



Usando Python para hacer “Data Science” en Seguridad Informática

Gaspar Modelo Howard

Científico Senior



Evento Mensual de Comunidad Dojo

Octubre 9, 2018

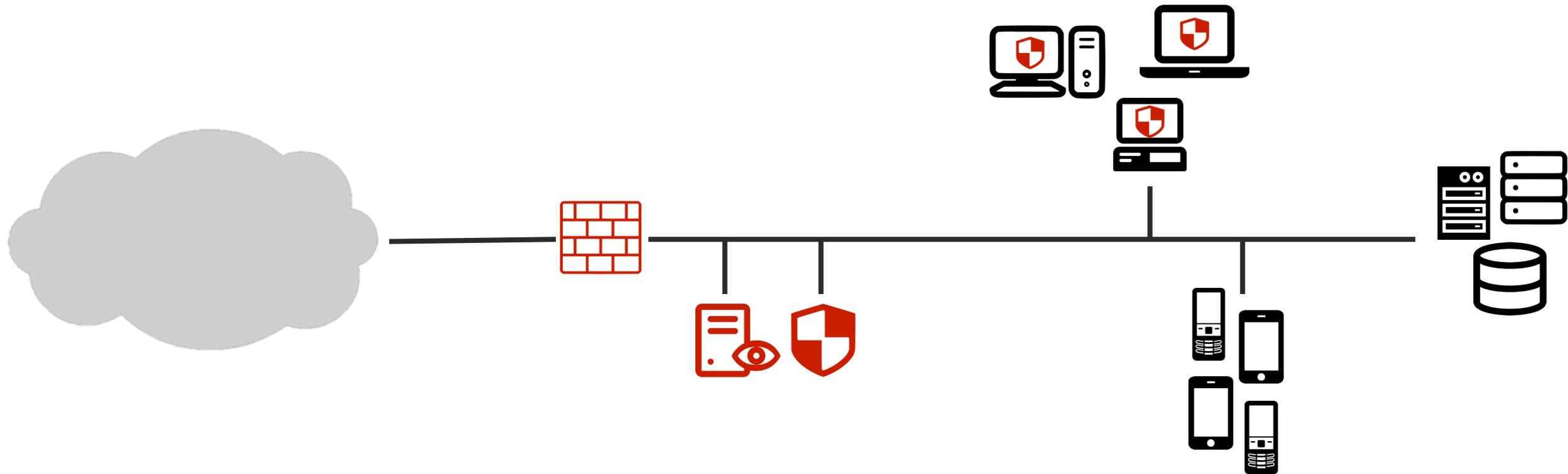
Encuesta

- ¿Sabes qué es Machine Learning (ML)?
- ¿Sabes qué es Data Science (DS)?
- ¿Utilizas DS / ML en tu trabajo (relacionado a seguridad o no)?
- ¿Utilizas Python?
- ¿Usas Python para asegurar tus sistemas?

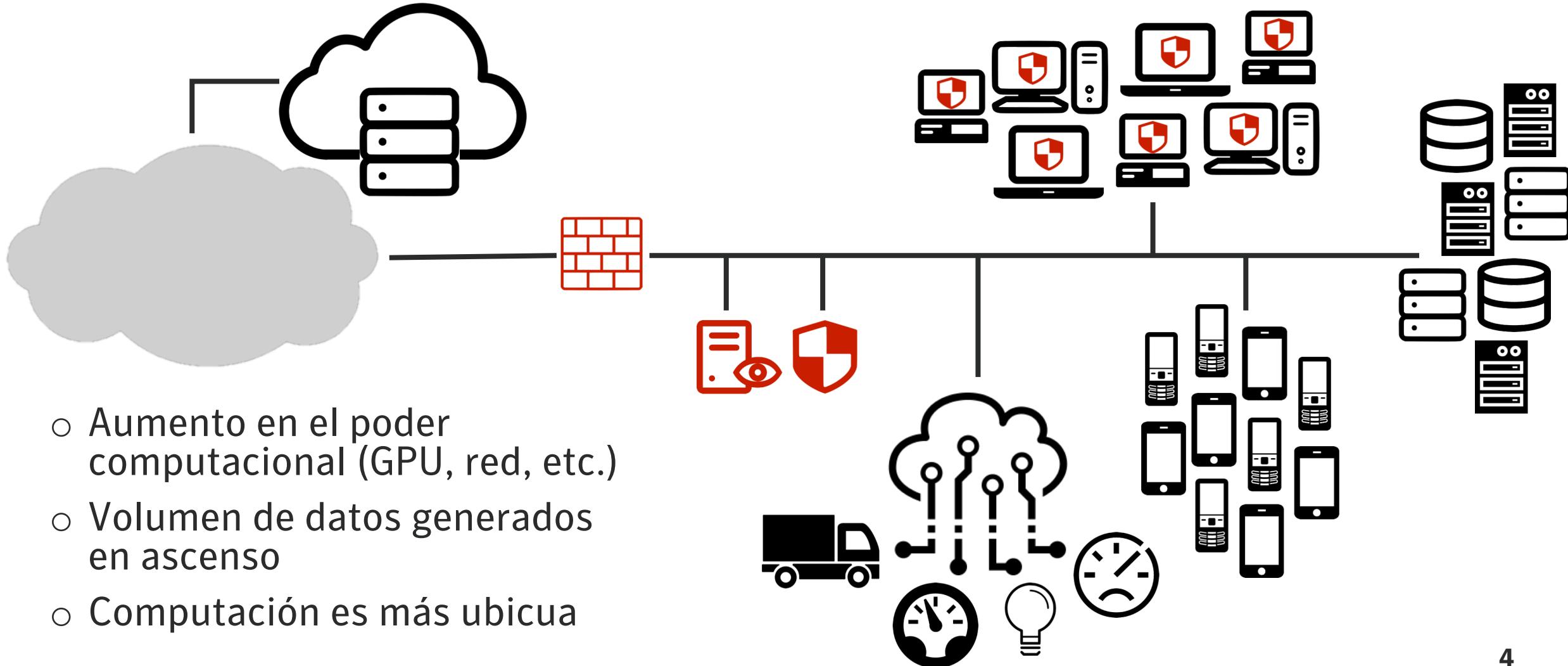
- ¿Tienes curiosidad sobre alguno de estos temas?

