Fast user classifying to establish forensic analysis priorities

A.Grillo, **A. Lentini**, G.Me, M.Ottoni

University of Rome “Tor Vergata”
Department of Computer Science, Systems and Production
lentini@disp.uniroma2.it

IMF 2009: 5th International Conference on IT Security Incident Management & IT Forensics

Outline

1. Introduction

2. The Methodology

3. Case Study

4. Conclusions

The problem.

Kryder's Law states that "the density of information on hard disks has been growing at an even faster rate, increasing by a factor of 1000 in 10.5 years, which corresponds to a doubling roughly every 13 months".

1. The seized material could be very ***huge***.
2. Only few devices are considered ***relevant*** for the investigation.

The target.

Setting priorities for the analysis investigation could be definitively important in order to dramatically cut the response time of the analysis.

This work presents a methodology and a tool in order to profile in a fast way the computer's user via a classification into forensic operator predefined categories.

The *target* is to obtain a fast response of the typology of the user of the seized computer, in order to efficient-schedule the analysis of the seized material.

State of the art

There are two categories of computer forensics tools: translation tools and presentation tools.

Many computer forensic tools suites implement both categories of tools within the same package:

- EnCase® Forensic Guidance Software
- Forensic Toolkit® Access Data
- Helix E-Fense

Guidance® SOFTWARE

AccessData® A Pioneer in Digital Investigations Since 1987

e-fense® CARPE DATUM

State of the art

What are the limitations of existing solutions?

- The analysis must wait the processing of all the extracted files
- The effective analysis of information extracted is completely in charge of the forensic operator and his/her experience
- The software tools (as well as the methodology) are "static" or there is no mechanism of evolution based on the experience gained by the forensic analyst

The ***novelty*** of our methodology:

- establish an analysis priority schedule of seized hard drive based on the computer user profiling via machine learning techniques;
- reduce the time to select appropriate hard drive to examine more accurately (with the existing tool) and to discard data not useful in the investigation process.

Outline

1. Introduction
2. The Methodology
3. Case Study
4. Conclusions

The methodology

Preliminary Phase:

Definition of target classes:

- identification and choice of interesting users ***profiles***;

- identification of attributes (called ***features***) that characterize the identified user profiles to be classified

This stage can also be a refinement of the definitions (profiles and features) already identified in previous investigations

Operational Phase:

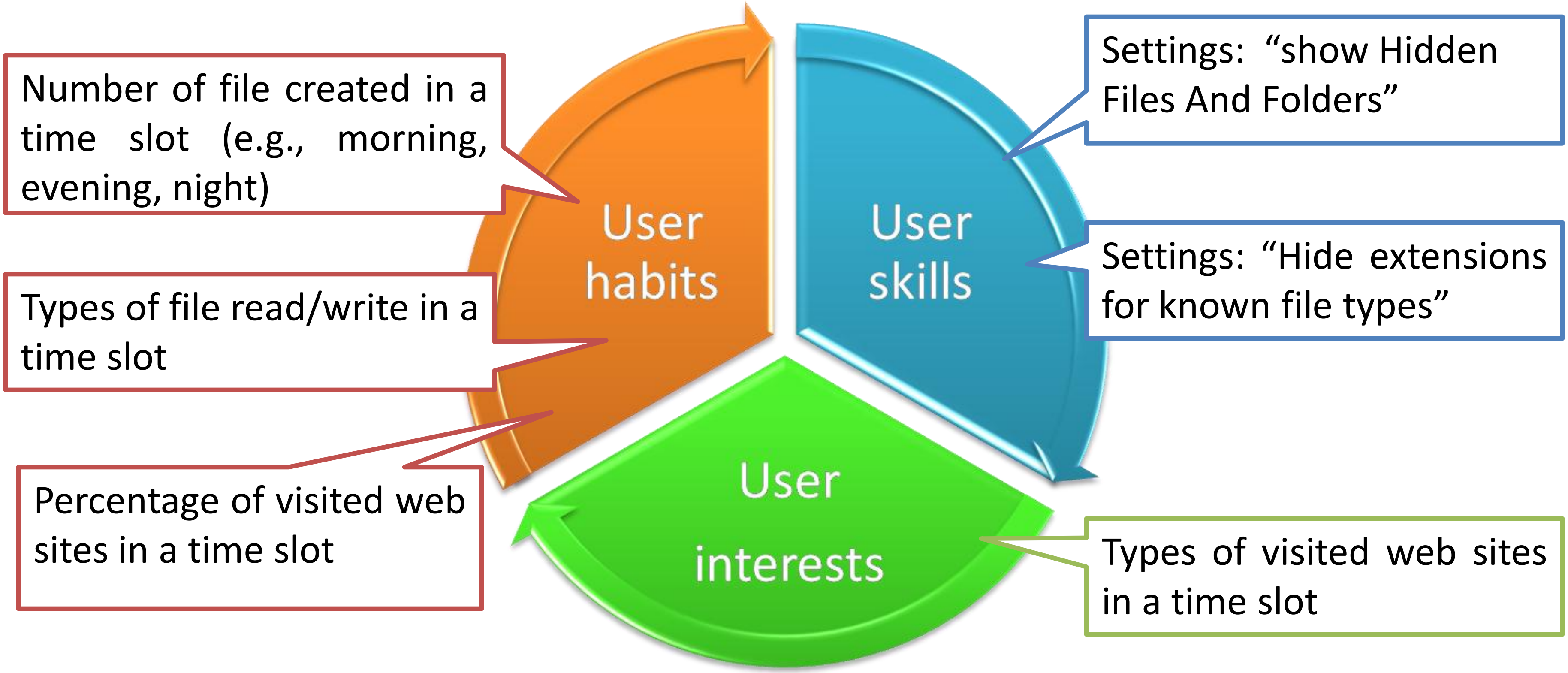
1. Automatic ***extraction*** of relevant information
2. Automatic ***elaboration*** of relevant information extracted
3. Automatic ***classification*** of the user profile

The methodology - Preliminary Phase

Preliminary Phase:

Definition of target classes:

- identification and choice of interesting users ***profiles***;
- identification of attributes (called ***features***) that characterize the identified user profiles to be classified



Number of file created in a time slot (e.g., morning, evening, night)

Types of file read/write in a time slot

Percentage of visited web sites in a time slot

Settings: “show Hidden Files And Folders”

