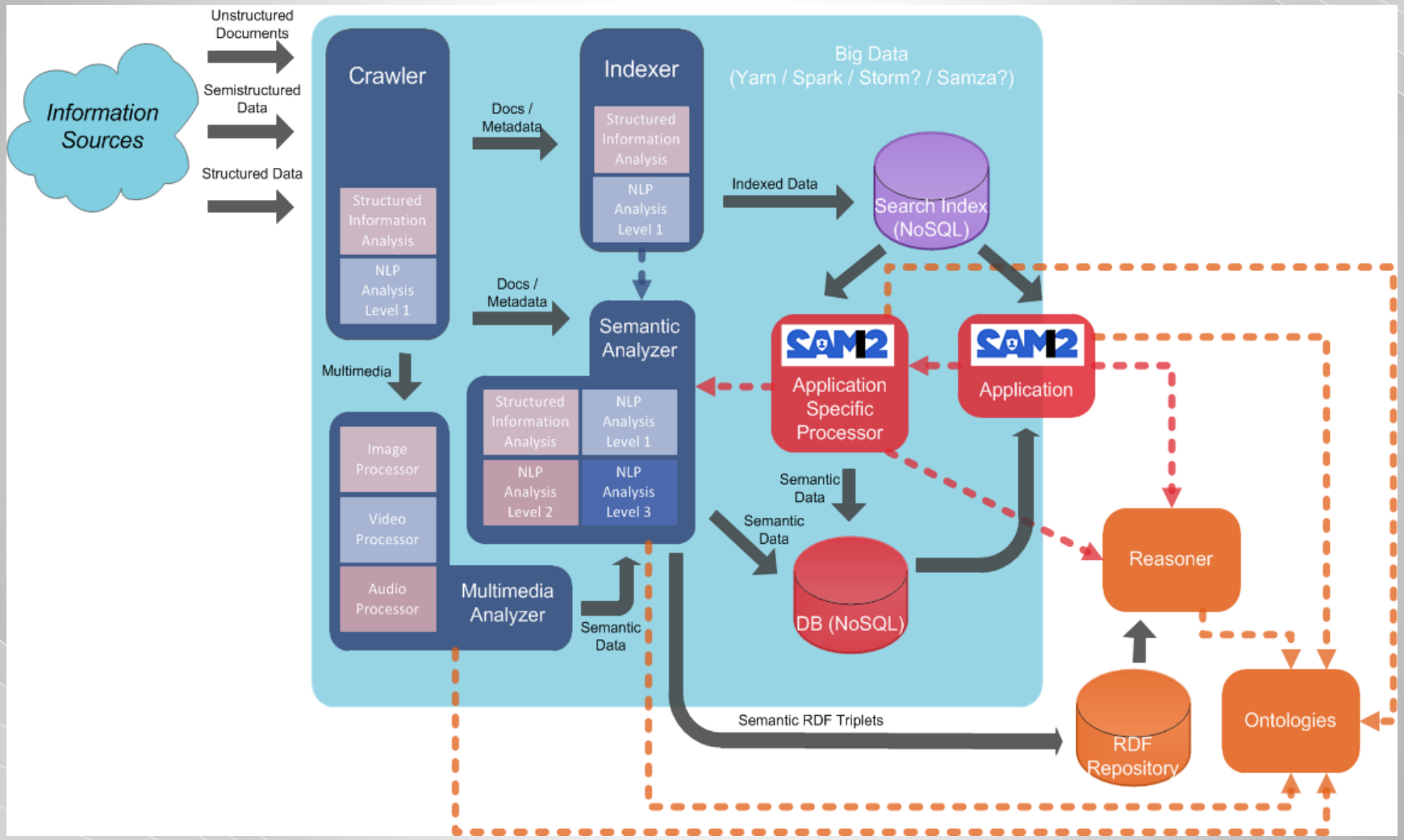
- **Extensible** architecture able to add new processing capabilities
  - Raw Multimedia Sources: Audio, Video, Audio
  - Capability of integration with other OSNs, e.g. Facebook, Instagram
  - Increasing the processing capability adding more machines
  - Other semantic processing levels:
    - Social Analysis: Relationships between persons involved in the messages
    - Sentimental Analysis: positivity or negativity of the feelings involved

Characteristics

- Architecture based on **open-source** technologies under:
  - The Big Data paradigm
  - Natural Language processing framework
  - Machine learning
- Specific use of **semantics** for **Data Mining**
- Well-documented **open** and **extensible interfaces** to connect this information with the information systems currently used by security forces
- **Practical**. Users have been involved in the whole project cycle to achieve tangible results.

