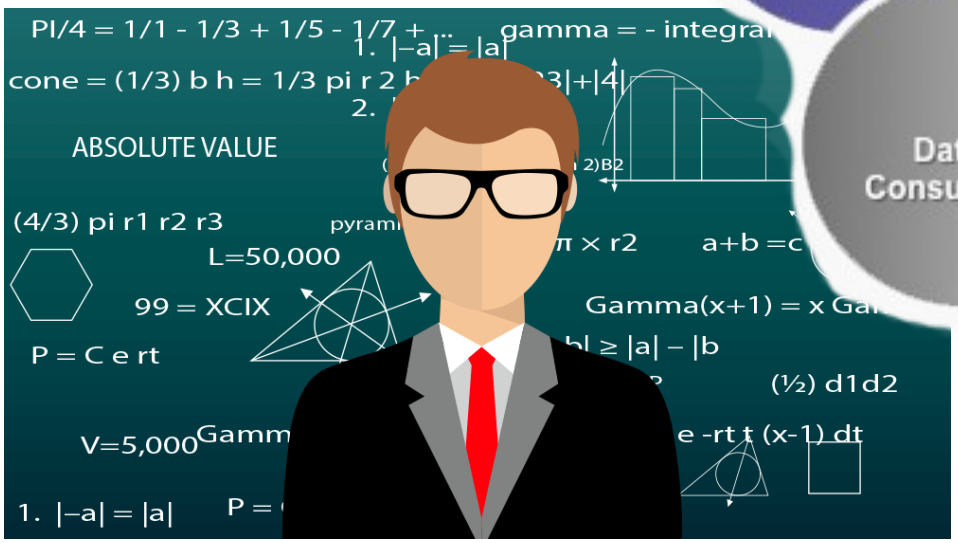

# Data science, pero, ¿por dónde empiezo?

Víctor Valero
Big Data Scientist en StratioDB

# Presentación




# ¿Qué NO es Data Science?

- No es Big Data
- No es estadística
- No es programar con datos
- No es Inteligencia Artificial ni Machine Learning

# ¿Qué es Data Science?

Campo multidisciplinar que se encarga de la obtención, transformación y explotación de los datos.

Parte de los datos para extraer información y convertirlo en conocimiento.

Disciplinas :

- Estadística
- Desarrollo de software
- Machine learning
- Visualización de datos
- Experiencia en el área de negocio

# Roles del mundo de los datos

- Arquitecto de datos: Diseña y construye las plataformas de procesamiento de datos. Integra tecnologías para dar soporte a los requisitos que se deben cubrir.

- Ingeniero de datos: Prepara y cruza los datos, ofrece soluciones para dar eficiencia a sus procesos. Construye los datos que serán consumidos.

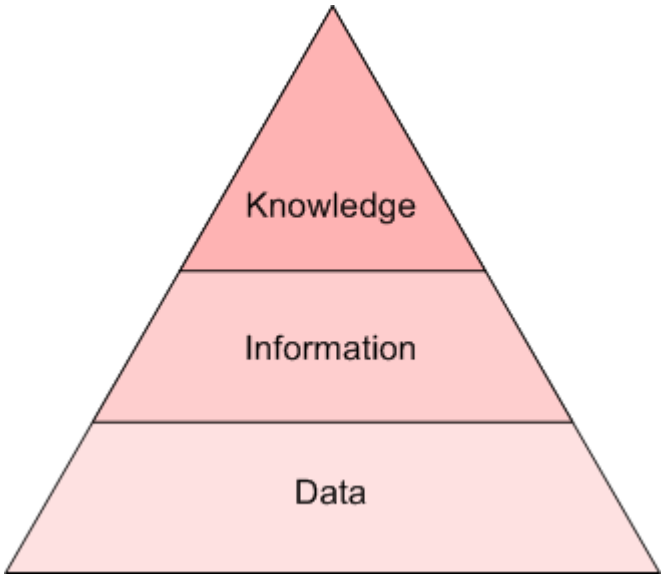
- Analista de datos: Transformación y limpieza de datos, realiza análisis descriptivo y visualizaciones.