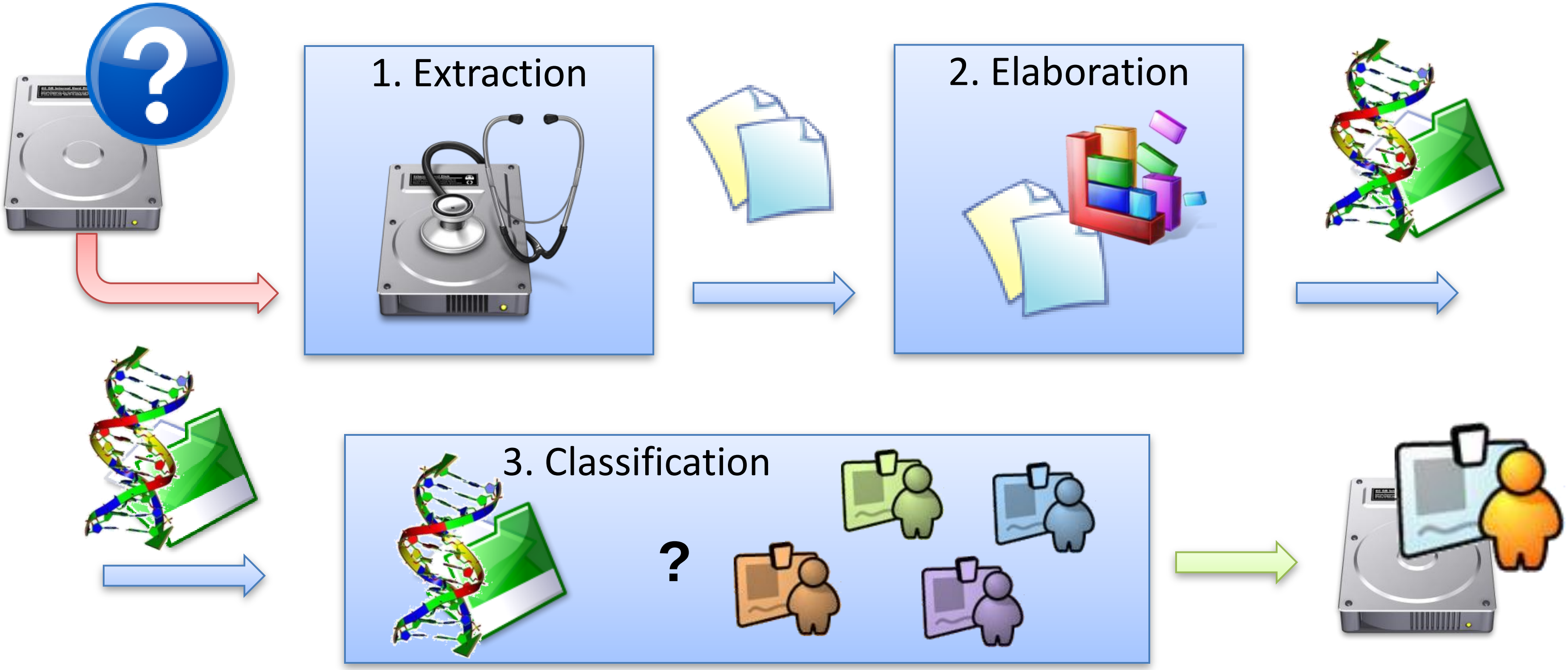
Settings: “Hide extensions for known file types”

Types of visited web sites in a time slot

The methodology – Operational Phase

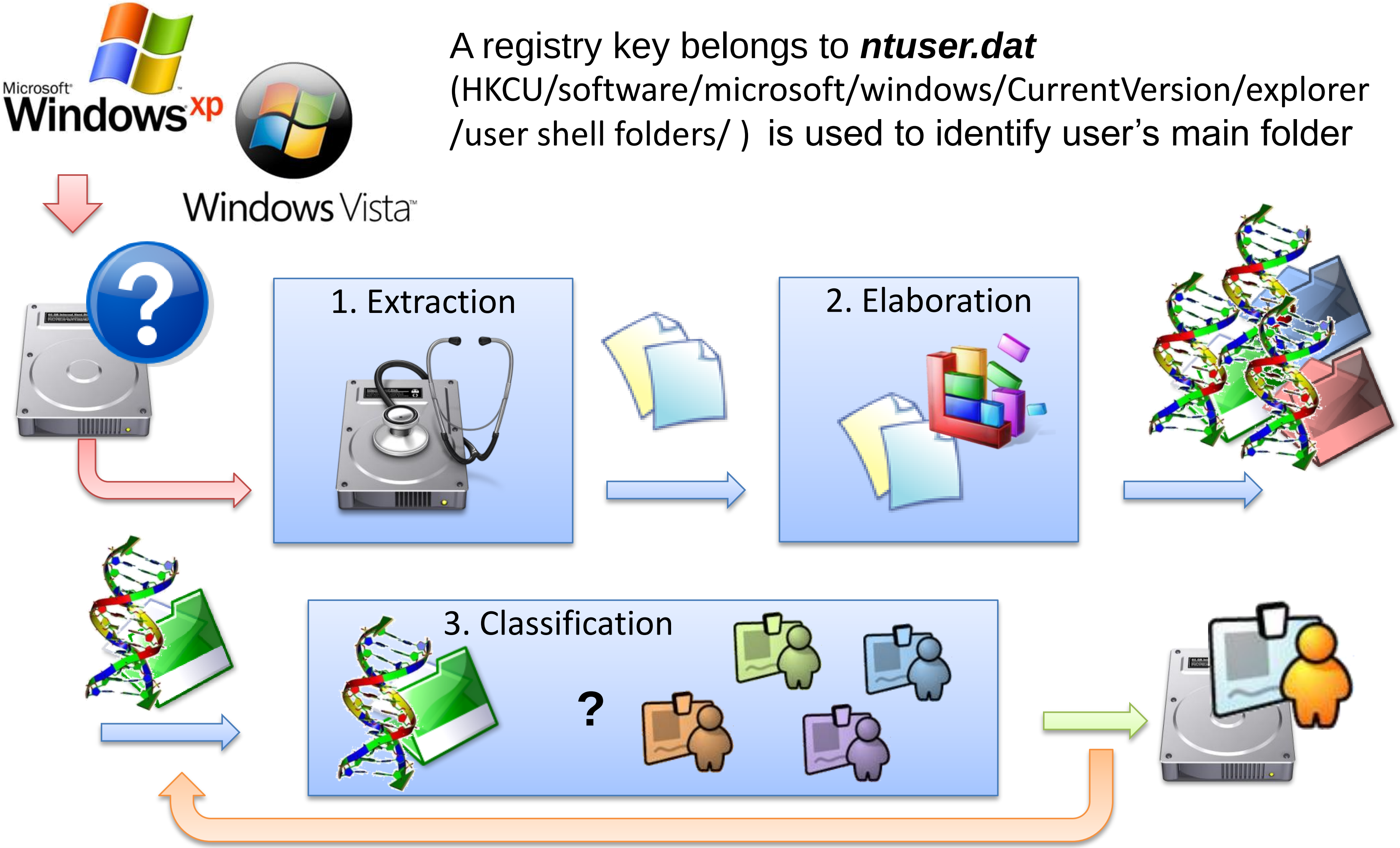
Operational Phase:

1. Automatic ***extraction*** of relevant information
2. Automatic ***elaboration*** of relevant information extracted
3. Automatic ***classification*** of the user profile



