

# Evaluación de algoritmos de fuzzy hashing para similitud entre procesos

**Iñaki Abadía Osta**

Director: Ricardo J. Rodríguez

Ponente: José Merseguer Hernáiz

Septiembre de 2017

Curso 16/17

Trabajo Fin de Grado – Grado en Ingeniería Informática

**Escuela de Ingeniería y Arquitectura**

Universidad de Zaragoza

# Contenidos

- 1 Introducción
- 2 Ejecutables y procesos Windows
- 3
- 4 Herramienta ProcessFuzzyHash
- 5 Experimentación
- 6 Trabajo relacionado
- 7 Conclusiones y líneas futuras

# Contenidos

- 1 Introducción
- 2 Ejecutables y procesos Windows
- 3
- 4 Herramienta ProcessFuzzyHash
- 5 Experimentación
- 6 Trabajo relacionado
- 7 Conclusiones y líneas futuras

Situación actual malware  
Funciones de hash criptográfico  
Funciones de fuzzy hash  
Contribución

# Contenidos

- 1 Introducción
- 2 Ejecutables y procesos Windows**
- 3
- 4 Herramienta ProcessFuzzyHash
- 5 Experimentación
- 6 Trabajo relacionado
- 7 Conclusiones y líneas futuras

Ejecutables Windows  
Procesos Windows

# Contenidos

- 1 Introducción
- 2 Ejecutables y procesos Windows
- 3
- 4 Herramienta ProcessFuzzyHash**
- 5 Experimentación
- 6 Trabajo relacionado
- 7 Conclusiones y líneas futuras

Arquitectura

Ejecución



**Universidad**  
Zaragoza

# Contenidos

- 1 Introducción
- 2 Ejecutables y procesos Windows
- 3
- 4 Herramienta ProcessFuzzyHash
- 5 Experimentación**
- 6 Trabajo relacionado
- 7 Conclusiones y líneas futuras

Entorno de pruebas

Resultados

# Contenidos

- 1 Introducción
- 2 Ejecutables y procesos Windows
- 3
- 4 Herramienta ProcessFuzzyHash
- 5 Experimentación
- 6 Trabajo relacionado**
- 7 Conclusiones y líneas futuras

Herramientas

Publicaciones



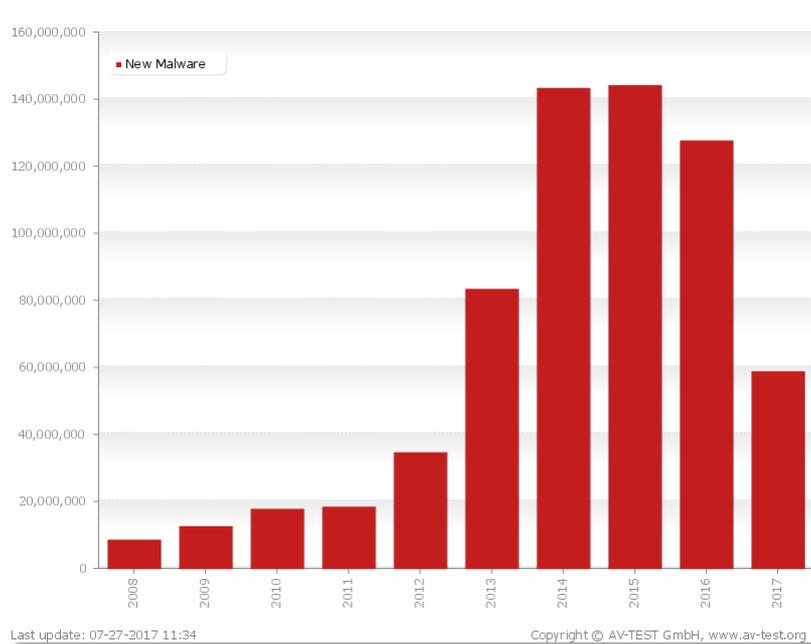
**Universidad**  
Zaragoza

# Contenidos

- 1 Introducción
- 2 Ejecutables y procesos Windows
- 3
- 4 Herramienta ProcessFuzzyHash
- 5 Experimentación
- 6 Trabajo relacionado
- 7 Conclusiones y líneas futuras

Conclusiones  
Trabajo futuro

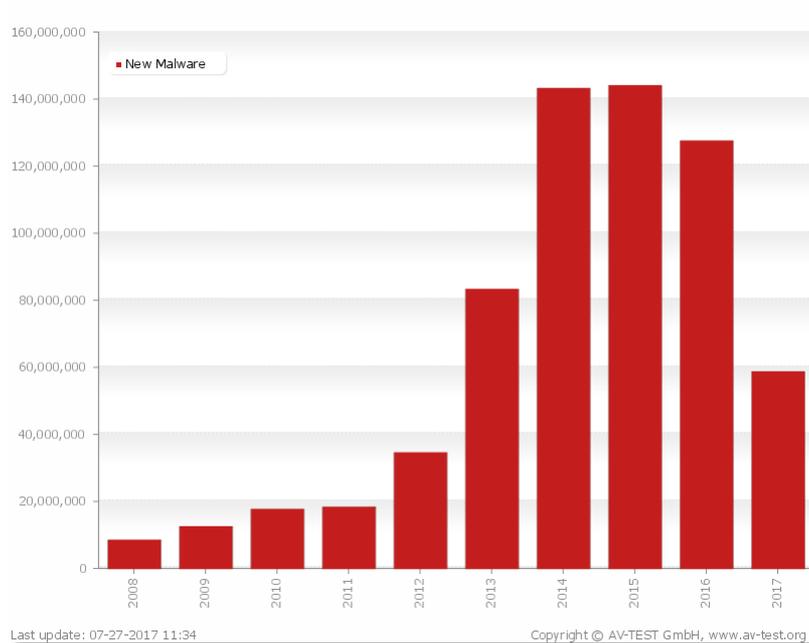
# Situación actual software malicioso (malware)



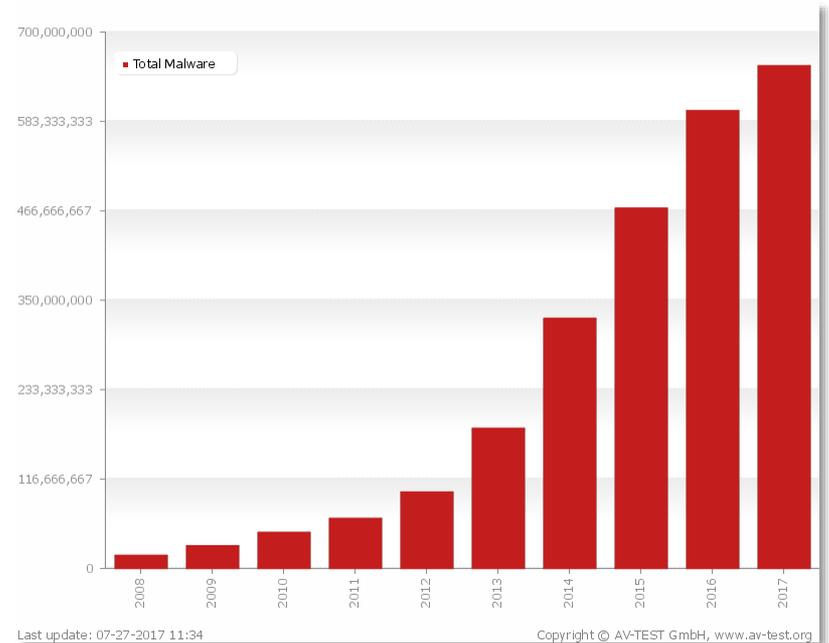
## Malware nuevo 2008-2017

\* Imágenes tomadas de [AT17]

# Situación actual software malicioso (malware)

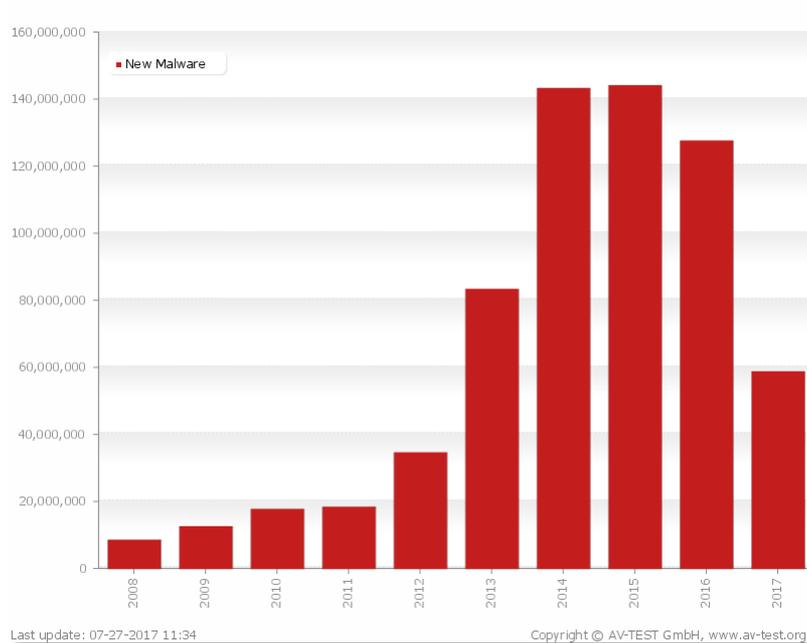


**Malware nuevo  
2008-2017**

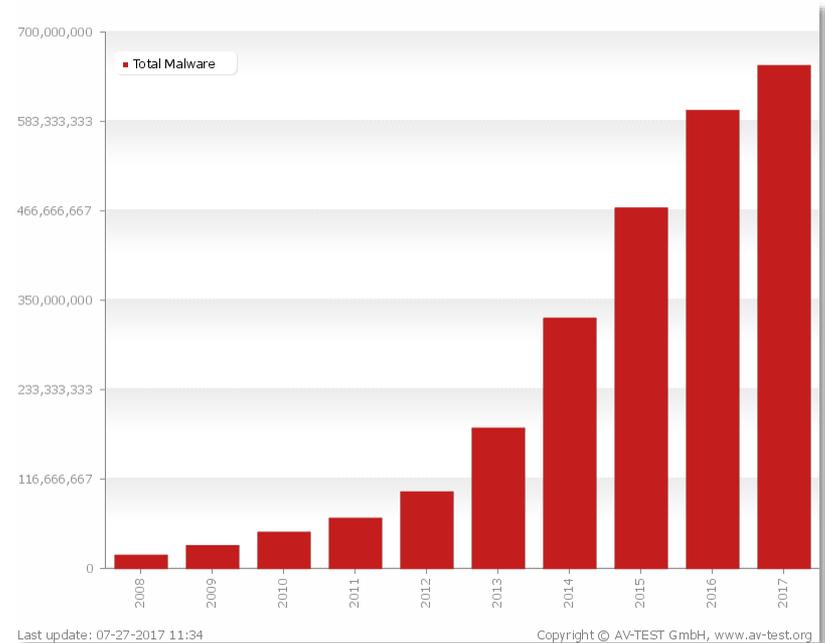


**Malware en circulación  
2008-2017**

# Situación actual software malicioso (malware)



**Malware nuevo  
2008-2017**



**Malware en circulación  
2008-2017**

**85% Windows**

# Respuesta a incidentes

# Respuesta a incidentes

## Curso habitual de los eventos

# Respuesta a incidentes

## Curso habitual de los eventos

- Alguien se da cuenta de que hay una máquina comprometida

\* Imágenes tomadas de [AT17]

# Respuesta a incidentes

## Curso habitual de los eventos

- Alguien se da cuenta de que hay una máquina comprometida
- Se da la alerta

# Respuesta a incidentes

## Curso habitual de los eventos

- Alguien se da cuenta de que hay una máquina comprometida
- Se da la alerta
- Necesidad de averiguar si la máquina está realmente comprometida

\* Imágenes tomadas de [AT17]

# Respuesta a incidentes

## Curso habitual de los eventos

- Alguien se da cuenta de que hay una máquina comprometida
- Se da la alerta
- Necesidad de averiguar si la máquina está realmente comprometida
- **Análisis de la máquina**

\* Imágenes tomadas de [AT17]