- Data Scientist: Evolución del analista de datos, aplica métodos estadísticos en profundidad a sus análisis, crea modelos de ML para aportar un análisis avanzado.

# Pasos en Data Science

Todo proyecto en Data Science debe contener:

- Fase de obtención de datos
- Fase de exploración
- Fase de transformación
- Fase de análisis de datos
- Fase de visualización



# Obtención de datos

- En esta fase debemos recoger los datos o diseñar la toma de muestras de las distintas fuentes.

- Se deben responder preguntas sobre la temporalidad de los datos. ¿Cada cuanto tiempo recojo datos?, ¿Cuantos años de datos necesito?

- Se deben explorar datos complementarios que aporten valor. Ej: Datos meteorológicos

- Se debe valorar si los datos son dinámicos o estáticos. Ej: CSV es estático, API es dinámico.

# Exploración de datos

- En esta fase se centra en establecer los preguntas a las que queremos responder.

- Puede realizarse antes de la obtención pero aquí ya tendremos los datos para ver la viabilidad de los objetivos.

- Investigaremos la necesidad de transformar los datos con los que contamos. ¿Tenemos todo tal y como lo necesitamos?

# Procesado de datos

- Fase en la cual realizamos transformaciones en los datos para adaptarlos a nuestras necesidades.

- Esta parte incluye la limpieza del dato. En general los datos suelen estar bastante “sucios” y es necesario limpiarlos. Ej: Fechas mal formateadas, campos vacíos, outliers.

- Cruces de datos de diferentes orígenes.

# Procesado de datos: Big Data

- Si la volumetría o velocidad que manejamos al procesar datos sobrepasa las capacidades de nuestra máquina quizás sea buena idea emplear Big Data.

- Se basa en repartir la carga de trabajo entre varias máquinas.

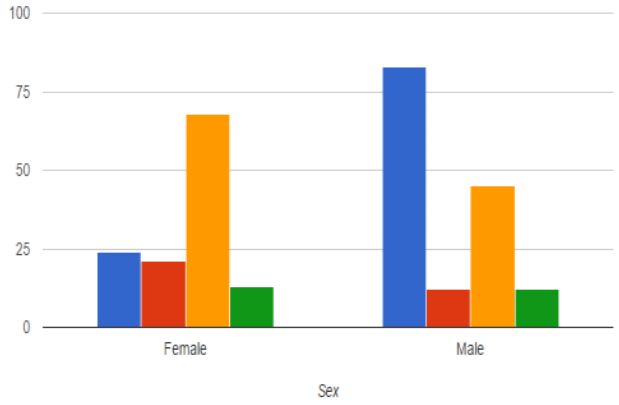
- Valorar si realmente hace falta Big Data en nuestra solución (complica las cosas).

# Análisis de datos

- Fase donde resaltamos la información encontrada en los datos.
- Debe aportar conclusiones sobre lo que estamos investigando.
- Se aplican métodos estadísticos para analizar nuestro objetivo.

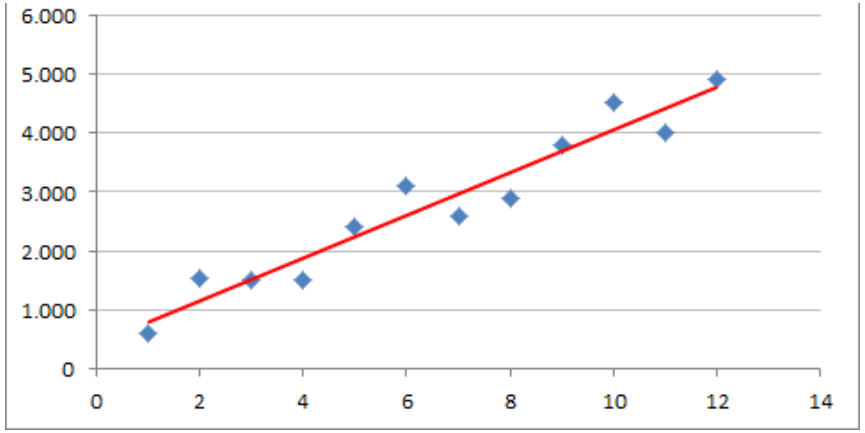
# Visualización

- Fase en la cual se grafican y se evidencian los resultados encontrados.
- Las gráficas deben ser fácilmente entendibles por cualquier usuario.
- Deben contener información útil que hayamos descubierto en la fase de análisis.

