OWASP
The Open Web Application Security Project

Introducción a *Machine Learning* para Seguridad Informática

Frans van Dunné

San José, Abril 30, 2018

¡Hola OWASP!

OWASP
The Open Web Application Security Project

Frans van Dunné, *PhD*
Chief Data Officer

ixpantia

- Estrategia e **innovación basada en datos**
- Modelado de **procesos y gestión de datos**
- Diseño e implementación de **algoritmos y dataductos**
- **Interoperabilidad** de datos
- Arquitecturas de **microservicios**
- Industrias diversas (privado, gobiernos, ONG's)

@fransvandunne

Introducción

OWASP
The Open Web Application Security Project

Definiciones

- Tipos de Datos
- Métodos Supervisados y No Supervisados
- Dataductos

Ejemplos

- PCA - reducción de dimensiones
- Random Forest

Discusión

- Resumen y discusión

Detección

Anomalias

OWASP
The Open Web Application Security Project

Alguien hizo login desde Managua y Ciudad de Panamá a la misma vez.

Alguien está bajando todos los archivos de la jefatura de finanzas.

Un usuario está haciendo login cada 5 minutos durante 24 horas

OWASP
The Open Web Application Security Project

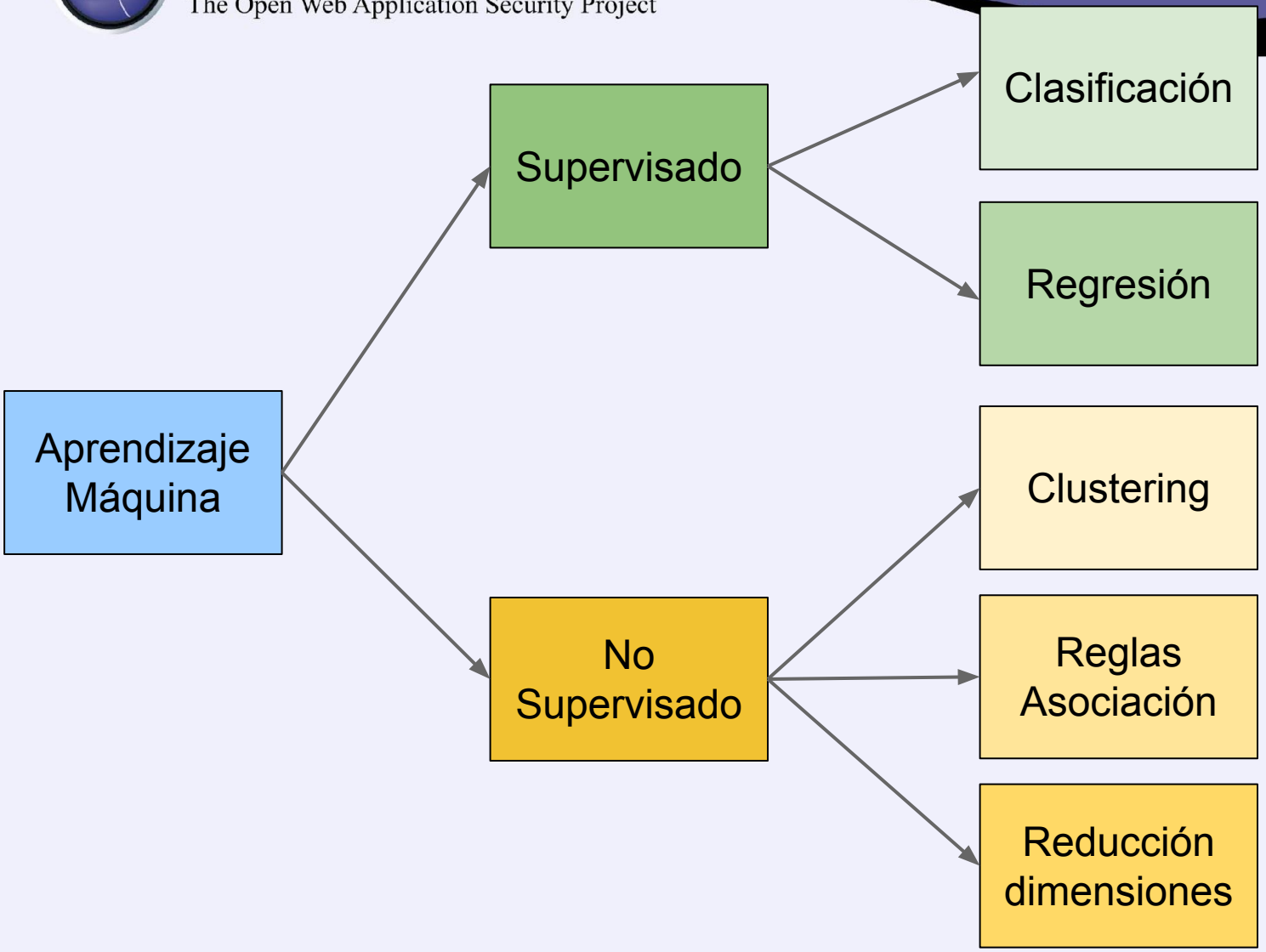
- Tipos de datos
- Actualidad de datos
- Veracidad de datos
- Velocidad de datos
- Variabilidad de datos

Log Files

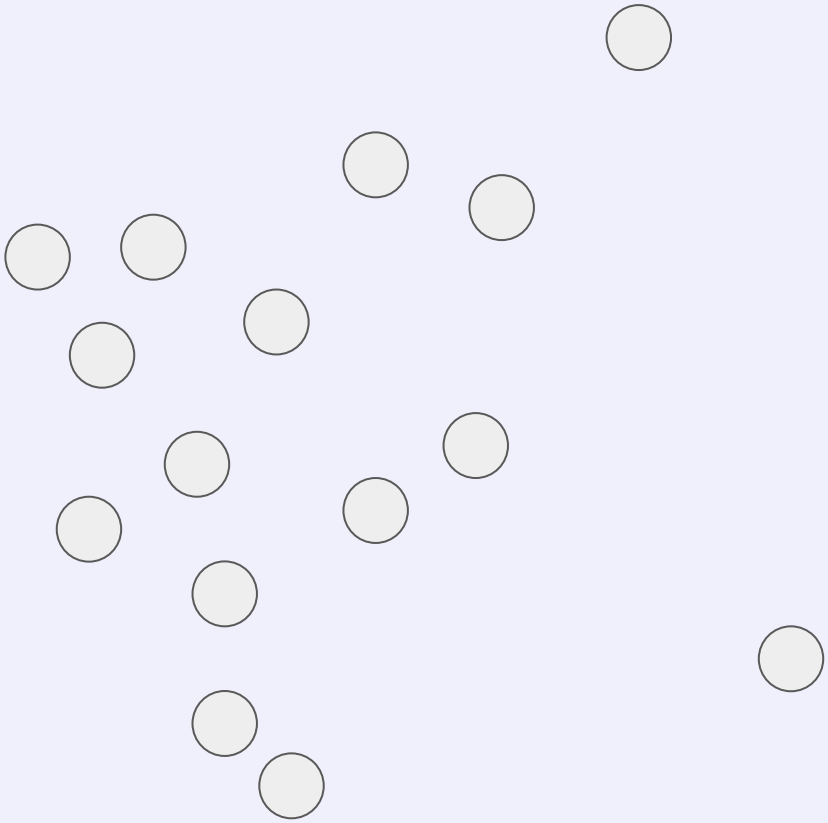
OWASP
The Open Web Application Security Project

- Alto volumen de datos
- Registros históricos
- Alta Velocidad de actualización
- Muy pocos datos etiquetados (*labeled*)
- Poca Variabilidad

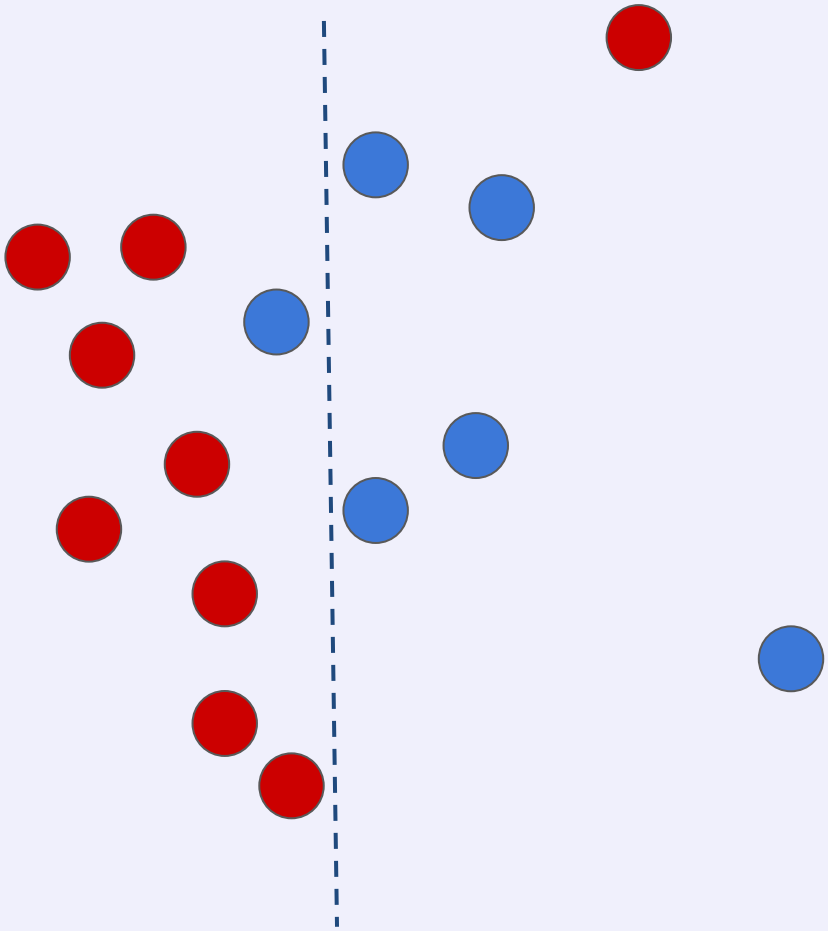
→ Campo para sentido común vs comportamiento observado!