# Respuesta a incidentes

## Curso habitual de los eventos

- Alguien se da cuenta de que hay una máquina comprometida
- Se da la alerta
- Necesidad de averiguar si la máquina está realmente comprometida
- **Análisis de la máquina**

## Análisis de la máquina

# Respuesta a incidentes

## Curso habitual de los eventos

- Alguien se da cuenta de que hay una máquina comprometida
- Se da la alerta
- Necesidad de averiguar si la máquina está realmente comprometida
- **Análisis de la máquina**

## Análisis de la máquina

- Extracción del volcado de memoria

\* Imágenes tomadas de [AT17]

# Respuesta a incidentes

## Curso habitual de los eventos

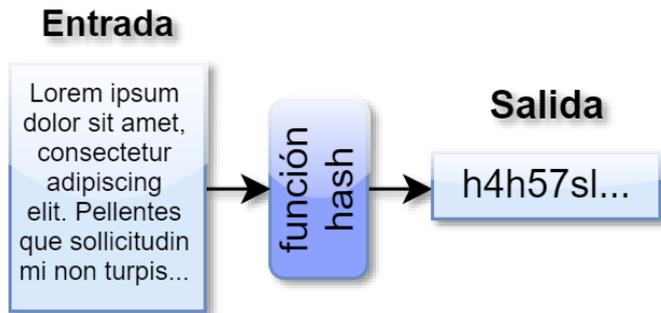
- Alguien se da cuenta de que hay una máquina comprometida
- Se da la alerta
- Necesidad de averiguar si la máquina está realmente comprometida
- **Análisis de la máquina**

## Análisis de la máquina

- Extracción del volcado de memoria
- Análisis del volcado
  - **Cálculo de hashes**

\* Imágenes tomadas de [AT17]

# Funciones de hash criptográfico



# Funciones de hash criptográfico



# Funciones de hash criptográfico



## Ejemplos

MD5	2ae4f65fc262c0120b037438d05883f8
SHA1	fcbab2d91923161bc46022a2b16ce50e8420fdec
Tiger	df81bf2a909edc6e76f0118a372a19f3f89233f4a5c5a3f6
RipeMD128	b259ecc284326e3773c596e14bff2d0675d4320a

# Funciones de fuzzy hash

## Propiedades

- **Evitan efecto cascada**
- Equivalencia = colisiones

## Uso

- Detección de *spam*
- Copyright

## Ejemplos

- ssdeep
- nilsimsa

# Funciones de fuzzy hash

## Propiedades

- **Evitan efecto cascada**
- Equivalencia = colisiones

## Uso

- Detección de *spam*
- Copyright

## Ejemplos

- ssdeep
- nilsimsa

## MD5

# Funciones de fuzzy hash

## Propiedades

- **Evitan efecto cascada**
- Equivalencia = colisiones

## Uso

- Detección de *spam*
- Copyright

## Ejemplos

- ssdeep
- nilsimsa

## MD5

**Original:** 2ae4f65fc262c0120b037438d05883f8

# Funciones de fuzzy hash

## Propiedades

- **Evitan efecto cascada**
- Equivalencia = colisiones

## Uso

- Detección de *spam*
- Copyright

## Ejemplos

- ssdeep
- nilsimsa

## MD5

**Original:** 2ae4f65fc262c0120b037438d05883f8

**Alterado:** faf3e0b7e3008463714c7bf779a014a8

# Funciones de fuzzy hash

## Propiedades

- **Evitan efecto cascada**
- Equivalencia = colisiones

## Uso

- Detección de *spam*
- Copyright

## Ejemplos

- ssdeep
- nilsimsa

## MD5

**Original:** 2ae4f65fc262c0120b037438d05883f8

**Alterado:** faf3e0b7e3008463714c7bf779a014a8

## ssdeep



# Funciones de fuzzy hash

## Propiedades

- **Evitan efecto cascada**
- Equivalencia = colisiones

## Uso

- Detección de *spam*
- Copyright

## Ejemplos

- ssdeep
- nilsimsa

## MD5

**Original:** 2ae4f65fc262c0120b037438d05883f8

**Alterado:** faf3e0b7e3008463714c7bf779a014a8

## ssdeep

**Original:**

49152:oKPDou/1fcThWebCqvWb/zIEqqG1H8ATR0qeuZEqv8eGYWb/zIEJGd8A90KZE2vJKI



# Funciones de fuzzy hash

## Propiedades

- **Evitan efecto cascada**
- Equivalencia = colisiones

## Uso

- Detección de *spam*
- Copyright

## Ejemplos

- ssdeep
- nilsimsa

## MD5

**Original:** 2ae4f65fc262c0120b037438d05883f8

**Alterado:** faf3e0b7e3008463714c7bf779a014a8

## ssdeep

**Original:**

49152:oKPDou/1fcThWebCqvWb/zIEqqG1H8ATR0qeuZEqv8eGYWb/zIEJGd8A90KZE2vJKI

**Alterado:**

49152:gKPDou/1fcThWebCqvWb/zIEqqG1H8ATR0qeuZEqv8eGYWb/zIEJGd8A90KZE2vJKo



# Funciones de fuzzy hash

## Propiedades

- **Evitan efecto cascada**
- Equivalencia = colisiones

## Uso

- Detección de *spam*
- Copyright

## Ejemplos

- ssdeep
- nilsimsa

## MD5

**Original:** 2ae4f65fc262c0120b037438d05883f8

**Alterado:** faf3e0b7e3008463714c7bf779a014a8

## ssdeep

**Original:**

49152:oKPDou/1fcThWebCqvWb/zIEqqG1H8ATR0qeuZEqv8eGYWb/zIEJGd8A90KZE2vJKI

**Alterado:**

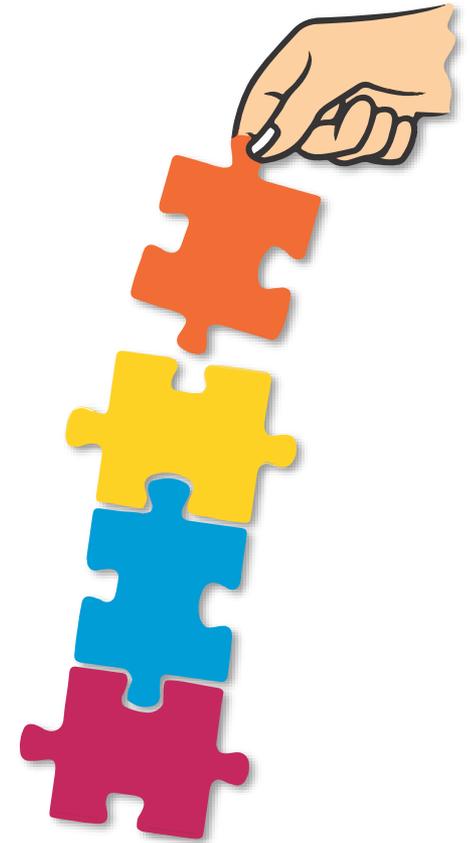
49152:gKPDou/1fcThWebCqvWb/zIEqqG1H8ATR0qeuZEqv8eGYWb/zIEJGd8A90KZE2vJKo

**98%**



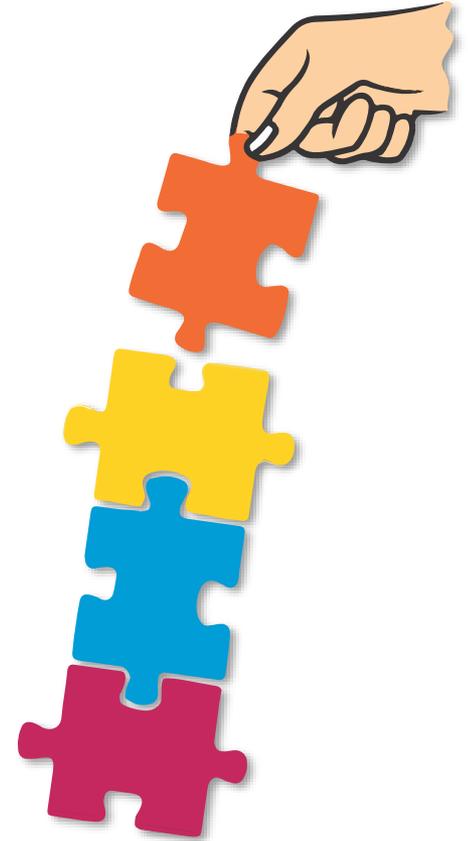
**Universidad**  
Zaragoza

# Contribución



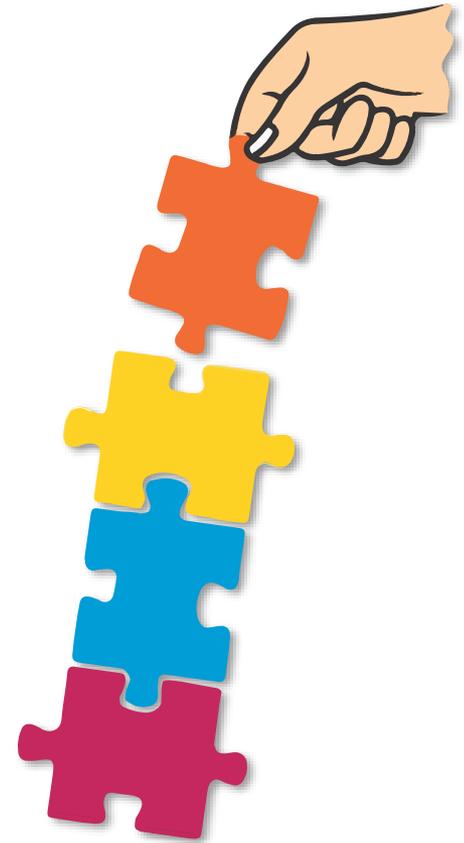
# Contribución

- Herramienta ProcessFuzzyHash
  - Calcular fuzzy hashes de procesos
  - Comparar fuzzy hashes de procesos



# Contribución

- Herramienta ProcessFuzzyHash
  - Calcular fuzzy hashes de procesos
  - Comparar fuzzy hashes de procesos
- Estudio y evaluación de algoritmos de fuzzy hashing
  - **Centrado en Windows**



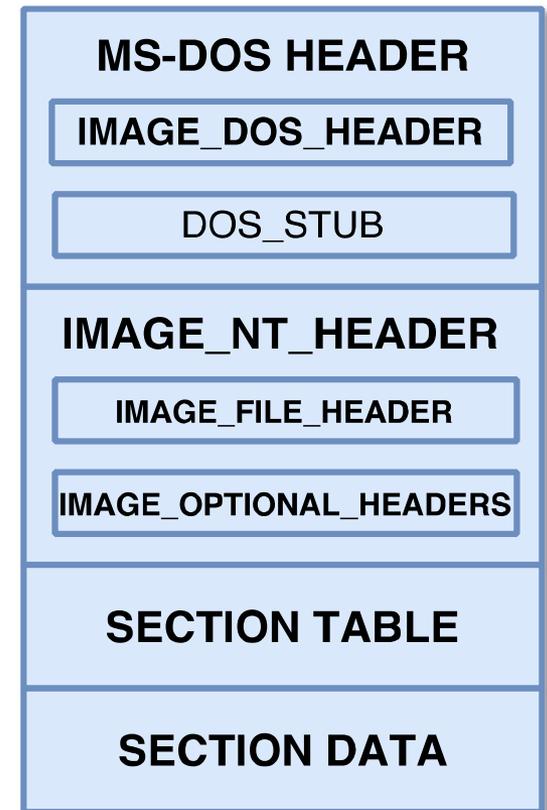
# Ejecutables Windows

# Ejecutables Windows

- **Formato PE**
  - Estándar de cabeceras de ficheros ejecutables (EXE, OBJ, SCR, etc.)

