ESCUELA DE INGENIERÍA
FACULTAD DE INGENIERÍA

EDUCACIÓN PROFESIONAL

Introducción y Motivación

Gestión de Datos

Maximiliano Arancibia

Educación Profesional - Escuela de Ingeniería

Tabla de Contenidos

Tabla de Contenidos

Bienvenida

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 Publicas.

Programa

Información general:

- Clases:
  - Viernes, 18:30-21:30
  - Sábados, 9:00 - 12:15
- Talleres:
  - José Antonio Délano
  - Sábados (algunos): 9:00 - 12:30
  - jose.delano@uc.cl

Programa

Información general:

- Clases:
  - Viernes, 18:30-21:30
  - Sábados, 9:00 - 12:15
- Talleres:
  - José Antonio Délano
  - Sábados (algunos): 9:00 - 12:30
  - jose.delano@uc.cl

Sitio oficial del curso:

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

Programa

Información general:

- Clases:
  - Viernes, 18:30-21:30
  - Sábados, 9:00 - 12:15
- Talleres:
  - José Antonio Délano
  - Sábados (algunos): 9:00 - 12:30
  - jose.delano@uc.cl

Sitio oficial del curso:

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

Atención alumnos:

- Vía mail
- Después de clases

Programa

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 actividades individuales y grupales. Sin nota.

Software: En esta clase y taller programaremos en R, para hacerlo se usará la plataforma Kaggle (*www.kaggle.com*). Además a futuro veremos el lenguaje SQL enlazado con R.

Condiciones de aprobación:

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

ESCUELA DE INGENIERÍA
FACULTAD DE INGENIERÍA

Programa

Bibliografía Básica

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

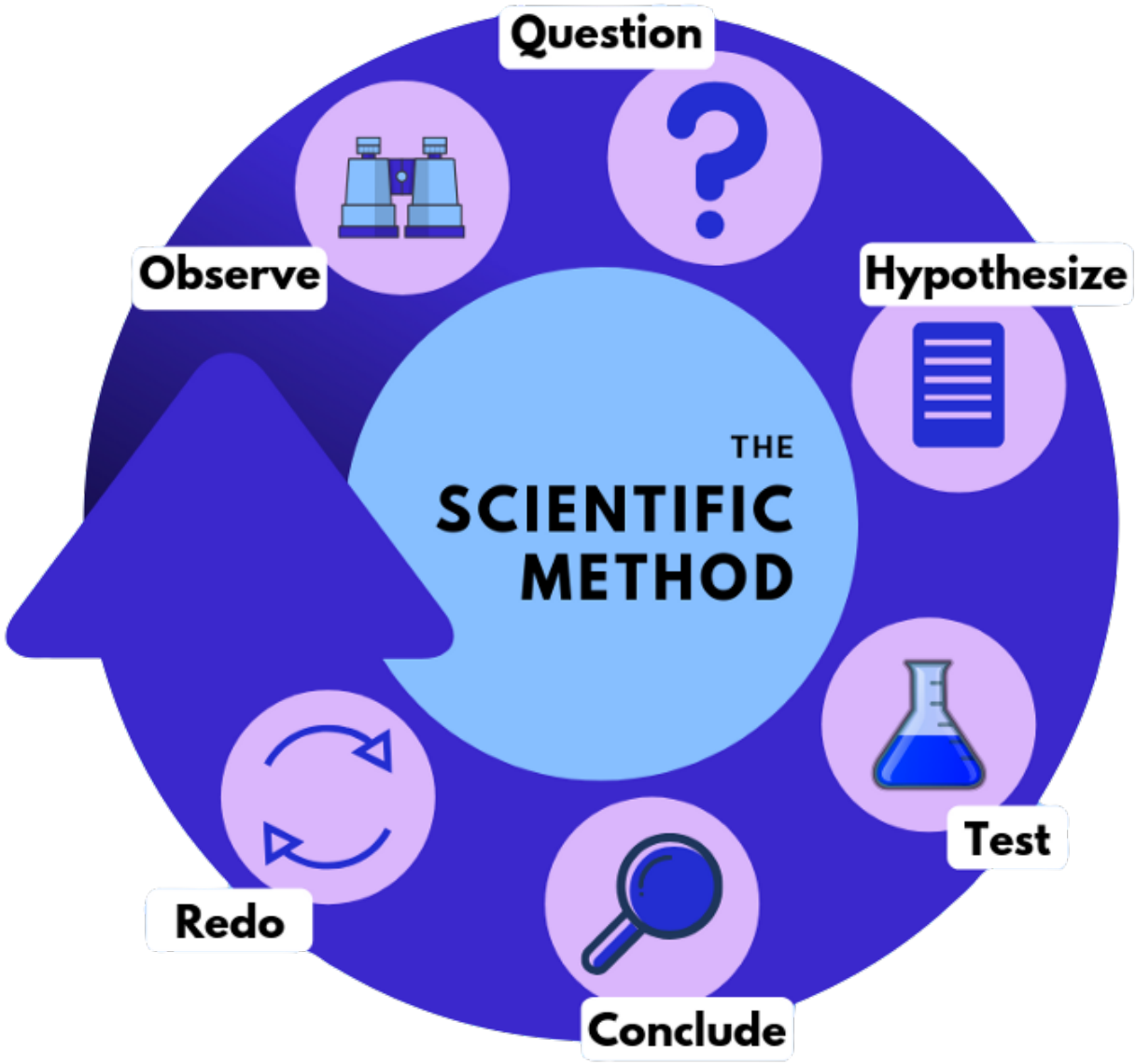
Tabla de Contenidos

ESCUELA DE INGENIERÍA
FACULTAD DE INGENIERÍA

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

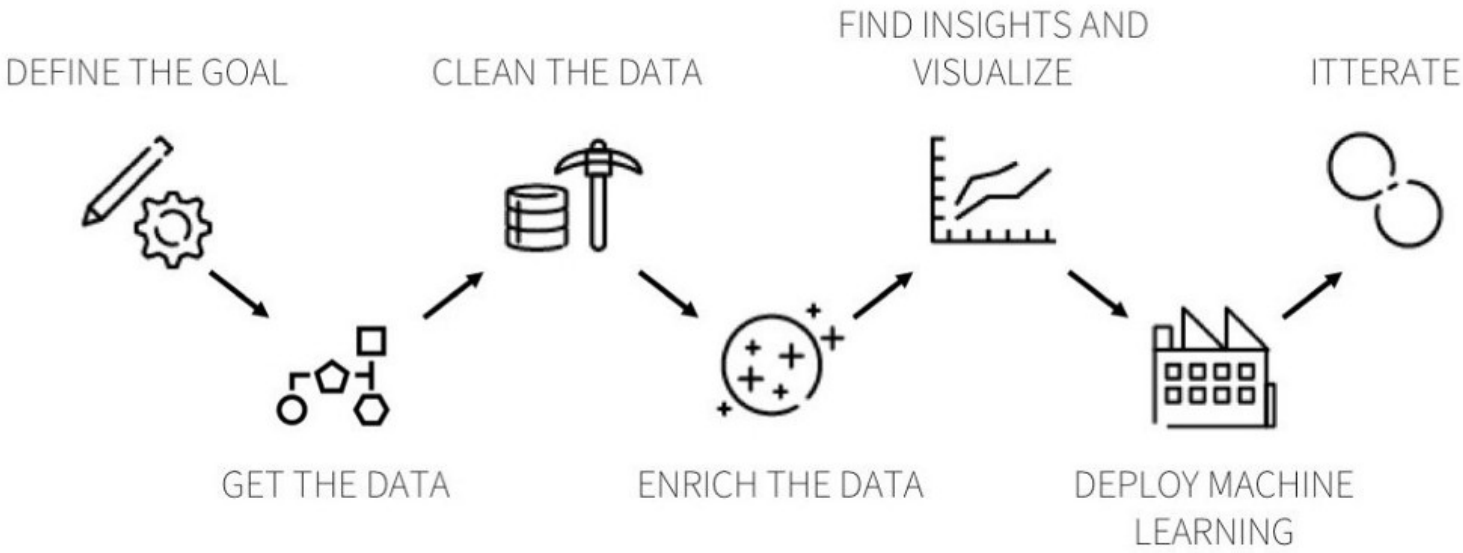


Tabla de Contenidos

Introducción

Primero que todo...

Introducción

Primero que todo...

¿Qué son los datos?

Introducción

Primero que todo...

¿Qué son los datos?

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

Introducción

Primero que todo...

¿Qué son los datos?

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

$\Rightarrow$ 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
- ...

Ciclo de vida de datos en investigación

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

Datos

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

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.

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.

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.

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.

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

ESCUELA DE INGENIERÍA
FACULTAD DE INGENIERÍA

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.

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.