Aprendizaje Máquina
Supervisado
No Supervisado
Clasificación
Regresión
Clustering
Reglas Asociación
Reducción dimensiones

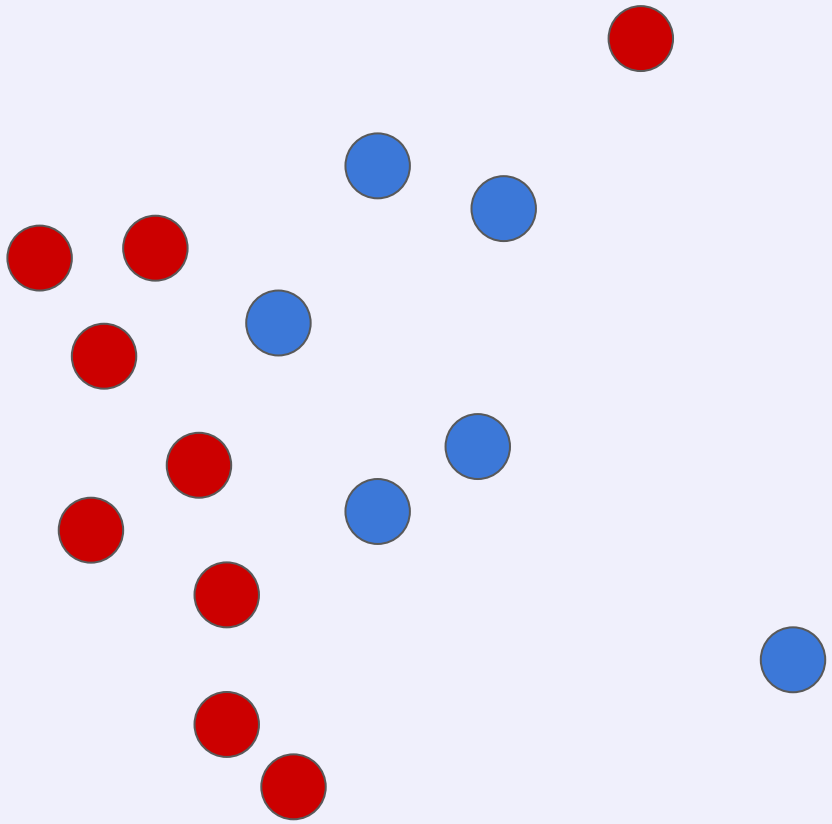
No Supervisado



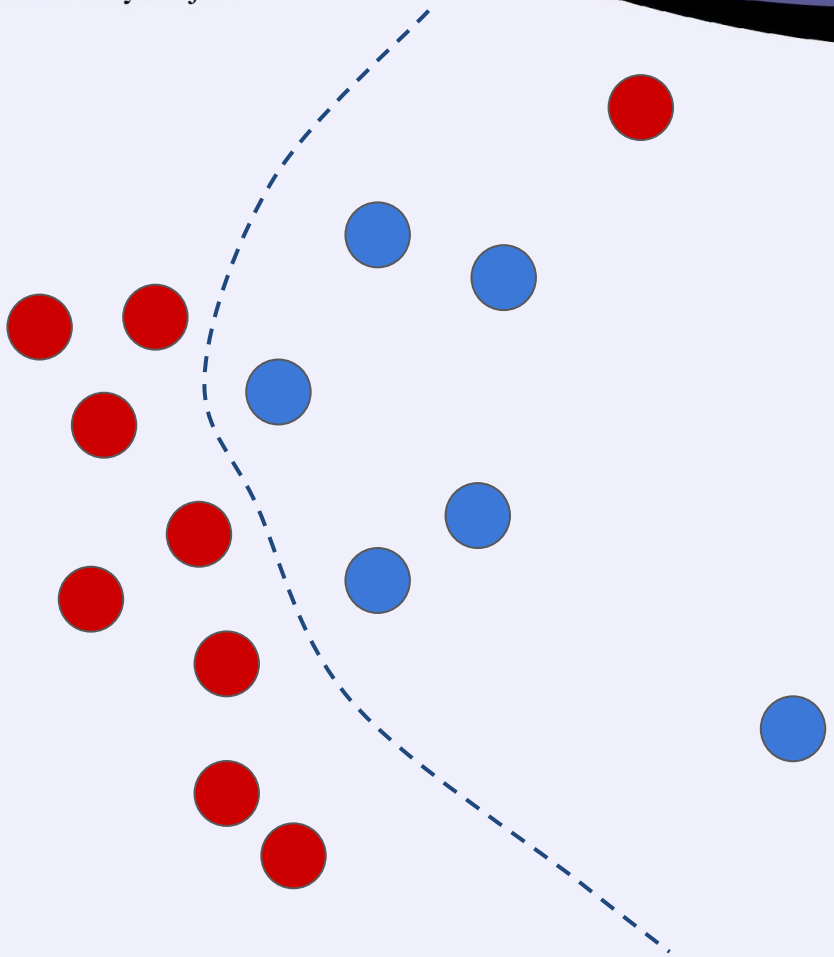
Clustering



Supervisado

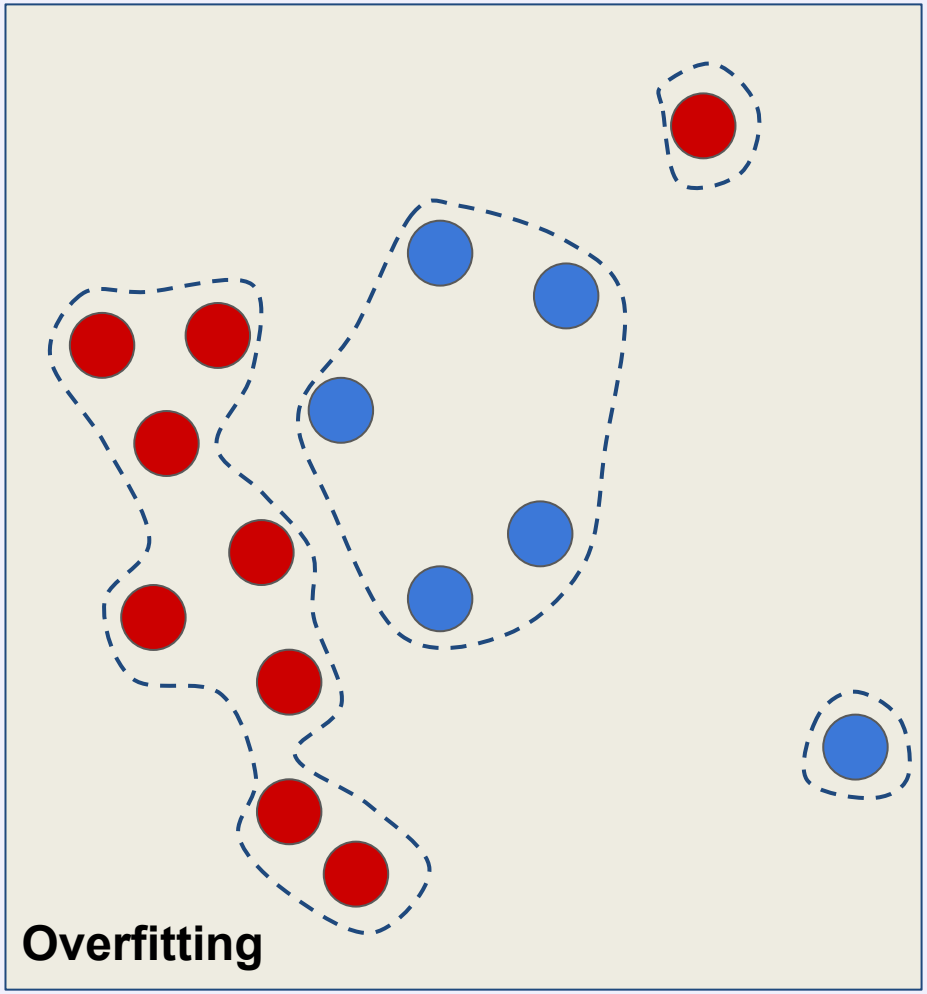


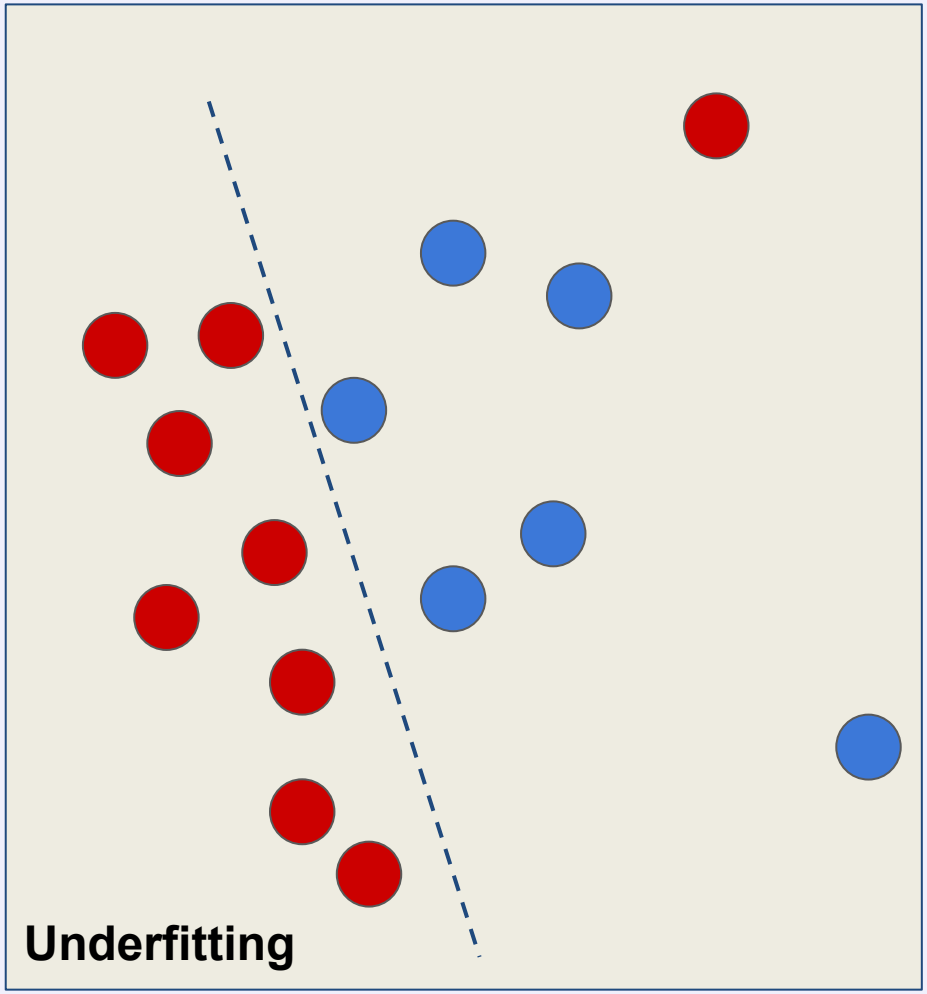
Supervisado



Labeled Data

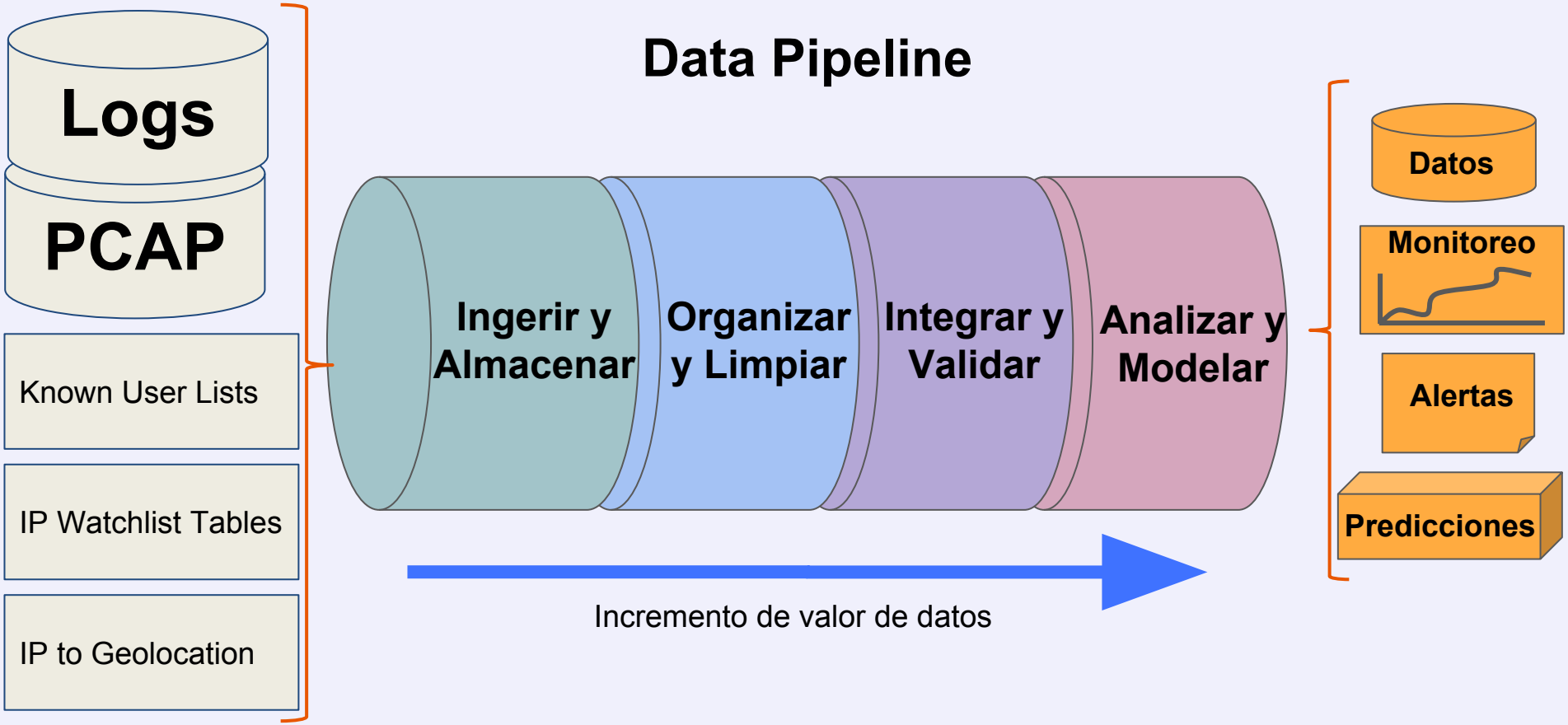
OWASP
The Open Web Application Security Project



