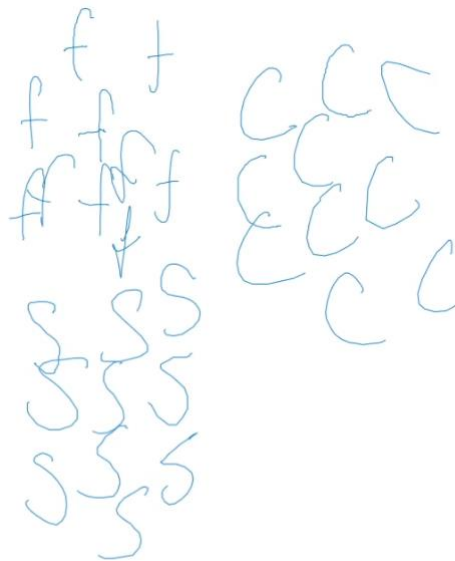


ORGANIZACIÓN DE ARCHIVOS DE DATOS

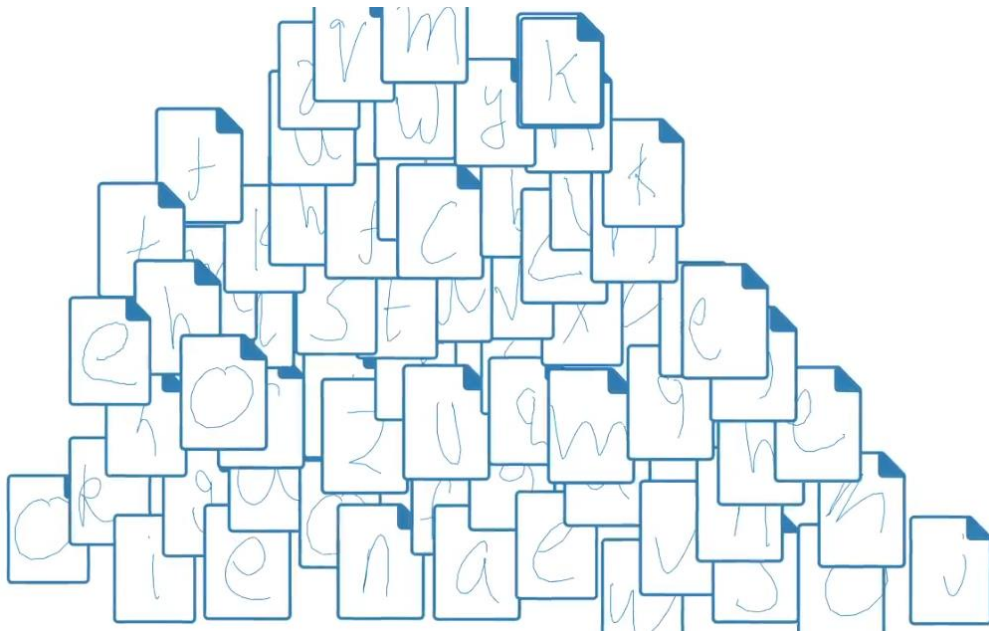
Para usar machine learning, se necesitan datos.

Si quiere que un modelo clasifique 26 caracteres distintos, necesitará varios ejemplos de cada letra. Esto implica cientos e incluso miles de observaciones individuales. ¿Cómo se almacenan estas observaciones?