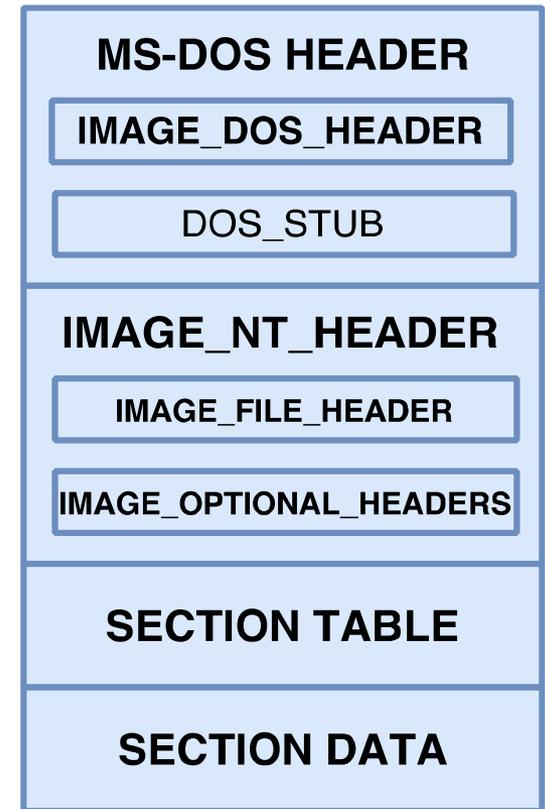
# Ejecutables Windows

- **Formato PE**
  - Estándar de cabeceras de ficheros ejecutables (EXE, OBJ, SCR, etc.)
- **Estructura**
  - Cabeceras
  - Tabla de secciones
  - Espacio de secciones



# Ejecutables Windows

- **Formato PE**
  - Estándar de cabeceras de ficheros ejecutables (EXE, OBJ, SCR, etc.)
- **Estructura**
  - Cabeceras
  - Tabla de secciones
  - Espacio de secciones
- **Secciones**
  - Código (R)
  - Datos (R)
  - Datos (R/W)
  - Recursos (R)
  - ...



# Procesos Windows

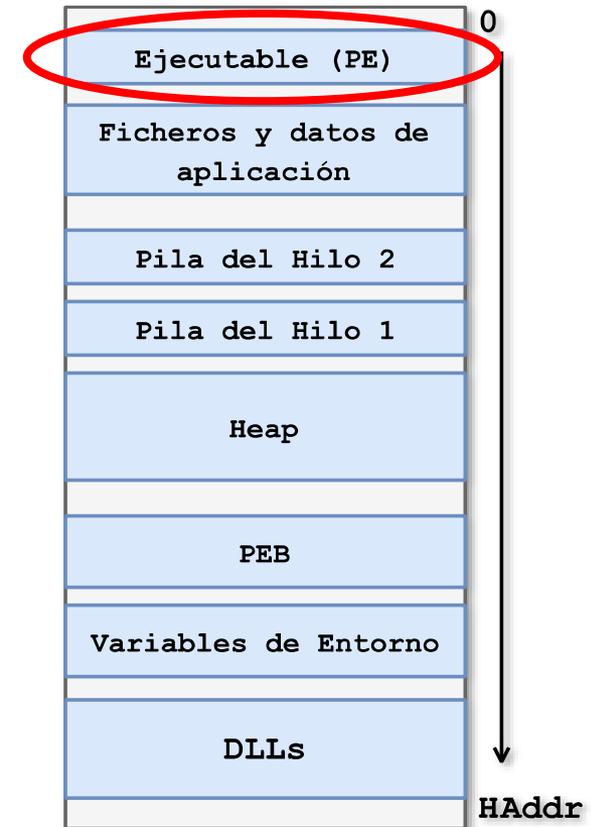
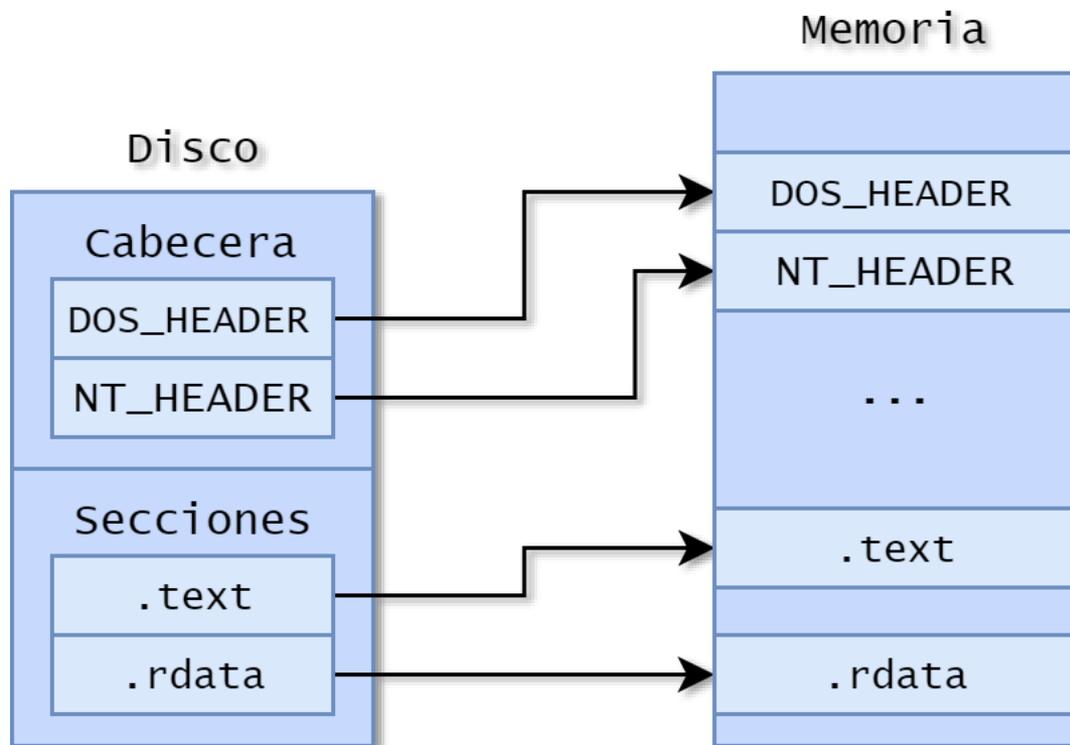
# Procesos Windows

- Contenedor de hilos



# Procesos Windows

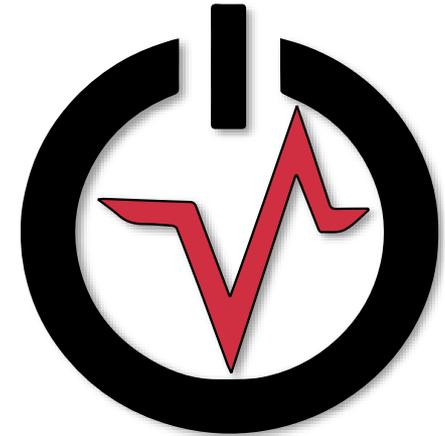
- Contenedor de hilos
- Ejecutable disco  $\neq$  memoria
  - Carga en memoria
  - Address Space Layout Randomization



# Plugin de Volatility

# Plugin de Volatility

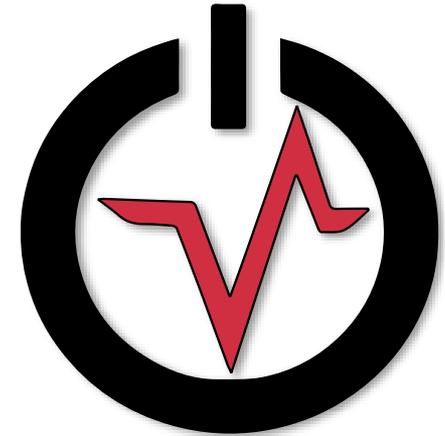
Volatility



# Plugin de Volatility

## Volatility

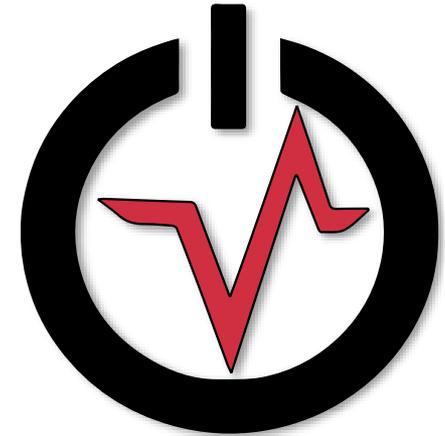
- Python 2



# Plugin de Volatility

## Volatility

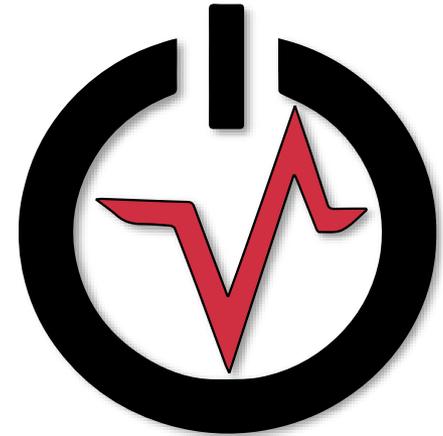
- Python 2
- Framework para análisis de memoria volátil
  - The Volatility Foundation
  - Comunidad



# Plugin de Volatility

## Volatility

- Python 2
- Framework para análisis de memoria volátil
  - The Volatility Foundation
  - Comunidad

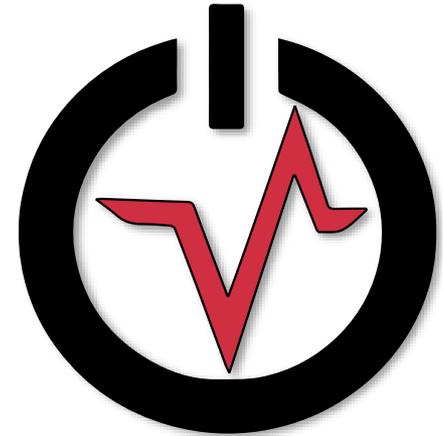


## Algunos plugins de Volatility

# Plugin de Volatility

## Volatility

- Python 2
- Framework para análisis de memoria volátil
  - The Volatility Foundation
  - Comunidad



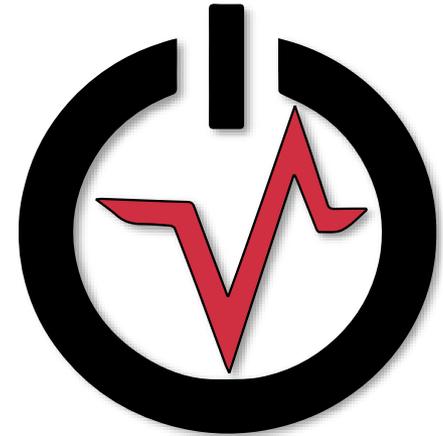
## Algunos plugins de Volatility

- **pslist**: Muestra el listado de procesos en ejecución

# Plugin de Volatility

## Volatility

- Python 2
- Framework para análisis de memoria volátil
  - The Volatility Foundation
  - Comunidad



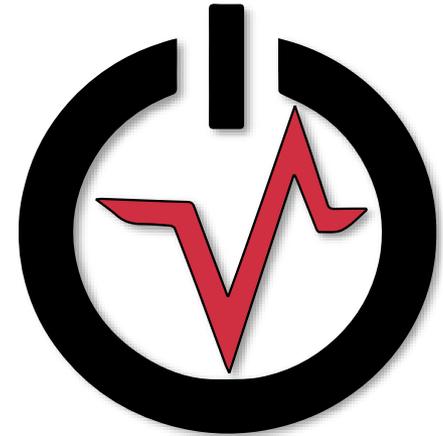
## Algunos plugins de Volatility

- **pslist**: Muestra el listado de procesos en ejecución
- **dumpregistry**: Extrae un volcado del registro Windows

