

PYTHON FUNDAMENTALS FOR DATA SCIENCE

Capítulo 3: Carga de datos en Python





OBJETIVOS

- Procesar diversos tipos de datos.





AGENDA

1. Definición de Data Science.
2. Archivos.
3. SQL.
4. CSV.
5. JSON.
6. MongoDB.
7. Beautiful Soup.





1. DEFINICIÓN DE DATA SCIENCE



La ciencia de datos es un campo interdisciplinario que involucra métodos científicos, procesos y sistemas para extraer conocimiento o un mejor entendimiento de datos en sus diferentes formas, ya sea estructurados o no estructurados.



La ciencia de datos es un campo interdisciplinario que combina Machine Learning, estadísticas, análisis avanzado y programación. Es una nueva forma de arte que revela información oculta y saca el máximo partido de los datos en la era cognitiva.



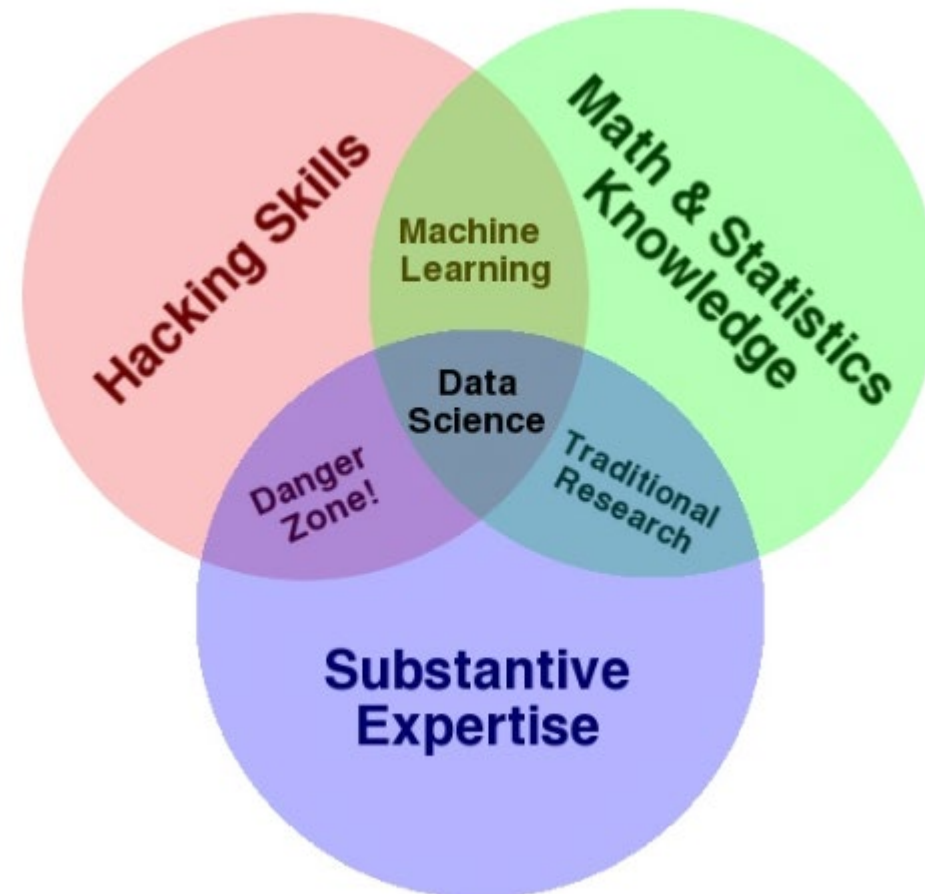
La ciencia de datos involucra métodos automatizados para analizar datos y extraer conocimiento de estos.





