Universidad de Sonora

Reconocimiento de patrones: Preparación de los datos

Ramón Soto C. (rsotoc@moviquest.com)

Usabilidad de los datos

Los datos son la base de la nueva economía de la información y cada día se incrementa la cantidad de información. Para el año 2020 se espera que hayamos superado el zettabyte de datos producidos por día (1 ZB = 10^{21} bytes = 10^{9} TB... mh! almacenados en discos SSD de 1TB, en 45 días se formaría una torre de la tierra a la luna, suponiendo que son incompresibles). Estos datos son generados por sensores, GPSs, redes sociales, mensajes electrónicos, transacciones comerciales, publicaciones regulares, etc. (aquí un interesante *infográfico* al respecto). Esta cantidad de datos permite generar una gran cantidad de información para analizar virtualmente cualquier problema. Sin embargo, antes de poder explotar la información contenida en ellos y antes de poder generar conocimiento de utilidad para la toma de decisiones, es necesario garantizar que los datos se encuentren en 'buenas condiciones'.

La **usabilidad de los datos** es un término utilizado para especificar qué tan fácil es para el usuario, procesar y analizar un determinado conjunto de datos. Aunque la mayoría de las organizaciones cuenta con una gran cantidad de datos, tales datos a menudo están desorganizados, o no pueden ser analizados de manera efectiva para producir información útil para apoyar la toma de decisiones.

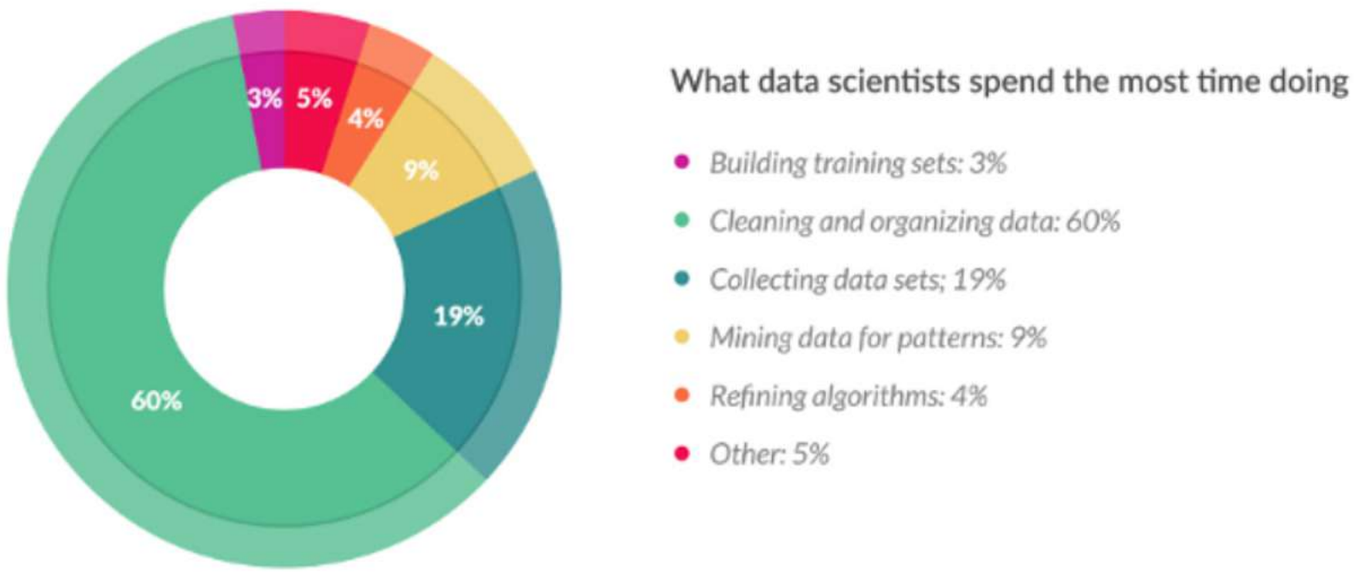
Existen varias dimensiones para definir la usabilidad de los datos, entre las que se destacan las siguientes:

- **Relevancia**: Se refiere al grado de consistencia entre el contenido de los datos y el área de interés del usuario.
- **Calidad**: La calidad de los datos es una percepción o una evaluación de la idoneidad de los datos para cumplir su propósito en un contexto determinado. En muchas ocasiones, este concepto es el que se utiliza como "paraguas" que engloba las características de los datos (como la relevancia).
- **Cobertura** y **granularidad**: Qué tan amplia es la exploración del universo de elementos y qué tan detallada es el análisis.
- **Accesibilidad** y **documentación**: Facilidad para adquirir y procesar los datos e informaiccón acerca de cómo está conformada (nombres de las variables, su significado, especificaciones de su captura, etc.).
- **Facilidad de análisis**: La capacidad para utilizar técnicas "convencionales" de procesamiento de datos.

