



Introducción y Motivación

Gestión de Datos

Maximiliano Arancibia

Educación Profesional - Escuela de Ingeniería

Tabla de Contenidos

Introducción

Proyectos de ciencia de datos

Ciclo de vida de los datos

Datos

Gestión

¿Cómo trabajar con datos?

Bases de datos y DBMS

¿Por qué DBMS?



Tabla de Contenidos

Introducción

Proyectos de ciencia de datos

Datos

¿Cómo trabajar con datos?



Gestión de datos - Maximiliano Arancibia

- Ingeniero Civil Industrial, Mención Matemática UC.
- Magíster en Ciencias de la Ingeniería, Mención Eléctrica/Matemática UC.
- Mi mail : mgarancibia@uc.cl



Área de investigación: Deep Reinforcement Learning, Optimización, Problemas de decisión estocástico, Análisis Estocástico, Políticas Públicas.



Gestión de datos - Matías Toro

- Ingeniero Civil en Computación, Universidad de Chile
- Doctor en Ciencias mención Computación, Universidad de Chile
- Mi mail : matias.toro@uc.cl



Área de investigación: Lenguajes de programación, Verificación de programas, Privacidad en bases de datos, Privacidad diferencial



Información general:

- Clases: Viernes, 18:30-21:30
- Talleres:
 - Francisco Alessandri
 - Sabados (algunos): 9:30 - 13:30
 - fdalessandri@uc.cl



Información general:

- Clases: Viernes, 18:30-21:30
- Talleres:
 - Francisco Alessandri
 - Sabados (algunos): 9:30 - 13:30
 - fdalessandri@uc.cl

Sitio oficial del curso:

Moodle, se publicarán los anuncios y el material del curso.



Información general:

- Clases: Viernes, 18:30-21:30
- Talleres:
 - Francisco Alessandri
 - Sabados (algunos): 9:30 - 13:30
 - fdalessandri@uc.cl

Sitio oficial del curso:

Moodle, se publicarán los anuncios y el material del curso.

Atención alumnos:

- Vía mail
- Después de clases



Estructura de las evaluaciones:

- Dos tareas individuales (Nota (NT_i))
- Tarea grupal (Nota (NG))

Las tareas se darán a las 12:00 de los días de Taller 2 y 3. Se leerá en conjunto y luego podrán desarrollarla como una actividad en clase.

Actividades. Se realizaran controles, actividades individuales y grupales para saber si se entienden los contenidos. Sin nota.

Software: En esta clase y taller programaremos en R, para hacerlo se usará R Colab (www.colab.tor/r). Además a futuro veremos el lenguaje SQL enlazado con R.

Condiciones de aprobación:

- Asistencia $\geq 75\%$
- Nota final: $NF = \frac{NT_1 + NT_2 + NG}{3} \geq 4.0$



Bibliografía Básica

- Raghu Ramakrishnan, Johannes Gehrke. *Database Management Systems* (2000). McGraw-Hill Companies.
- Nicholas J. Horton, Ken Kleinman. *Using R and RStudio for Data Management, Statistical Analysis and Graphics* (2015). Chapman and Hall.
- Bradley C. Boehmke. *Data Wrangling with R* (2016). Springer.
- Hadley Wickham, Garrett Grolemund. *R for Data Science: Import, Tidy, Transform, Visualize, and Model Data* (2017). O'Reilly Media.
- Victor Eijkhout, Edmond Chow, Robert van de Geijn. *Introduction to High Performance Scientific Computing* (Revision 2015).



Tabla de Contenidos

Introducción

Proyectos de ciencia de datos

Ciclo de vida de los datos

Datos

¿Cómo trabajar con datos?