1. DEFINICIÓN DE DATA SCIENCE

- El Data Science es la intersección de distintas materias -



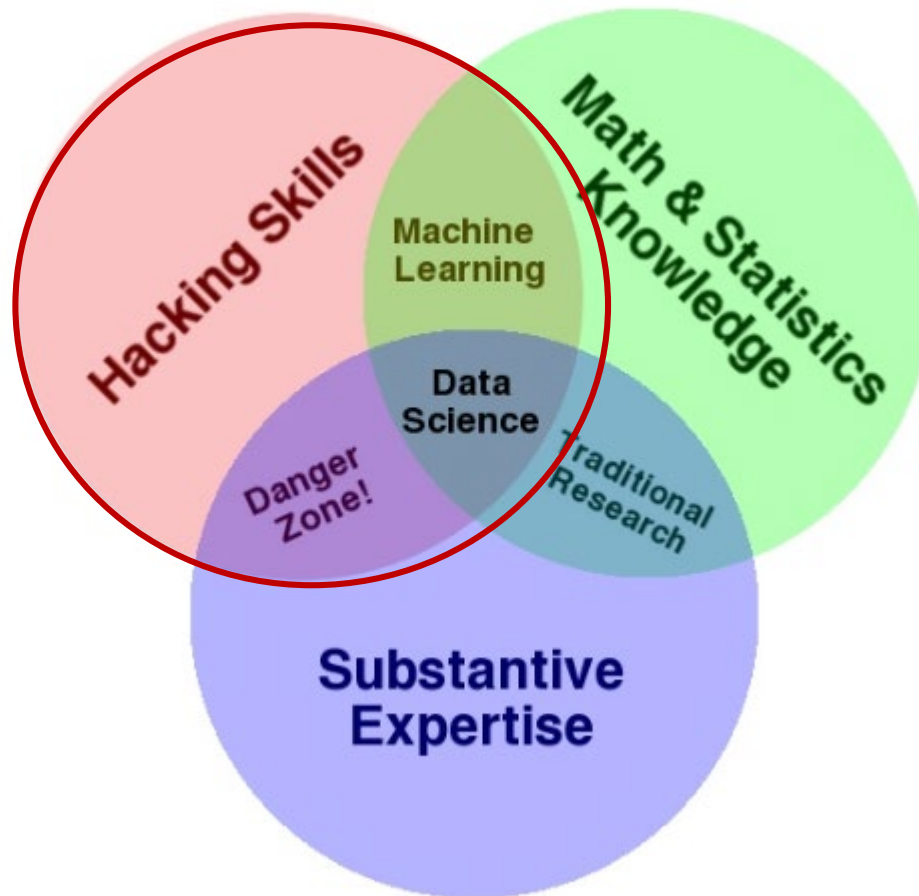
Fuente: drewconway.com





1. DEFINICIÓN DE DATA SCIENCE

- Hacking Skills -



Programación que incluye entre otros lenguajes:

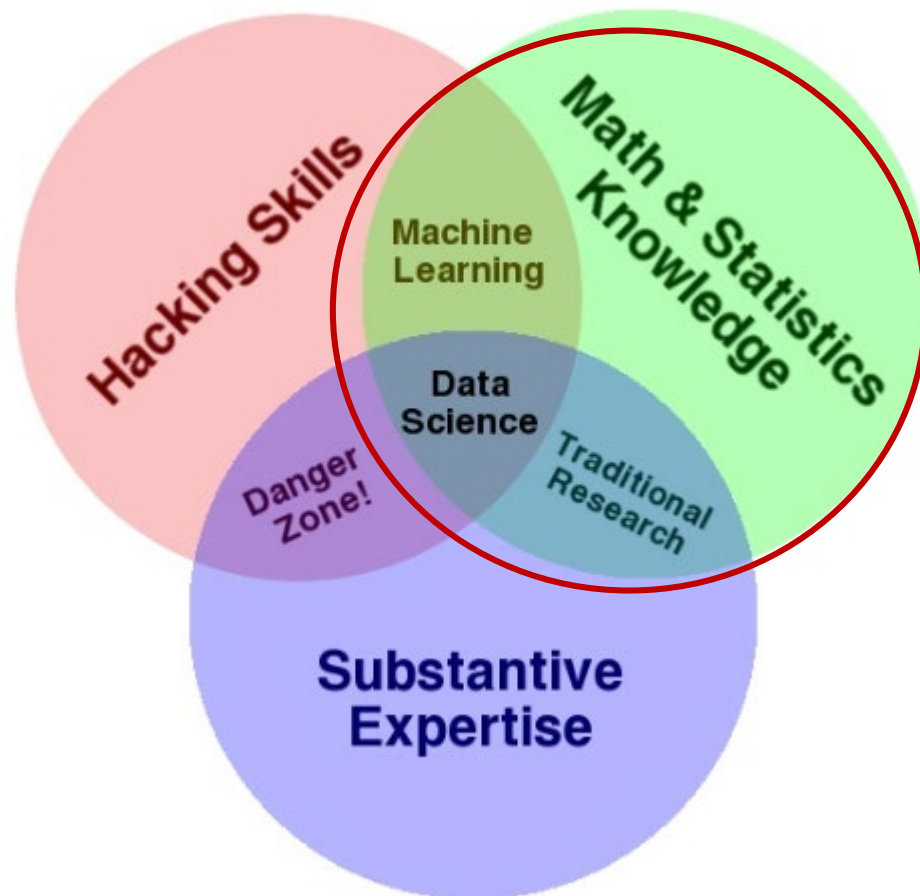
- C
- C++
- Java
- .Net
- Spark
- Scala
- Javascript
- R
- Python