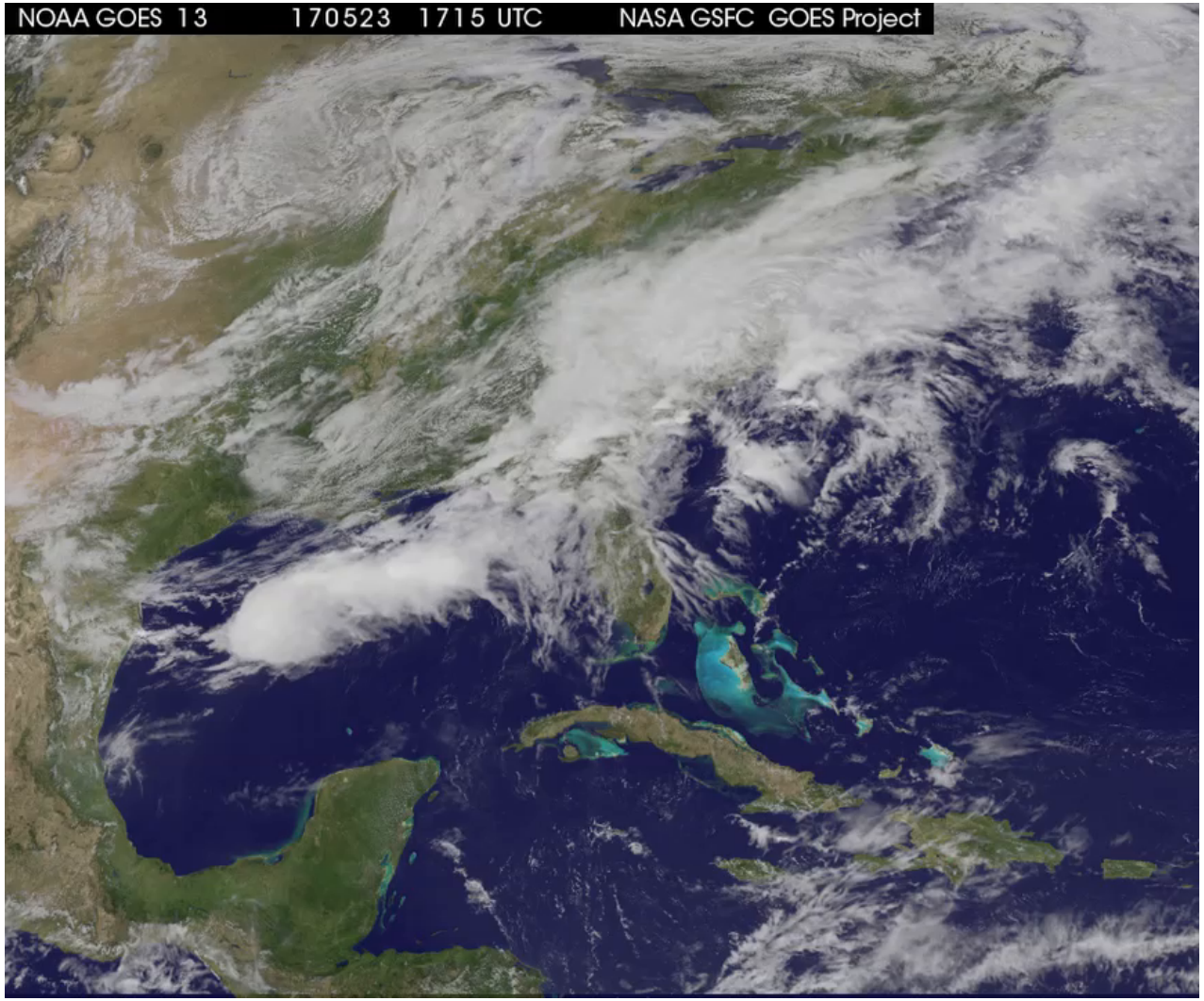
Visualización: Formas de comunicar la información recolectada. Puede ser a través de gráficos, tablas, imágenes, etc.

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.