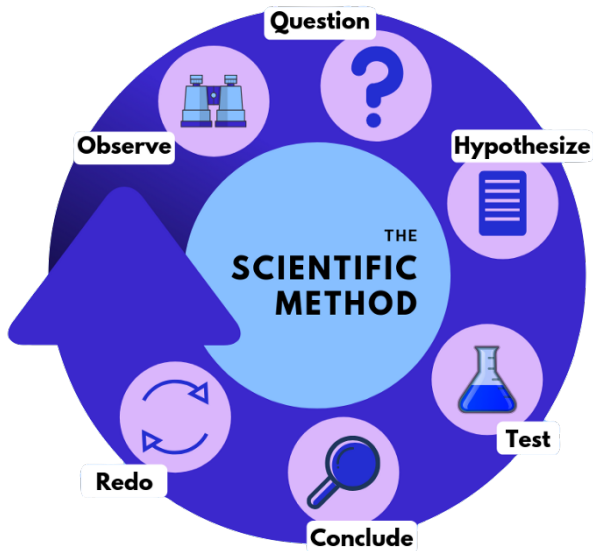


Un proyecto de ciencia de datos siempre partirá con una pregunta o un objetivo:

- ¿Se relaciona la asistencia a clases con el promedio de un alumno?
- ¿Podemos relacionar el flujo de datos en Twitter con la propagación de daños en los desastres naturales?
- Predecir velocidad de propagación de COVID en base a datos de geolocalización.
- Etc...



Ciclo de vida de datos en investigación



Ciclo de vida de datos en investigación

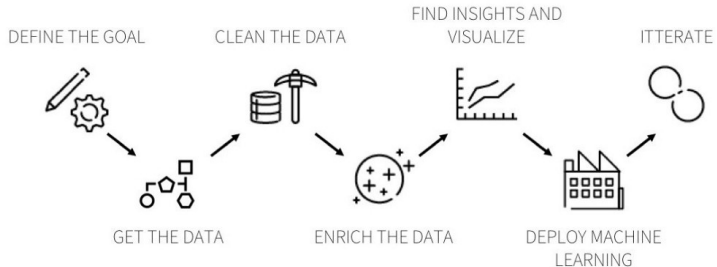


Tabla de Contenidos

Introducción

Proyectos de ciencia de datos

Datos

Gestión

¿Cómo trabajar con datos?



Primero que todo...



Primero que todo...

¿Qué son los datos?



Primero que todo...

¿Qué son los datos?

- Información
- Colección de hechos
- Representación de algo que captura algunas características e ignora otras.



Primero que todo...

¿Qué son los datos?

- Información
- Colección de hechos
- Representación de algo que captura algunas características e ignora otras.

⇒ Todo en la vida puede ser datos



Introducción

Datos pueden ser clasificados dependiendo de una variedad de criterios:

- Contenido
Numérico, texto, multimedia...
- Formato:
Hojas de calculo, bases de datos, imágenes, texto...
- Método de recolección:
Experimentación, observación, simulación, derivado...
- Naturaleza:
Analógico o digital.
- Fuente:
Primaria (i.e. creada con un objetivo) o secundaria (reutilizada)
- Procesamiento:
Datos crudos o procesados



La gestión de los datos tiene principalmente el objetivo de hacer que el proceso de extracción y almacenamiento de datos sea:

- Seguro
- Sostenible en el tiempo
- Fáciles de encontrar
- Fácil de entender
- Reutilizable



Ciclo de vida de datos en investigación



Ciclo de vida de datos en investigación



Ciclo de vida de datos en investigación



Ciclo de vida de datos en investigación



Ciclo de vida de datos en investigación



Ciclo de vida de datos en investigación



Generación:

Cualquier cambio
puede generar nueva
información, incluso
el cambios en el
observador.



Generación:

Cualquier cambio puede generar nueva información, incluso el cambios en el observador.

Recolectar: Adquirir la información de una o mas fuentes.
Sensores, encuestas, scrapping, etc.



Datos

Generación:

Cualquier cambio puede generar nueva información, incluso el cambios en el observador.