1. DEFINICIÓN DE DATA SCIENCE

- Matemáticas y Estadística -



Matemáticas:

- Álgebra lineal
- Cálculo
- Probabilidades

Estadística:

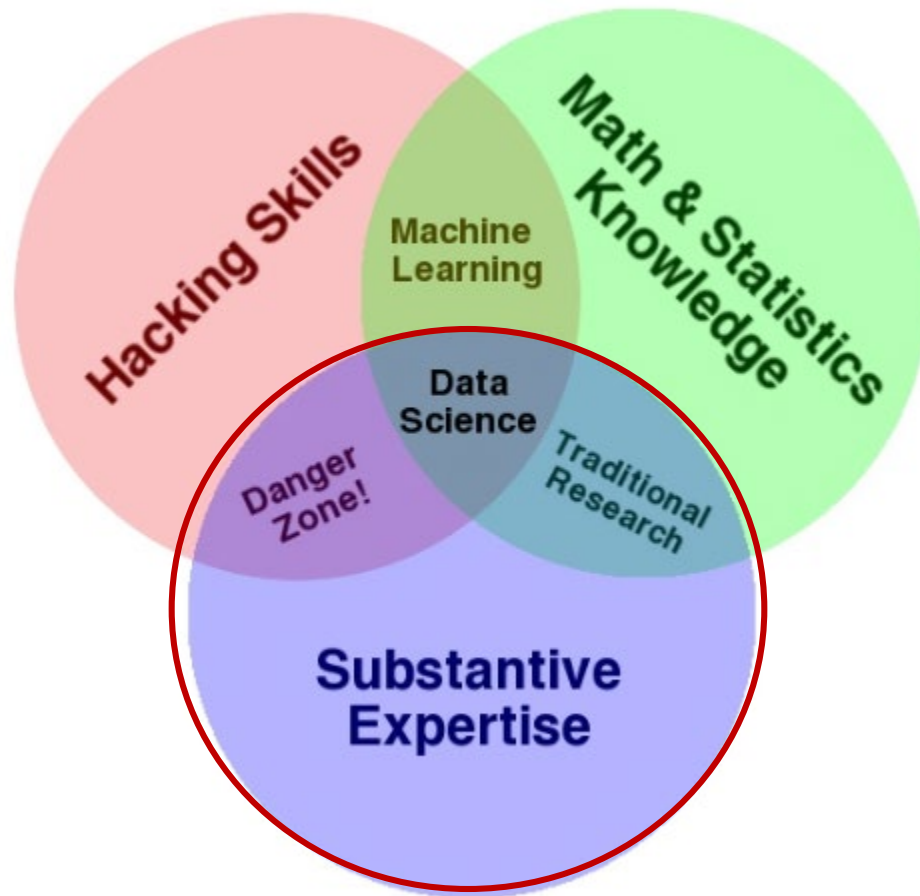
- Descriptiva
- Inferencia
- Distribuciones
- Correlación
- Test de hipótesis
- Etc.