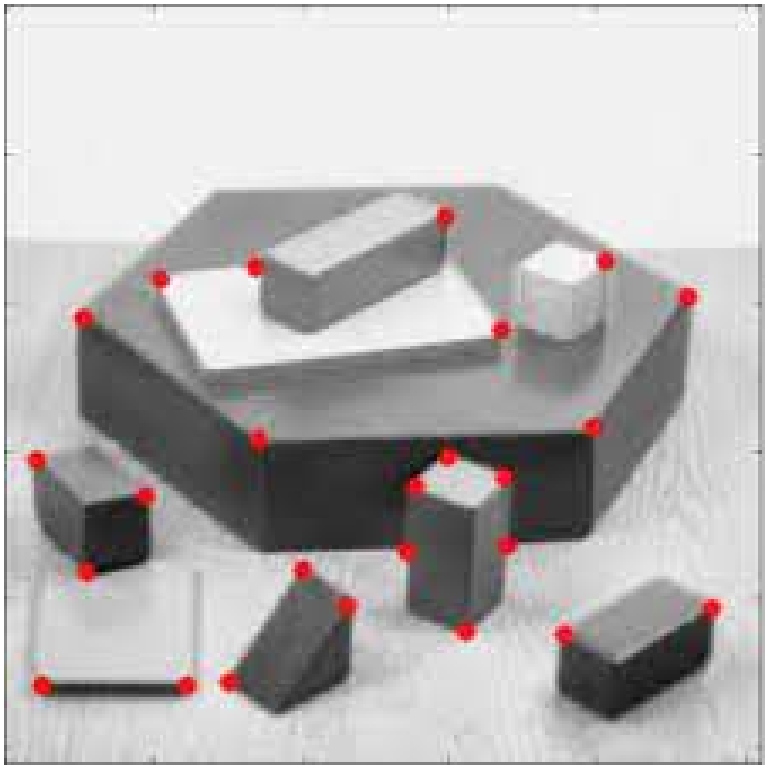
Datos

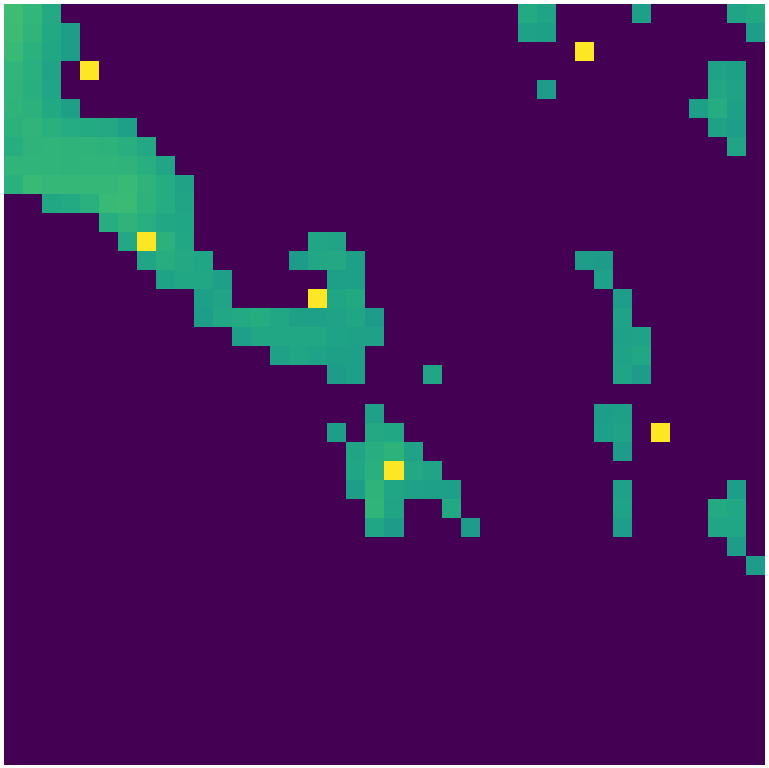
Generación: NASA



Datos

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





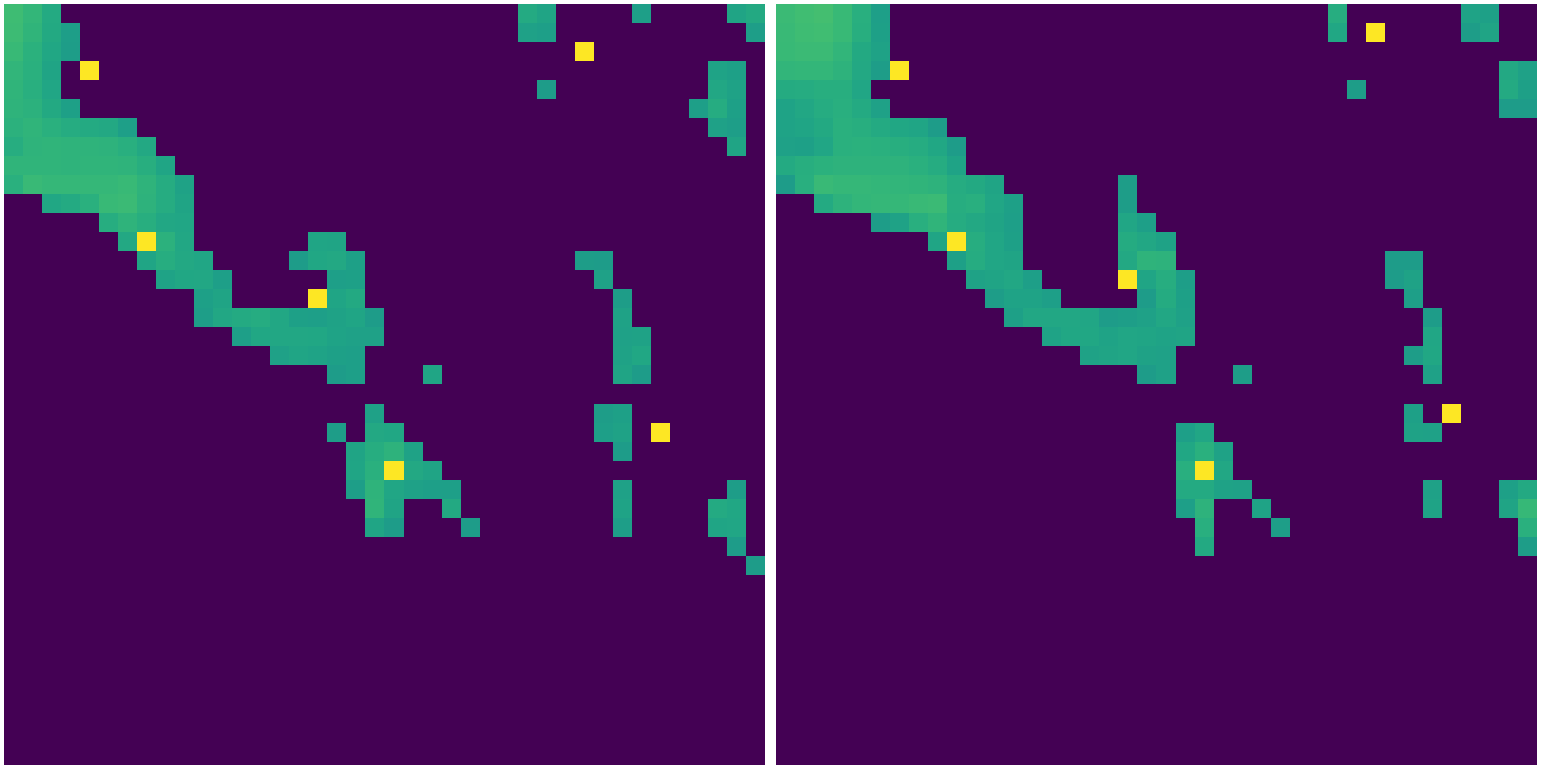
Datos

Almacenamiento: base de datos de características.