Recolectar: Adquirir la información de una o mas fuentes.

Sensores, encuestas, scrapping, etc.

Almacenamiento:

usualmente en discos duros o en la nube pero puede ser incluso una hoja de papel o la memoria.



Datos

Generación:

Cualquier cambio puede generar nueva información, incluso el cambios en el observador.

Recolectar: Adquirir la información de una o mas fuentes. Sensores, encuestas, scrapping, etc.

Almacenamiento: usualmente en discos duros o en la nube pero puede ser incluso una hoja de papel o la memoria.

Visualización: Formas de comunicar la información recolectada. Puede ser a través de gráficos, tablas, imágenes, etc.



Generación:

Cualquier cambio puede generar nueva información, incluso el cambios en el observador.

Visualización: Formas de comunicar la información recolectada. Puede ser a través de gráficos, tablas, imágenes, etc.

Recolectar: Adquirir la información de una o mas fuentes. Sensores, encuestas, scrapping, etc.

Análisis: Extraer información útil y condensada para convertirla en inteligencia.

Almacenamiento: usualmente en discos duros o en la nube pero puede ser incluso una hoja de papel o la memoria.



Generación:

Cualquier cambio puede generar nueva información, incluso el cambios en el observador.

Recolectar: Adquirir la información de una o mas fuentes.

Sensores, encuestas, scrapping, etc.

Almacenamiento:

usualmente en discos duros o en la nube pero puede ser incluso una hoja de papel o la memoria.

Visualización: Formas de comunicar la información recolectada. Puede ser a través de gráficos, tablas, imágenes, etc.

Análisis: Extraer información útil y condensada para convertirla en inteligencia.

Acción: usar la inteligencia para proponer cambios y optimizar procesos. Esto también involucra generación y recolección de nuevos datos.



Generación:

Cualquier cambio puede generar nueva información, incluso el cambios en el observador.

Recolectar: Adquirir la información de una o mas fuentes.

Sensores, encuestas, scrapping, etc.

Almacenamiento:

usualmente en discos duros o en la nube pero puede ser incluso una hoja de papel o la memoria.

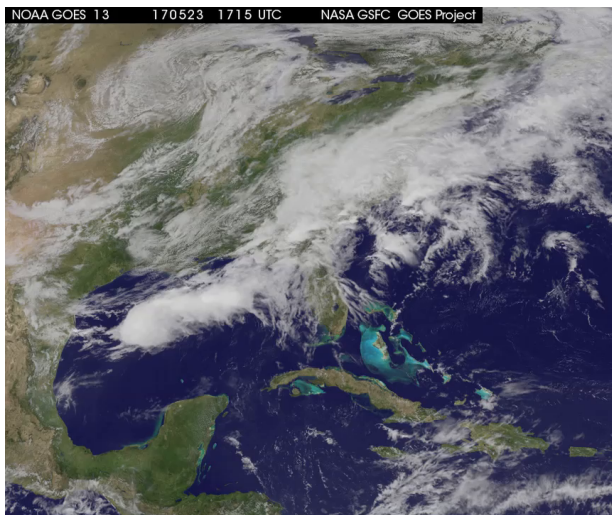
Visualización: Formas de comunicar la información recolectada. Puede ser a través de gráficos, tablas, imágenes, etc.

Análisis: Extraer información útil y condensada para convertirla en inteligencia.