HI iberia Middlesex University SAM2

Architecture (I)



HI iberia Middlesex University SAM2

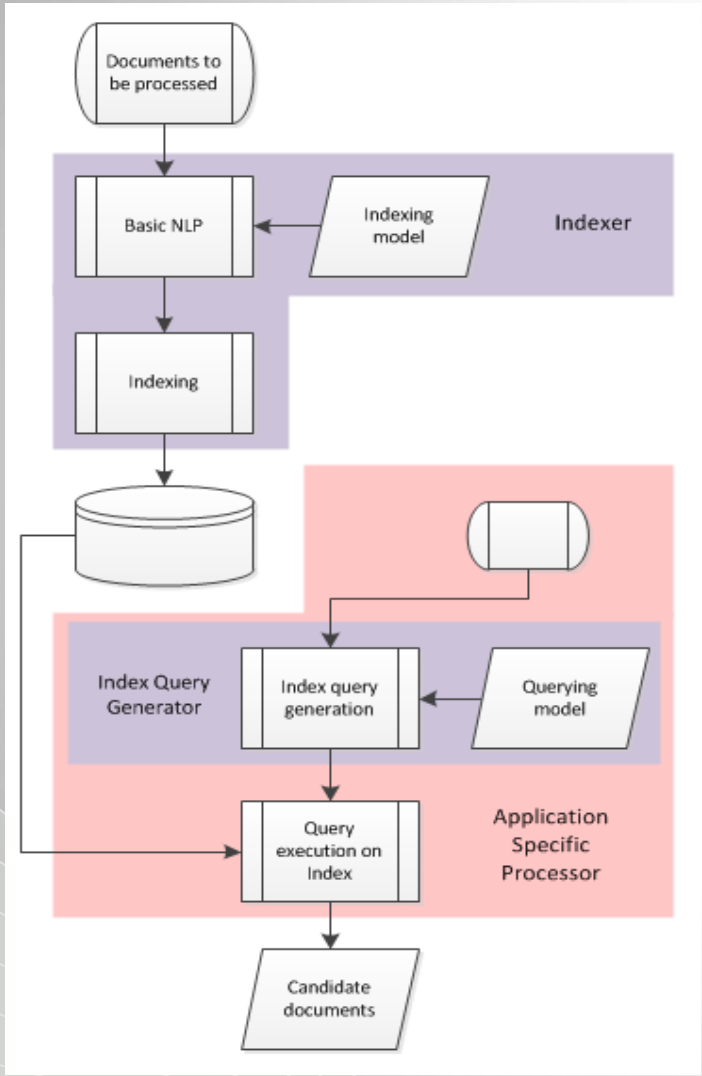
Architecture (II)

- **Crawler:** message/document adaptation and retrieval.
- **Indexer:** Optimize access to messages/documents. Enable *a posteriori* analysis. First filtering.
- **Anonymizer:** Ensure privacy in processed data.
- **Semantic Analyzer:** Deep natural Language Processing. Extraction of general and domain concepts, semantic relations, relevant individuals. Network structures. Sentiment analysis.
- **Multimedia Analyzer:** Extraction of usable information from multimedia elements (future): image indexing, object recognition, etc.
- **Semantic Framework:** General and domain specific ontologies. Reasoning.

Architecture (III)

- **Application Specific Processor:** Specific high level processing tasks for every application domain.
- **Machine Learning Module:** provides learning capabilities to the other modules (neural networks, SVMs, etc.)
- **Application:** Provide any further adaptation to the final application domain
- **Functional components deployed over a scalable Big Data based cluster:**
  - Indexing based on Apache Solr
  - Semantic processing distributing on an Apache Spark cluster
  - Distributed data storage based on MongoDB

Processing Flow (I)