1. DEFINICIÓN DE DATA SCIENCE

- Conocimiento del negocio -



- Consultoría
- Negociación
- Storytelling
- Presentación
- Visualización de datos
- Gestión de proyectos
- Finanzas





2. ARCHIVOS

Características

- Incluido en Python.
- Soporte para archivos de texto y binarios.
- Lectura, escritura y actualización.

```
#iterate opcion 3 - most efficient
with open('dog_breeds.txt', 'r') as reader:
    # Read and print the entire file line by line
    for line in reader:
        print(line, end='')
```

Pug
Jack Russell Terrier
English Springer Spaniel
German Shepherd
Staffordshire Bull Terrier
Cavalier King Charles Spaniel
Golden Retriever
West Highland White Terrier
Boxer
Border Terrier





3. SQL

PostgreSQL



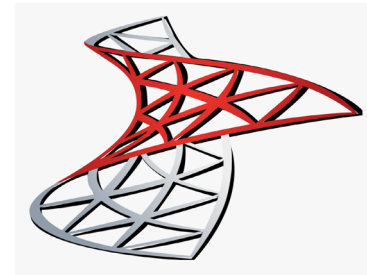
ORACLE®

SQLAlchemy



ORM

SQL





4. CSV

Características

- Un formato muy popular.
- CSV = comma separated values.
- Se puede visualizar en Excel.

```
name,department,birthday month  
John Smith,Accounting,November  
Erica Meyers,IT,March
```

birthday.csv



Excel			
A	B	C	D
name	department	birthday month	
John Smith	Accounting	November	
Erica Meyers	IT	March	

Excel





5. JSON

Características

- Es muy popular.
- Se parece a un diccionario en Python.
- Semiestructurado.

```
[
  {
    "completed": true,
    "userId": 5,
    "id": 81,
    "title": "suscipit qui totam"
  },
  {
    "completed": true,
    "userId": 5,
    "id": 83,
    "title": "quidem at rerum quis ex aut sit quam"
  },
  {
    "completed": true,
    "userId": 5,
    "id": 85,
    "title": "et quia ad iste a"
  },
  {
    "completed": true,
    "userId": 5,
    "id": 86,
    "title": "incidunt ut saepe autem"
  }
]
```





LABORATORIO N° 3: ARCHIVOS Y SQL

Al finalizar el laboratorio, el alumno logrará:

- Manejar archivos de texto.
- Interactuar con bases de datos SQL.
- Manejar archivos CSV.
- Manejar archivos JSON.





TAREA Nº 3: ARCHIVOS Y SQL

- Resolver los ejercicios en el Notebook Jupyter compartido.
- Enviar en **Notebook Jupyter** por correo al instructor.





6. MONGODB

Base de datos NoSQL

Documentos en lugar de tablas

Esquema tipo JSON

Flexible

Estructuras complejas

Escalamiento bueno horizontal



mongo
DB



7. BEAUTIFUL SOUP

- Scraper versátil y simple de usar.
- Convierte HTML -> Datos
- Identifica elementos y extrae los datos seleccionados.



LABORATORIO N° 4: MONGODB Y SCRAPING



Al finalizar el laboratorio, el alumno logrará:

- Manejar MongoDB.
- Realizar Scraping con BeautifulSoup.





TAREA Nº 4: MONGODB Y SCRAPING

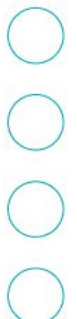
- Resolver los ejercicios en el Notebook Jupyter compartido.
- Enviar en **Notebook Jupyter** por correo al instructor.





RESUMEN

En este capítulo, usted aprendió:

- Que existe una amplia variedad de datos que van desde el procesamiento de archivos de texto, CSV o JSON, hasta bases de datos relacionales y no relacionales como MongoDB. 
- El Scraping es otra herramienta poderosa para la obtención de datos web.
- Python posee diversas herramientas para manejar estos datos.





BIBLIOGRAFÍA

- Python. Python for beginners.
<https://www.python.org/doc/>
- Scikit-learn. Biblioteca de aprendizaje automático.
<https://scikit-learn.org/stable/>
- TensorFlow. Crea modelos de aprendizaje automático.
<https://www.tensorflow.org/?hl=es-419>
- Kaggle. Comunidad de científicos de datos del aprendizaje automático.
<https://www.kaggle.com/>