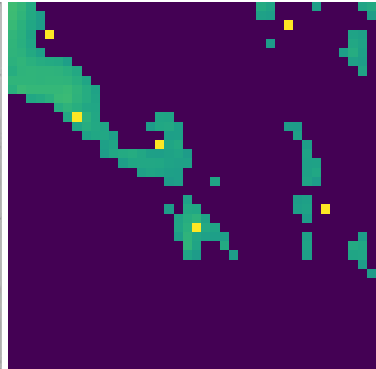
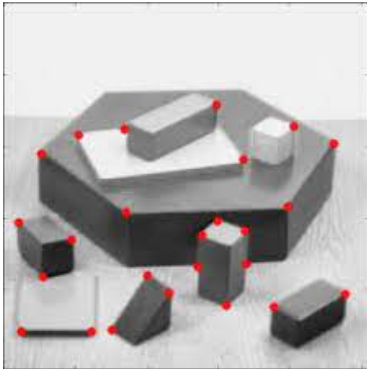
Acción: usar la inteligencia para proponer cambios y optimizar procesos. Esto también involucra generación y recolección de nuevos datos.



Generación: NASA



Recolección: Método de *thresholding* + Shi-Tomasi, detector de esquinas.

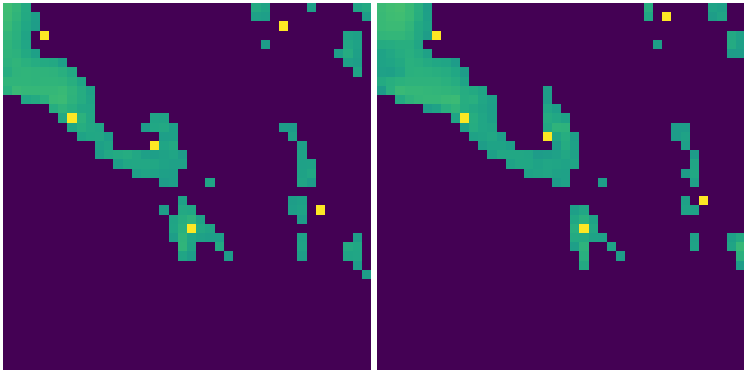


Almacenamiento: base de datos de características. ¿Está bien?

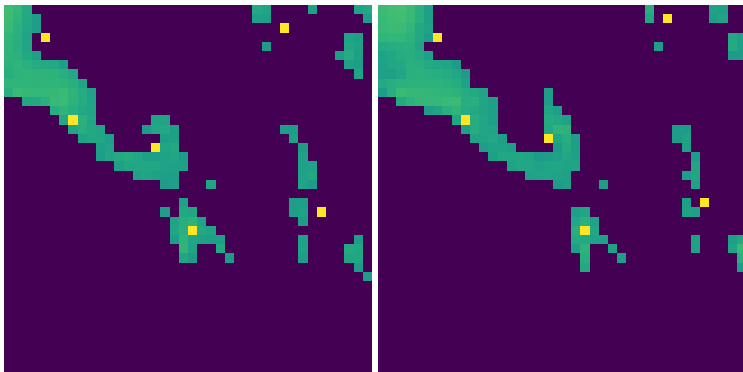
| | x | y |
|---|-------|------|
| 1 | 1.26 | 1.37 |
| 2 | -0.78 | 4.00 |
| 3 | 4.01 | 0.24 |
| 4 | 2.58 | 0.77 |
| 5 | 3.91 | 4.17 |
| 6 | -0.10 | 4.43 |



Visualización: ¿Como se comportan dos instantes de tiempo distinto?



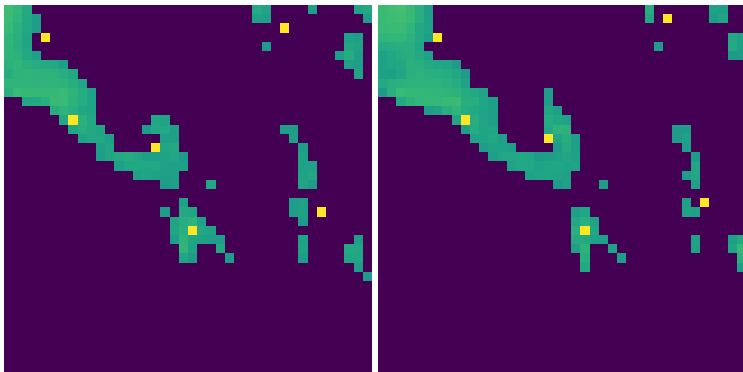
Visualización: ¿Como se comportan dos instantes de tiempo distinto?