# Plugin de Volatility

## Volatility

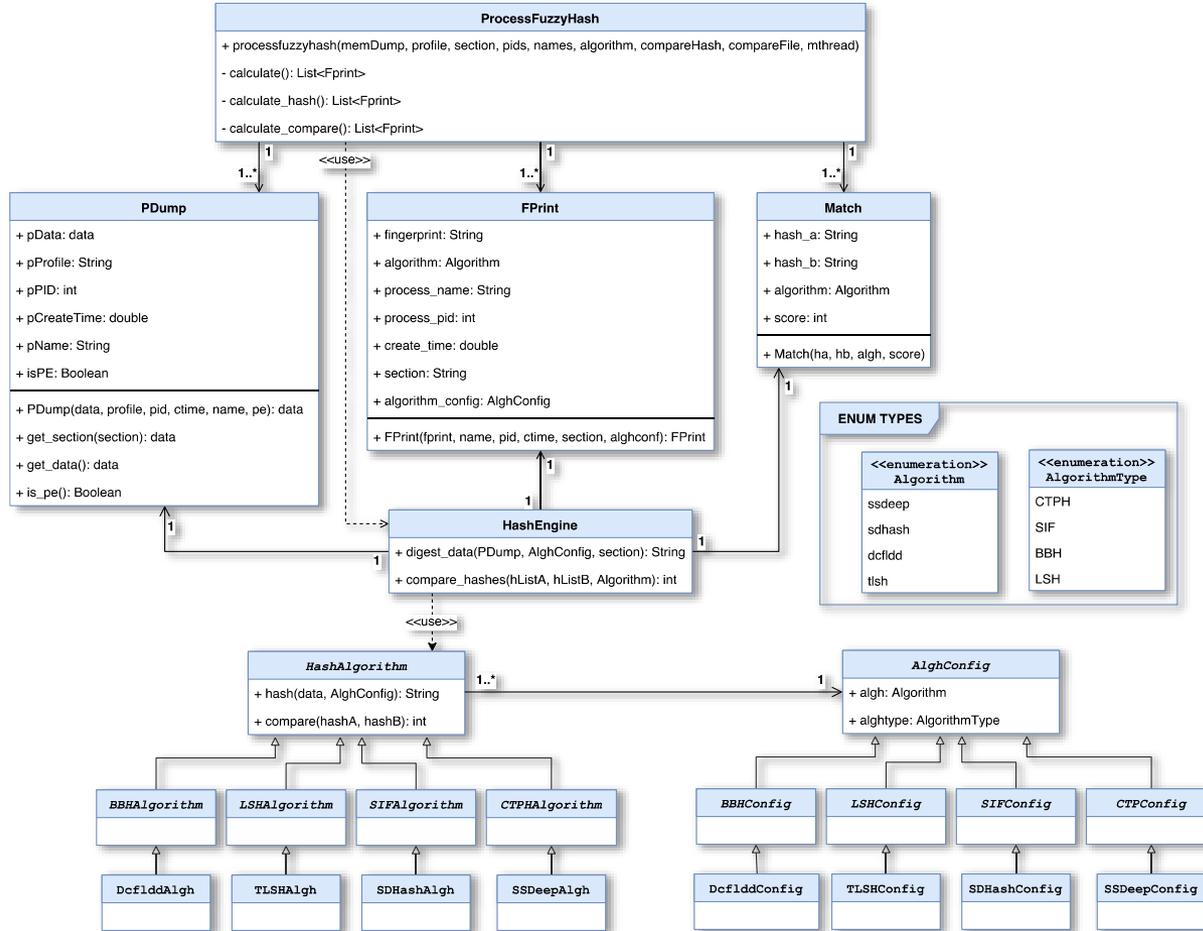
- Python 2
- Framework para análisis de memoria volátil
  - The Volatility Foundation
  - Comunidad



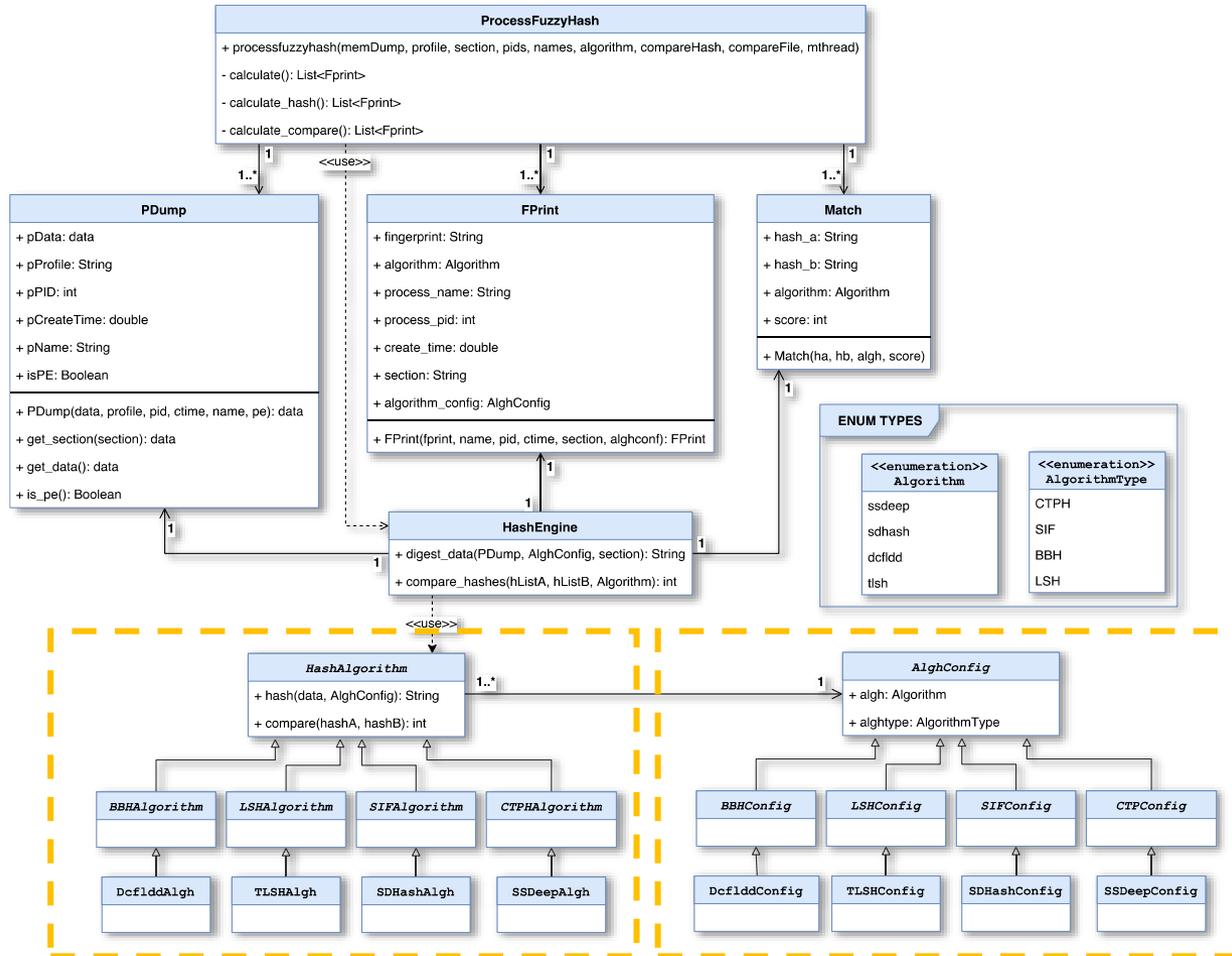
## Algunos plugins de Volatility

- **pslist**: Muestra el listado de procesos en ejecución
- **dumpregistry**: Extrae un volcado del registro Windows
- **vboxinfo**: Muestra información de un volcado de una VM VBox

# Arquitectura – diagrama de clases



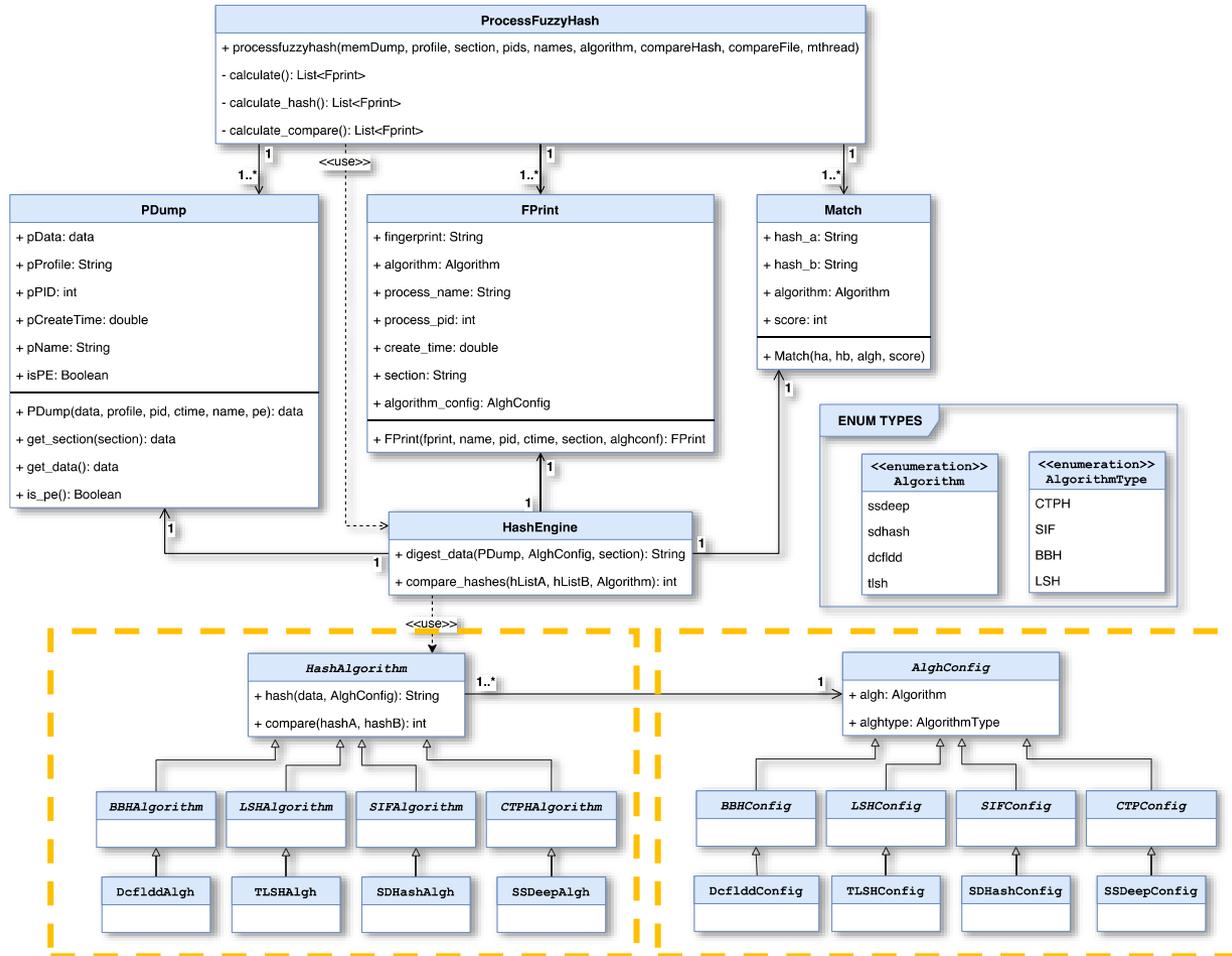
# Arquitectura – diagrama de clases



## Facade

- Dcfldd (BBH)
- Sdhash (SIF)
- TLSh (LSH)
- Ssdeep (CTPH)

# Arquitectura – diagrama de clases



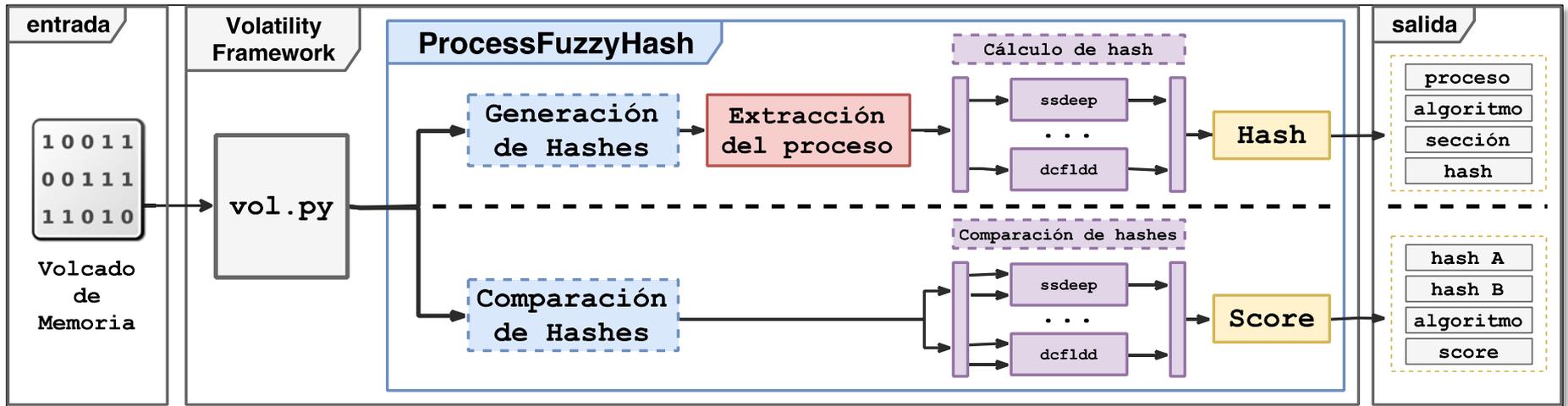
## Facade

- Dcfldd (BBH)
- Sdhash (SIF)
- TLSh (LSH)
- Ssdeep (CTPH)