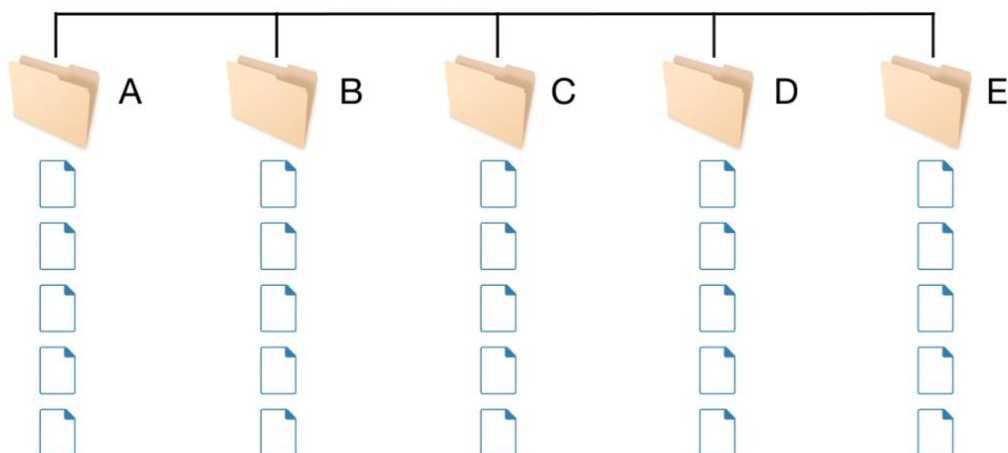


A
B
C
D
E
F
G
H
I
J
K
L
M
N
O
P
Q
R
S
T
U
V
W
X
Y
Z

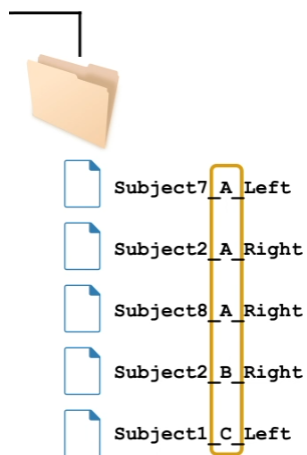
Es poco habitual que todos los datos estén en un archivo de gran tamaño. En el ejemplo de la escritura manual, cada muestra de cada letra se almacena en un archivo independiente. Pero ¿cómo se organizan esos cientos de archivos? ¿Y cómo se sabe qué letra se representa en cada archivo?



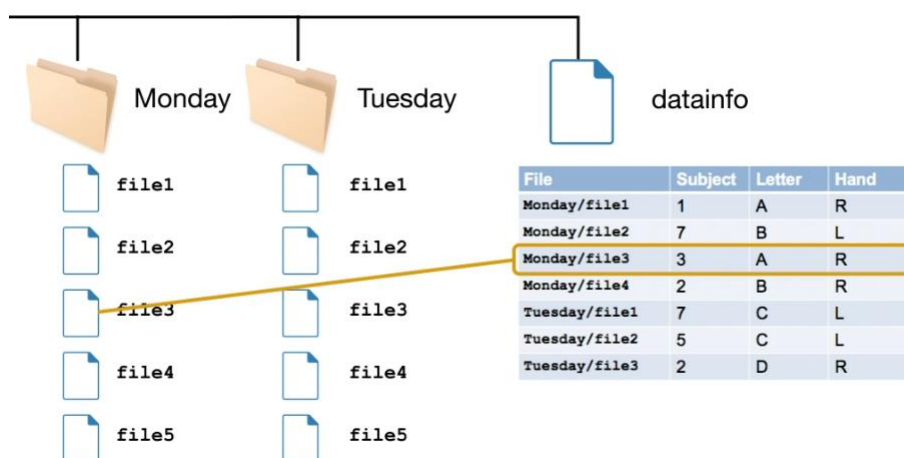
Algunas opciones son tener una carpeta para cada letra,



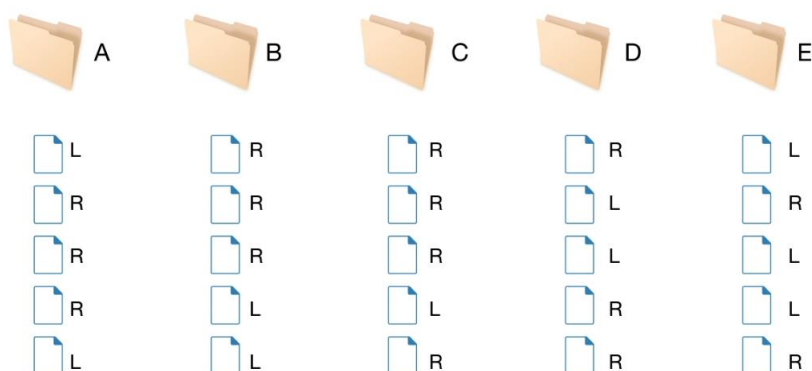
o una carpeta con todos los archivos con la letra en el nombre del archivo,



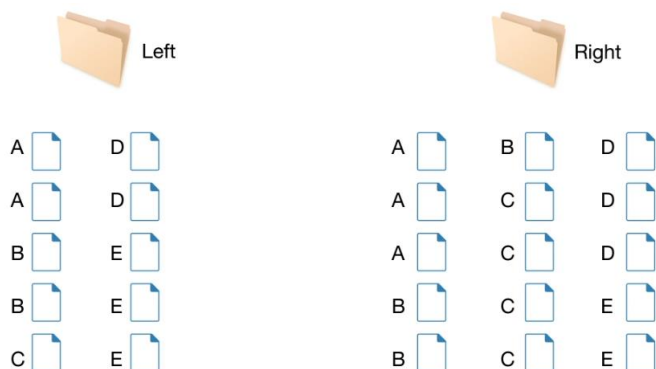
o un archivo independiente con una tabla de búsqueda de los archivos y las letras que contienen.