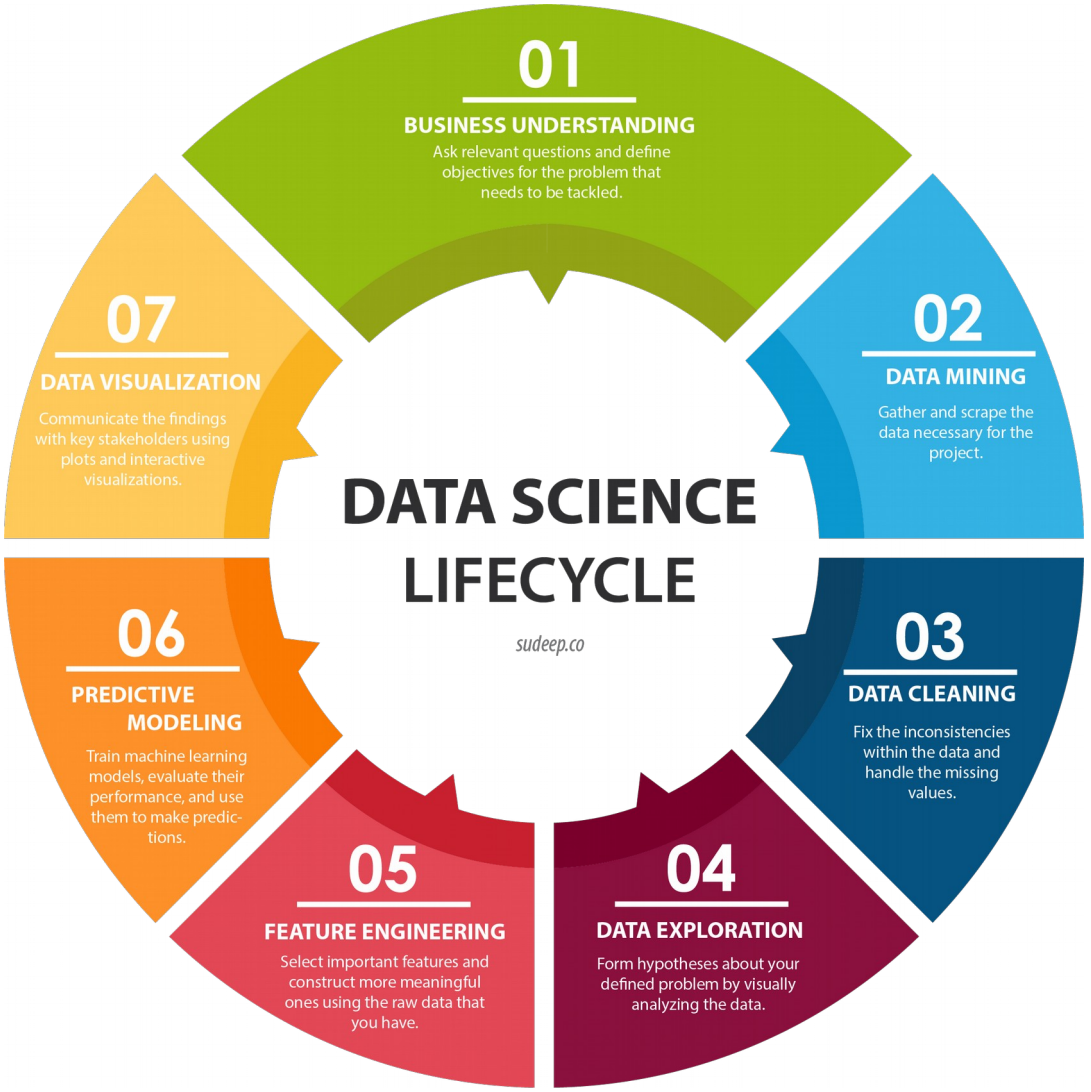


# Analítica avanzada: Machine Learning

- Rama de la Inteligencia Artificial que crea sistemas que aprenden de manera automática.
- En esta fase se modelan los datos mediante diferentes algoritmos.
- Se crean modelos que se intentan adaptar a los datos.
- Cuando se generen datos nuevos el modelo intentará predecir el resultado.