Se mueven para distintos lados unas de otras.



Visualización: ¿Como se comportan dos instantes de tiempo distinto?

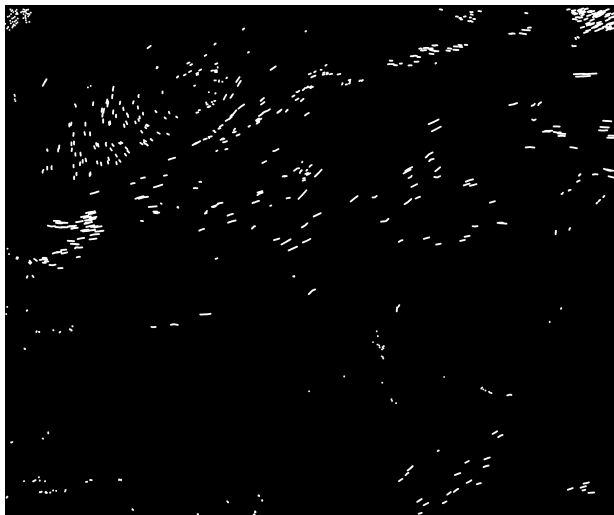


Se mueven para distintos lados unas de otras.

¿Y si vemos el viaje de mas de una partícula?



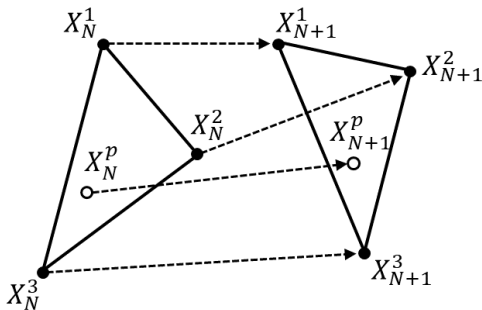
Visualización: No se mueven tan aleatoriamente en esta imagen.



Análisis: Las variación en la posición de las partículas en cada instante de tiempo (dX_t) se mueve siguiendo el movimiento medio a nivel de atmosfera ($m(t)$) pero alterado por un ruido con memoria (U_t). Este actúa como ráfagas" de viento por sector que persisten en el tiempo pero son absorbidas por el movimiento a mesoescala.

$$dX_t = [U_t + m(t)]dt$$

$$dU_t = -\beta U_t dt + \sigma dW_t$$



Estimación de parámetros, validación del modelo, etc.

Acción: Tomar el conocimiento y aplicarlo

- Predictor climático de mediano plazo.
- Se puede utilizar para predecir producción de energía solar
- Ayudar a la integración de las energías limpias



Acción: Tomar el conocimiento y aplicarlo

- Predictor climático de mediano plazo.
- Se puede utilizar para predecir producción de energía solar
- Ayudar a la integración de las energías limpias