Advertencia: Opiniones y comentarios son personales, no de Symantec

Escenario Informático (2000s)



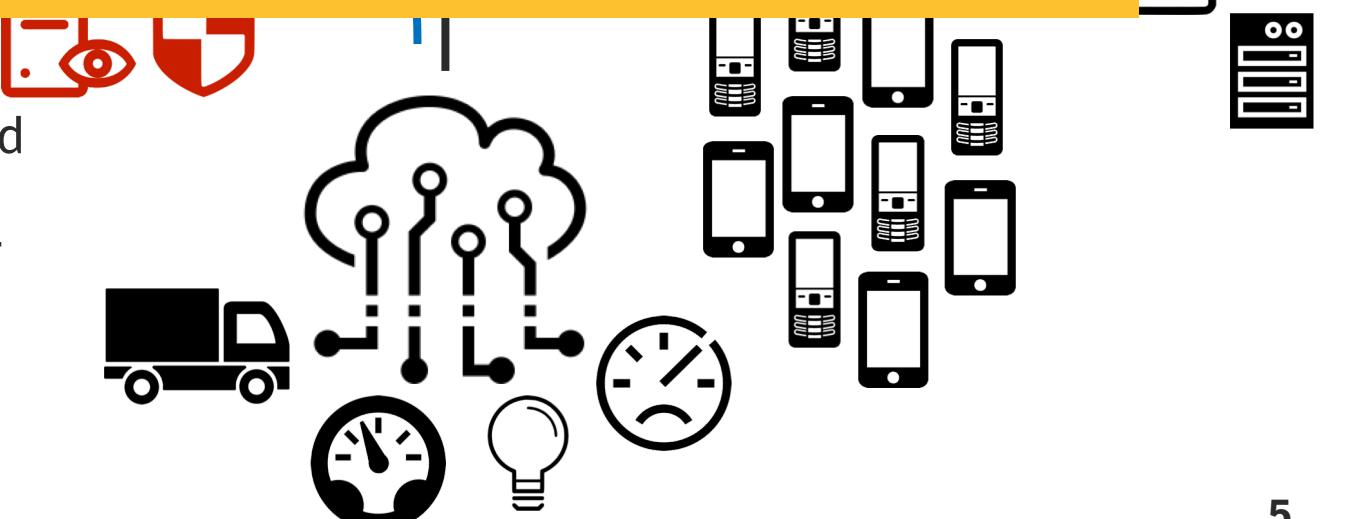
Escenario Informático (2020)



Escenario Informático (2020s)

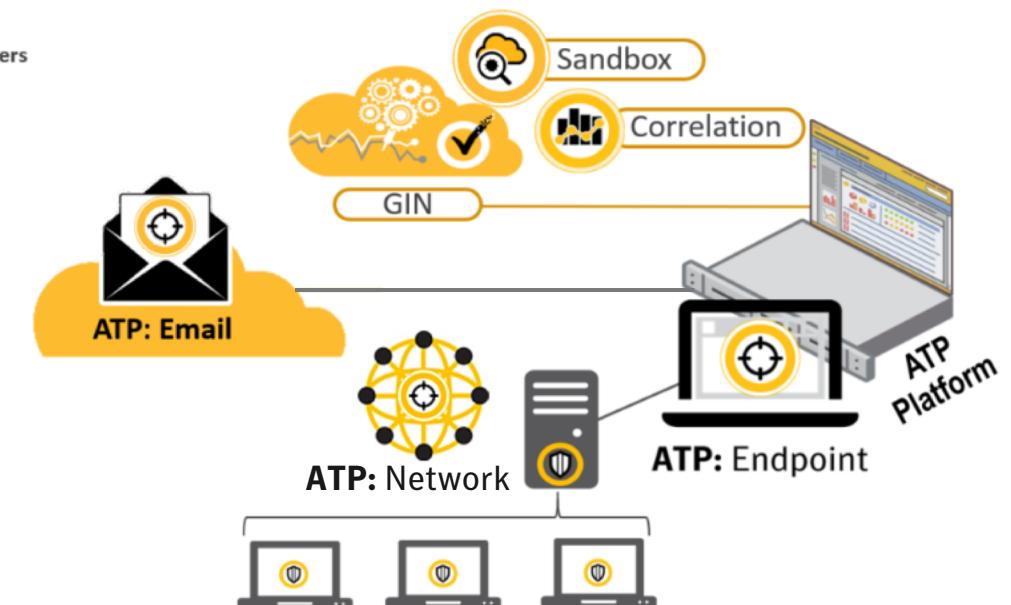
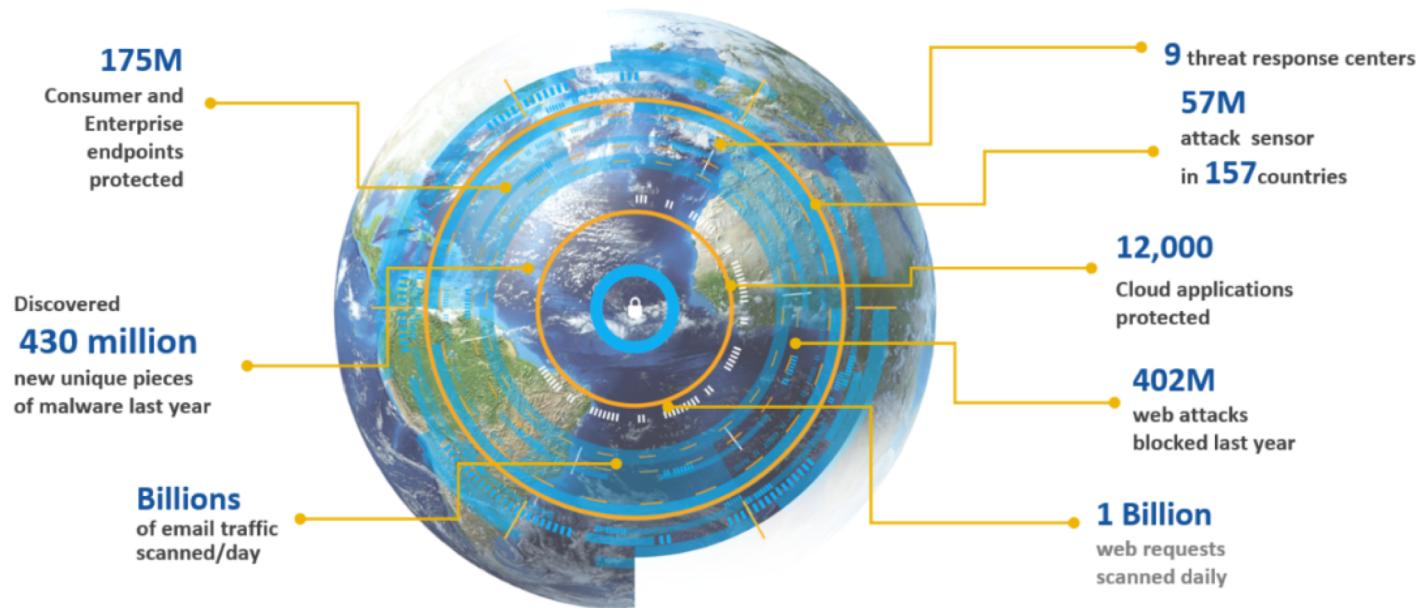
Data Science al rescate!