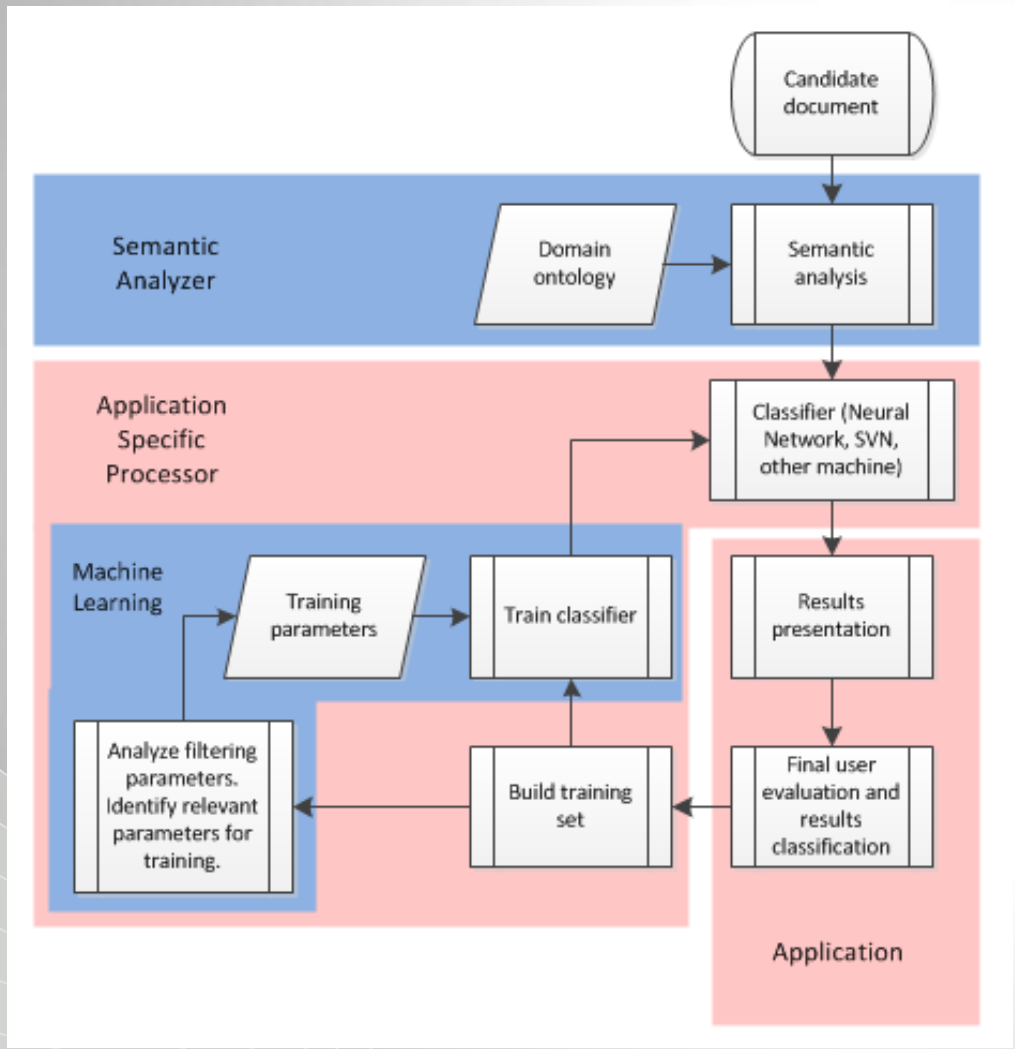
First processing level:

- Basic NLP (Spanish)

- Improved indexing of Spanish messages, including lexical analysis and specific domain dictionaries (manage Twitter slang). In fact, this involves a shallow semantic analysis.

- First level of filtering, aimed to limit the number of messages to be deeply semantically analyzed.