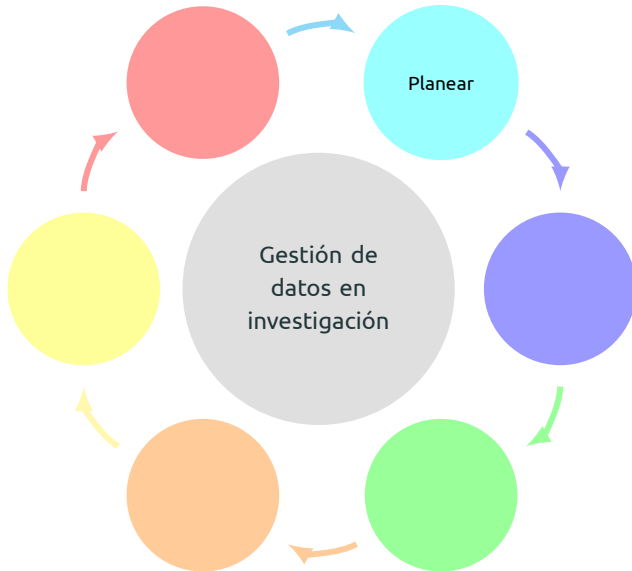
Pero **¿y la parte de gestión?**



Ciclo de vida de datos en investigación



Ciclo de vida de datos en investigación



Ciclo de vida de datos en investigación



Ciclo de vida de datos en investigación



Ciclo de vida de datos en investigación



Ciclo de vida de datos en investigación



Planear: Anticipar recursos, buscar bases de datos, seguridad, planear esquemas relacionales y formas de recolección.



Planear: Anticipar recursos, buscar bases de datos, seguridad, planear esquemas relacionales y formas de recolección.

Recolectar: Recolección de datos, creación de metadatos, documentación, medidas de calidad de datos. .



Planear: Anticipar recursos, buscar bases de datos, seguridad, planear esquemas relacionales y formas de recolección.

Recolectar: Recolección de datos, creación de metadatos, documentación, medidas de calidad de datos. .

Procesar y analizar: Procesar datos, anonimizar, consolidar datos bajo un formato, validación, analisis.



Datos

Planear: Anticipar recursos, buscar bases de datos, seguridad, planear esquemas relacionales y formas de recolección.

Recolectar: Recolección de datos, creación de metadatos, documentación, medidas de calidad de datos. .

Procesar y analizar: Procesar datos, anonimizar, consolidar datos bajo un formato, validación, analisis.

Preservar: Seleccionar datos, migrar a un formato adecuado, almacenar datos en lugar adecuado



Datos

Planear: Anticipar recursos, buscar bases de datos, seguridad, planear esquemas relacionales y formas de recolección.

Preservar:
Seleccionar datos, migrar a un formato adecuado, almacenar datos en lugar adecuado

Recolectar:
Recolección de datos, creación de metadatos, documentación, medidas de calidad de datos. .

Compartir: Escoger repositorio, entregar documentación y metadatos, elegir una licencia adecuada

Procesar y analizar:
Procesar datos, anonimizar, consolidar datos bajo un formato, validación, análisis.



Planear: Anticipar recursos, buscar bases de datos, seguridad, planear esquemas relacionales y formas de recolección.

Preservar: Seleccionar datos, migrar a un formato adecuado, almacenar datos en lugar adecuado

Recolectar: Recolección de datos, creación de metadatos, documentación, medidas de calidad de datos. .

Compartir: Escoger repositorio, entregar documentación y metadatos, elegir una licencia adecuada

Procesar y analizar: Procesar datos, anonimizar, consolidar datos bajo un formato, validación, análisis.

Reusar: Buscar datos, conseguir permisos, citar los datos.



Planear: Anticipar recursos, buscar bases de datos, seguridad, planear esquemas relacionales y formas de recolección.

Preservar: Seleccionar datos, migrar a un formato adecuado, almacenar datos en lugar adecuado

Recolectar: Recolección de datos, creación de metadatos, documentación, medidas de calidad de datos. .

Compartir: Escoger repositorio, entregar documentación y metadatos, elegir una licencia adecuada

Procesar y analizar: Procesar datos, anonimizar, consolidar datos bajo un formato, validación, análisis.

Reusar: Buscar datos, conseguir permisos, citar los datos.



En el curso nos centraremos en:

- Planificación de recursos y estructuras relacionales.
- Creación de metadatos, medidas de calidad de datos.
- Procesar, modificar, derivar, consolidar, analizar y validar los datos.
- Almacenamiento en lugar y formato adecuado.



En el curso nos centraremos en:

- Planificación de recursos y estructuras relacionales.
- Creación de metadatos, medidas de calidad de datos.
- Procesar, modificar, derivar, consolidar, analizar y validar los datos.
- Almacenamiento en lugar y formato adecuado.