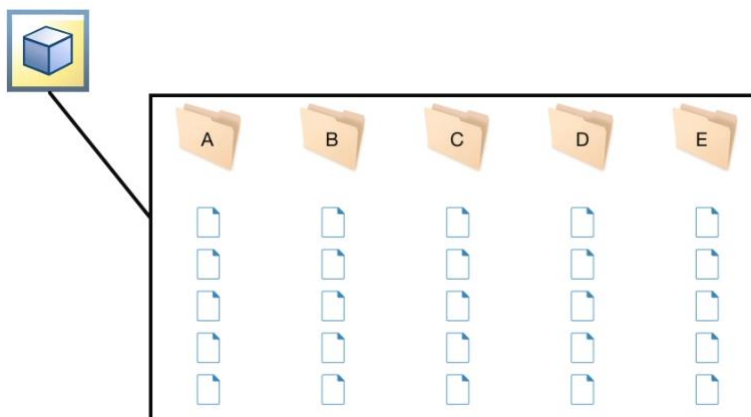
Si trabaja con un conjunto de datos existente, tiene que trabajar con los datos tal y como están almacenados. Pero, si es posible, debería pensar en la organización de los archivos antes de recopilar los datos para tener los datos en la disposición más conveniente. La mejor disposición depende del problema que esté estudiando. Si desea clasificar las letras, una carpeta para cada letra tiene sentido.



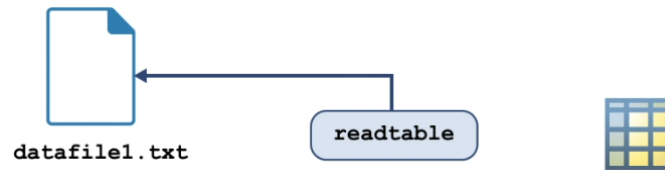
Pero, si desea estudiar, por ejemplo, a los zurdos frente a los diestros, probablemente le interesará otra disposición.



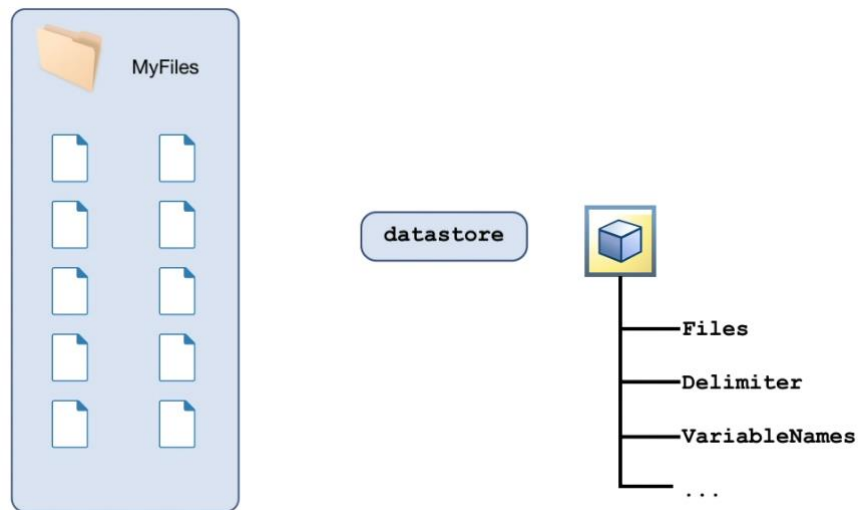
Independientemente de cómo estén organizados los archivos, los almacenes de datos ofrecen una forma cómoda de acceder a los datos almacenados en diversos archivos.



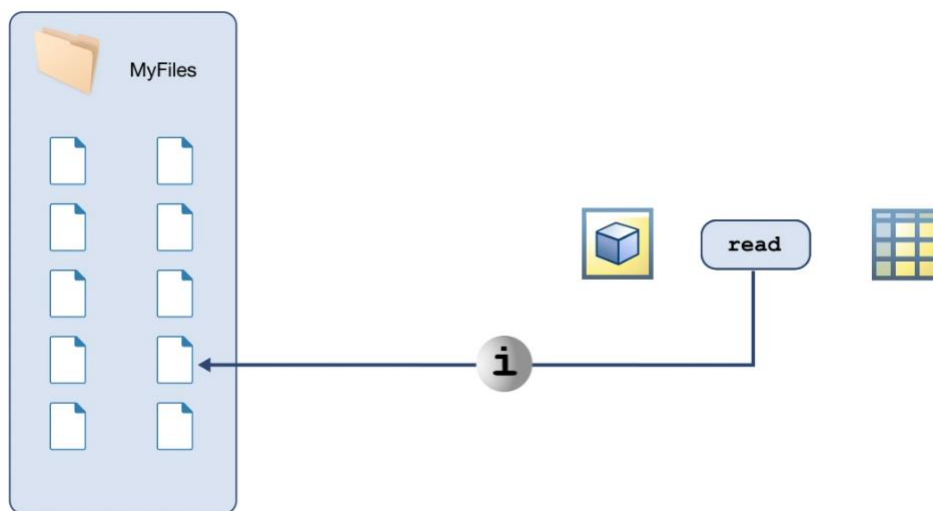
Con las funciones de importación, como `readtable`, se proporciona el nombre de un archivo y se obtienen los datos que contiene.