	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

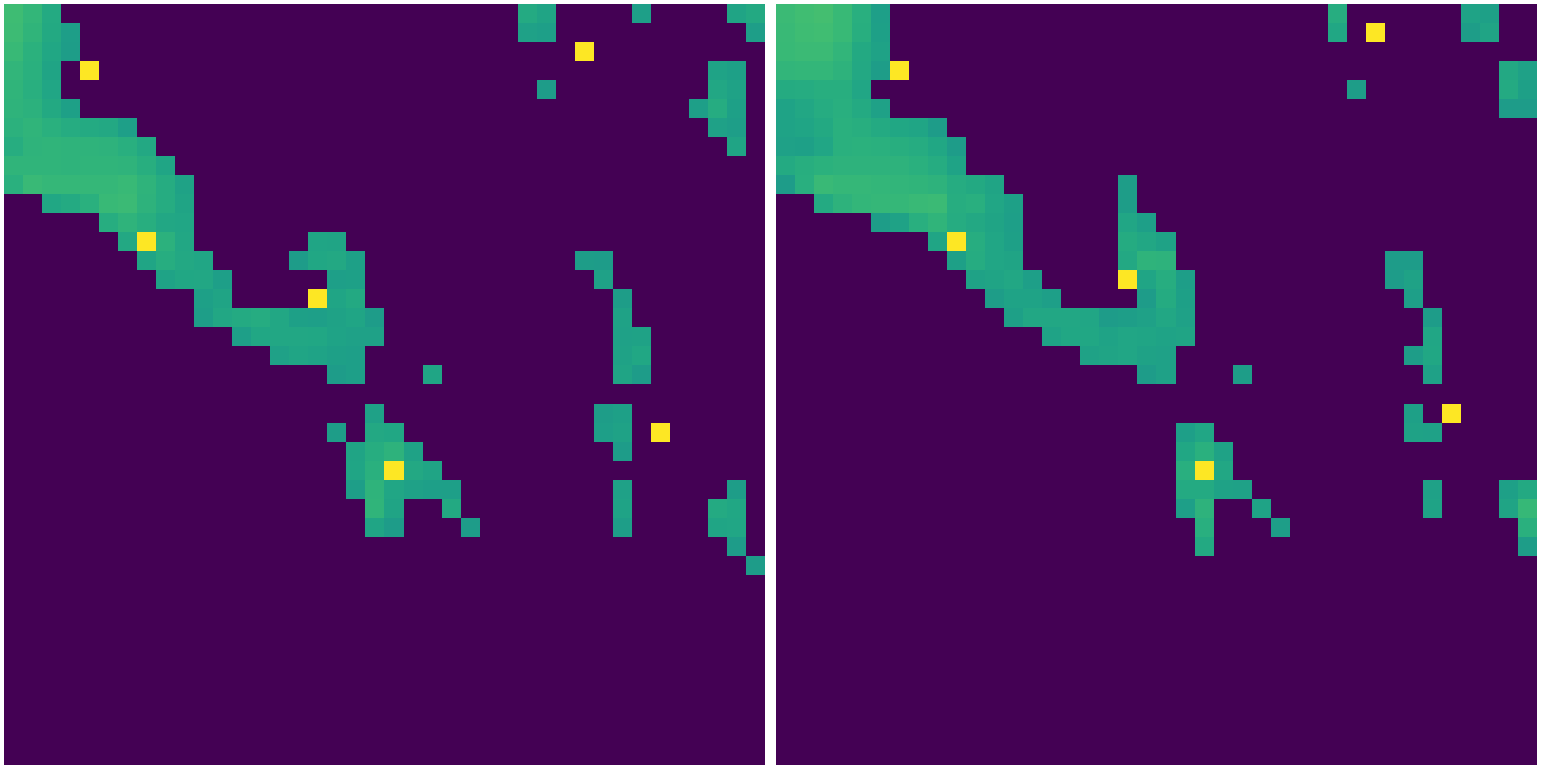
Datos

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