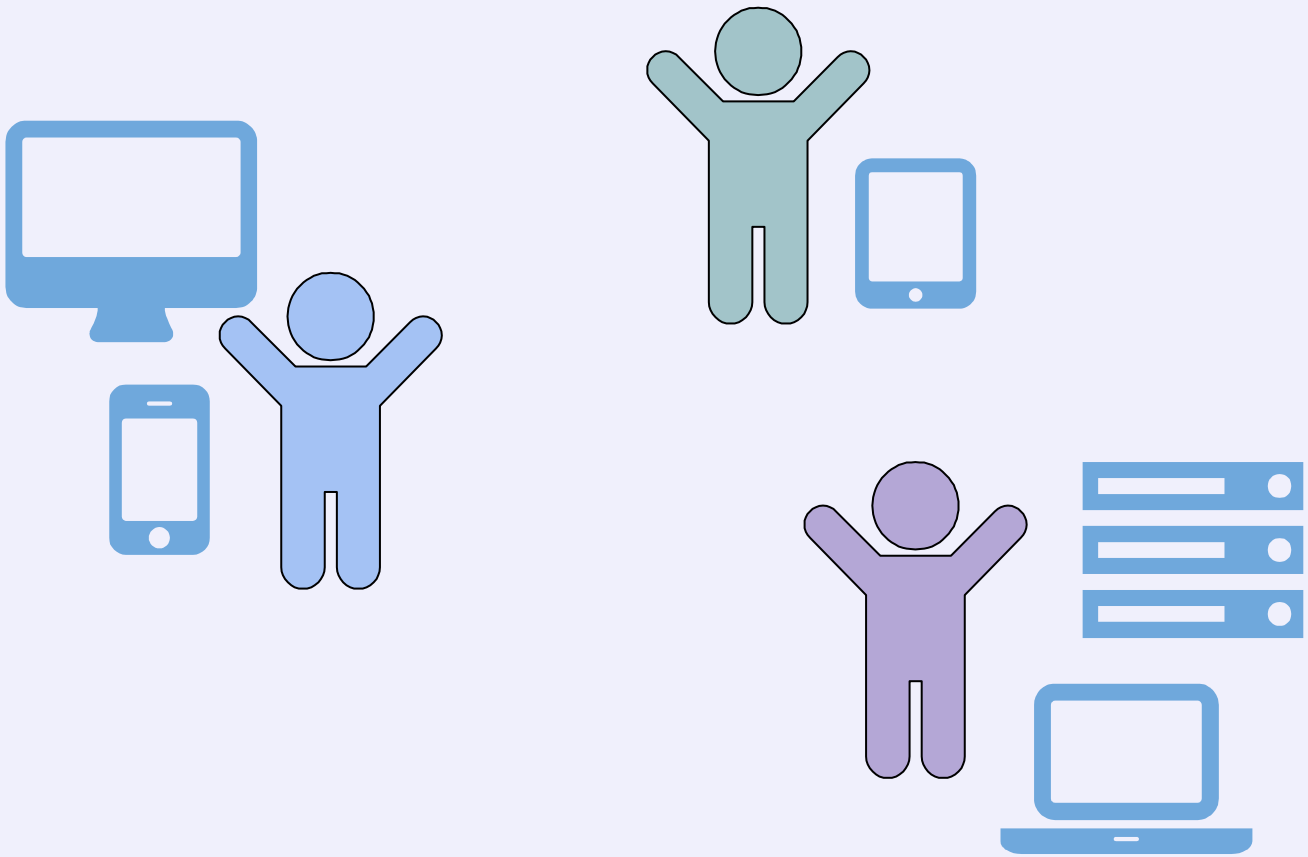


Feature Engineering

OWASP
The Open Web Application Security Project

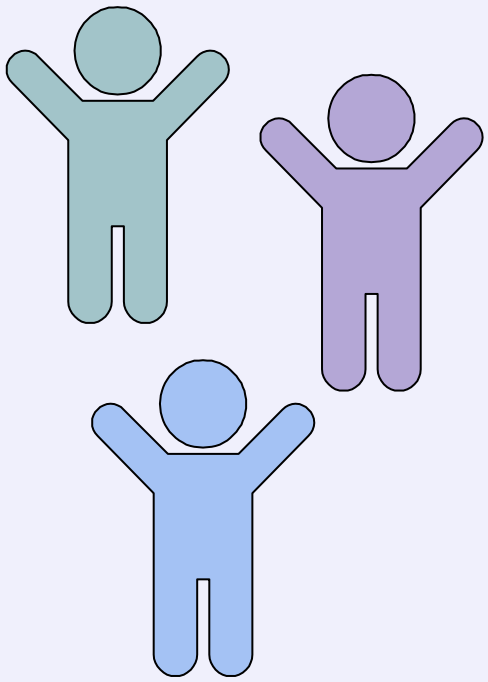


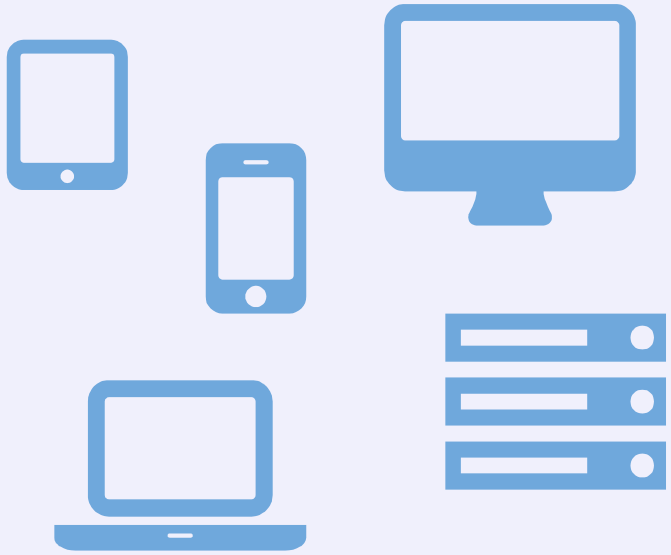
Comportamiento - de quien



Comportamiento - de quien

OWASP
The Open Web Application Security Project

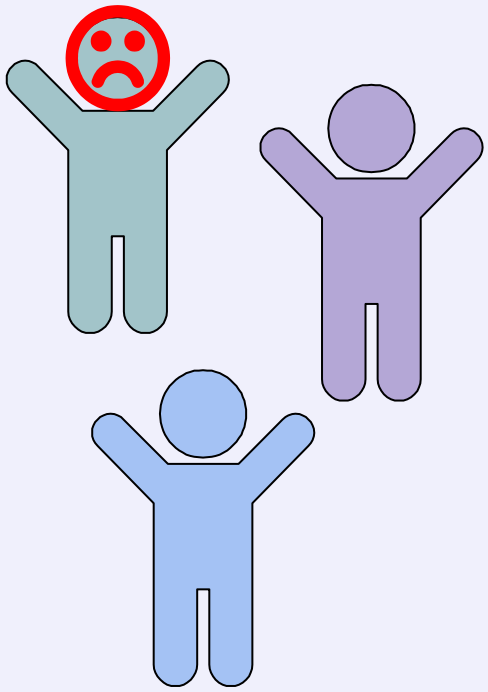


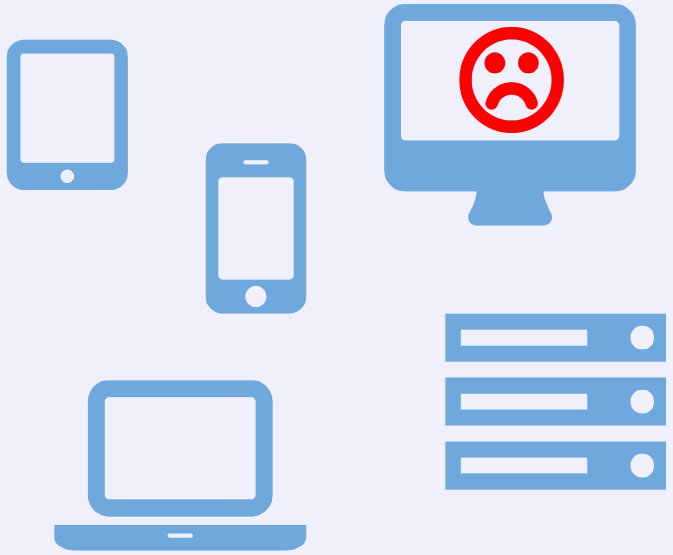


Unidad de Análisis

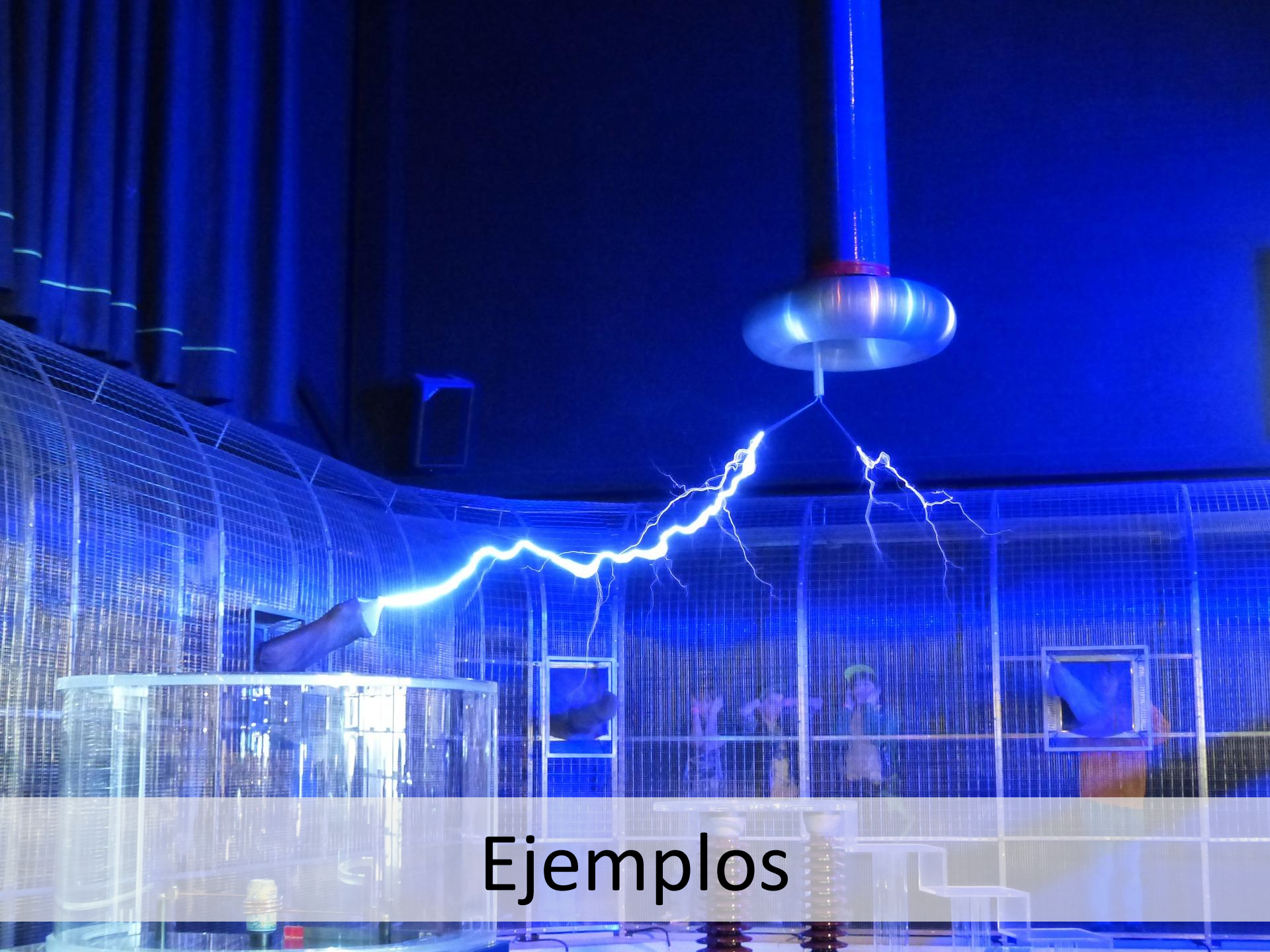
Comportamiento - de quien

OWASP
The Open Web Application Security Project





Class Imbalance



Ejemplos

Datos

OWASP
The Open Web Application Security Project

SecRepo.com - Samples of Security Related Data

Finding samples of various types of Security related can be a giant pain. This is my attempt to keep a somewhat curated list of Security related data I've found, created, or was pointed to. If you perform any kind of analysis with any of this data please let me know and I'd be happy to link it from here or host it here. Hopefully by looking at others research and analysis it will inspire people to add-on, improve, and create new ideas.