Operative systems

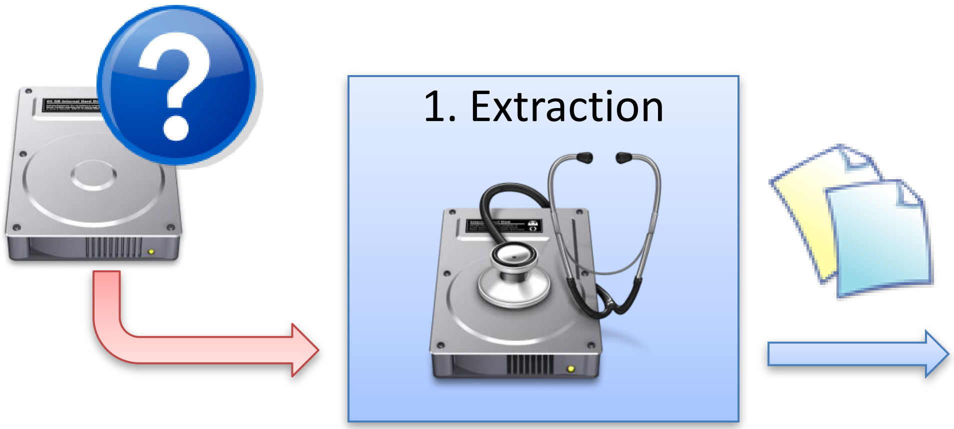
A registry key belongs to ***ntuser.dat*** (HKCU/software/microsoft/windows/CurrentVersion/explorer /user shell folders/) is used to identify user’s main folder



The methodology – Operational Phase (1)

Operational Phase:

1. Automatic ***extraction*** of relevant information



- **Internet Explorer:**
history, temporary files, bookmarks
- **Mozilla Firefox 2-3:**
history, downloaded files, bookmarks, search bar content
- **Windows registry:**
users of the system, installed sw
- **Statistics:**
    - number of files on filesystem
    - extensions of file on filesystem
    - timetable of creating files
    - timetable of web sites visited
    - list of sw programs installed

    …

Tools of SleuthKit© utilized:

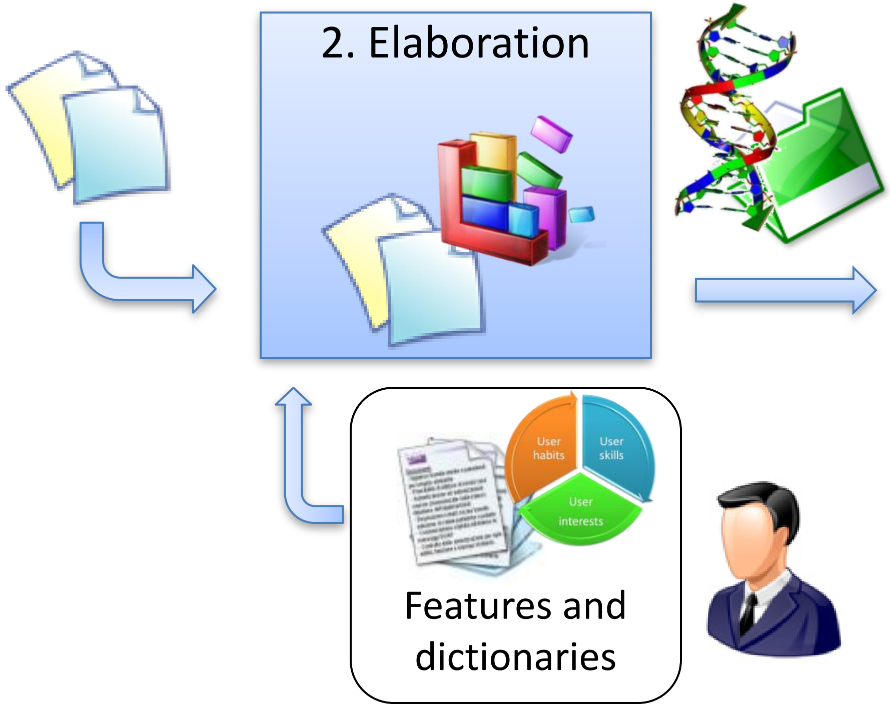
- fsstat
- fls
- ifind
- icat

RegLookup© used to extracted the keys from the register

The methodology – Operational Phase (2)

Operational Phase:

2. Automatic ***elaboration*** of relevant information extracted



2.1 The raw data extracted is processed through specific tools*

2.2 All the processed information is converted and stored in an output file with the same standard XML format;

2.3 The subset of most interesting information is selected as suggested by the forensic operator in the preliminary phase;

2.4 From the selected subset, the vector of features that characterize the system is populated; in this step the information is compared to the suggested dictionaries given by the operator.

*tools utilized:

- Pasco: to convert the internet explorer files (index.dat)
- Mork : to convert the Firefox files with extension .mork in textual format

The methodology – Operational Phase (3)

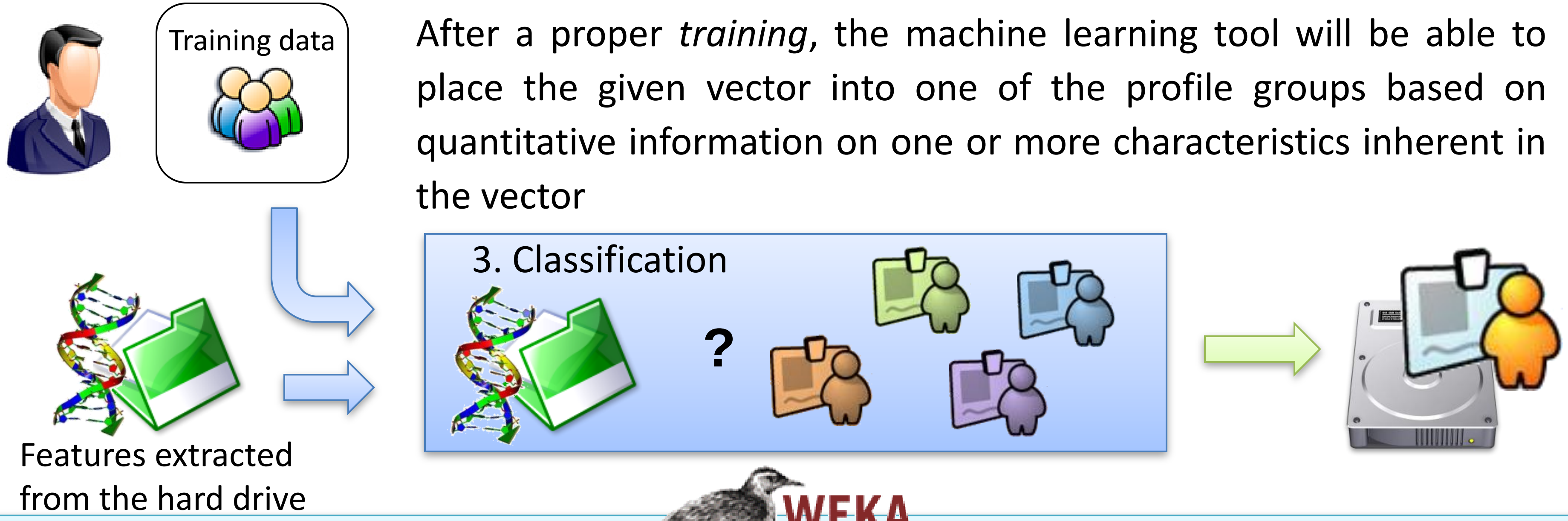
Operational Phase:

3. Automatic ***classification*** of the user profile

The classification follows an approach based on **Machine Learning**:

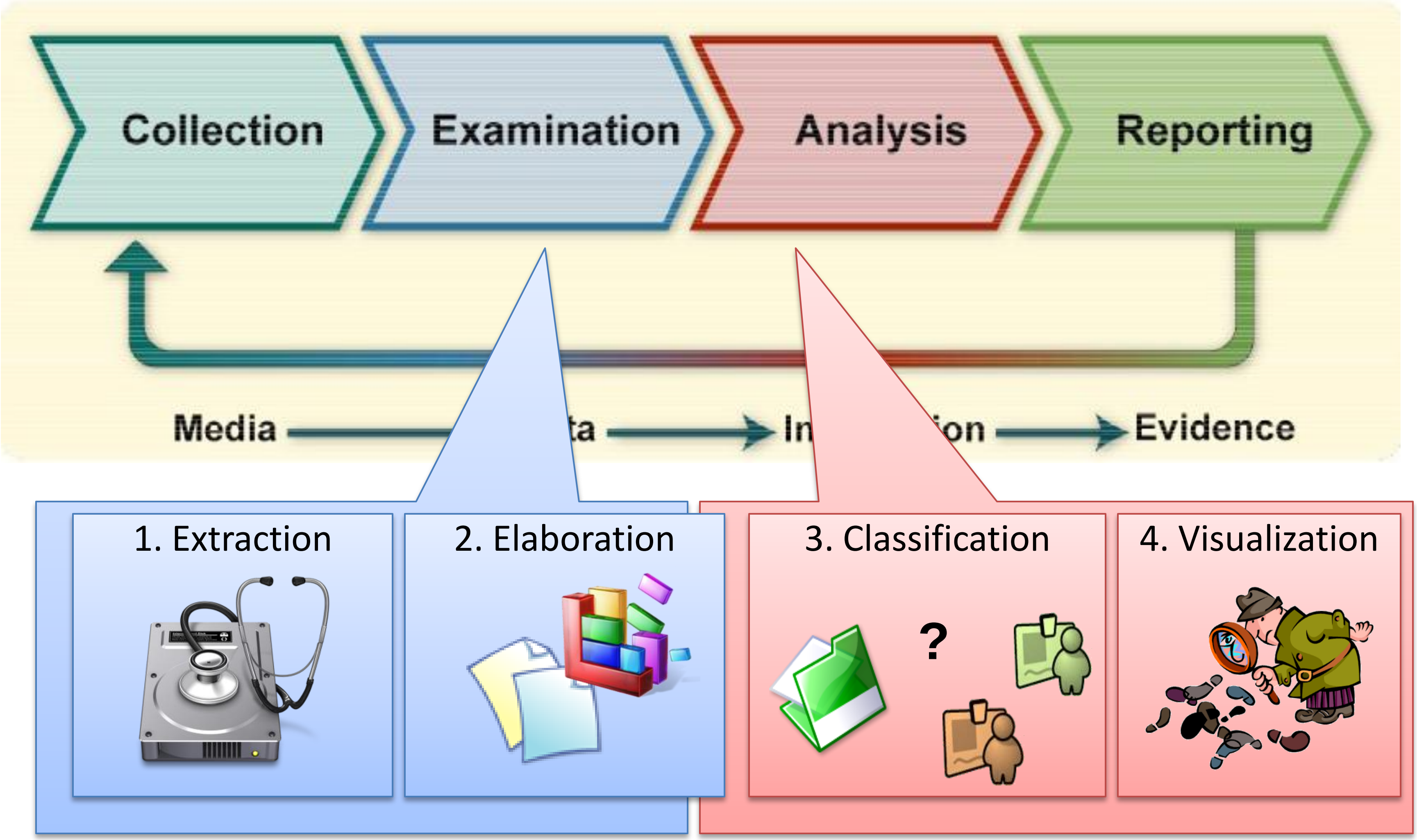
the relevant features characterizing a system user usage, extracted in the previous steps, produce a vector to be handled by the machine learning tool “Weka workbench”.

After a proper *training*, the machine learning tool will be able to place the given vector into one of the profile groups based on quantitative information on one or more characteristics inherent in the vector