- Amenazas creciendo en costo, complejidad y severidad
- Frecuencia con la cual se necesita realizar un análisis ad-hoc también en aumento
- Reportes estandarizados, aunque útiles, no son suficientes
- **¿Cómo analizar el diluvio de datos?
Imposible hacerlo de forma manual**



Plataforma Integrada de Seguridad

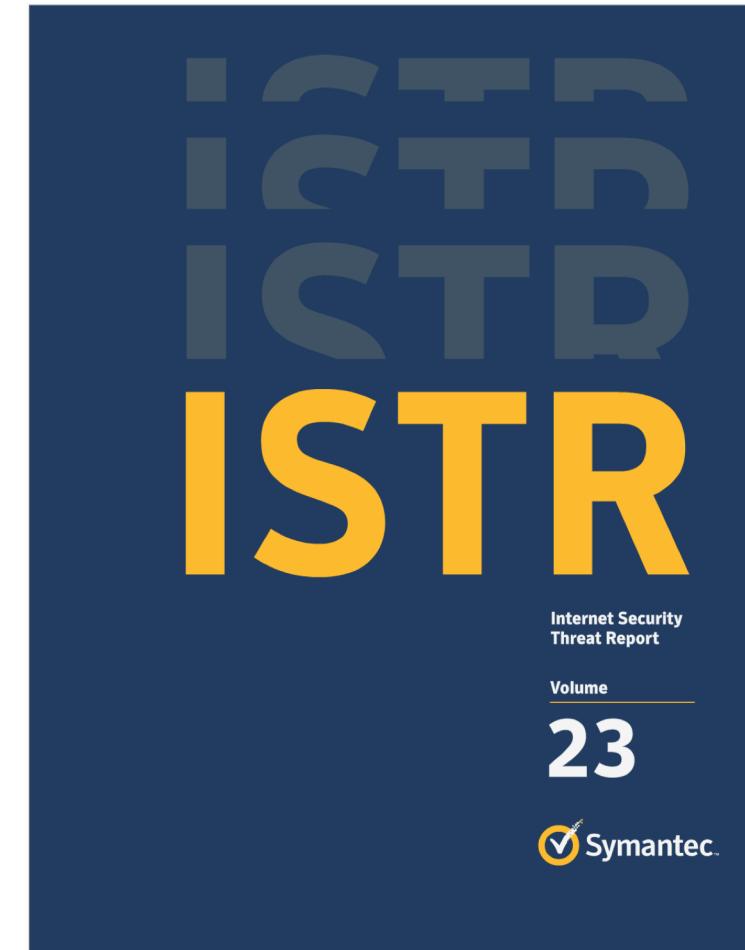
- Herramienta de *data science* para especialistas en seguridad
- Caso Symantec: Advanced Threat Protection (ATP)



- Todo el mundo lo está haciendo: Splunk, FireEye, Cisco, PAN, etc.

