TELEDUC
Educación a Distancia

CIENCIA DE DATOS

FELIPE LÓPEZ ROJAS

>>> Profesor del curso

Introducción

- Big Data y Ciencia de Datos.
- Archivos CSV.

>>> Motivación: contextualizando el curso

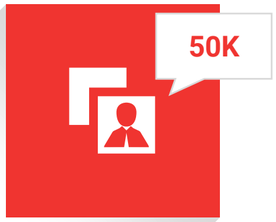
Reflexionemos sobre ¿cuántos datos se generan en internet?

Según estudio de DOMO Empresa mundial líder en tecnología digital y servicios web, **solo en 1 minuto**:

4 millones de búsquedas en Google.

Se ven apróx. **4,5 millones** de videos en Youtube.

Se mandan apróx. **12 millones** de SMS.



Suben casi **50 mil** fotos a Instagram.

Reflexionemos sobre ¿cuántos datos se generan en internet?

¿Por qué es relevante?

La cantidad de datos generados es enorme y aumenta.

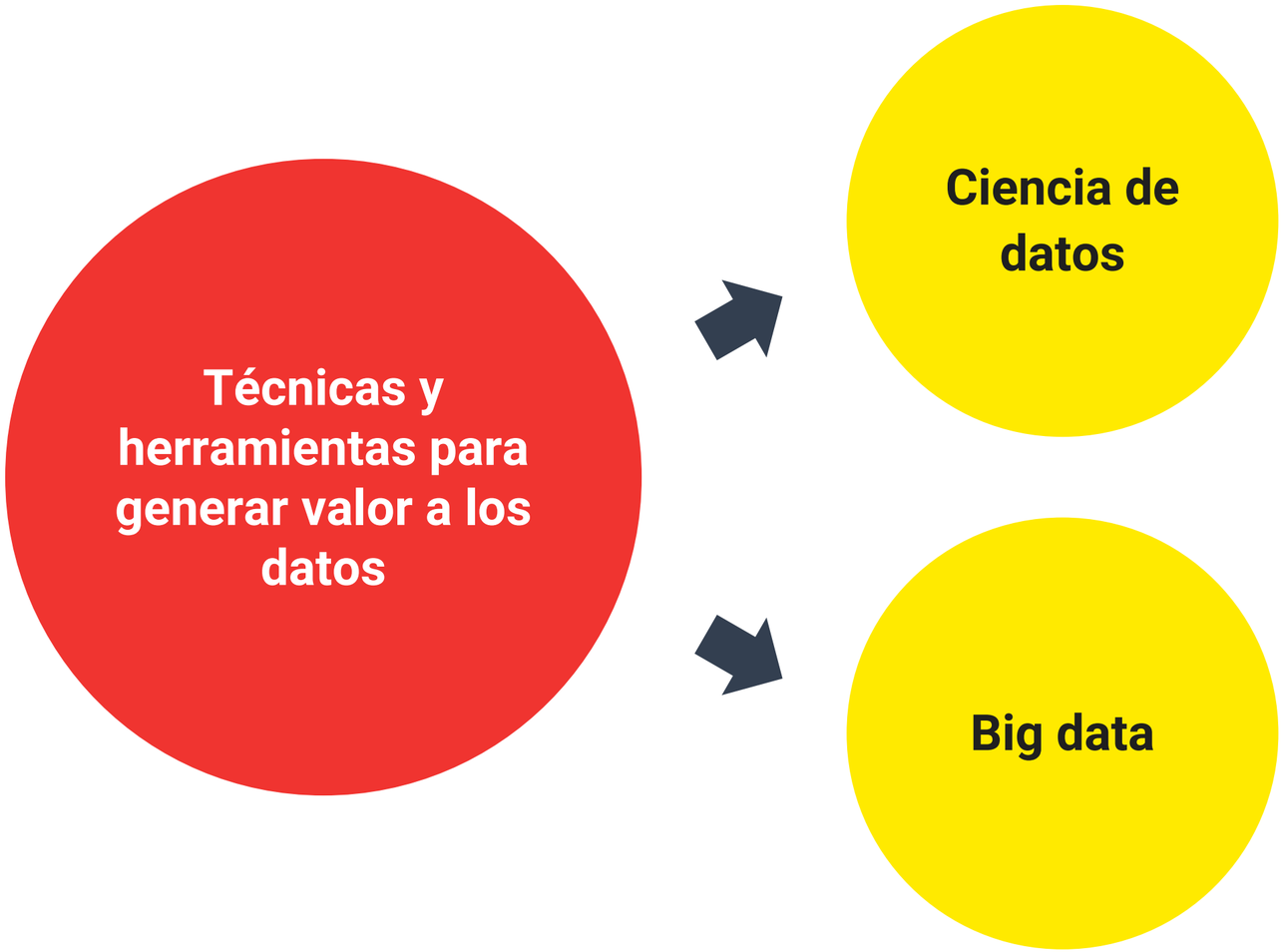
¿Y en las empresas?