WEKA
The University
of Waikato

The forensics process



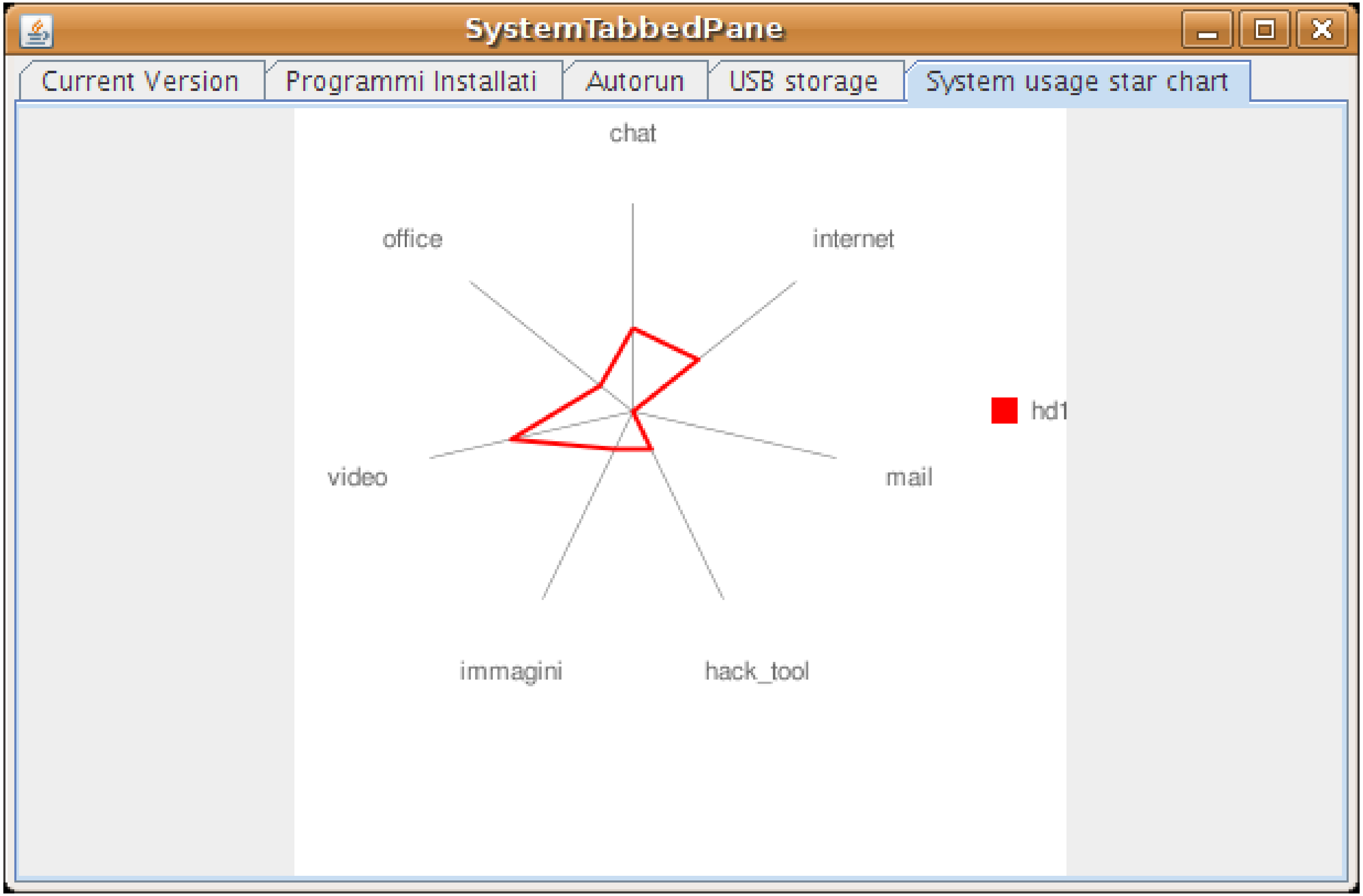
1. Extraction

2. Elaboration

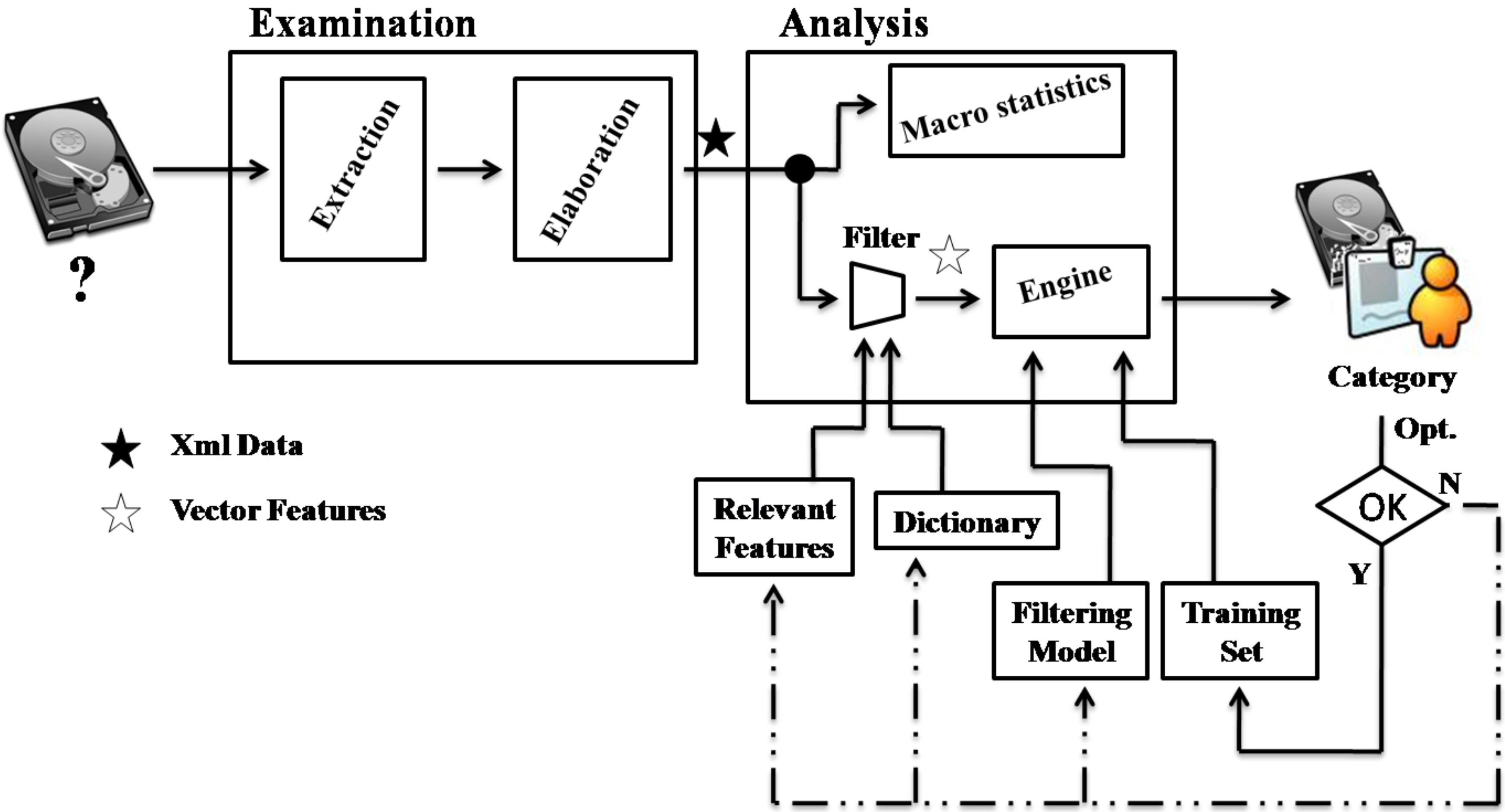
3. Classification

4. Visualization

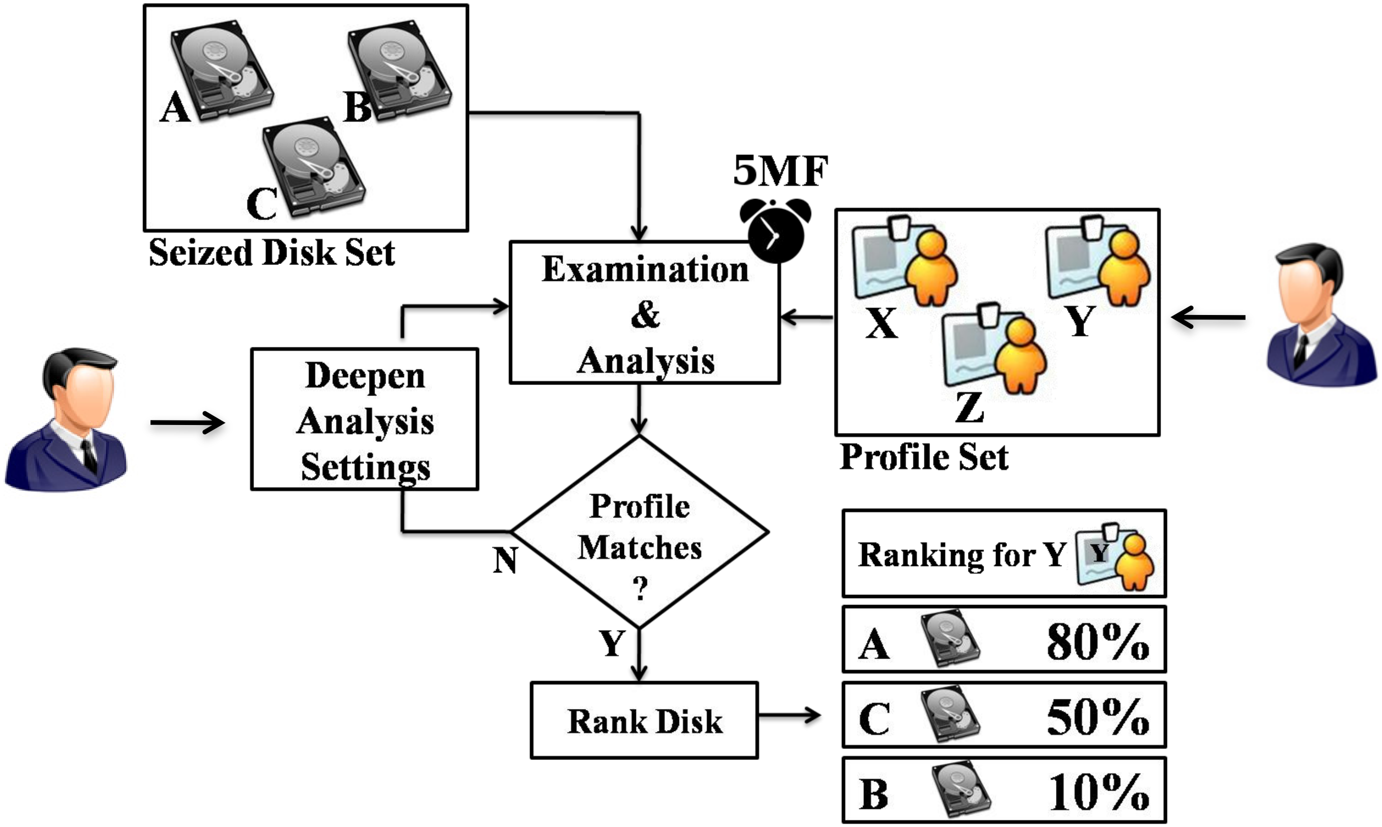
4. Visualization



The methodology



The methodology – Execution Time



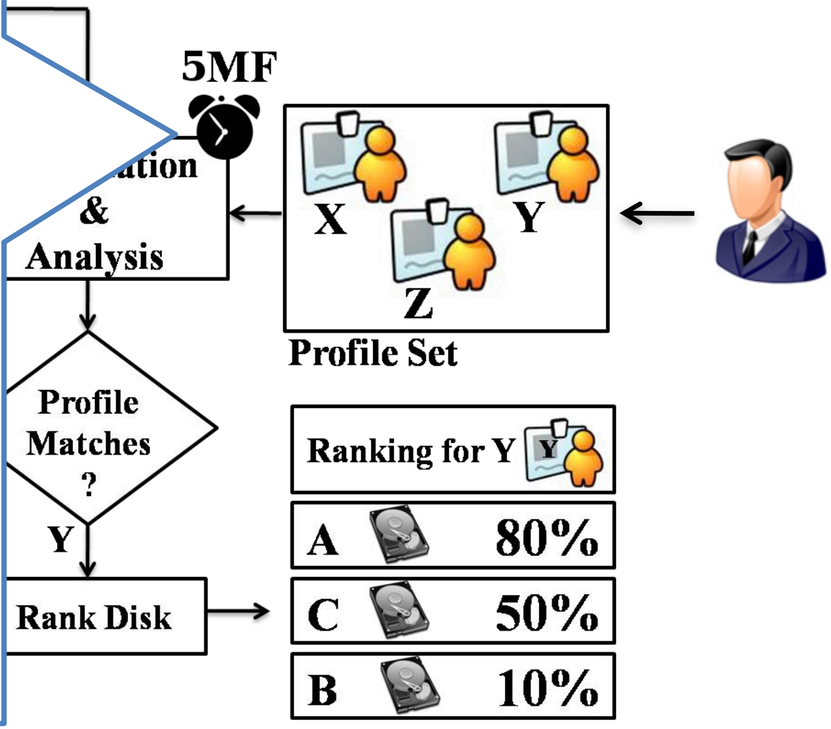
The methodology allows to compare via a machine learning approach the seized disk set to a profile set defined by the investigator. This could suggest a list of priorities for disks analysis.

In the case of single seized disk can address the subsequent analysis suggesting a user profile

The methodology – Execution Time

Image Size	68 GB
Image Files	70622
Image O.S.	Win XP
Platform	Pentium M 1,4 GHz RAM 1 GB
Extraction	< 10 sec.
Elaboration	< 38 sec.
Analysis	< 60 sec.

5 Minute Forensic: The tool we developed carries out the examination and analysis in less than 5 minute



The methodology allows to compare via a machine learning approach the seized disk set to a profile set defined by the investigator. This could suggest a list of priorities for disks analysis.

In the case of single seized disk can address the subsequent analysis suggesting a user profile

Outline

1. Introduction
2. The Methodology
3. Case Study