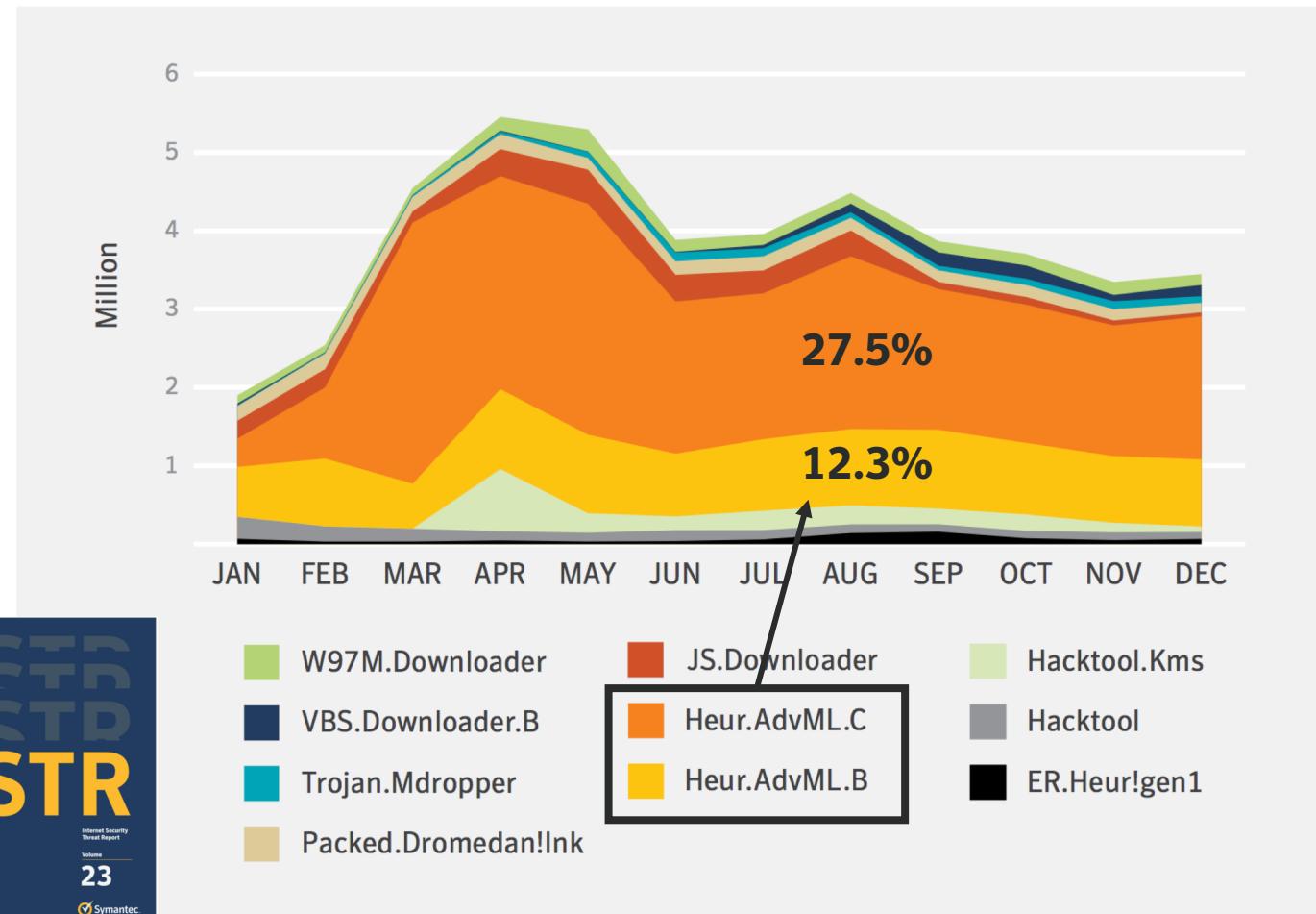
Reporte de Seguridad 2018

- Symantec publica anualmente un resumen de la actividad maliciosa observada, las nuevas tendencias y algunos pronósticos
- <https://symc.ly/2DS11HR>



Top 10 Detecciones de Malware 2017

- La lista está dominada por mecanismos de ML, las cuales detectan nuevas formas de malware genérico
 - ~40% de detecciones
 - Top 10 representa el ~54%
- Symantec detectó WannaCry semanas antes, gracias a módulo de ML



En Conclusión...

- Data Science es parte de la caja de herramientas de una estrategia de seguridad completa
 - No es posible analizar manualmente los datos a nuestro alcance
- Data Science incluye Machine Learning
 - Necesario familiarizarte con esta nueva herramienta



Agenda

- 1** ¿Qué es machine learning y data science?
- 2** Rol de ML+DS en Seguridad
- 3** Ecosistema de data science en Python
- 4** Demo
- 5** Conclusiones

¿Qué es Aprendizaje de Máquinas (ML)?

- Usemos un ejemplo: ¿cómo identificar una cara?



¿Qué es Aprendizaje de Máquinas (ML)?

- Usemos un ejemplo: ¿cómo identificar una cara?



Irving Saladino

- Solución equivocada: programar a mano (definir que es cada objeto)
- **Solución correcta:** Hacer que el programa aprenda,
mostrándole muchos ejemplos de lo que queremos
- En realidad, se escribe la estructura del programa y durante el aprendizaje
se ajustan los parámetros internos (θ)

} Estrategia de
aprendizaje de A.I.

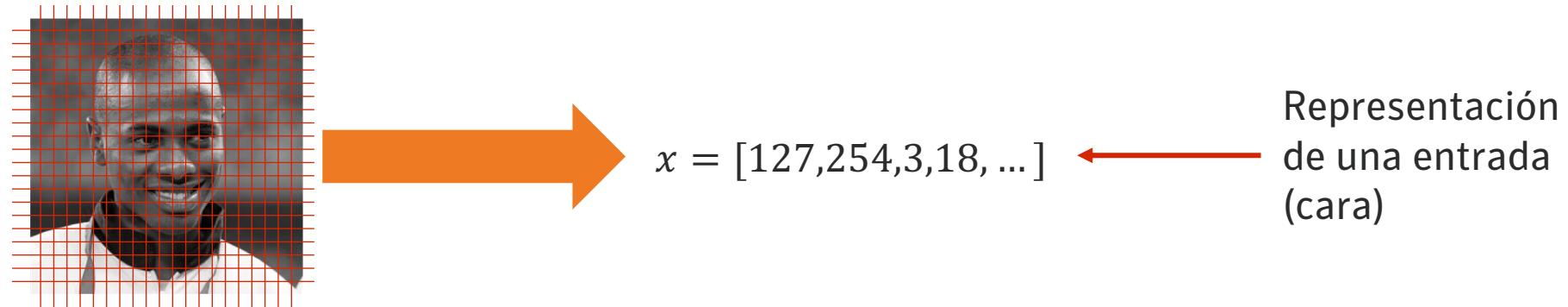
Funciones Canónicas de ML

- X = Entradas, Y = respuestas correctas, Z = salida del programa

Estrategia de Aprendizaje	Descripción	
Supervisado (Supervised Learning)	Dado ejemplos de entradas (X) y su correspondiente salidas (Y), predecir salidas de futuras entradas Ej.: clasificación, regresión	Principales estrategias de ML
No supervisado (Unsupervised Learning)	Dado solamente entradas (X), descubrir estructura, características, etc. Ej.: clustering, outlier detection, compresión	
Descubrimiento de Reglas (Rule Learning)	Descubrir subconjuntos comunes de mediciones	
Con Reforzamiento (Reinforcement Learning)	Maximizar recompensa en base a acciones	

¿Qué es Aprendizaje de Máquinas (ML)?