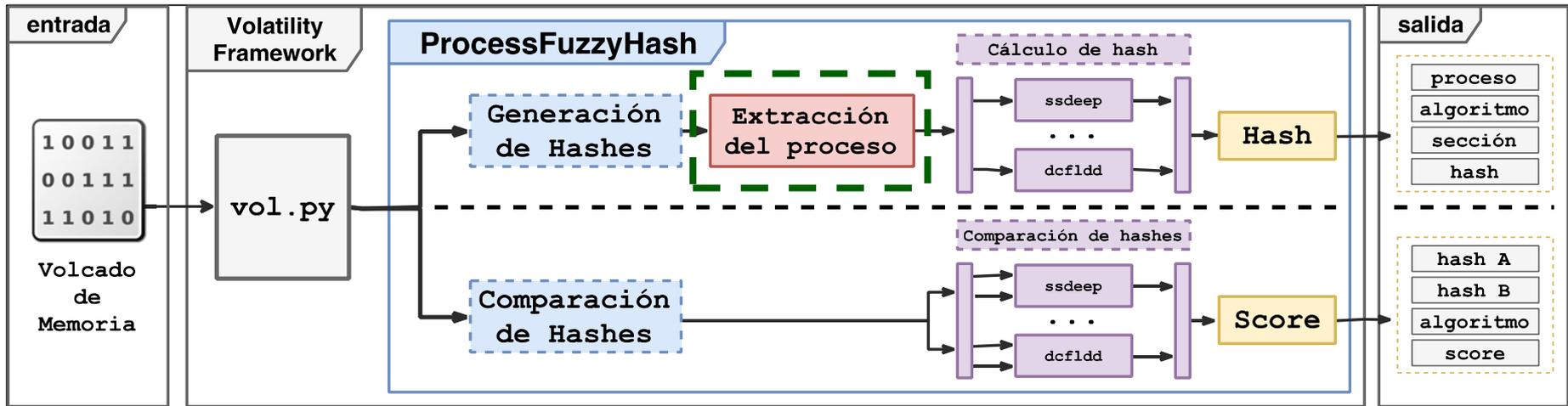
## Criterio

- Mayor diversidad
- Implementación en Python

# Arquitectura – diagrama de sistema



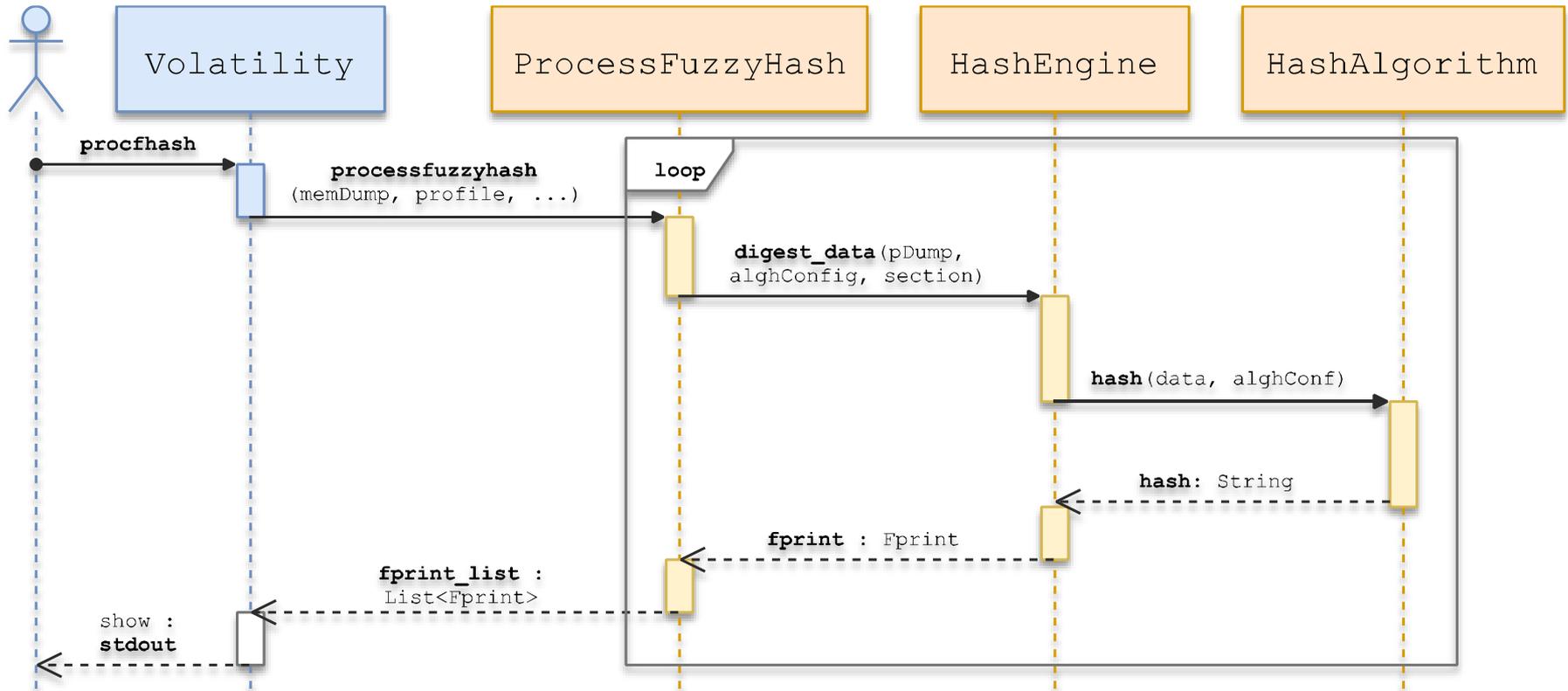
# Arquitectura – diagrama de sistema



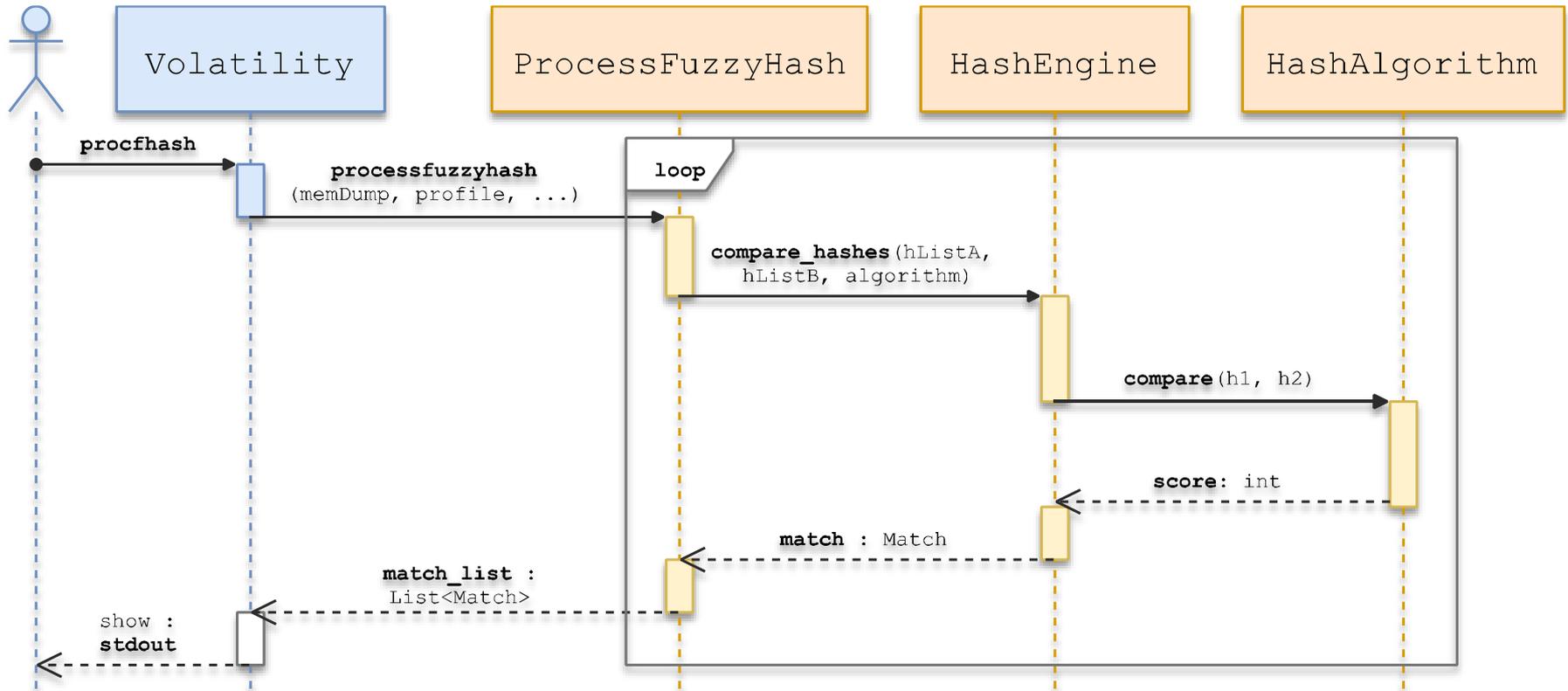
## Extracción de procesos

- Apoyo en plugins:
  - **procdump:** Extrae procesos de memoria
  - **memdump:** Extrae todas las páginas de memoria en uso

# Arquitectura – flujo generación



# Arquitectura – flujo comparación



# Ejecución – salida de generación

# Ejecución – salida de generación

## Ejecución

```
$ python vol.py --plugins=ProcessFuzzyHash/ -f vmcore.elf \  
> --profile=Win10x86_15063 processfuzzyhash -A ssdeep -S pe \  
> -N VBoxService,winlogon,services
```

# Ejecución – salida de generación

## Ejecución

```
$ python vol.py --plugins=ProcessFuzzyHash/ -f vmcore.elf \
> --profile=Win10x86_15063 processfuzzyhash -A ssdeep -S pe \
> -N VBoxService,winlogon,services
```

## Salida

Volatility Foundation Volatility Framework 2.6

Name	PID	Create Time	Section	Algorithm	Hash
winlogon.exe	500	131483892000	pe	SSDeep	6144:pzP/qv...8ciJQdsyJqj
services.exe	544	131483892003	pe	SSDeep	6144:Q/6kXE...jXd5
VBoxService	1060	131483892039	pe	SSDeep	12288:K/oDR...CxuexSQ

# Ejecución – salida de comparación

# Ejecución – salida de comparación

## Ejecución

```
$ python vol.py --plugins=ProcessFuzzyHash/ -f vmcore.elf \  
> --profile=Win10x86_15063 processfuzzyhash -A ssdeep -S pe -N svchost \  
> -c '768:9n3SsSfvr0t0HW4C05LTiMRMxVKPhPDjRWWm:d3BGr0t02N05LTiqUVKP5/zm'
```

# Ejecución – salida de comparación

## Ejecución

```
$ python vol.py --plugins=ProcessFuzzyHash/ -f vmcore.elf \
> --profile=Win10x86_15063 processfuzzyhash -A ssdeep -S pe -N svchost \
> -c '768:9n3SsSfvr0t0HW4C05LTiMRMxVKPhPDjRWWm:d3BGr0t02N05LTiqUVKP5/zm'
```