Para esto estudiaremos las **bases de datos** (*DataBase*, DB), los **marcos de datos** (*DataFrames*, DF) y los **sistemas de gestión** de estas (*Database Management System*, DBMS)



Tabla de Contenidos

Introducción

Proyectos de ciencia de datos

Datos

¿Cómo trabajar con datos?

Bases de datos y DBMS

¿Por qué DBMS?



¿Cómo trabajar con datos?

Por lo general se manipulan los datos a 2 escalas:

- Bases de datos, con DBMS.
- Marcos de datos, con Python, R o cualquier lenguaje de programación.

Para un proyecto de investigación es común partir de una base de datos grande y tomar una muestra que se trabaja como marco de datos y luego expandirla a toda la DB.



¿Cómo trabajar con datos?

Ojo: que los marcos de datos sean más pequeños no significa que sean menos problemáticos. De hecho es común que en un proyecto interactúen en un 99 % solo con marcos de datos, lo que hace prioritario un buen manejo de estos.

La materia de este curso y los ejemplos se centraran en bases de datos, sin embargo los algoritmos para marcos de datos (que no veremos) son muy similares y requieren el mismo razonamiento para entenderlos.



Las **bases de datos** son una forma de almacenar datos, el objetivo:

- Procesamiento de forma eficiente.
- Que no tenga que programar cada detalle.
- Que sea portable.



Database Management System - DBMS

- Programa que facilite el manejo de grandes volúmenes de datos
- Datos se almacenan en disco
- Pero los usuarios interactúan con una capa lógica (ej. tablas)

Ojo: Los más clásicos son los motores SQL, pero estos trabajan sobre un modelo relacional.



¿Cómo funcionaban las bibliotecas en los años 90?



¿Cómo funcionaban las bibliotecas en los años 90?



Pensemos en el proceso de buscar un libro.



Buscando Winds of Winter de George R.R. Martin

- Buscarlo en los ficheros bibliográficos.



Buscando Winds of Winter de George R.R. Martin

- Buscarlo en los ficheros bibliográficos.
 - Irse al cajón de las "M" encuentras la primera ubicación de "Martin"



Buscando Winds of Winter de George R.R. Martin

- Buscarlo en los ficheros bibliográficos.
 - Irse al cajón de las "M" encuentras la primera ubicación de "Martin"
 - Empiezas a leer las fichas buscando ahora el título "Winds of Winter"



Buscando Winds of Winter de George R.R. Martin

- Buscarlo en los ficheros bibliográficos.
 - Irse al cajón de las "M" encuentras la primera ubicación de "Martin"
 - Empiezas a leer las fichas buscando ahora el titulo "Winds of Winter"
- Obtienes y anotas código de búsqueda en papel.



Buscando Winds of Winter de George R.R. Martin

- Buscarlo en los ficheros bibliográficos.
 - Irse al cajón de las "M" encuentras la primera ubicación de "Martin"
 - Empiezas a leer las fichas buscando ahora el título "Winds of Winter"
- Obtienes y anotas código de búsqueda en papel.
- Le pasas el código al bibliotecario.



Buscando Winds of Winter de George R.R. Martin

- Buscarlo en los ficheros bibliográficos.
 - Irse al cajón de las "M" encuentras la primera ubicación de "Martin"
 - Empiezas a leer las fichas buscando ahora el titulo "Winds of Winter"
- Obtienes y anotas código de búsqueda en papel.
- Le pasas el código al bibliotecario.
- El lo encuentra rápidamente y te retorna el libro.



Buscando Winds of Winter de George R.R. Martin

- Buscarlo en los ficheros bibliográficos.
 - Irse al cajón de las "M" encuentras la primera ubicación de "Martin"
 - Empiezas a leer las fichas buscando ahora el titulo "Winds of Winter"
- Obtienes y anotas código de búsqueda en papel.
- Le pasas el código al bibliotecario.
- El lo encuentra rápidamente y te retorna el libro.