- ¿Cómo representar la información? Necesitamos seleccionar un método que identifique las regularidades o estructura de la data



- Nuestro programa necesita entonces:

1. Función matemática para calcular salida: $Z = f(X, \theta)$
2. Función de Error: $L(X, Y, Z)$

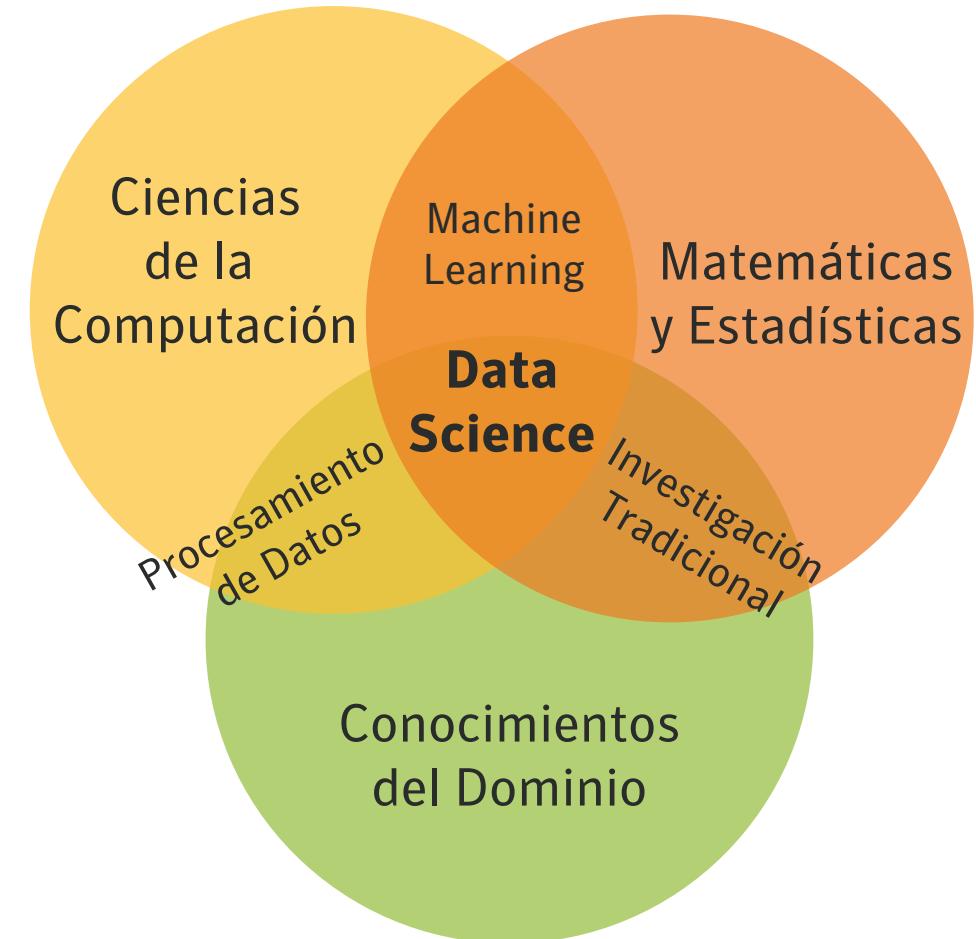
- Reto es decidir como representar X, Y , y $f()$

θ = parámetros a ajustar

Cuan bien estamos aprendiendo

¿Qué es Data Science?

- Es un campo interdisciplinario para:
 - recolectar, organizar, analizar, interpretar y representar datos,
 - usando ciencias de la computación,
 - con un conocimiento del área de interés



Fuente: <http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram>

Rol de Data Science en Seguridad

- ¿Cómo detectar fallas de seguridad (vulnerabilidades o intrusiones) en mi sistema?