Q: How do you give without having to do anything?
A: Simply visit this site.
I've decided that I'm going to start posting the logs from this site to the site. It's a great way to open source some data, and after a few discussions I don't think any privacy will be violated. If I receive a lot of backlash about this decision perhaps I'll reverse it, but until further notice web logs for this domain will be available here.

http://www.secrepo.com

Ejemplo Log Files

OWASP
The Open Web Application Security Project

```
GET
http://localhost:8080/tienda1/publico/anadir.jsp?id=2&nombre=Jam%F3n+Ib%E9rico&precio=85&cantidad=%27%3B+DROP+TABLE+usuarios%3B+SELECT+*+FROM+datos+WHERE+nombre
+LIKE+%27%25&B1=A%F1adir+al+carrito HTTP/1.1
User-Agent: Mozilla/5.0 (compatible; Konqueror/3.5; Linux) KHTML/3.5.8 (like Gecko)
Pragma: no-cache
Cache-control: no-cache
Accept: text/xml,application/xml,application/xhtml+xml,text/html;q=0.9,text/plain;q=0.8,image/png,*/*;q=0.5
Accept-Encoding: x-gzip, x-deflate, gzip, deflate
Accept-Charset: utf-8, utf-8;q=0.5, *;q=0.5
Accept-Language: en
Host: localhost:8080
Cookie: JSESSIONID=B92A8B48B9008CD29F622A994E0F650D
Connection: close


POST http://localhost:8080/tienda1/publico/anadir.jsp HTTP/1.1
User-Agent: Mozilla/5.0 (compatible; Konqueror/3.5; Linux) KHTML/3.5.8 (like Gecko)
Pragma: no-cache
Cache-control: no-cache
Accept: text/xml,application/xml,application/xhtml+xml,text/html;q=0.9,text/plain;q=0.8,image/png,*/*;q=0.5
Accept-Encoding: x-gzip, x-deflate, gzip, deflate
Accept-Charset: utf-8, utf-8;q=0.5, *;q=0.5
Accept-Language: en
Host: localhost:8080
Cookie: JSESSIONID=AE29AEEBDE479D5E1A18B4108C8E3CE0
Content-Type: application/x-www-form-urlencoded
Connection: close
Content-Length: 146

id=2&nombre=Jam%F3n+Ib%E9rico&precio=85&cantidad=%27%3B+DROP+TABLE+usuarios%3B+SELECT+*+FROM+datos+WHERE+nombre+LIKE+%27%25&B1=A%F1adir+al+carrito

GET http://localhost:8080/tienda1/publico/anadir.jsp?id=2%3B&nombre=Jam%F3n+Ib%E9rico&precio=85&cantidad=49&B1=A%F1adir+al+carrito HTTP/1.1
User-Agent: Mozilla/5.0 (compatible; Konqueror/3.5; Linux) KHTML/3.5.8 (like Gecko)
Pragma: no-cache
Cache-control: no-cache
Accept: text/xml,application/xml,application/xhtml+xml,text/html;q=0.9,text/plain;q=0.8,image/png,*/*;q=0.5
Accept-Encoding: x-gzip, x-deflate, gzip, deflate
Accept-Charset: utf-8, utf-8;q=0.5, *;q=0.5
Accept-Language: en
Host: localhost:8080
Cookie: JSESSIONID=F563B5262843F12ECAE41815ABDEEA54
Connection: close
```

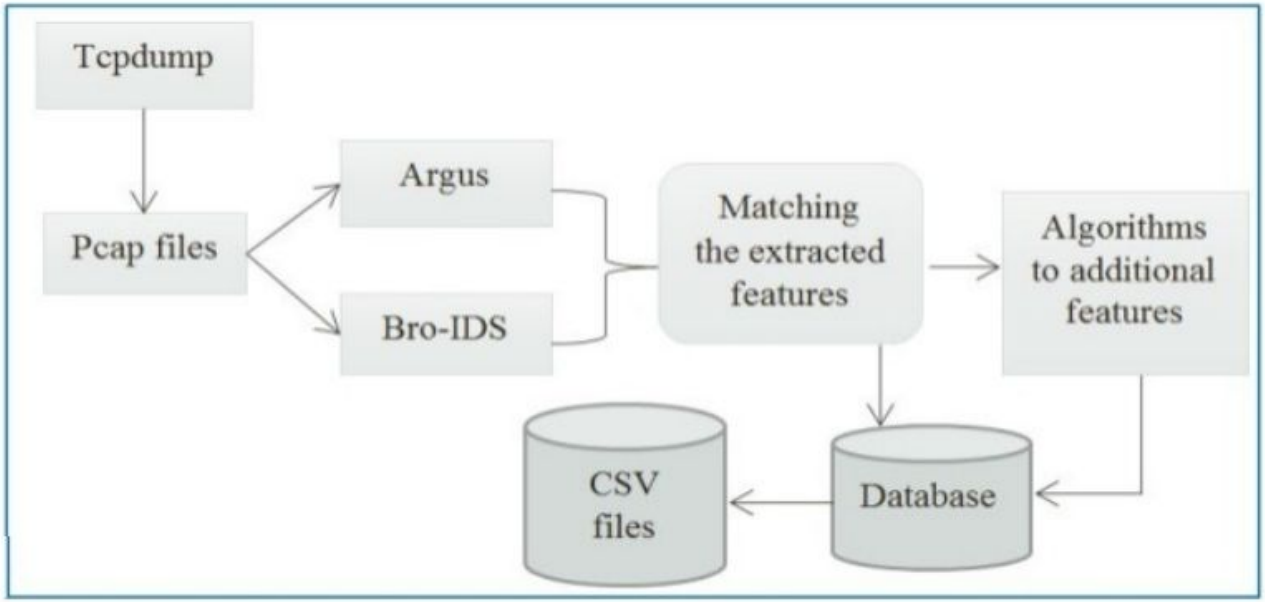

GET http://localhost:8080/tienda1/publico/anadir.jsp?id=2
User-Agent: Mozilla/5.0 (compatible; Konqueror/3.5; Linux
Pragma: no-cache
Cache-control: no-cache
Accept: text/xml,application/xml,application/xhtml+xml,te
Accept-Encoding: x-gzip, x-deflate, gzip, deflate
Accept-Charset: utf-8, utf-8;q=0.5, *;q=0.5
Accept-Language: en
Host: localhost:8080
Cookie: JSESSIONID=B92A8B48B9008CD29F622A994E0F650D
Connection: close

UNSW-NB15: A Comprehensive Data set for Network Intrusion Detection systems

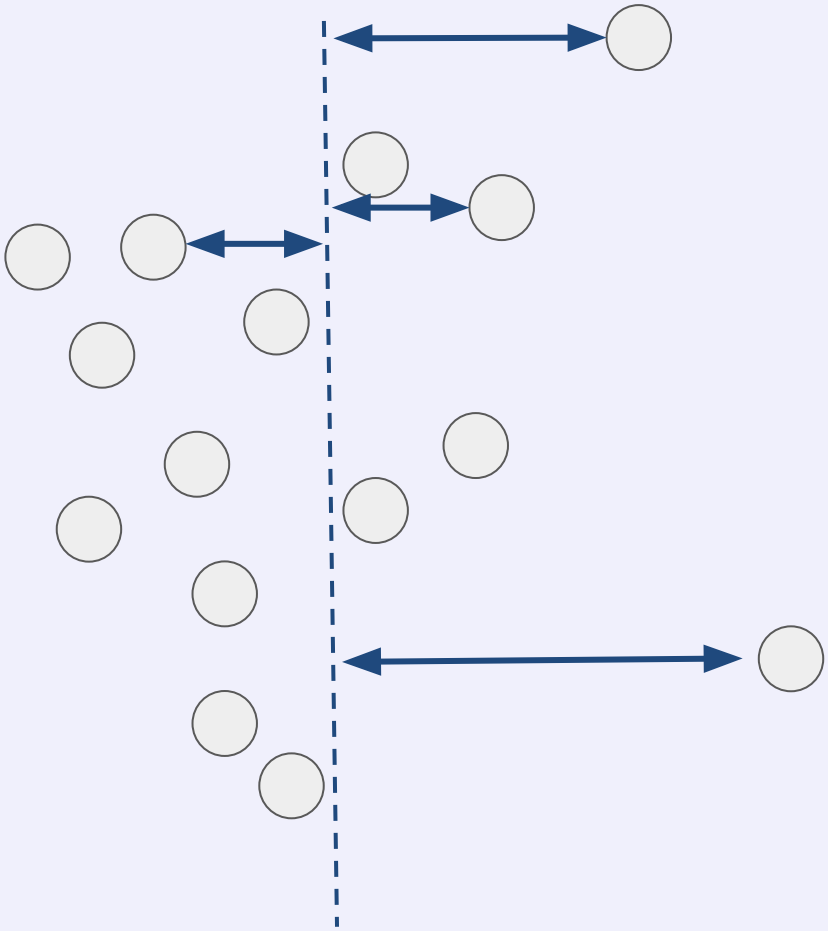
Nour Moustafa, IEEE student Member, Jill Slay