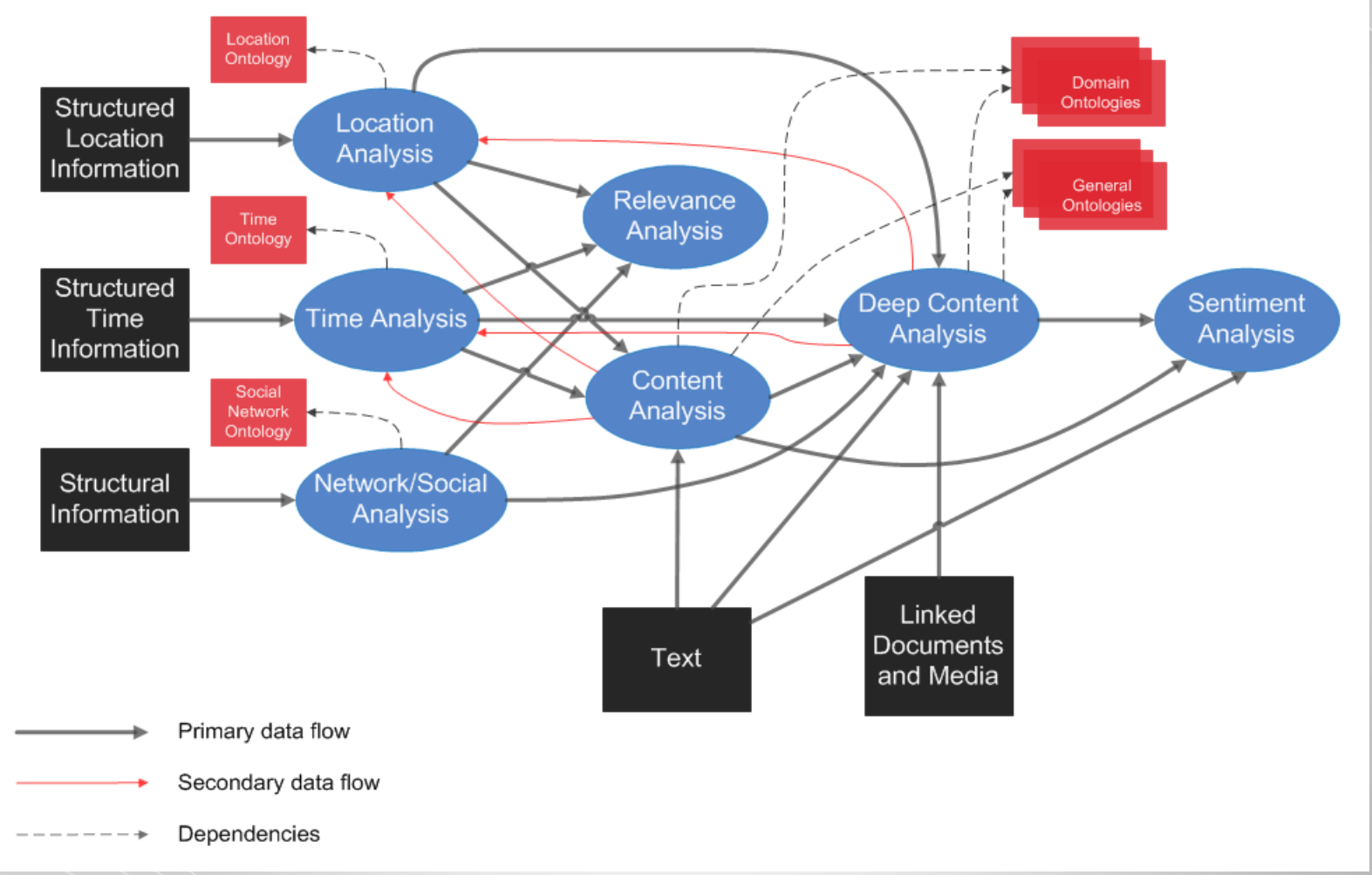
Processing Flow (II)