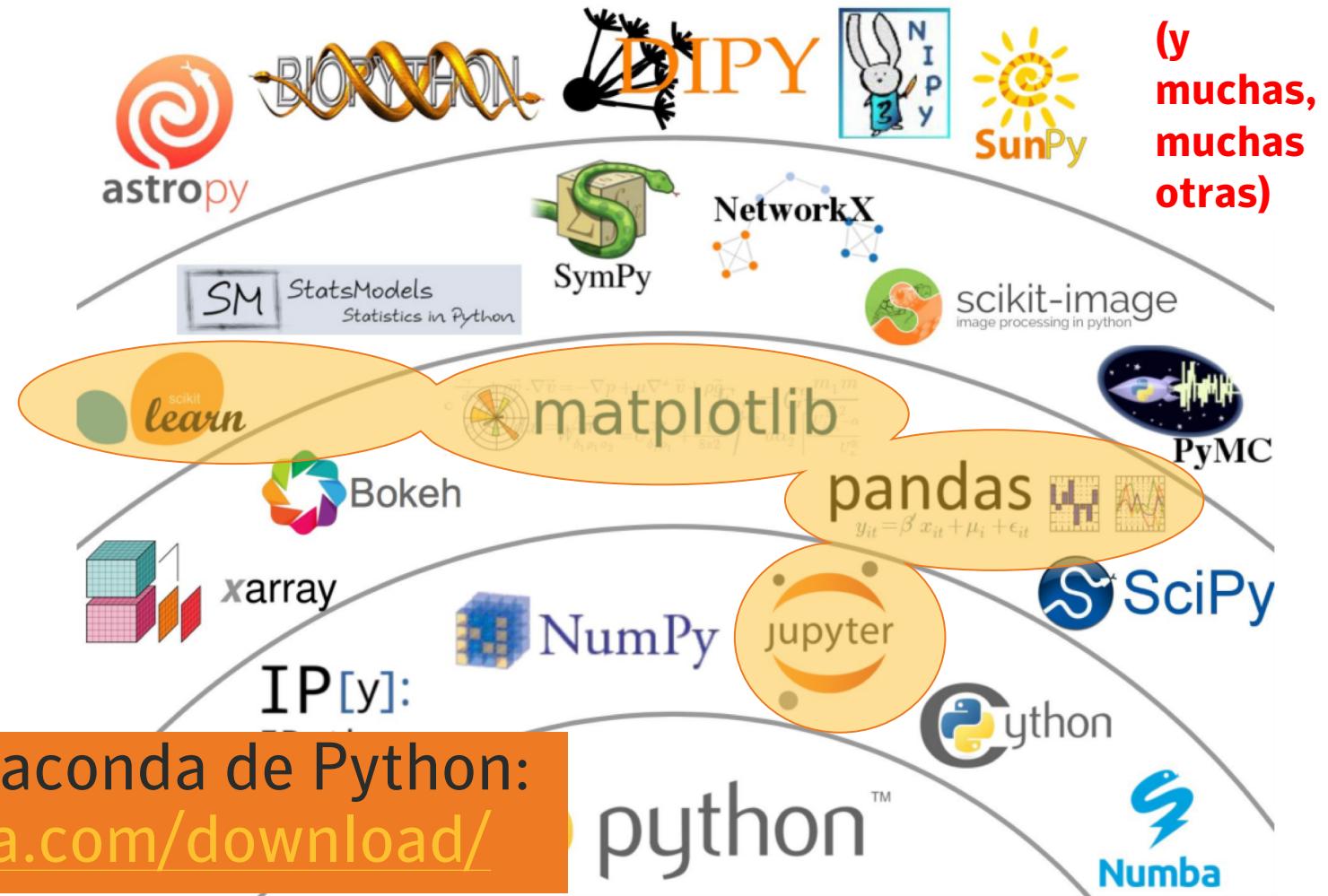
Python



- Lenguaje de programación orientado a objetos
- Funciona para “scripting” y operaciones de misión critica
- Interpretado pero compilado bytecode
- Sintaxis simple pero robusta
- Lenguaje “fácil de leer”
- “Baterías incluidas”: sin número de librerías disponibles
- Versiones: 2.7 (EOL: 1/1/2020) y 3.x

```
print("Hello World!")
```

Ecosistema Científico de Python

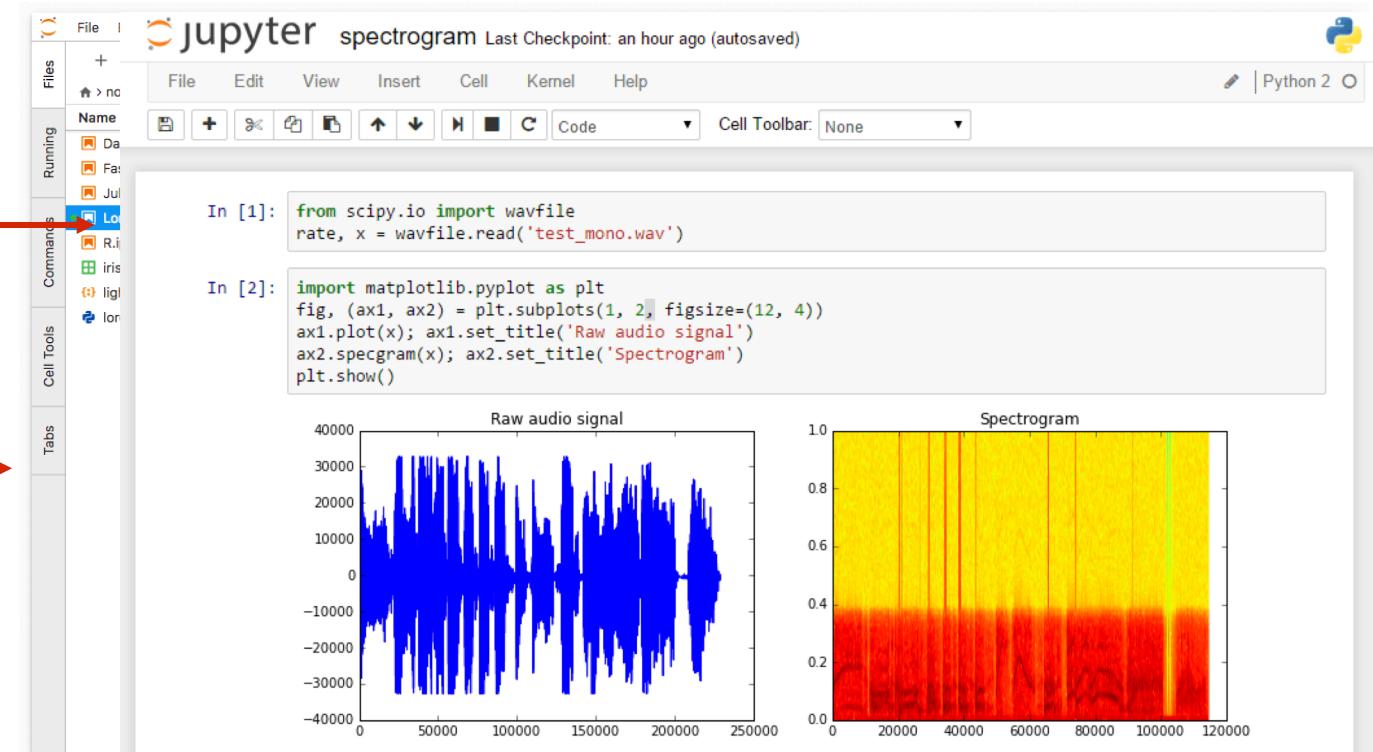


Instalar distribución Anaconda de Python:
<https://www.anaconda.com/download/>

Jupyter Notebooks



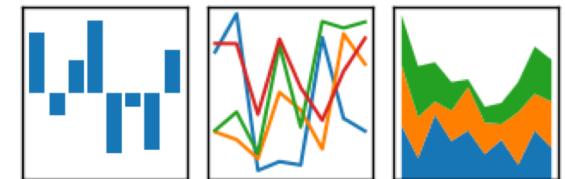
- Herramienta para computación interactiva
 - Conversación con una computadora sobre la data
- Ecosistema Jupyter
 - Ipython
 - Jupyter Notebook
 - Widgets
 - JupyterHub
 - JupyterLab