## Salida

```
Volatility Foundation Volatility Framework 2.6
Hash A                Hash B                Algorithm Score
768:9n3Ss...qUVKP5/zm 768:9n3SsS...qDVKP5/0m  ssdeep    94
768:9n3Ss...qUVKP5/zm 768:9n3SsS...q5VKP5/0m  ssdeep    94
768:9n3Ss...qUVKP5/zm 768:9n3SsS...qUVKP5/zm  ssdeep   100
768:9n3Ss...qUVKP5/zm 768:9n3SsS...qFVKP5/zm  ssdeep    97
768:9n3Ss...qUVKP5/zm 768:9n3SsS...qMVKP5/zm  ssdeep   100
768:9n3Ss...qUVKP5/zm 768:9n3SsS...qAVKP5/zm  ssdeep    97
...
```



# Entorno de pruebas



# Entorno de pruebas

## Máquina host

- Ubuntu 16.04 x64
- Intel Core i7-4720HQ @ 2.60GHz
- 12GiB RAM
- SSD SATA3
- HDD @ 5400 RPM USB 3.0



# Entorno de pruebas

## Máquina host

- Ubuntu 16.04 x64
- Intel Core i7-4720HQ @ 2.60GHz
- 12GiB RAM
- SSD SATA3
- HDD @ 5400 RPM USB 3.0

## Máquinas Virtuales

- Windows 7 32 (x86) y 64 bits (x64)
- Windows 10 32 (x86) y 64 bits (x64)



# Entorno de pruebas

## Máquina host

- Ubuntu 16.04 x64
- Intel Core i7-4720HQ @ 2.60GHz
- 12GiB RAM
- SSD SATA3
- HDD @ 5400 RPM USB 3.0

## Máquinas Virtuales

- Windows 7 32 (x86) y 64 bits (x64)
- Windows 10 32 (x86) y 64 bits (x64)

## Procesos

- Sistema operativo (winlogon, svchost, explorer)
- Software adicional:
  - Navegador Web (IE, Edge, Chrome, Firefox)
  - Lector de PDF (Acrobat Reader)
- Malware:
  - POS RAM Scrapper (ALINA)



# Entorno de pruebas

## Máquina host

- Ubuntu 16.04 x64
- Intel Core i7-4720HQ @ 2.60GHz
- 12GiB RAM
- SSD SATA3
- HDD @ 5400 RPM USB 3.0

## Máquinas Virtuales

- Windows 7 32 (x86) y 64 bits (x64)
- Windows 10 32 (x86) y 64 bits (x64)

## Procesos

- Sistema operativo (winlogon, svchost, explorer)
- Software adicional:
  - Navegador Web (IE, Edge, Chrome, Firefox)
  - Lector de PDF (Acrobat Reader)
- Malware:
  - POS RAM Scrapper (ALINA)



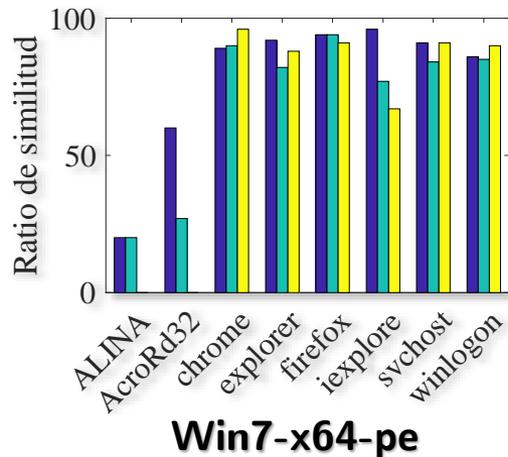
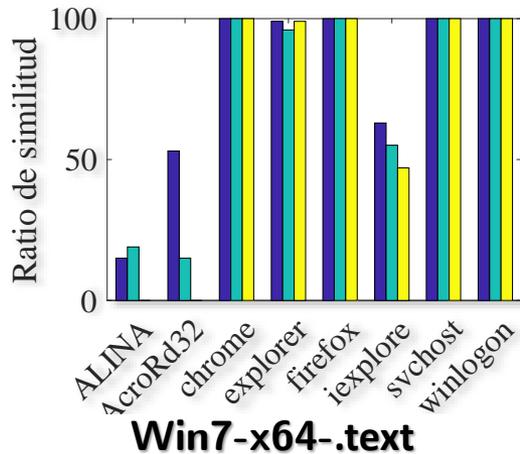
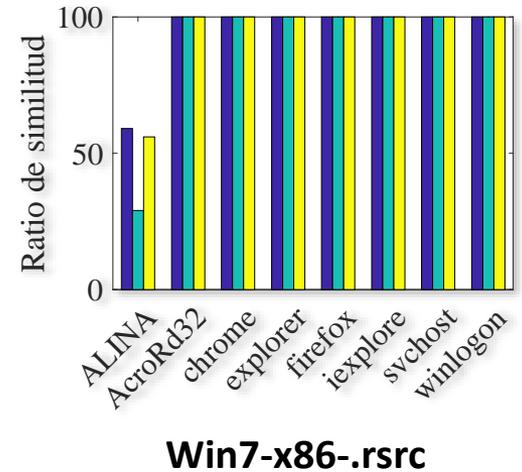
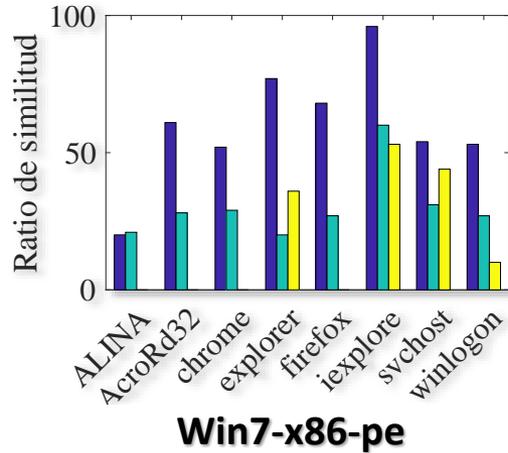
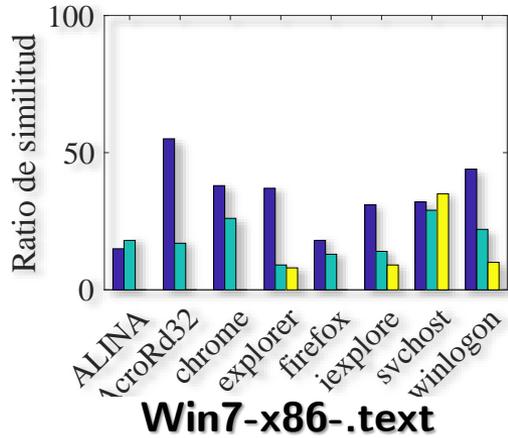
**Se van a considerar 4 casos**



**Universidad  
Zaragoza**

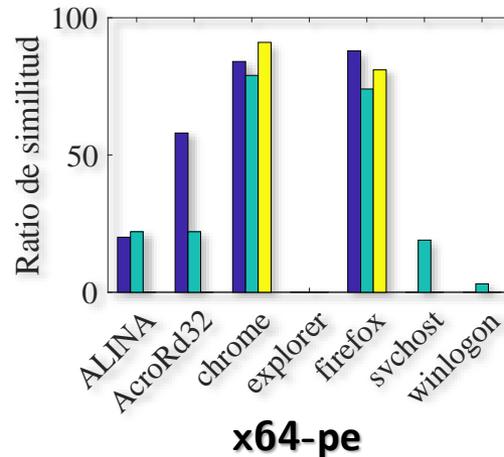
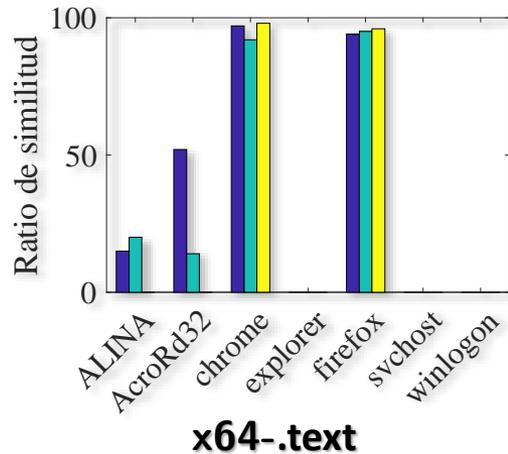
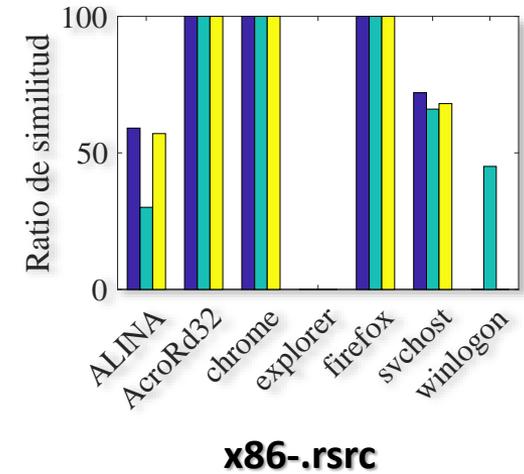
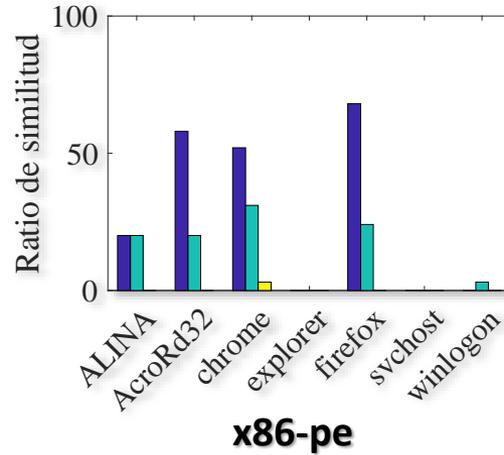
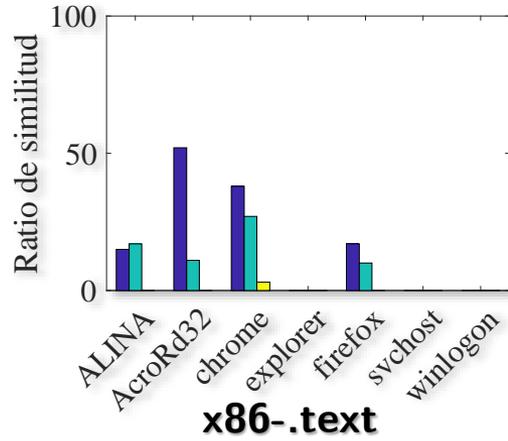
# Resultados – caso A