Datos

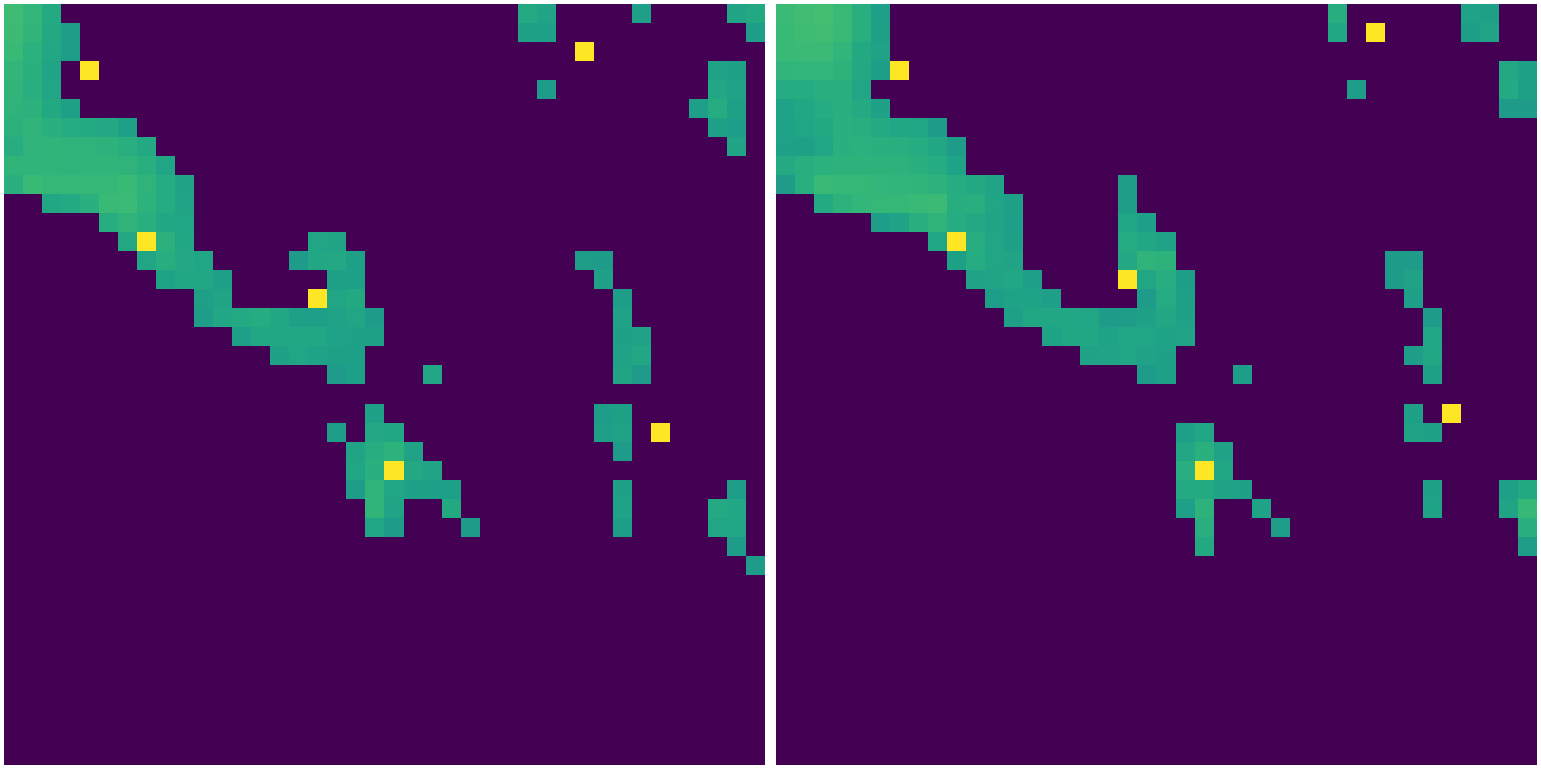
Visualización: ¿Como se comportan dos instantes de tiempo distinto?



Se mueven para distintos lados unas de otras.