4. Conclusions

Case study- Preliminary Phase

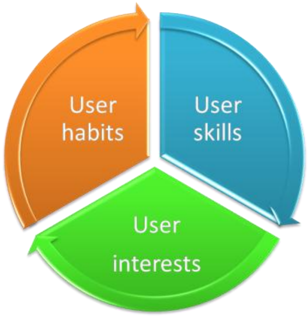
Preliminary Phase:

Definition of target classes:

- identification and choice of interesting users ***profiles***;
- identification of attributes (called ***features***) that characterize the identified user profiles to be classified

5 sample categories of interesting users profiles:

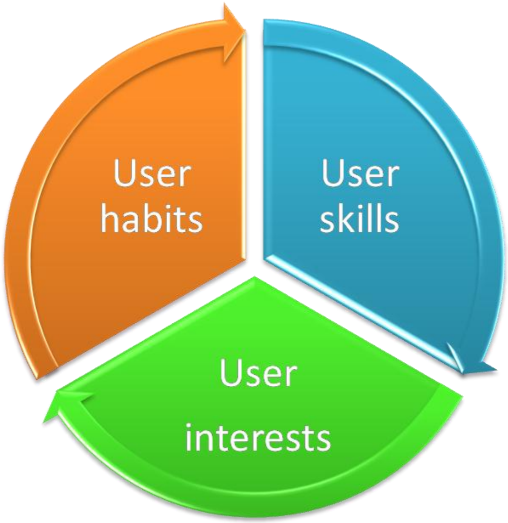
- “***Occasional user***” uses the computer for simple sporadic tasks;
- “***Web user***“ uses the PC to surf the web and chat with friend;
- “***Office worker user***” uses the PC to create and edit documents;
- "***Experienced user***“ has some advanced skill and able to change system settings ;
- “***Hacker user***" uses advanced tools and visits websites highly specialized.



100 sample features considered to represents a user profile

Case study- Preliminary Phase

Sample ***features*** considered to represents a user profile :



User skill:

- CV_SP3_present {true,false} (service pack 3 installed?)
- hide_extension {true,false} (Hide extensions for known file types setting)
- hide_system_file {true, false} (show system Files And Folders)
- hide_hidden_file {true, false} (show Hidden Files And Folders)
- multi_user {true,false} (more than one user account?)