TABLE I. DA

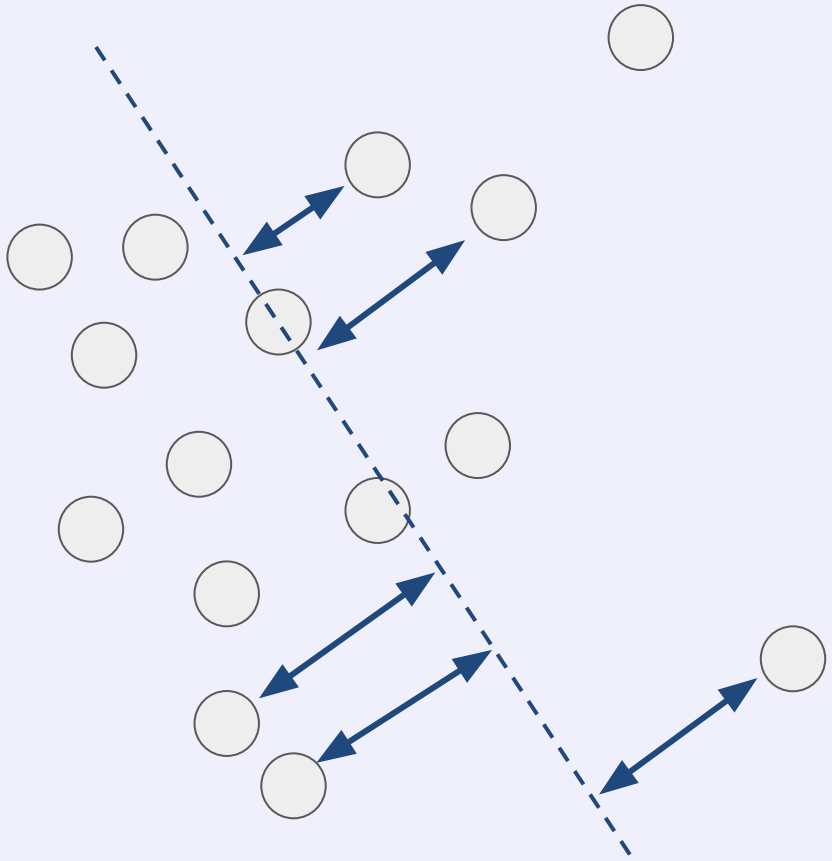
Statistical features			
No._of_flows		987	
Src_bytes		4,8	
Des_bytes		44,	
Src_Pkts		41,	
Dst_pkts		53,	
Protocol types	TCP	771	
	UDP	301	
	ICMP	150	
	Others	150	
Label	Normal	1,064,987	1,153,774
	Attack	22,215	299,068
Unique	Src_ip	40	41
	Dst_ip	44	45