Buscando Winds of Winter de George R.R. Martin

- Buscarlo en los ficheros bibliográficos.
 - Irse al cajón de las "M" encuentras la primera ubicación de "Martin"
 - Empiezas a leer las fichas buscando ahora el título "Winds of Winter"
- Obtienes y anotas código de búsqueda en papel.
- Le pasas el código al bibliotecario.
- El lo encuentra rápidamente y te retorna el libro.

Este es el mismo proceso que ocurre con las bases de datos, pero el sistema involucra mucho más.



En este pequeño ejercicio hemos pasado por alto las preguntas:

- ¿Como se organizan los libros?



En este pequeño ejercicio hemos pasado por alto las preguntas:

- ¿Como se organizan los libros?
- ¿Como se organizan los ficheros?



En este pequeño ejercicio hemos pasado por alto las preguntas:

- ¿Como se organizan los libros?
- ¿Como se organizan los ficheros?
- ¿Por qué no puedo ir a buscar yo el libro?



En este pequeño ejercicio hemos pasado por alto las preguntas:

- ¿Como se organizan los libros?
- ¿Como se organizan los ficheros?
- ¿Por qué no puedo ir a buscar yo el libro?
- ¿Debería poner las fichas de las revistas en los mismos ficheros?



En este pequeño ejercicio hemos pasado por alto las preguntas:

- ¿Como se organizan los libros?
- ¿Como se organizan los ficheros?
- ¿Por qué no puedo ir a buscar yo el libro?
- ¿Debería poner las fichas de las revistas en los mismos ficheros?
- ¿Que pasa si es que yo me se el nombre del libro y no el autor?



Las bases de datos en este ejemplo son un análogo a los libros.

El administrador de bases de datos (DBMS) es la biblioteca en si y su estructura.

- La ficha representa la llave en una base de datos.



Las bases de datos en este ejemplo son un análogo a los libros.

El administrador de bases de datos (DBMS) es la biblioteca en si y su estructura.

- La ficha representa la llave en una base de datos.
 - ¿Diferencias?



Las bases de datos en este ejemplo son un análogo a los libros.

El administrador de bases de datos (DBMS) es la biblioteca en si y su estructura.

- La ficha representa la llave en una base de datos.
 - ¿Diferencias?
- El código representa la ubicación en disco duro.



Las bases de datos en este ejemplo son un análogo a los libros.

El administrador de bases de datos (DBMS) es la biblioteca en si y su estructura.

- La ficha representa la llave en una base de datos.
 - ¿Diferencias?
- El código representa la ubicación en disco duro.
 - ¿Por qué es mejor generar un orden y no apilar los libros según el orden de incorporación a la biblioteca?



Las bases de datos en este ejemplo son un análogo a los libros.

El administrador de bases de datos (DBMS) es la biblioteca en si y su estructura.

- La ficha representa la llave en una base de datos.
 - ¿Diferencias?
- El código representa la ubicación en disco duro.
 - ¿Por qué es mejor generar un orden y no apilar los libros según el orden de incorporación a la biblioteca?
- El bibliotecario es el motor de búsqueda, sus conocimientos son los algoritmos.



¿Por qué DBMS?

¿Que podemos hacer en DBMS?

- Almacenar datos (insertar)
- Encontrar datos (búsquedas y consultas)
- Modificar datos (update)
- Asegurar la consistencia de los datos
- Seguridad y privacidad de los datos



¿Por qué DBMS?

Pregunta obvia: **¿Por que no puedo crear un sistema de bases de datos a mi medida en R o Python?**

- Persistencia de datos
- Disponibilidad de datos
- Comunidad
- Limitaciones tecnológicas
- Velocidad de obtención