User habits:

- USB_count (number of usb external device connected)
- 5 file_percent (percentage of audio/video/image/document/exe files)
- 5 file_average_size (average size of audio/video/image/doc/exe files)
- 5 file_creation_time_slot_1 (percentage of file created during the morning time)
- 5 file_creation_time_slot_2 (percentage of file created during the afternoon time)
- 5 file_creation_time_slot_3 (percentage of file created during the evening time)
- 5 file_creation_time_slot_4 (percentage of file created during the nigth time)

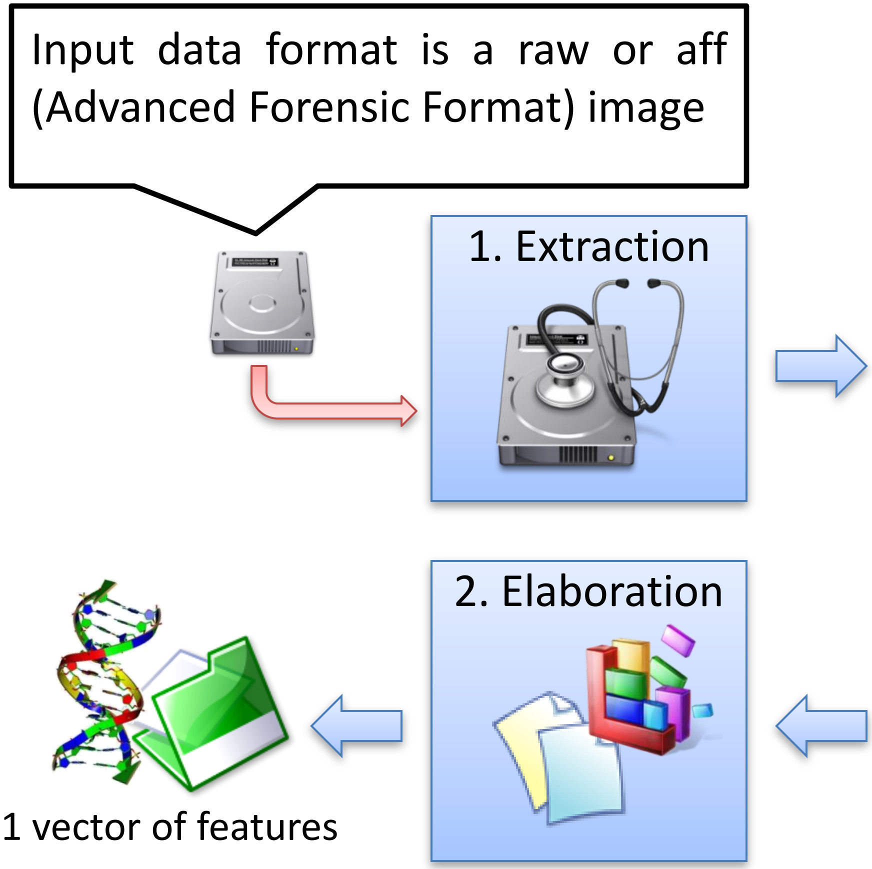
User interests:

- 7 sw_kind (number of sw installed: chat, image tool, browser, email client, office tool, hacker tool…)
- 7 ie_url_kind_visited (number of web site visited divided by category: news, free time, hot…)
- 4*7 ie_url_kind_time_slot_visited (percentage of file created during the 4 slot time divided by site kind)
- 7 ff_url_kind_visited (number of web site visited divided by category: news, free time, hot…)
- 4*7 ff_url_kind_time_slot_visited (percentage of file created during the 4 slot time divided by site kind)

Case study – Operational Phase (1-2)

Operational Phase:

1. Automatic ***extraction*** of relevant information
2. Automatic ***elaboration*** of relevant information extracted

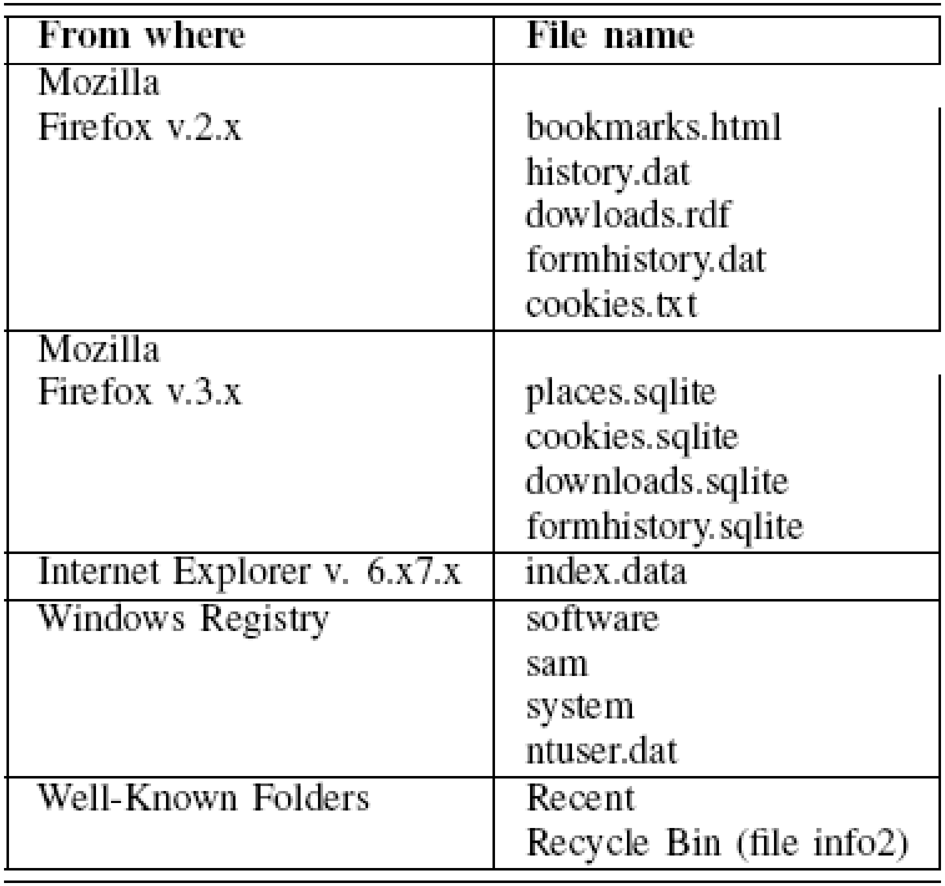


Input data format is a raw or aff (Advanced Forensic Format) image

1. Extraction

2. Elaboration

1 vector of features



From where	File name
Mozilla Firefox v.2.x	bookmarks.html history.dat dowloads.rdf formhistory.dat cookies.txt
Mozilla Firefox v.3.x	places.sqlite cookies.sqlite downloads.sqlite formhistory.sqlite
Internet Explorer v. 6.x7.x	index.data
Windows Registry	software sam system ntuser.dat
Well-Known Folders	Recent Recycle Bin (file info2)

Most relevant files.

Two main categories:

- information extracted from registry files;
- information extracted from the file cached by the browsers.

HKEY_USERS , HKEY_CURRENT_USER, HKEY_LOCAL_MACHINE...

- MRU
- SOFTWARE (CurrentVersion CurrentVersion\Run …)
  <Programs> (Main, TypedURLs, SearchMRU..)
- SYSTEM(Tcpip\Parameters\Interfaces, USBSTOR …)

Browser

- bookmarks: site URL;
- history: site URL, number of visits, typed URL with the date/time;
- form history: strings typed into the search box of Firefox;
- downloads: URL of the downloaded resources and the destination path.