Pandas

pandas

$$y_{it} = \beta' x_{it} + \mu_i + \epsilon_{it}$$



- Provee estructuras de datos (matriz) y herramientas de manipulación
- Basado en Numpy, otra librería muy útil
- Capacidades similares a hojas de cálculo y bases de datos relacionales
- Robusta funcionalidad de indexado que permite fácil manejo de datos

	<code>_id</code>	<code>_index</code>	<code>_score</code>	<code>_source.category_id</code>	<code>_source.type_id</code>	<code>_source.user_name</code>	<code>_source.version</code>	<code>_type</code>
0	51d7f0a- 53c1-11e8- d3fa- 00000018db99	atp_cmd_results-fdrfulldump- 5a99025495f04e6a95a2c34eb139306b- 2018-05-23	1.0	5	8002	SYSTEM	1.0.0	event
1	51f93440- 53c1-11e8- c6d6- 00000018db9a	atp_cmd_results-fdrfulldump- 5a99025495f04e6a95a2c34eb139306b- 2018-05-23	1.0	5	8007	SYSTEM	1.0.0	event

Sklearn



- Librería de machine learning, incluye versiones eficientes de un gran número de algoritmos comunes
- Posee un API uniforme y claro y muy buena documentación

```
model.fit(X_train, y_train)  
model.predict(X_test)
```



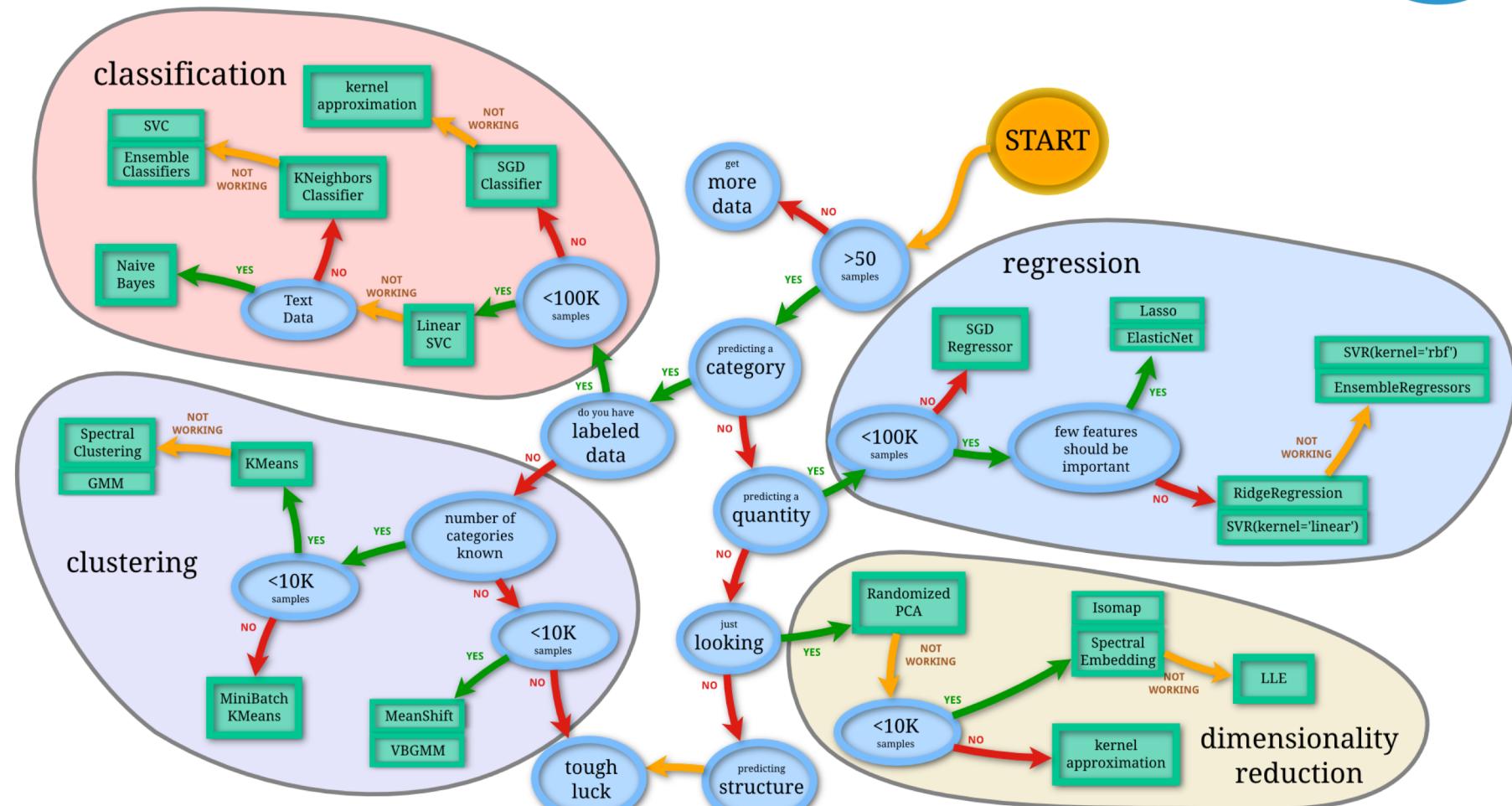
- Produce un nuevo vector de predicciones (y)
- Aplica para algoritmos de clasificación, regresión y clustering

```
model.transform(X_test)
```



- Produce una nueva representación de X
- Aplica para algoritmos de pre-procesamiento, reducción de dimensiones, selección de características