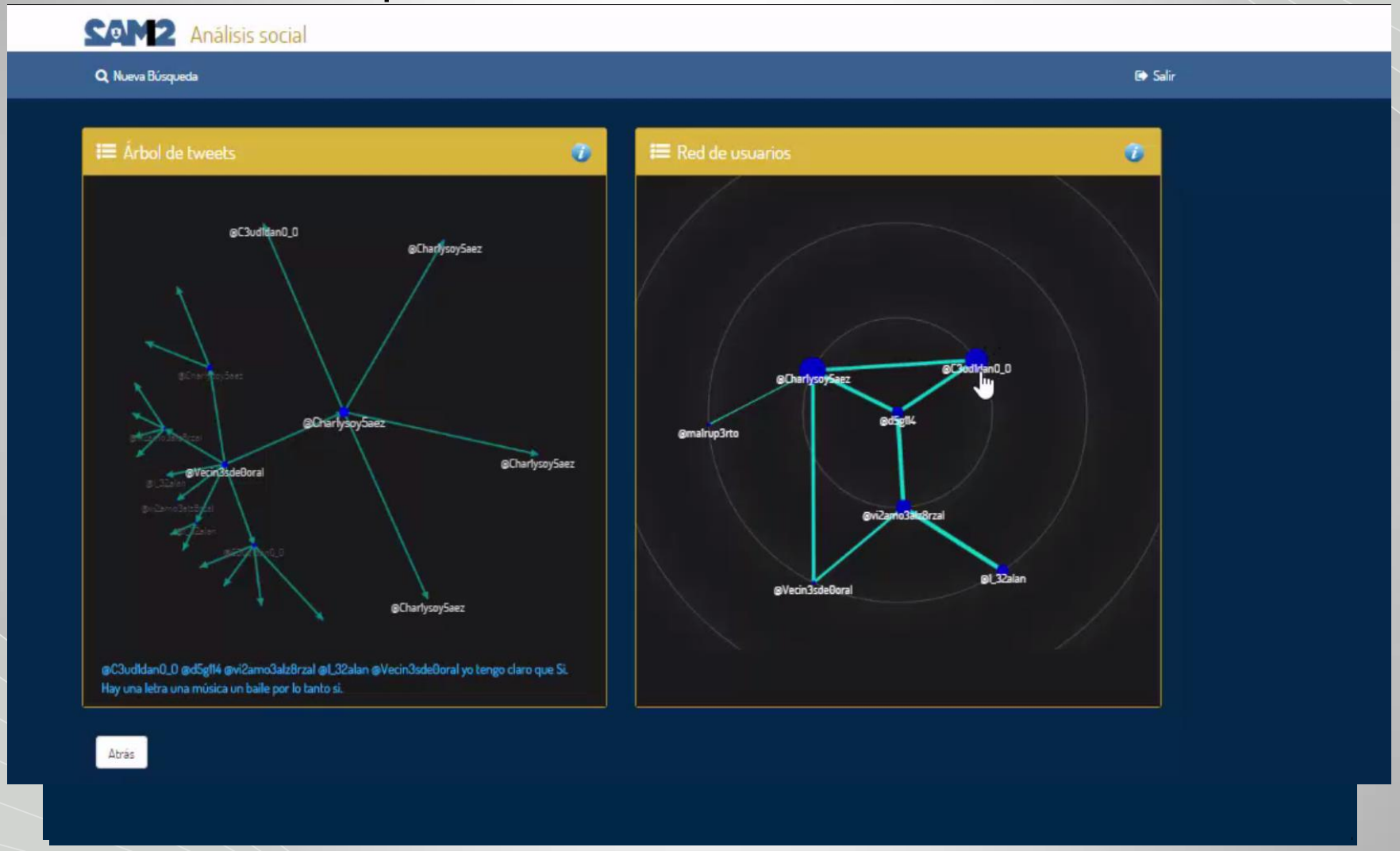
Second processing level:

- Deep semantic analysis of the messages identified in the previous steps.
- Machine learning
- Specific application processing