Datos

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

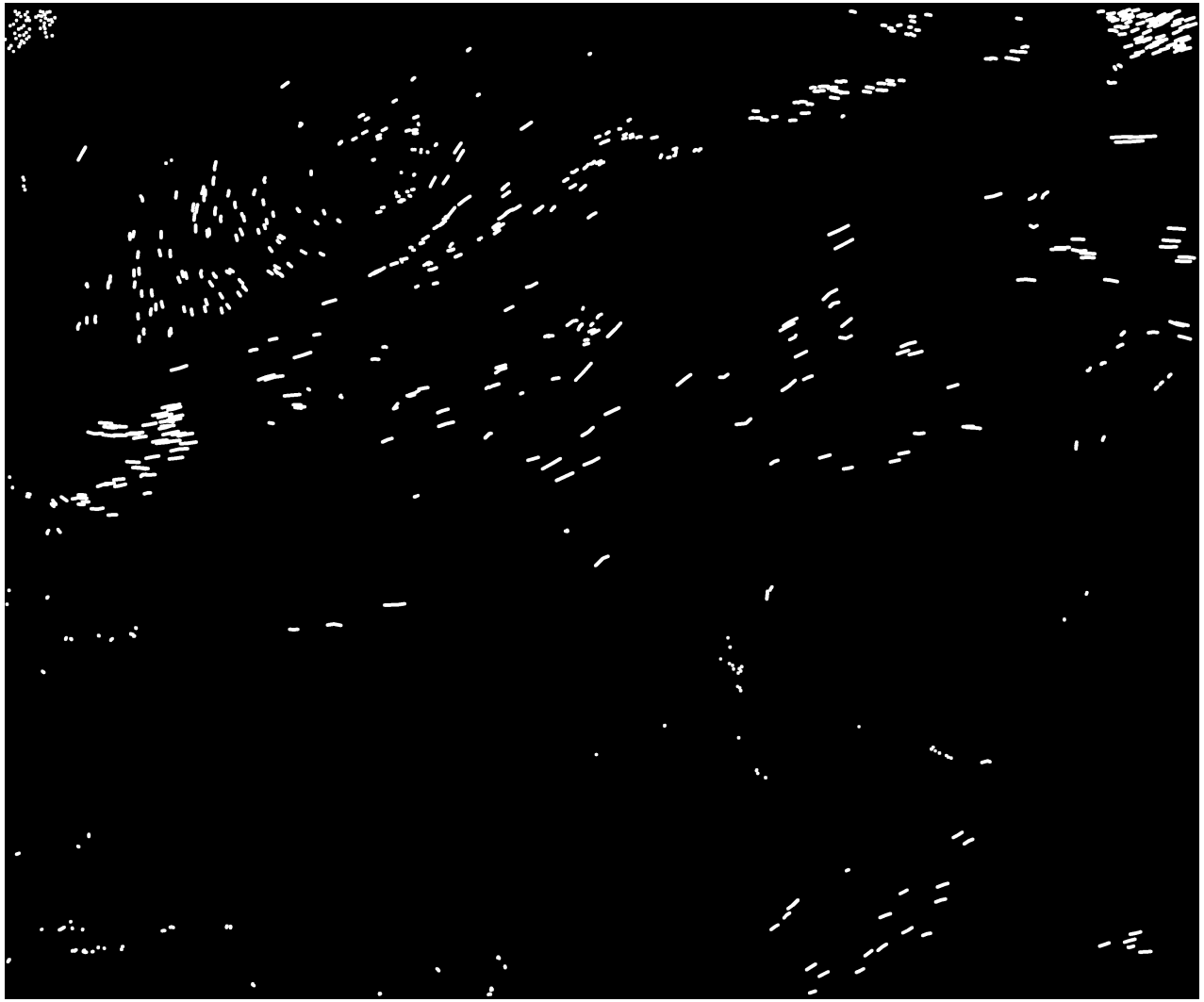


Se mueven para distintos lados unas de otras.

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

Datos

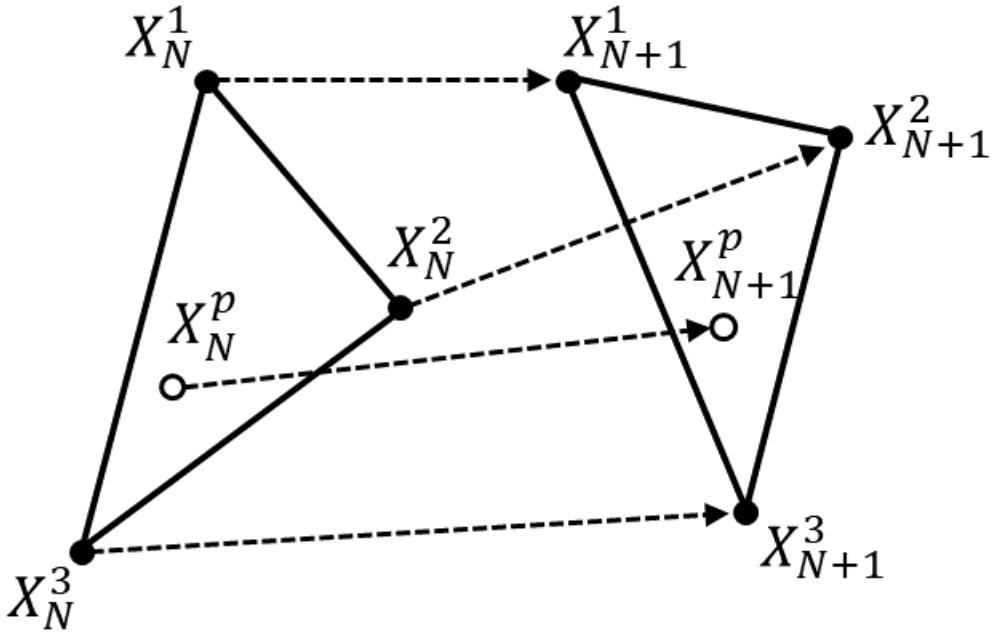
Visualización: No se mueven tan aleatoriamente en esta imagen.



Datos

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 atmósfera ($m(t)$) 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.

Datos

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

Datos

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?**

Gestión de datos

Recolectar y almacenar

- **Planear**, recursos, seguridad, esquemas relacionales.

Gestión de datos

Recolectar y almacenar

- **Planear**, recursos, seguridad, esquemas relacionales.
- **Recolectar**, metadatos, documentación, calidad de datos.

Gestión de datos

Recolectar y almacenar

- **Planear**, recursos, seguridad, esquemas relacionales.
- **Recolectar**, metadatos, documentación, calidad de datos.
- **Procesar y analizar**, anonimizar, consolidar, validar.

Gestión de datos

Recolectar y almacenar

- **Planear**, recursos, seguridad, esquemas relacionales.
- **Recolectar**, metadatos, documentación, calidad de datos.
- **Procesar y analizar**, anonimizar, consolidar, validar.
- **Preservar**, seleccionar, migrar, almacenar.