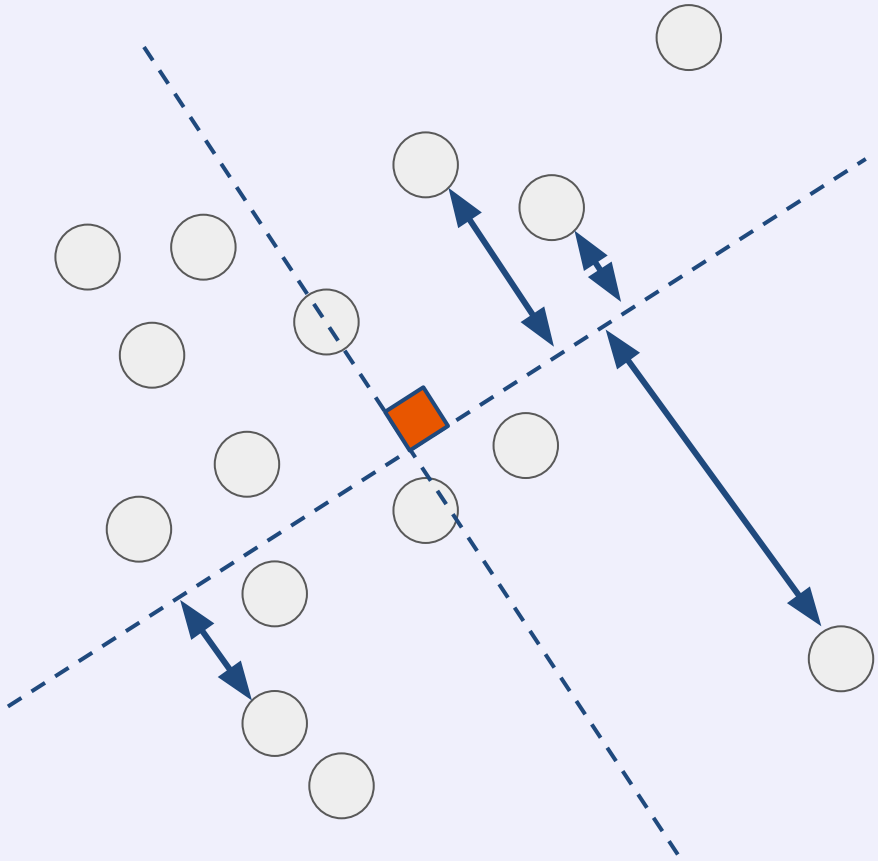
PCA Reducción de Dimensiones



PCA Reducción de Dimensiones

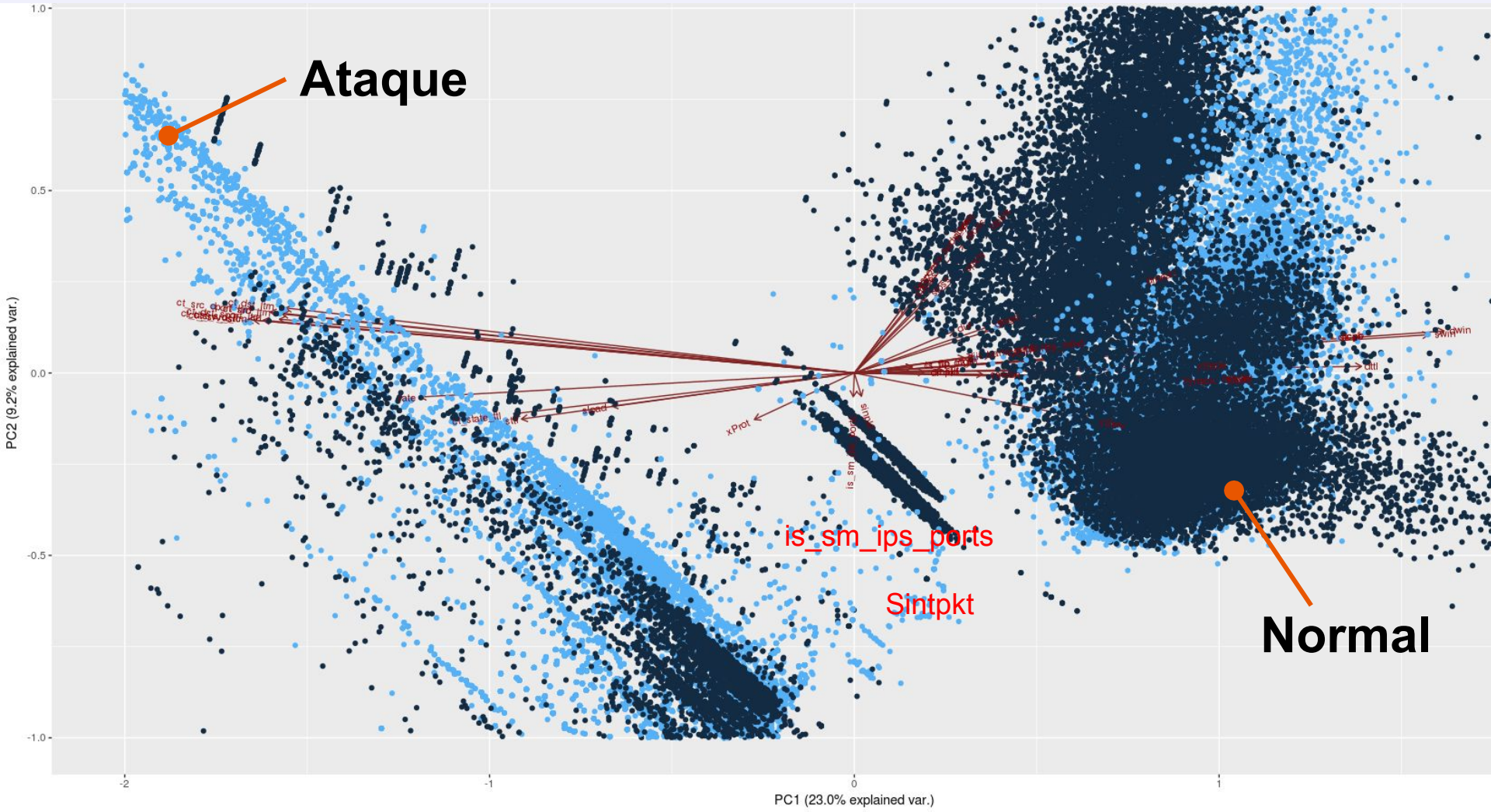


PCA Reducción de Dimensiones

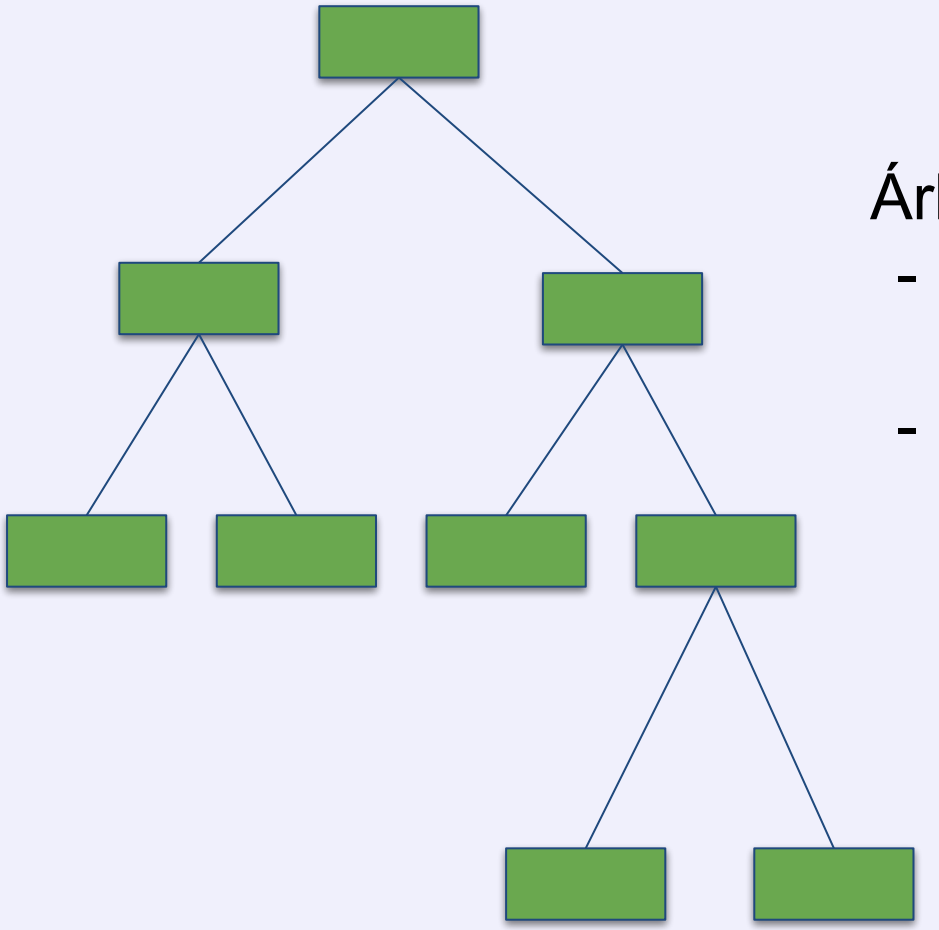


PCA

OWASP