## Misma máquina, mismo ejecutable, distinta ejecución



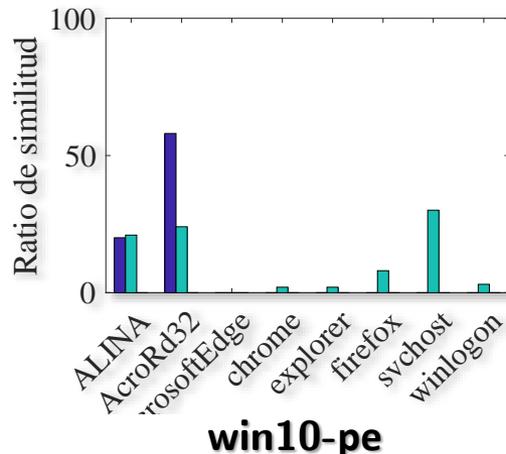
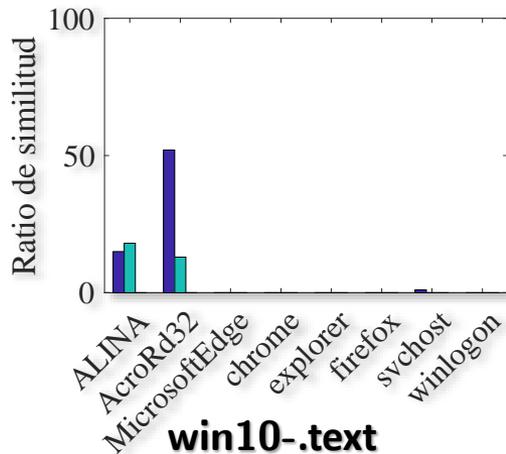
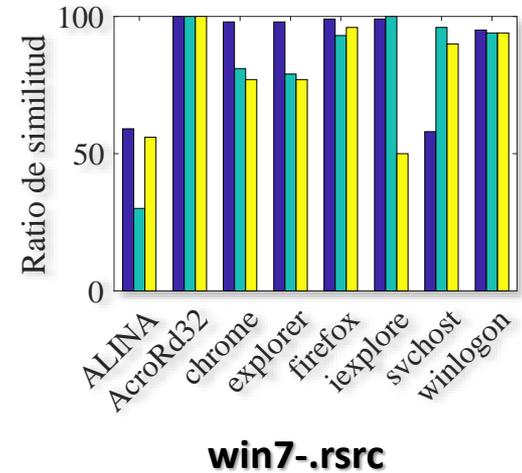
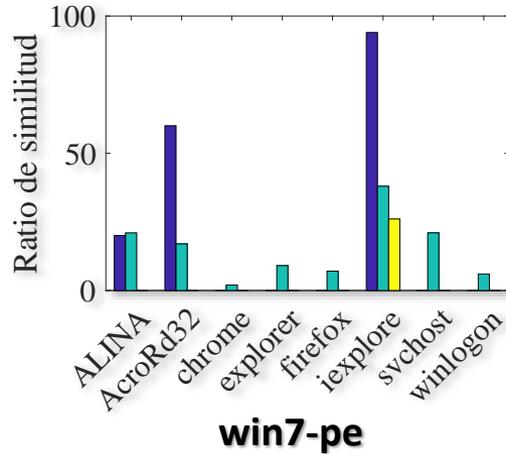
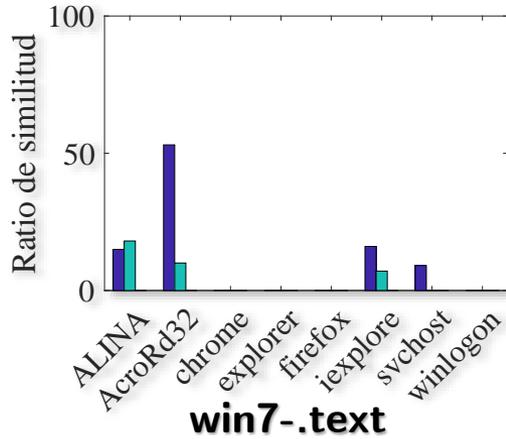
# Resultados – caso B

## Mismo programa y arquitectura, distinto SO



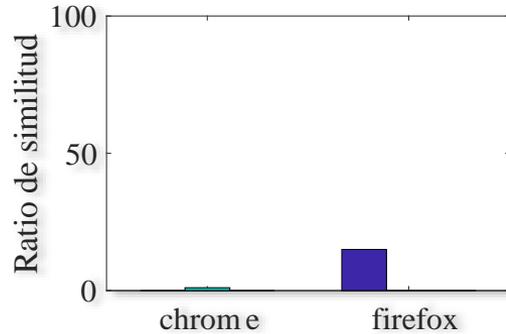
# Resultados – caso C

## Mismo programa y SO, distinta arquitectura

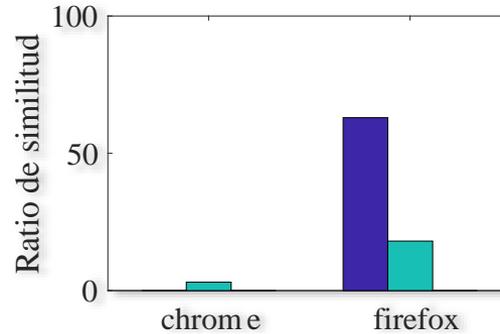


# Resultados – caso D

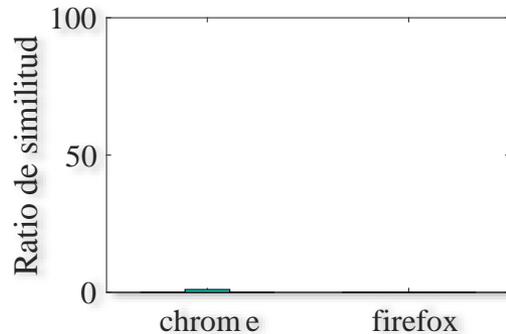
## Mismo programa y misma máquina, distinta versión



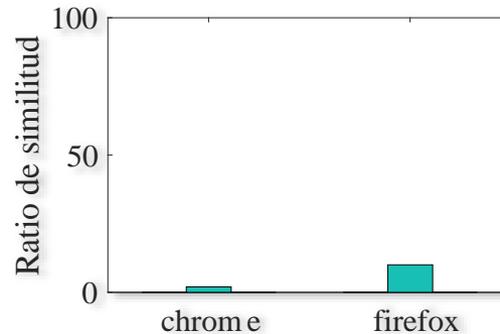
**Win7-x86-.text**



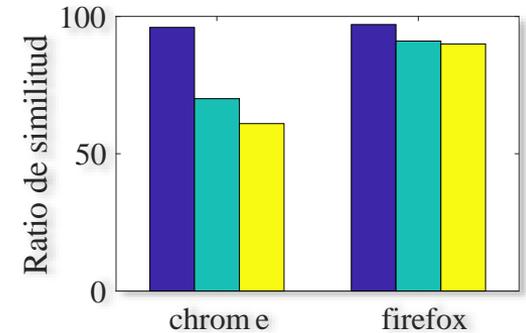
**Win7-x86-pe**



**Win7-x64-.text**



**Win7-x64-pe**



**Win7-x86-.rsrc**



# Herramientas y publicaciones

# Herramientas y publicaciones

## Herramientas

- Malfunction: Análisis de funciones
- Binwally: Detección de cambios entre versiones de software

# Herramientas y publicaciones

## Herramientas

- Malfunction: Análisis de funciones
- Binwally: Detección de cambios entre versiones de software

## Publicaciones

- **[SBAAN16]** Evaluación de algoritmos fuzzy hash
- **[NMAM+16]** Análisis estático de ejecutables
- **[AS15]** Malware → JPEG. Clusterización.

# Conclusiones



# Conclusiones

- Mejor algoritmo: **dcfldd (BBH)**
  - TLSH (LSH) irregular
  - Sdhash (SIF) y ssdeep (CTPH) similar (ssdeep más ceros)



# Conclusiones

- Mejor algoritmo: **dcfldd (BBH)**
  - TLSH (LSH) irregular
  - Sdhash (SIF) y ssdeep (CTPH) similar (ssdeep más ceros)
- Mejor sección: **datos sólo lectura**
  - Sección de código en segundo lugar, ejecutable completo último
  - Sección de código en x64 casi como datos lectura
  - Variaciones en código muy dependiente del programa

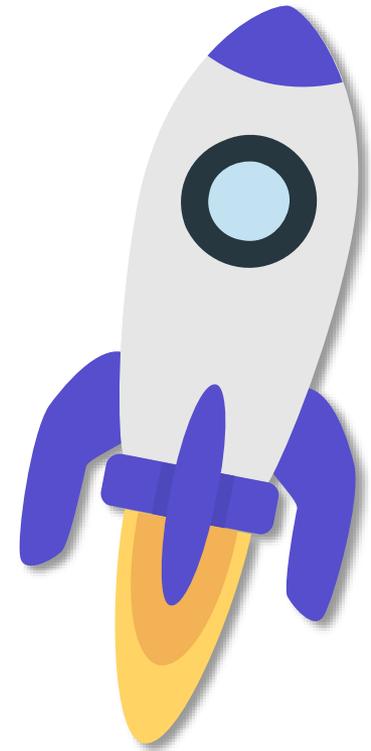


# Conclusiones

- Mejor algoritmo: **dcfldd (BBH)**
  - TLSH (LSH) irregular
  - Sdhash (SIF) y ssdeep (CTPH) similar (ssdeep más ceros)
- Mejor sección: **datos sólo lectura**
  - Sección de código en segundo lugar, ejecutable completo último
  - Sección de código en x64 casi como datos lectura
  - Variaciones en código muy dependiente del programa
- Programas
  - ALINA no supera el **50%**
  - **Procesos de sistema** dan resultados **excelentes** caso A
  - Versiones del mismo software varían mucho salvo en datos de lectura

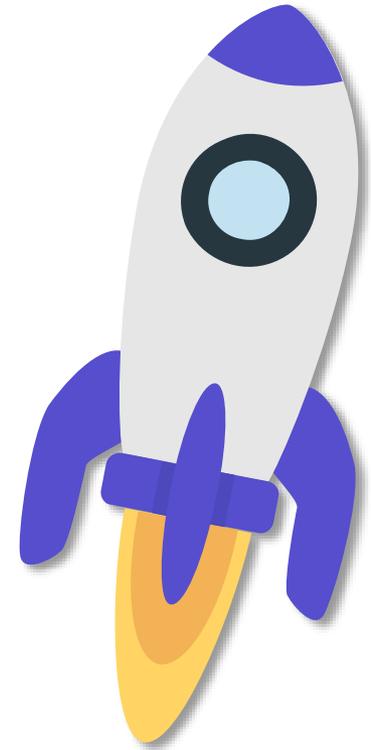


# Líneas futuras



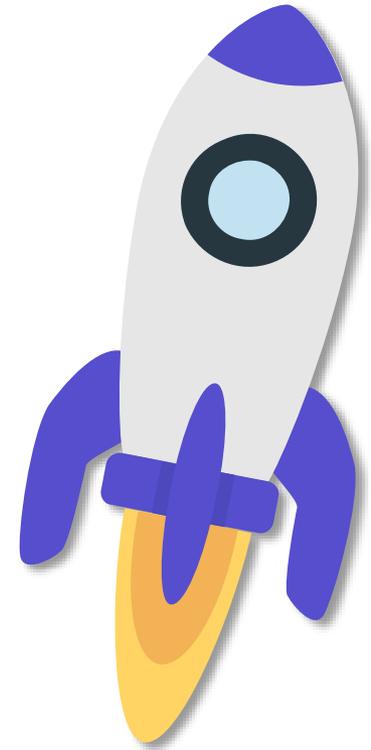
# Líneas futuras

- Independencia de otros plugins



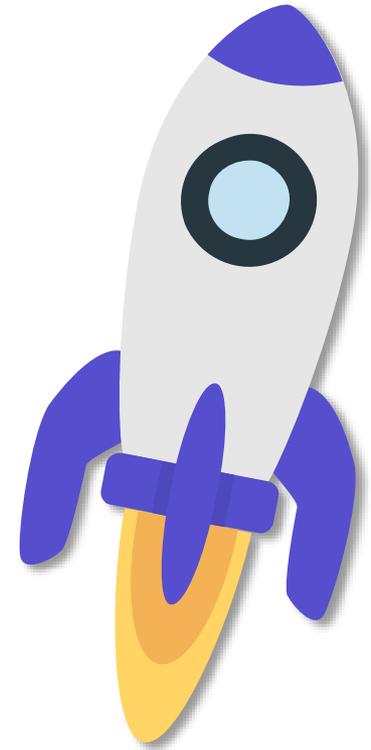
# Líneas futuras

- Independencia de otros plugins
- Paralelización completa



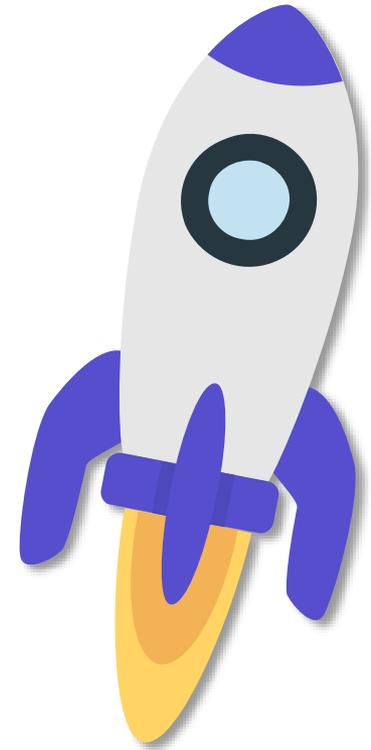
# Líneas futuras

- Independencia de otros plugins
- Paralelización completa
- Implementación de más algoritmos