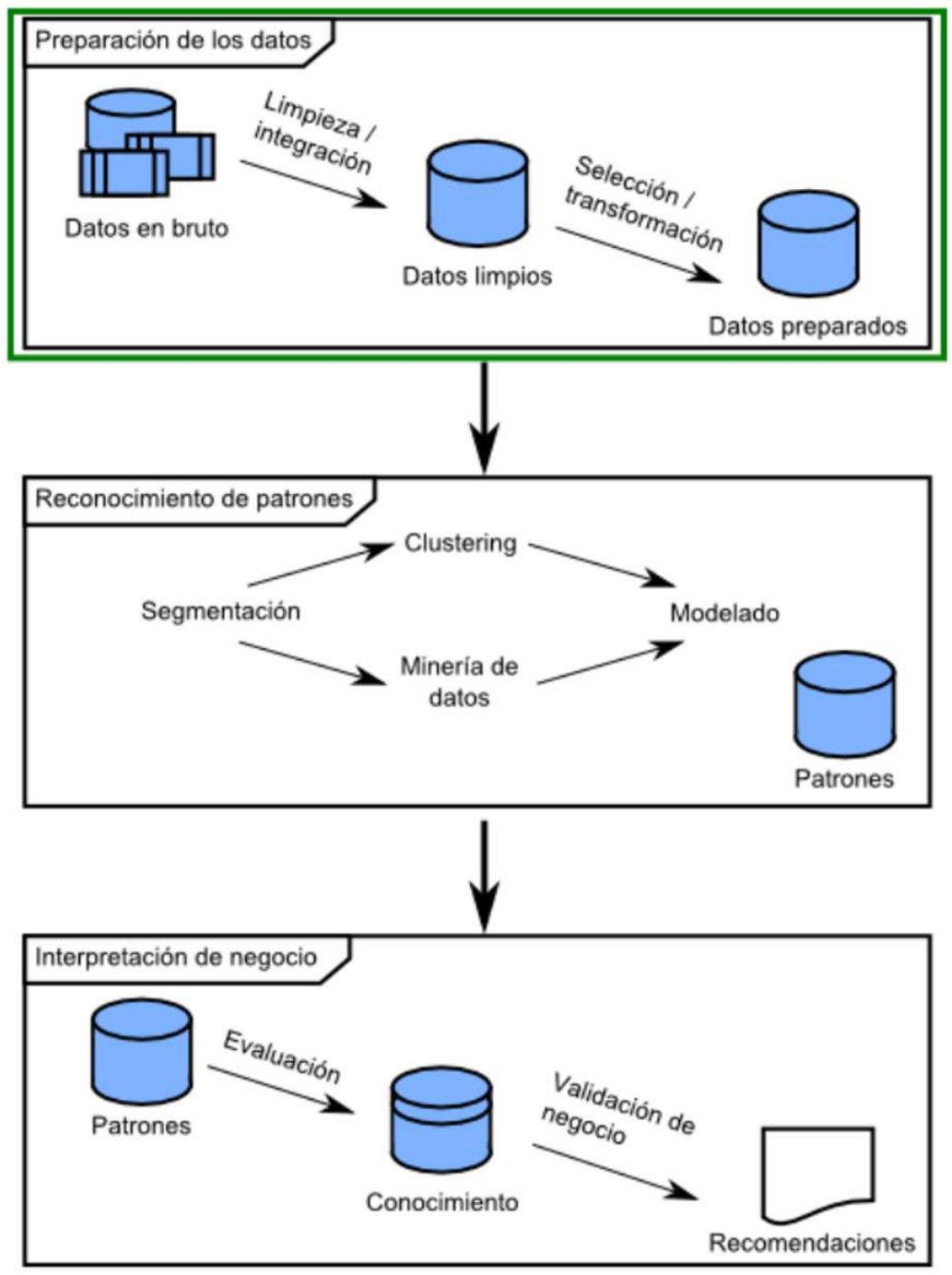
Para garantizar tales aspectos, suele ser necesario un procedimiento de preparación de los datos.

La fase de preparación de los datos

Es una estimación bien conocida entre los científicos de datos que el 80% del tiempo dedicado a la solución de un problema se invierte en la preparación de los datos:



El proceso de mejoramiento de los datos es lo que se denomina **preparación de los datos**.



- La *limpieza de datos* consiste en rellenar valores faltantes, suavizar datos con ruido, identificar y remover valores atípicos y resolver inconsistencias.
- La *integración de datos* es la integración de diversas fuentes de datos: bases de datos, cubos de datos o archivos.
- La *selección de datos* consiste en seleccionar el conjunto de datos adecuado para analizar el sistema, incluyendo el muestreo.
- La *selección de características* es un proceso mediante el cual se analizan las variables determinantes para describir los datos.
- La *transformación de datos* incluye operaciones como normalización, agregación, codificación.

Conclusiones

- El éxito de un proyecto de reconocimiento de patrones (minería de datos, ciencia de los datos) depende de contar con un conjunto de datos de alta calidad.
- Para asegurar la calidad de los datos, se requiere de un análisis exploratorio para detectar posibles deficiencias y corregirlas.
- La etapa de preparación de los datos es una de las más costosas en cualquier proyecto de ciencias de los datos.