El desafío es darle valor a los datos disponibles.

¿Cómo dar valor a los datos de una empresa?

>>> Big data y Ciencia de Datos

Big data y Ciencia de Datos



Big data


Colección masiva de datos
Diversos formatos y fuentes
BIG DATA
Datos no estructurados
Datos Semi estructurados
Datos estructurados

Algunas definiciones sobre Ciencia de Datos

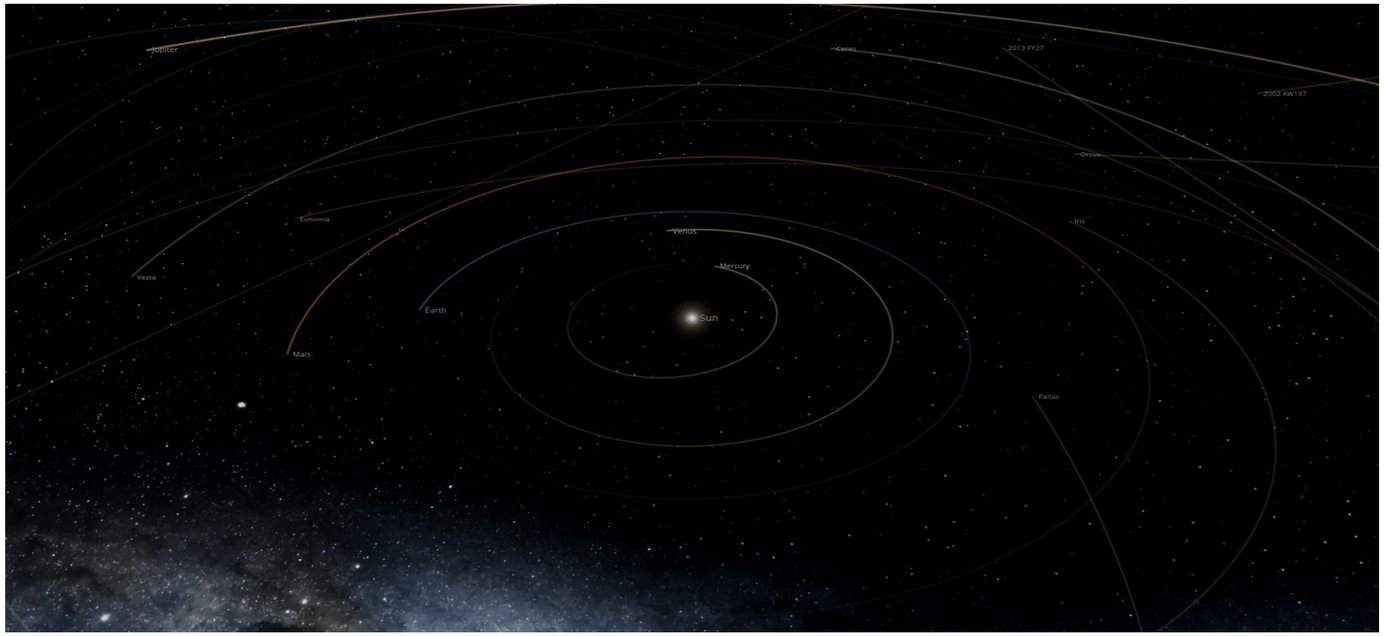
> **Es la disciplina que permite encontrar patrones predecibles en sets de datos estructurados y no estructurados.**
>
> *Vasant Dhar, 2013*

Algunas definiciones sobre Ciencia de Datos

> **Es un concepto que busca unificar estadísticas, análisis de datos, inteligencia artificial y cualquier otro método similar para poder entender y analizar fenómenos con los datos.**
>
> *Chikio Hayashi, 1996.*

Ciencia de Datos

Puede complementarse con otras ciencias que requieran un análisis de patrones en grandes cantidades de datos. Por ejemplo la **Astronomía**.