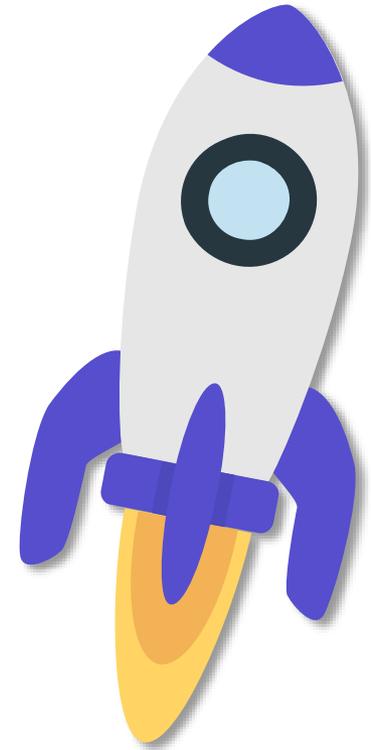
# Líneas futuras

- Independencia de otros plugins
- Paralelización completa
- Implementación de más algoritmos
- Extender herramienta para hash de otras partes de la cabecera



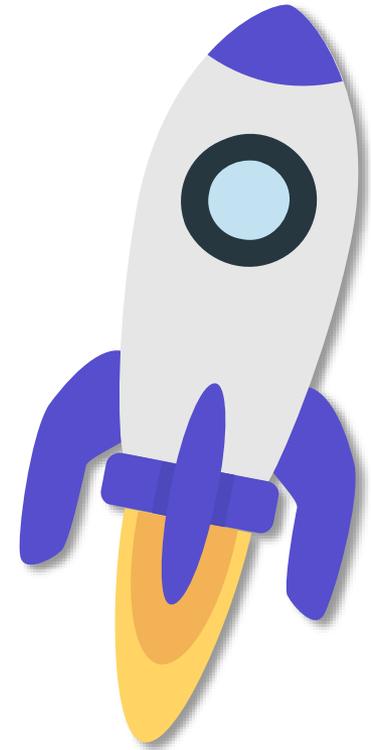
# Líneas futuras

- Independencia de otros plugins
- Paralelización completa
- Implementación de más algoritmos
- Extender herramienta para hash de otras partes de la cabecera
- Estudiar otras partes de un proceso



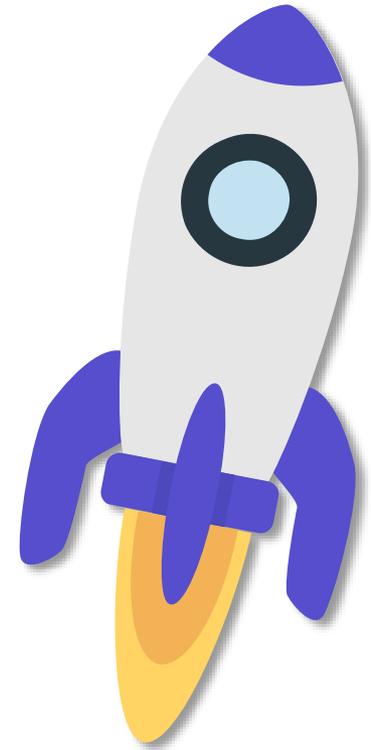
# Líneas futuras

- Independencia de otros plugins
- Paralelización completa
- Implementación de más algoritmos
- Extender herramienta para hash de otras partes de la cabecera
- Estudiar otras partes de un proceso
- Evaluación de rendimiento de algoritmos



# Líneas futuras

- Independencia de otros plugins
- Paralelización completa
- Implementación de más algoritmos
- Extender herramienta para hash de otras partes de la cabecera
- Estudiar otras partes de un proceso
- Evaluación de rendimiento de algoritmos
- Ampliar a más tipos de malware



# Evaluación de algoritmos de fuzzy hashing para similitud entre procesos

**Iñaki Abadía Osta**

Director: Ricardo J. Rodríguez

Ponente: José Merseguer Hernáiz

Septiembre de 2017

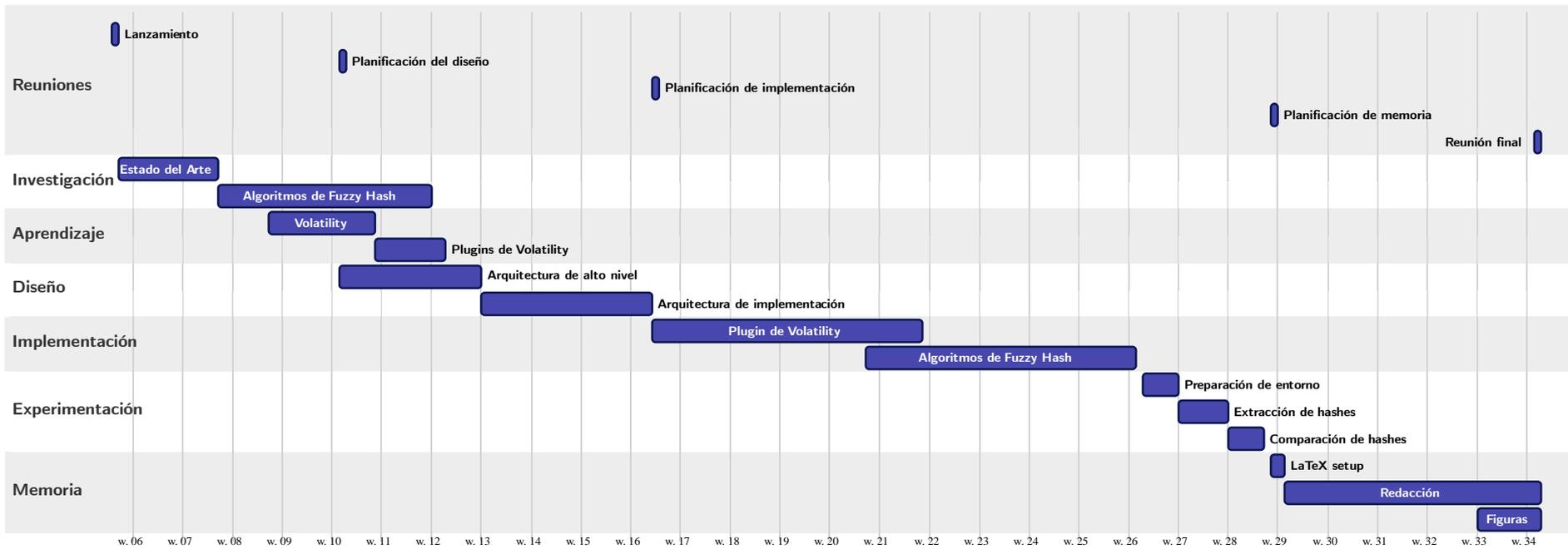
Curso 16/17

Trabajo Fin de Grado – Grado en Ingeniería Informática

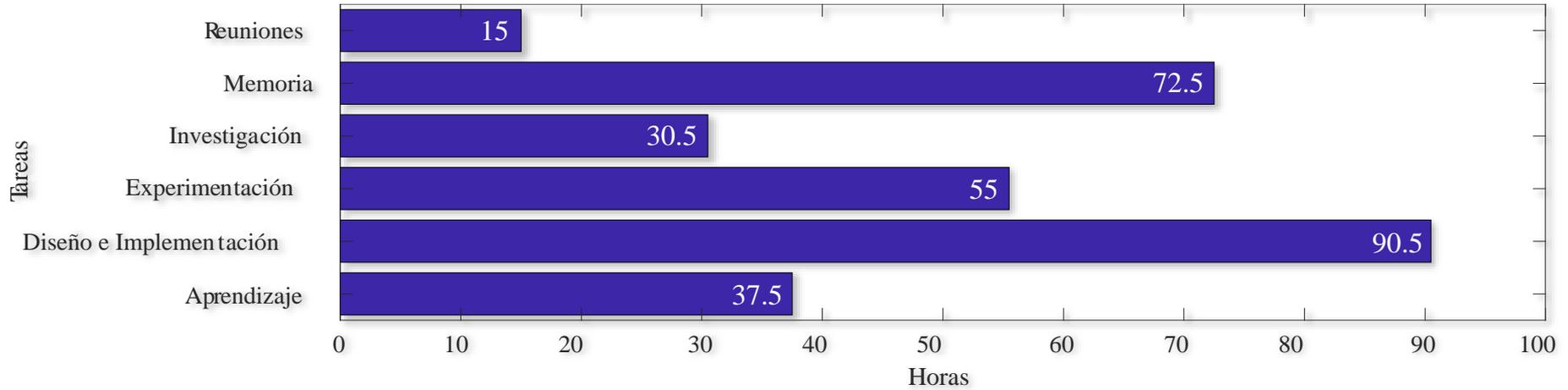
**Escuela de Ingeniería y Arquitectura**

Universidad de Zaragoza

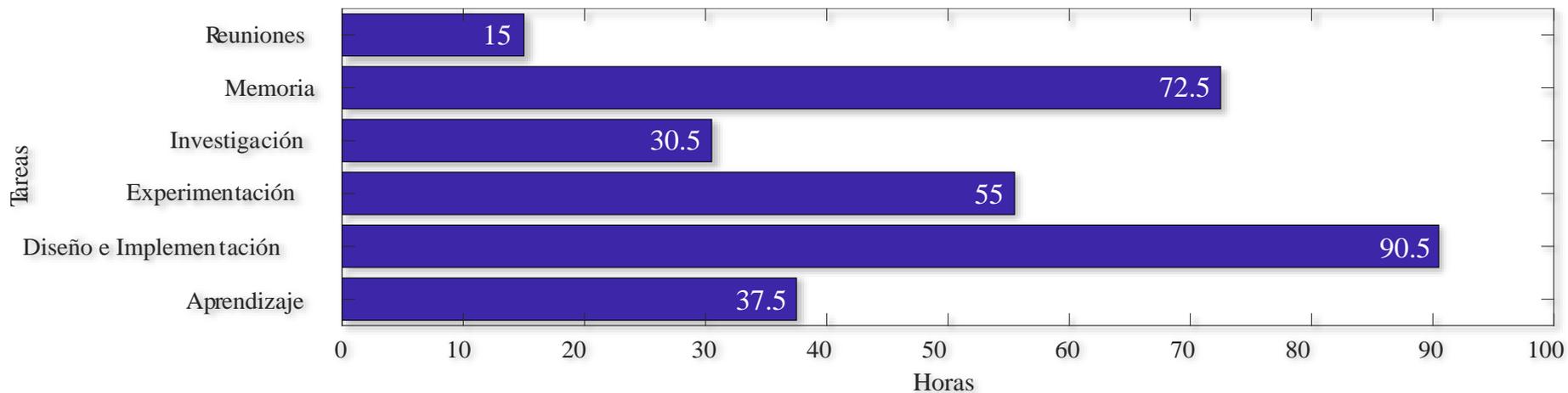
# Horas de trabajo



# Horas de trabajo



# Horas de trabajo



Total: **301** horas

# Bibliografía I

- 📄 **[AT17]** AV-TEST. Malware statistics. <https://www.av-test.org/en/statistics/malware/>, 2017. [Online; accedido 26 de Julio de 2017].
- 📄 **[SBAAN16]** N. Sarantinos, C. Benzaïd, O. Arabiat, and A. Al-Nemrat. Forensic Malware Analysis: The Value of Fuzzy Hashing Algorithms in Identifying Similarities. In *2016 IEEE Trustcom/BigDataSE/ISPA*, pages 1782-1787, Aug 2016.
- 📄 **[NMAM<sup>+</sup>16]** A. P. Namanya, Q. K. A. Mirza, H. Al-Mohannadi, I. U. Awan, and J. F. P. Disso. Detection of Malicious Portable Executables Using Evidence Combinational Theory with Fuzzy Hashing. In *2016 IEEE 4th International Conference on Future Internet of Things and Cloud (FiCloud)*, pages 91-98, Aug 2016.
- 📄 **[AS15]** M. Arefkhani and M. Soryani. Malware clustering using image processing hashes. In *2015 9th Iranian Conference on Machine Vision and Image Processing (MVIP)*, pages 214-218, Nov 2015.

# Evaluación de algoritmos de fuzzy hashing para similitud entre procesos

**Iñaki Abadía Osta**

Director: Ricardo J. Rodríguez

Ponente: José Merseguer Hernáiz

Septiembre de 2017

Curso 16/17

Trabajo Fin de Grado – Grado en Ingeniería Informática

**Escuela de Ingeniería y Arquitectura**

Universidad de Zaragoza