Pero, cuando se crea un almacén de datos, se proporciona la ubicación de los datos. Entonces, MATLAB mira todos los archivos de datos de la ubicación y devuelve una variable que contiene información sobre los archivos y el formato de su contenido.

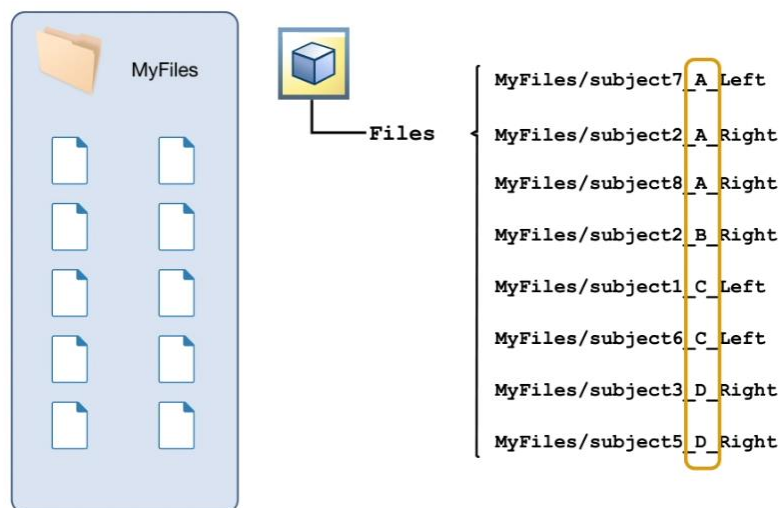


Como podrían existir muchos datos, los datos no se importan hasta que se solicita. De esta forma, se puede controlar lo que se obtiene y cuándo se obtiene.

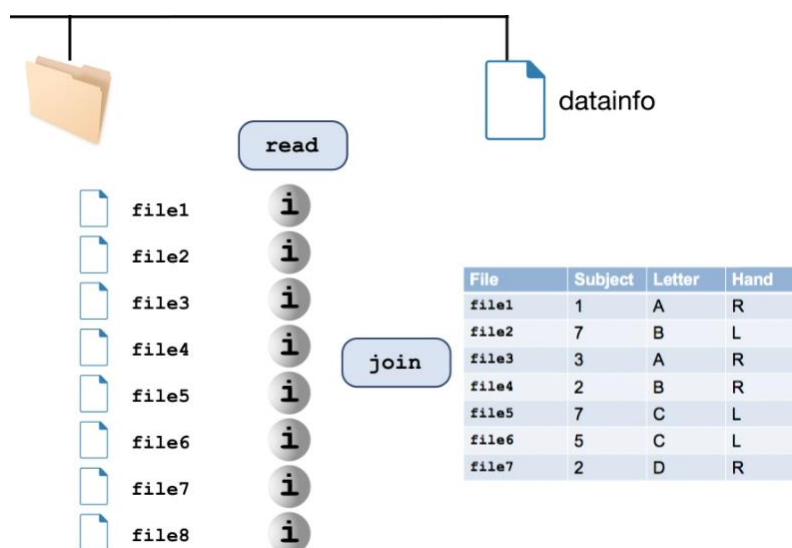


La propiedad `Files` de un almacén de datos contiene la lista de los nombres de archivo completos de todos los archivos de datos. Muchas veces, la información sobre la observación forma parte de

la ruta o el nombre del archivo. En tal caso, puede usar funciones de cadena como `extractBetween` para localizar y extraer esa información.






Si la información sobre las observaciones se almacena como una tabla de búsqueda independiente, normalmente usará una función como `readtable` para importarla y, después, la unirá a los datos asociados.



Los datos de escritura manual se almacenaron en una carpeta, con información en los nombres de archivo. En las siguientes secciones, creará un almacén de datos y aprenderá cómo importar y procesar los datos de estos archivos, de manera que esté listo para extraer la información de los datos y crear modelos predictivos.



-  user001_M_1.txt
-  user001_V_1.txt
-  user002_M_1.txt
-  user002_V_1.txt
-  user003_M_1.txt
-  user003_M_2.txt
-  user003_V_1.txt
-  user003_V_2.txt

read

preprocess