Case study – Operational Phase (3)

Operational Phase:

3. Automatic ***classification*** of the user profile

Validation of the features and training of the machine learning tool:

- 25 features vectors of users profile (5 for each sample categories)

- ten-fold cross validation technique;

Tested different algorithms of classification:

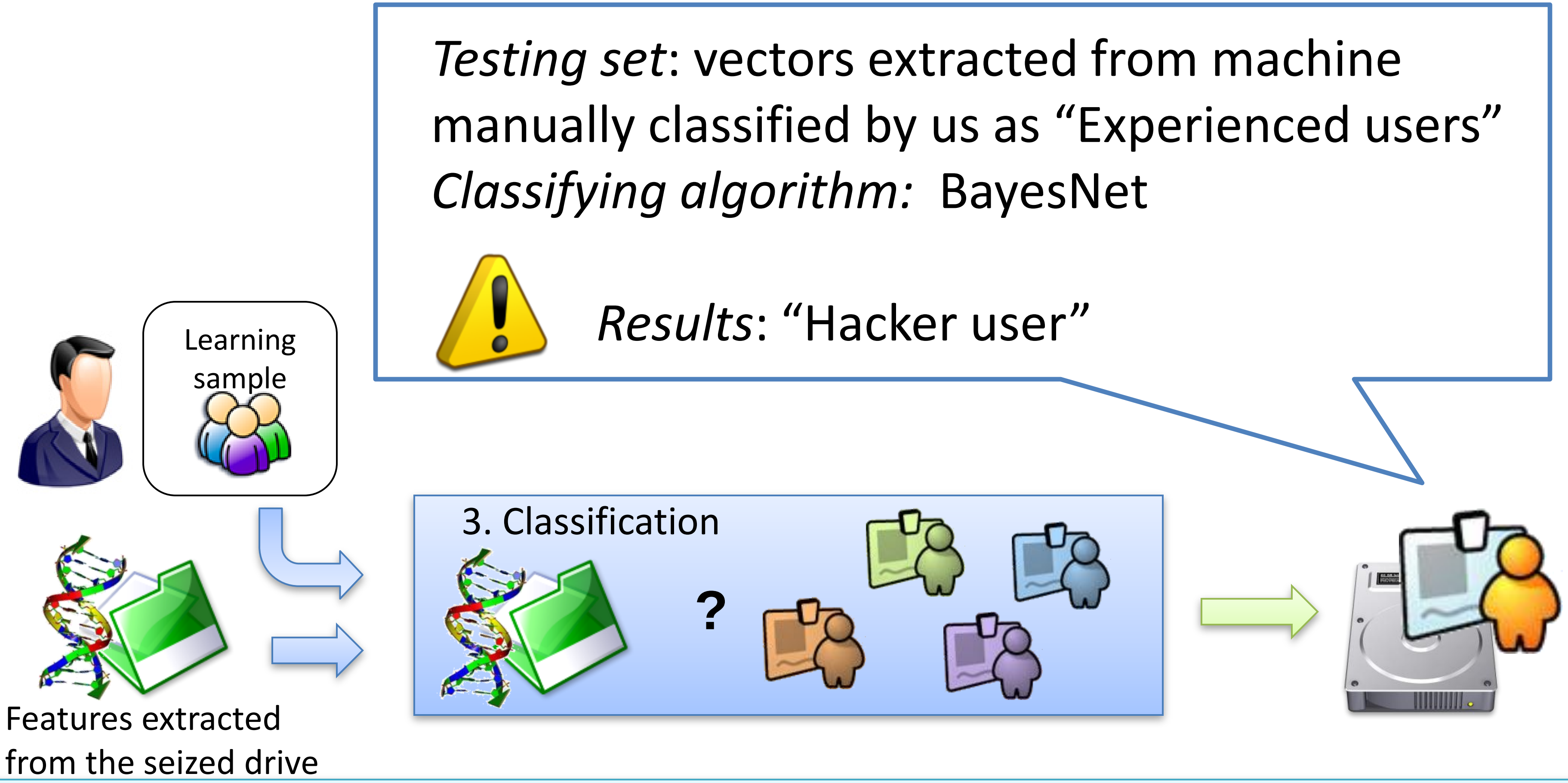
- BayesNet algorithm obtained the 100% of instances correctly classified
- NaiveBayes algorithm scored 92% of instances correctly classified.
- The tree classifier J48 algorithm obtained 84% of instances correctly classified

Training data
3. Classification
?

Case study – Operational Phase (3)

Operational Phase:

3. Automatic ***classification*** of the user profile



Testing set: vectors extracted from machine manually classified by us as “Experienced users”

Classifying algorithm: BayesNet

⚠ *Results*: “Hacker user”

Case study – Operational Phase (3)

Operational Phase:

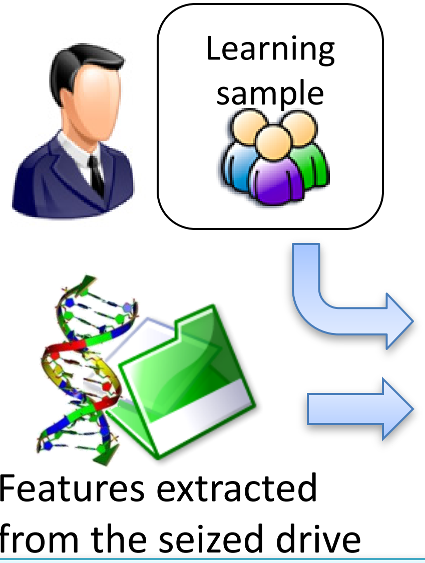
3. Automatic ***classification*** of the user profile

Testing set: vectors extracted from machine manually classified by us as “Experienced users”

Classifying algorithm: BayesNet

⚠ *Results*: “Hacker user”

- The category assigned is the nearest to the correct one.
- Lack of a significant number of examples.
- Lack of a precision in labelling the data sample.



Outline

1. Introduction
2. The Methodology
3. Case Study
4. Conclusions

Conclusion and future works

In this work we have shown a methodology with preliminary results of a computer user fast profiling in order to schedule efficiently the analysis of a huge amount of seized computers

We have a collaboration with the Italian police intelligence in order to fine tune the tool developed,

- a supervised contribution of the forensic operator in the start-up phase is needed to define the target of the investigation (user profiles of interest, characterizing features)
- an accurate definition of the dictionaries is needed to bring a particular instance (a website, a sw program ,...) to the category of membership
- An accurately and appropriately-sized training-set is needed to the classifier to learn the correct classification model.

In the near future, we plan to test further classification models, in order to identify the efficacy/efficiency of different algorithms in different scenarios.

Fast user classifying to establish forensic analysis priorities

A.Grillo, **A. Lentini**, G.Me, M.Ottoni

University of Rome “Tor Vergata”
Department of Computer Science, Systems and Production
lentini@disp.uniroma2.it

IMF 2009: 5th International Conference on IT Security Incident Management & IT Forensics