Ciencia de Datos

Requiere de conocimientos avanzados → **Estadística, computación, matemáticas, etc.**

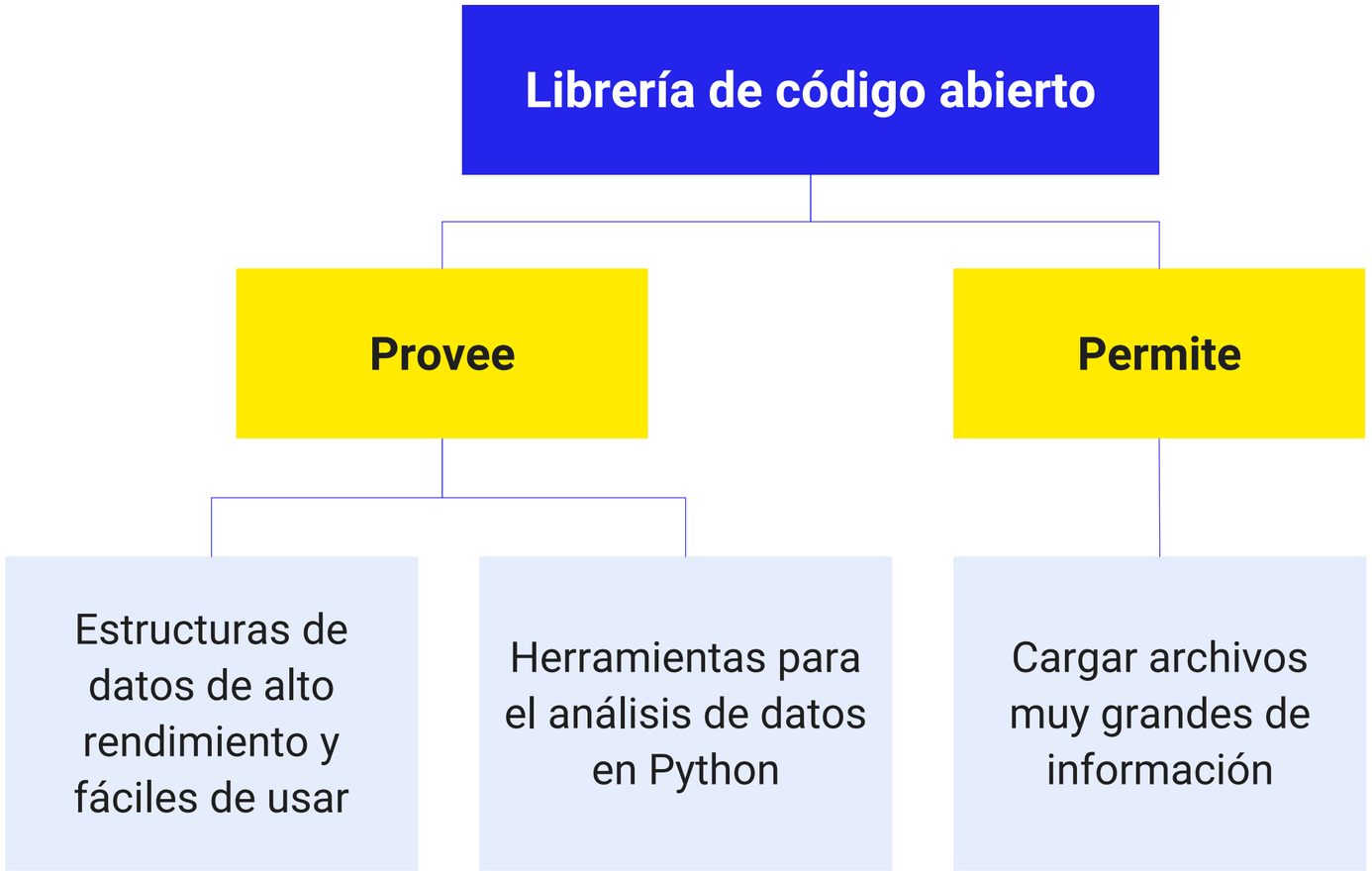
Lo importante son las herramientas

Lo científicos de datos lo usan para analizar e identificar patrones.