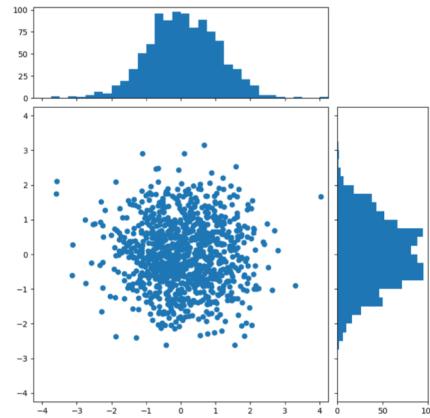
Scikit-learn



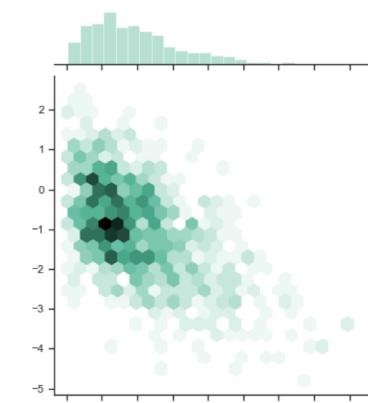
Matplotlib



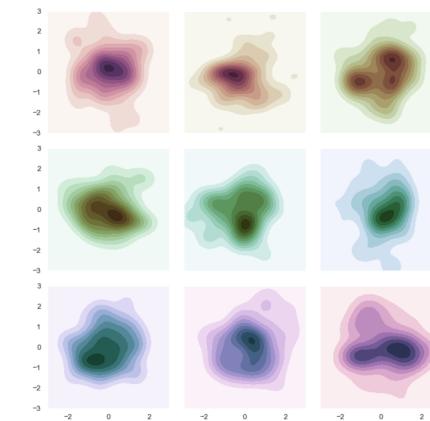
- Librería para la visualización 2D de datos (también 3D)
- Multi-plataforma, trabaja en scripts, Ipython, Jupyter, servidores web
- Posee módulo pyplot para interface tipo MATLAB
- Interface y estilo luce un poco desactualizado → Seaborn y Altair al rescate (otras librerías de visualización)



Matplotlib

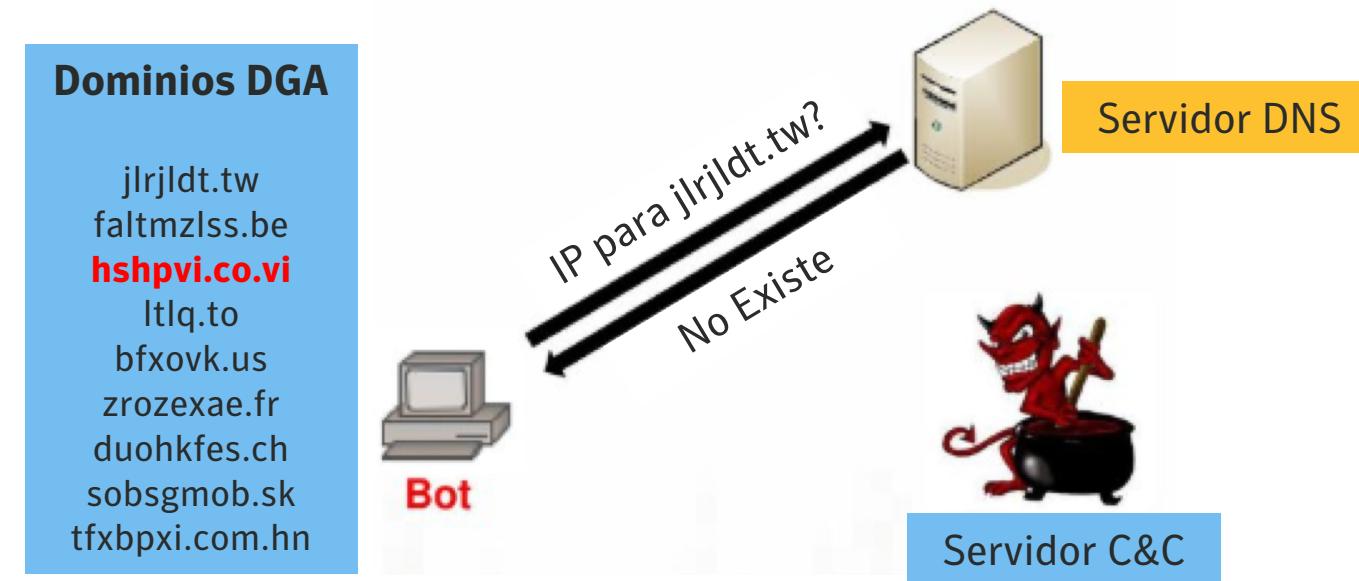


Seaborn



Demo con PyData Stack

- Jupyter Notebook para crear detector de dominios DGA
- Fuente: <https://github.com/gmhoward/dojo-pty-2018/>
 - Primer paso, instalar Python! (<https://www.anaconda.com/download>)



Conclusiones

- Data Science es parte de la caja de herramientas de una estrategia de seguridad completa
 - No es posible analizar manualmente los datos a nuestro alcance
 - Necesario familiarizarte con esta nueva herramienta
- Data Science incluye Machine Learning
- Python, buen lenguaje para aprender programación y data science

Referencias

- Jake Vanderplass, Python Data Science Handbook
- Tutoriales de PyData y PyCon → <https://pyvideo.org>
- 660 cursos gratuitos sobre programación y Python →
<http://bit.ly/2D05UkW> (<https://medium.freecodecamp.org>)
- Jupyter Notebooks sobre seguridad, por Click Security →
https://github.com/ClickSecurity/data_hacking
 - Versión en Español de caso DGA → <https://github.com/gmhoward/dojo-pt-2018/>
- Curso de Deep Learning → <http://www.fast.ai>
- Towards Data Science Blog → <http://towardsdatascience.com>



Muchas gracias!

Gaspar Modelo-Howard

 gaspar@acm.org

 [@gmodeloh](https://twitter.com/gmodeloh)