# Fases en Data Science: expectativas



# Fases en Data Science: realidad


# *¿Pero por dónde empiezo?*

# Herramientas de un Data Scientist I: SQL

- Lenguaje de queries
- Usado para manejar datos en BBDD relacionales

Usos en Data Science:

- Consultar datos en BBDD antes de trabajar con ellos
- Crear tablas donde dejar el resultado de nuestro trabajo
- Compartir queries con perfiles No Data Scientist

Qué mínimo debemos cubrir:

- Cláusulas : Select, where, joins, group by

# Herramientas de un Data Scientist I: SQL

```
SELECT
    trabajo,
    count (1) as asistentes
FROM meetup_data_science
WHERE trabajo = “Data Scientist”
GROUP BY  trabajo
```

| trabajo | asistentes |
|---|---|
| Data Scientist | 35 |

# Herramientas de un Data Scientist II: Programación R

- Lenguaje de programación Open Source
- Usado en computación estadística

Usos en Data Science:

- Obtener y transformar datos
- Analizar estadísticamente los datos
- Modelar datos
- Visualizar datos

Qué mínimo debemos cubrir:

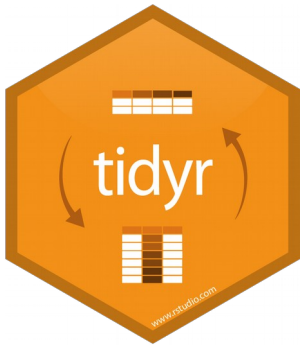
- Programación básica
- Librerías: tidyr, dplyr, ggplot2

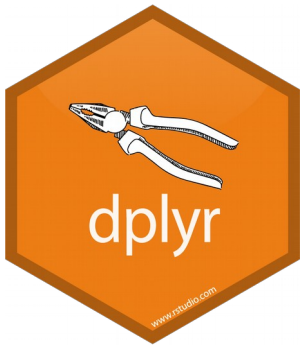
# Herramientas de un Data Scientist II: Programación R

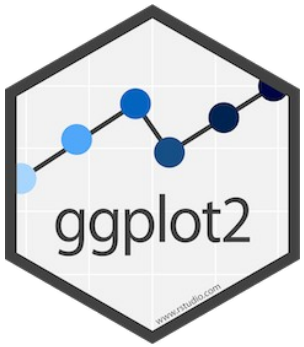
- IDE(Entorno de trabajo en R): R Studio

- Tidyr y Dplyr: librerías altamente eficientes para la transformación y manipulación de datos. Contienen gran cantidad de funciones optimizadas para filtrar, cruzar, limpiar crear y  seleccionar columnas.

- Ggplot2: librería de visualización, contiene gran cantidad de tipos de visualización y parametrizaciones



# Herramientas de un Data Scientist III: Python

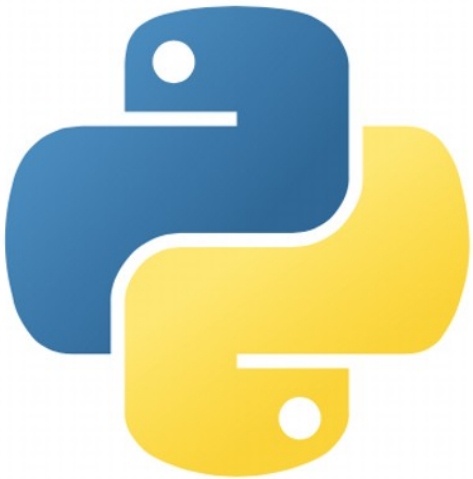
- Lenguaje de programación Open Source
- Usado en Web, procesado datos, scripting ...etc

Usos en Data Science:

- Obtener y transformar datos
- Analizar estadísticamente los datos
- Modelar datos
- Visualizar datos

Qué mínimo debemos cubrir:

- Programación básica
- Librerías: pandas, matplotlib, scikit



# Herramientas de un Data Scientist III: Python

- Pandas: librería para el procesamiento y análisis de datos.
- Matplotlib: librería de visualización con muchas visualizaciones y gráficos.
- Scikit Learn: librería que contiene gran cantidad de modelos y funcionalidades para implementar Machine Learning

matplotlib

scikit learn

<devAcademy>

# Herramientas de un Data Scientist IV: Notebooks

- Documentos que mezclan código, gráficos y texto para presentar conclusiones basadas en datos

Usos en Data Science:

- Analizar/modelar datos usando un lenguaje de programación
- Compartir los resultados con otros
- Preparar informes con gráficas

Qué mínimo debemos cubrir:

- Jupyter Notebooks (python)
- R Markdown (R)

# Herramientas de un Data Scientist IV: Notebooks

- R markdown: Documentos reproducibles e interactivos con código R, gráficas y texto.  Incluido en R studio

- Jupyter Notebooks: Aplicación web que permite crear notebooks reproducibles e interactivos con código python, gráficas y texto.


# Herramientas de un Data Scientist V: BBDD

- Conjunto de información relacionada
- Puede estar estructurada ( relacional o SQL) o no estructurada ( no relacional o noSQL)

Usos en Data Science:

- Obtener los datos de la BBDD
- Escribir los resultados en la BBDD

Qué mínimo debemos cubrir:

- Leer tablas mediante SQL
- Acceder a los datos de manera programática
- BBDD relacional (MySQL) y BBDD no relacional (MongoDB)

# Herramientas de un Data Scientist VI: Spark

- Framework de Big Data.
- Se basa en dividir los datos en diferentes servidores para repartir el trabajo.
- Se puede utilizar con Python y R.

Usos en Data Science:

- Transformar grandes volúmenes de datos
- Aplicar modelos de ML de manera distribuida

Qué mínimo debemos cubrir:

- PySpark SQL (python)
- SparklyR o SparkR (R)

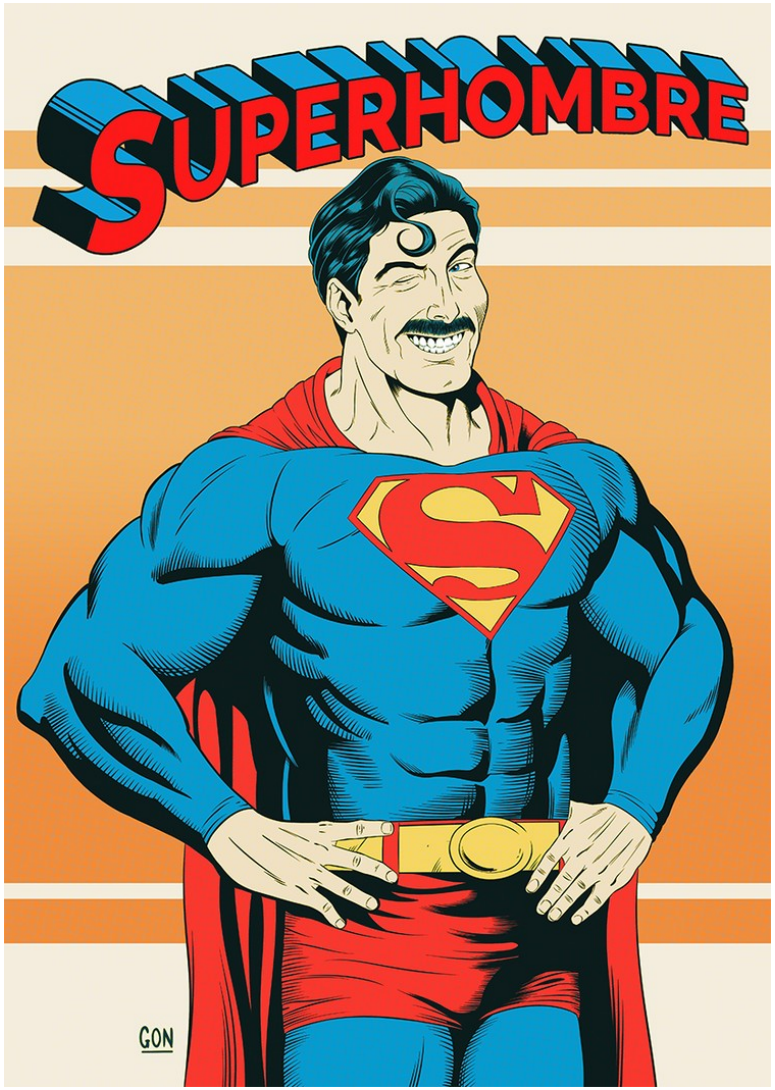
# Herramientas de un Data Scientist VI: Spark