En el caso del procesamiento de grandes volúmenes de datos en Python usan la librería llamada *Pandas*.

>>> Librería *Pandas*

¿Qué entendemos por librería *Pandas*?



¿Qué archivos ocuparemos en *Pandas*?

Texto plano y replican una matriz.

Archivo CSV

Por ejemplo tenemos una base de datos con nombre “**Funcionarios**” con los datos de los trabajadores de una empresa.

Nombre
Edad
Fecha de nacimiento
RUT

Ejemplo de base de datos

Nombre	Edad	Fecha de Nacimiento	RUT
Juan Pérez	27	31-01-1991	17.587.451-8
María Rojas	54	04-05-1964	9.475.362-4
Pedro Rodríguez	35	18-06-1983	13.748.645-2
Soledad Ríos	21	03-03-1997	20.471.472-1

¿Cómo se vería la base de datos en formato CSV?

El CSV es una representación de una tabla (base de datos)

CSV

```
Nombre;Edad;Fecha de Nacimiento;RUT
Juan Pérez;27;31-01-1991;17.587.451-8
María Rojas;54;04-05-1964;9.475.362-4
Pedro Rodríguez;35;18-06-1983;13.748.645-2
Soledad Ríos;21;03-03-1997;20.471.472-1
```

¿Cómo se vería la base de datos en formato CSV?

El CSV es una representación de una tabla (base de datos)

CSV

```
Nombre;Edad;Fecha de Nacimiento;RUT
Juan Pérez;27;31-01-1991;17.587.451-8
María Rojas;54;04-05-1964;9.475.362-4
Pedro Rodríguez;35;18-06-1983;13.748.645-2
Soledad Ríos;21;03-03-1997;20.471.472-1
```

La primera línea del texto corresponde a la primera fila de la tabla, y en general representa a los encabezados o bien los nombres de cada columna.

IMPORTANTE: No todos los archivos CSV tienen esta línea.

¿Cómo se vería la base de datos en formato CSV?

El CSV es una representación de una tabla (base de datos)

CSV

```
Nombre;Edad;Fecha de Nacimiento;RUT
Juan Pérez;27;31-01-1991;17.587.451-8
María Rojas;54;04-05-1964;9.475.362-4
Pedro Rodríguez;35;18-06-1983;13.748.645-2
Soledad Ríos;21;03-03-1997;20.471.472-1
```

Cada línea de un archivo CSV representa a una fila de la tabla.

¿Cómo se vería la base de datos en formato CSV?

El CSV es una representación de una tabla (base de datos)

CSV

```
Nombre;Edad;Fecha de Nacimiento;RUT
Juan Pérez;27;31-01-1991;17.587.451-8
María Rojas;54;04-05-1964;9.475.362-4
Pedro Rodríguez;35;18-06-1983;13.748.645-2
Soledad Ríos;21;03-03-1997;20.471.472-1
```

Dentro de cada línea del archivo, se separan las columnas con el carácter “;” o “,”.

Conclusiones

- Big Data y Ciencia de Datos son herramientas que permiten extraer información de grandes volúmenes de datos.
- La librería Pandas de Python es una de las más usadas para procesar grandes volúmenes de datos.
- El formato de archivos CSV es ampliamente usado como una forma de representar y almacenar a un bajo costo datos como una tabla.

Bibliografía

- Dhar, V. (2013). Data science and prediction. *Communications of the ACM*, 56(12), 64-73.
- Hayashi C. (1998) What is Data Science ? Fundamental Concepts and a Heuristic Example. In: Hayashi C., Yajima K., Bock HH.
- Ohsumi N., Tanaka Y., Baba Y. (eds) Data Science, Classification, and Related Methods. Studies in Classification, Data Analysis, and Knowledge Organization. Springer, Tokyo.

>>> Cierre

Has finalizado la revisión de los contenidos que corresponden a esta clase.

A continuación, te invitamos a estudiar la siguiente clase del módulo.