The Open Web Application Security Project



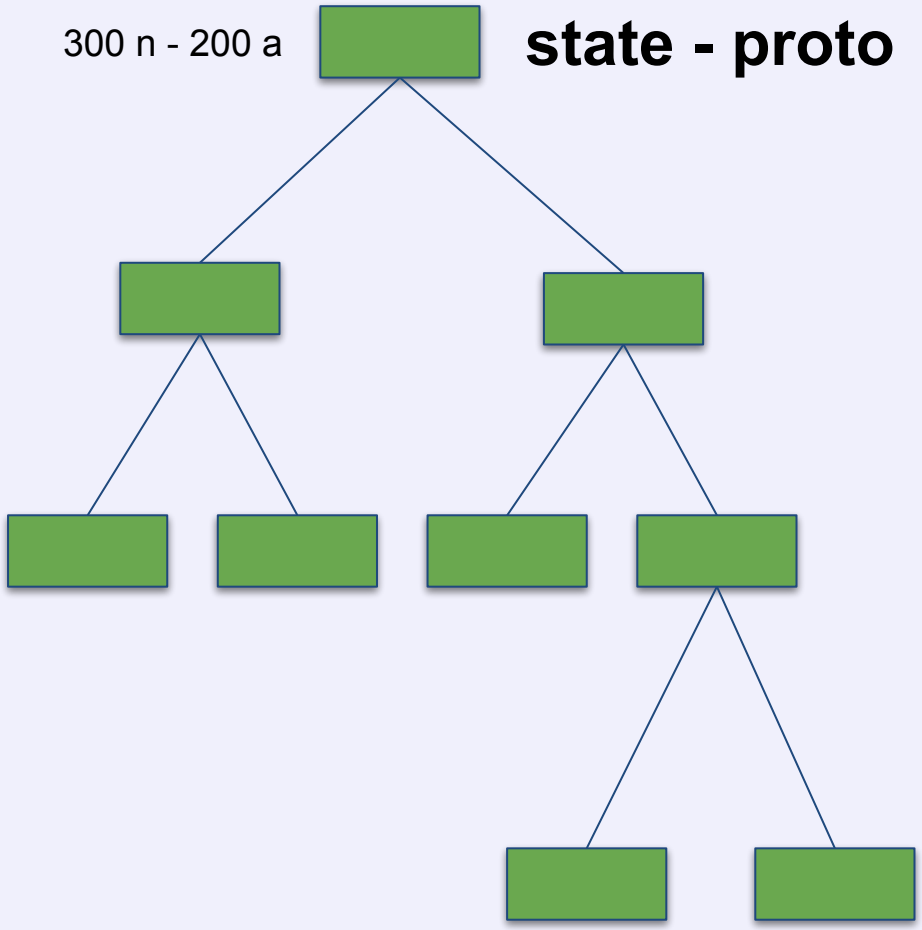
Random Forest



Árbol de decisión pero:

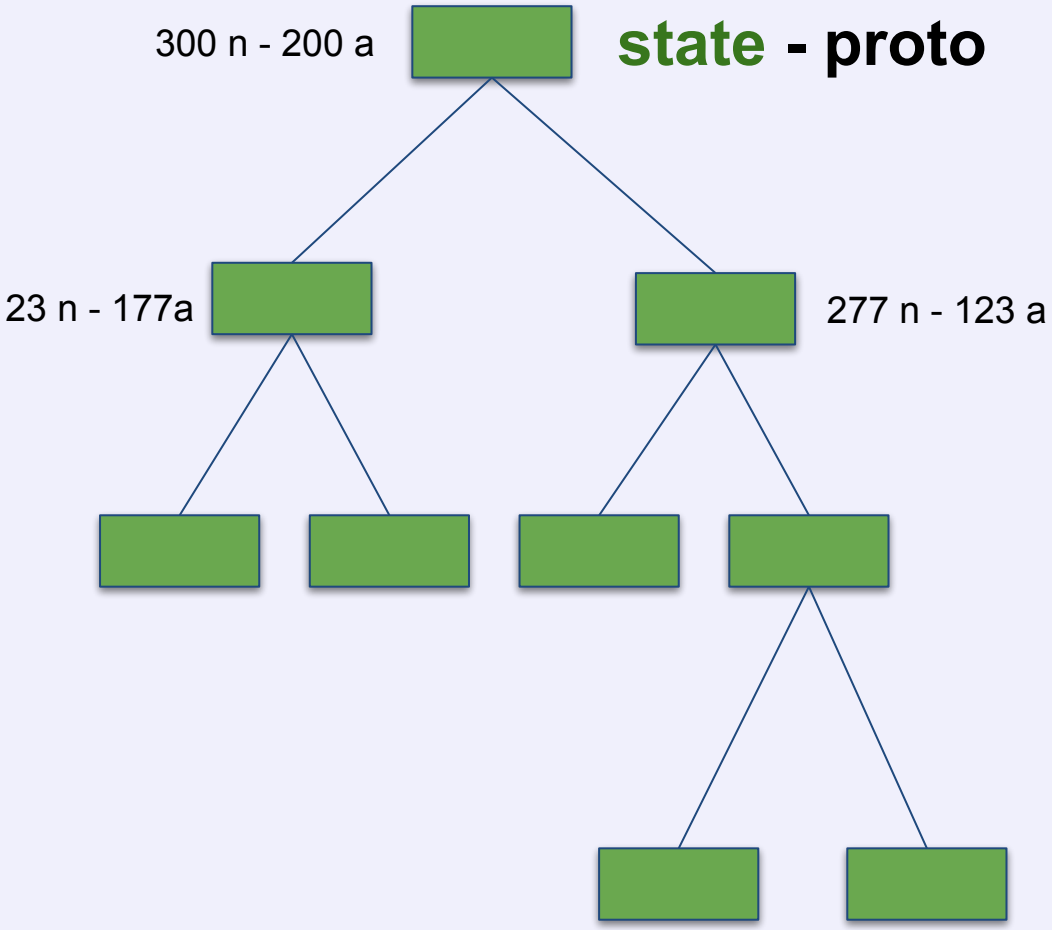
- Subconjunto aleatorio de registros
- Subconjunto aleatorio de variables para cada nodo