Gestión de datos

Recolectar y almacenar

- **Planear**, recursos, seguridad, esquemas relacionales.
- **Recolectar**, metadatos, documentación, calidad de datos.
- **Procesar y analizar**, anonimizar, consolidar, validar.
- **Preservar**, seleccionar, migrar, almacenar.
- **Compartir**, repositorios, licencia, liberar metadata.

Gestión de datos

Recolectar y almacenar

- **Planear**, recursos, seguridad, esquemas relacionales.
- **Recolectar**, metadatos, documentación, calidad de datos.
- **Procesar y analizar**, anonimizar, consolidar, validar.
- **Preservar**, seleccionar, migrar, almacenar.
- **Compartir**, repositorios, licencia, liberar metadata.
- **Reusar**: buscar, permisos, citar.

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.

Tabla de Contenidos

¿Cómo trabajar con datos?

Por lo general se manipulan los datos a 2 escalas:

- Dataframes (tablas, marcos de datos), con Python, R o cualquier lenguaje de programación.
- Bases de datos, a través de un administrador de bases de datos (DBMS).

Para un proyecto de investigación es común partir de una base de datos grande y tomar una muestra que se procesa para luego expandirla a todo el conjunto de datos.

¿Cómo funcionaban las bibliotecas antes de los computadores?

¿Cómo funcionaban las bibliotecas antes de los computadores?

Pensemos en el proceso de buscar un libro.

Buscando Las cronicas de Narnia de C. S. Lewis

- Buscarlo en los ficheros bibliográficos.

Buscando Las cronicas de Narnia de C. S. Lewis

- Buscarlo en los ficheros bibliográficos.
  - Irse al cajón de las ”L” encuentras la primera ubicación de ”Lewis”

Buscando Las cronicas de Narnia de C. S. Lewis

- Buscarlo en los ficheros bibliográficos.
  - Irse al cajón de las ”L” encuentras la primera ubicación de ”Lewis”
  - Empiezas a leer las fichas buscando ahora el titulo ”Las Crónicas de Narnia”

Buscando Las cronicas de Narnia de C. S. Lewis

- Buscarlo en los ficheros bibliográficos.
  - Irse al cajón de las ”L” encuentras la primera ubicación de ”Lewis”
  - Empiezas a leer las fichas buscando ahora el titulo ”Las Crónicas de Narnia”
  - De los resultados disponibles eliges el que más te gusta.

Buscando Las cronicas de Narnia de C. S. Lewis

- Buscarlo en los ficheros bibliográficos.
  - Irse al cajón de las ”L” encuentras la primera ubicación de ”Lewis”
  - Empiezas a leer las fichas buscando ahora el titulo ”Las Crónicas de Narnia”
  - De los resultados disponibles eliges el que más te gusta.
- Obtienes y anotas código de búsqueda en papel.

Buscando Las cronicas de Narnia de C. S. Lewis

- Buscarlo en los ficheros bibliográficos.
  - Irse al cajón de las ”L” encuentras la primera ubicación de ”Lewis”
  - Empiezas a leer las fichas buscando ahora el titulo ”Las Crónicas de Narnia”
  - De los resultados disponibles eliges el que más te gusta.
- Obtienes y anotas código de búsqueda en papel.
- Le pasas el código al bibliotecario.

Buscando Las cronicas de Narnia de C. S. Lewis

- Buscarlo en los ficheros bibliográficos.
  - Irse al cajón de las ”L” encuentras la primera ubicación de ”Lewis”
  - Empiezas a leer las fichas buscando ahora el titulo ”Las Crónicas de Narnia”
  - De los resultados disponibles eliges el que más te gusta.
- Obtienes y anotas código de búsqueda en papel.
- Le pasas el código al bibliotecario.
- El lo encuentra rápidamente y te retorna el libro.

Buscando Las cronicas de Narnia de C. S. Lewis

- Buscarlo en los ficheros bibliográficos.
  - Irse al cajón de las ”L” encuentras la primera ubicación de ”Lewis”
  - Empiezas a leer las fichas buscando ahora el titulo ”Las Crónicas de Narnia”
  - De los resultados disponibles eliges el que más te gusta.
- Obtienes y anotas código de búsqueda en papel.
- Le pasas el código al bibliotecario.
- El lo encuentra rápidamente y te retorna el libro.

Buscando Las cronicas de Narnia de C. S. Lewis

- Buscarlo en los ficheros bibliográficos.
  - Irse al cajón de las ”L” encuentras la primera ubicación de ”Lewis”
  - Empiezas a leer las fichas buscando ahora el titulo ”Las Crónicas de Narnia”
  - De los resultados disponibles eliges el que más te gusta.
- 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.

Preguntas

- ¿Como se organizan los libros?

Preguntas

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

Preguntas

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

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?

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?

Los datos en este ejemplo son un análogo a los libros.

Así como se generó una estructura para organizar los libros, existen formas de organizar los datos para que sea fácil y rápida su extracción.

Esto es una **base de datos**.

Bases de datos

Las **bases de datos** es un conjunto de información organizado de forma que:

- Tenga un sistema o método eficiente de almacenamiento y búsqueda.
- No tenga que especificar como se busca la información.

Sistema de base de datos

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.

¿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