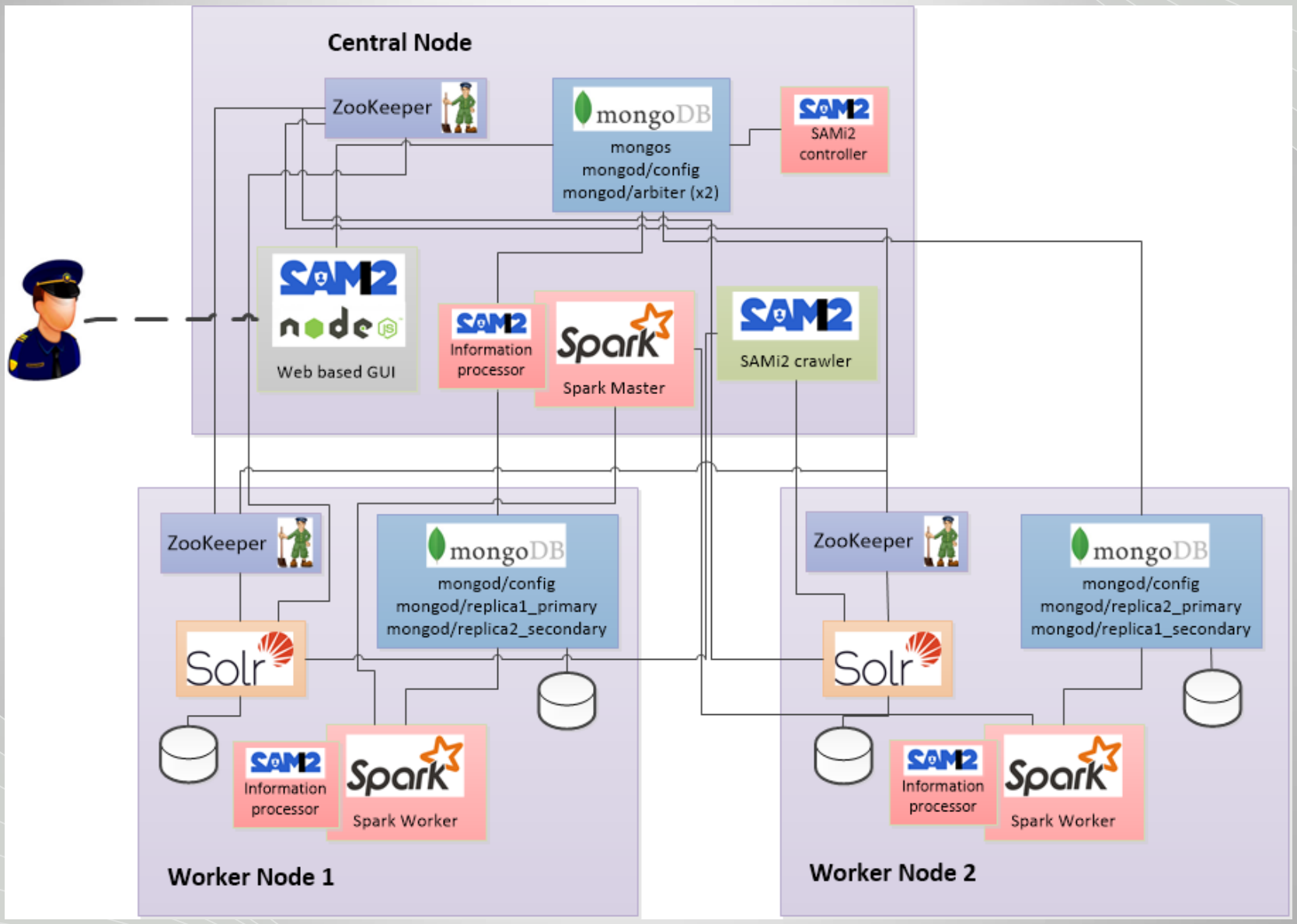
Semantic Processing



Graphical User Interface



Deployment


Central Node
ZooKeeper
mongoDB
mongos
mongod/config
mongod/arbiter (x2)
SAMi2
SAMi2 controller
SAMi2
node
Web based GUI
SAMi2
Information processor
Spark
Spark Master
SAMi2
SAMi2 crawler
ZooKeeper
mongoDB
mongod/config
mongod/replica1_primary
mongod/replica2_secondary
Solr
SAMi2
Information processor
Spark
Spark Worker
Worker Node 1
ZooKeeper
mongoDB
mongod/config
mongod/replica2_primary
mongod/replica1_secondary
Solr
SAMi2
Information processor
Spark
Spark Worker
Worker Node 2

Use Cases

- Different search profiles have been defined in SAMi2:
  - **Escrache.** SAMi2 is used to detect messages for the organization of demonstrations in which a group of activits go to the homes or workplaces of those who they want to condemn and publicly humiliate them in order to influence decision makers and governments into a certain course of action.
  - **Gangs.** SAMi2 is used to detect suspicious messages related to gangs' conflicts on the Internet.
  - **Illegal Events.** SAMi2 is used to detect and prevent the organization of illegal events during the day of reflection.
  - **Hate Speech.** SAMi2 is used to detect racist messages or hate speeches on the Internet.
  - **Post-mortem Analysis.** SAMi2 is used to detect if a conflict is postponed after it is stopped

Not-Use Cases

- **SAMi2** is designed for the prevention of illegal activities planned through OSNs, however, it **does not cover other** activities related with **cyber-crime**, such as:
  - Attacks against information systems
  - Information tampering
  - Theft of personal data
  - Identity theft

Trials

- The SAMi2 tool is being used and tested in realistic conditions by members of the Madrid City Council police
- They have provided their feedback along the whole project cycle
- The tool is envisaged to be made available to other security forces in later stages of R&D and product phases

Future lines

- Extensions with other analysis tools:
  - Crawling other OSNs (Facebook, Instagram) and deep/dark web
  - Sentiment analysis
  - Multimedia (i.e., text in pictures, faces) analysis
  - Extensions to other languages
- Added benefits from longer and more realistic operational runs:
  - Connection to full Twitter API (now just the public API)
  - Topics for search profiles expanded: new categories of end-users, learning through analysis of events.
  - Cross OSN profiles of persons

Thanks for your attention!

Video and poster available on the hall

Q+A

http://sami2-project.eu

10th December 2015

PSC Europe Forum Conference