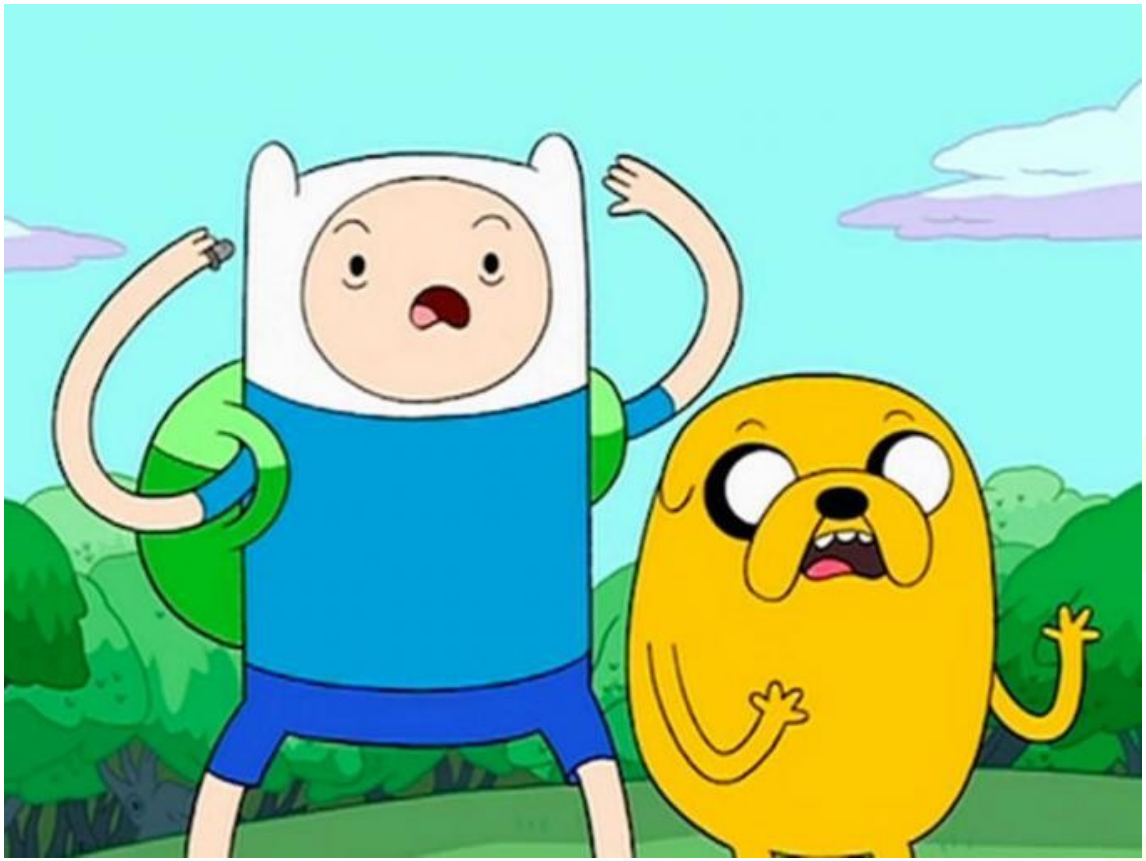
- SparklyR: librería de R para emplear Spark en los cálculos. Soporta sintaxis de dplyr.

- PySpark SQL: módulo de Spark para trabajar con datos estructurados en Big Data programando con python.


# Pasos para llegar a Data Scientist

- Ser proactivo
- Formación continua
- Trabajar en equipo
- Ser comunicativo
- Tener curiosidad





# *Demo time !!*


# *Preguntas y respuestas*


# *Gracias*