Random Forest



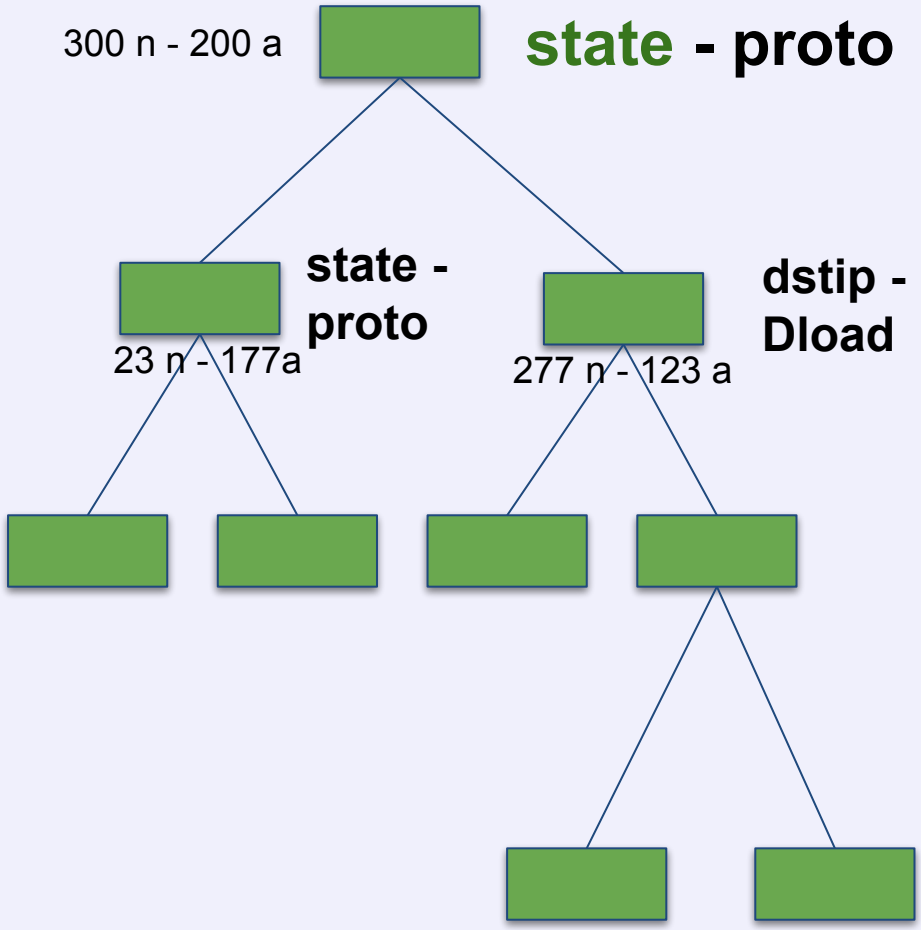
- state
- proto
- dbytes
- dstip
- Dload

Random Forest



- state
- proto
- dbytes
- dstip
- Dload

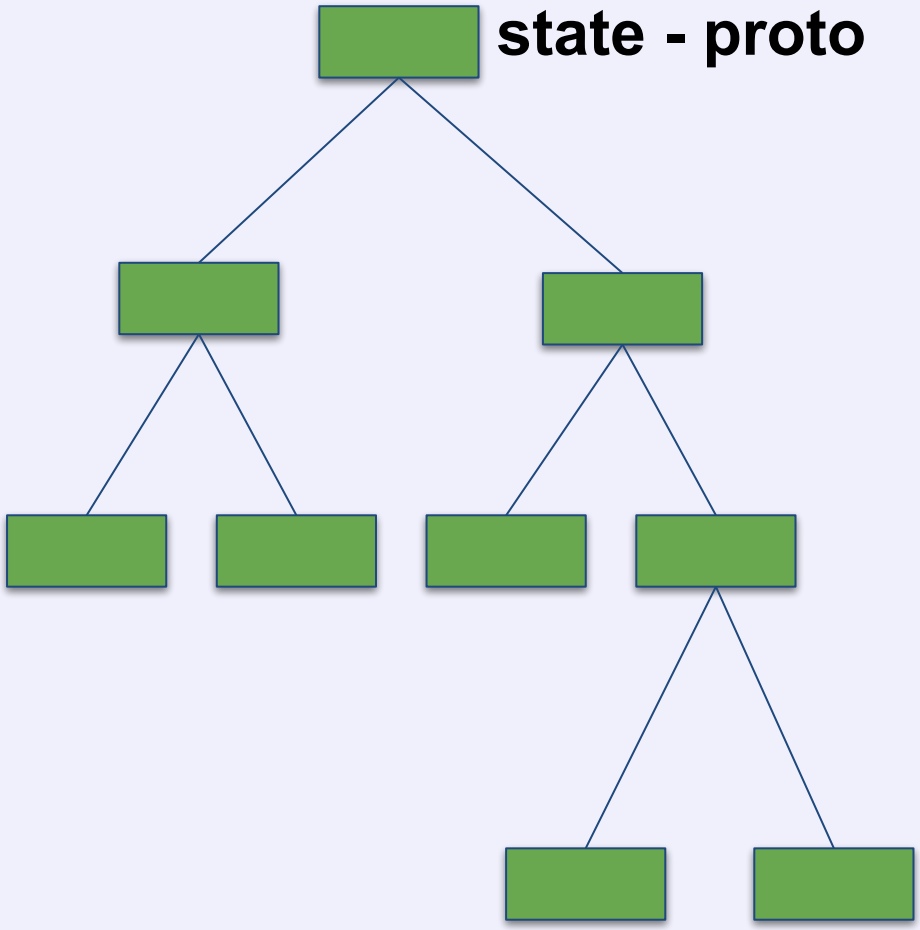
Random Forest

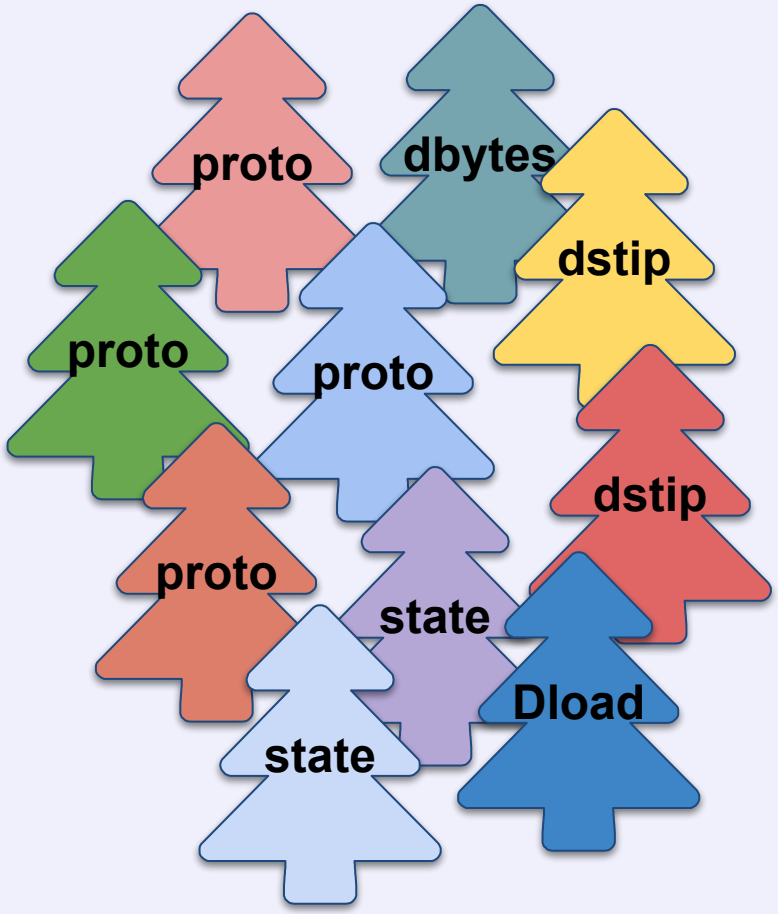


- state
- proto
- dbytes
- dstip
- Dload

Random Forest

- Crea arboles de decisión





Matriz de Confusion

OWASP
The Open Web Application Security Project

	Predicción	
Actual	**Normal**	**Ataque**
Normal	36302	698
Ataque	1155	44177

Matriz de Confusion

OWASP
The Open Web Application Security Project

	Predicción	
Actual	**Normal**	**Ataque**
Normal	**Positivo Real**	698
Ataque	1155	**Negativo Real**

Matriz de Confusion

OWASP
The Open Web Application Security Project

	Predicción	
Actual	**Normal**	**Ataque**
Normal	36302	**Positivo Falso**
Ataque	**Negativo Falso**	44177

Monitoreo

OWASP
The Open Web Application Security Project

- Dashboards
- Alertas
- Protocolos de toma de acción

Monitoreo afecta comportamiento

OWASP
The Open Web Application Security Project

- Los modelos necesitan incluir cambios, por ejemplo en:
  - En comportamiento
  - En volumen
  - En temporalidad

- Oportuna divulgación de resultados de monitoreo pueden ayudar a evitar fraude
- Cual es el costo de un falso positivo, y cual el de un falso negativo?

Los procesos de negocio dentro de los cuales se implementan modelos de detección de fraude necesitan controles y contrapesos adecuados.

Resumen y Discusión

OWASP
The Open Web Application Security Project

Definiciones

- Tipos de Datos
- Métodos Supervisados y No Supervisados
- Dataductos

Ejemplos

- PCA - reducción de dimensiones
- Random Forest

Discusión

- Resumen y discusión

Referencias

OWASP
The Open Web Application Security Project

SaiGanesh Gopalakrishnan, 2017. *Data Science & Machine Learning in Cybersecurity*. AT&T Business Report.

Marvin N. Wright y Andreas Ziegler. 2017. *Fast Implementation of Random Forests for High Dimensional Data in C++ and R.* Journal of Statistical Software 77:1.

Moustafa, Nour, and Jill Slay. UNSW-NB15: a *comprehensive data set for network intrusion detection systems (UNSW-NB15 network data set)*. Military Communications and Information Systems Conference (MilCIS), 2015. IEEE, 2015.

Moustafa, Nour, and Jill Slay. *The evaluation of Network Anomaly Detection Systems: Statistical analysis of the UNSW-NB15 data set and the comparison with the KDD99 data set*. Information Security Journal: A Global Perspective (2016): 1-14.

Botes, F., Leenen, L. and De La Harpe, R. (2017). *Ant Colony Induced Decision Trees for Intrusion Detection*. In: 16th European Conference on Cyber Warfare and Security. ACPI (June 12, 2017), pp.74-83.

¡Hola OWASP!

OWASP
The Open Web Application Security Project

@fransvandunne

ixpantia

www.ixpantia.com

Data Latam

www.datalatam.com

OWASP

The Open Web Application Security Project