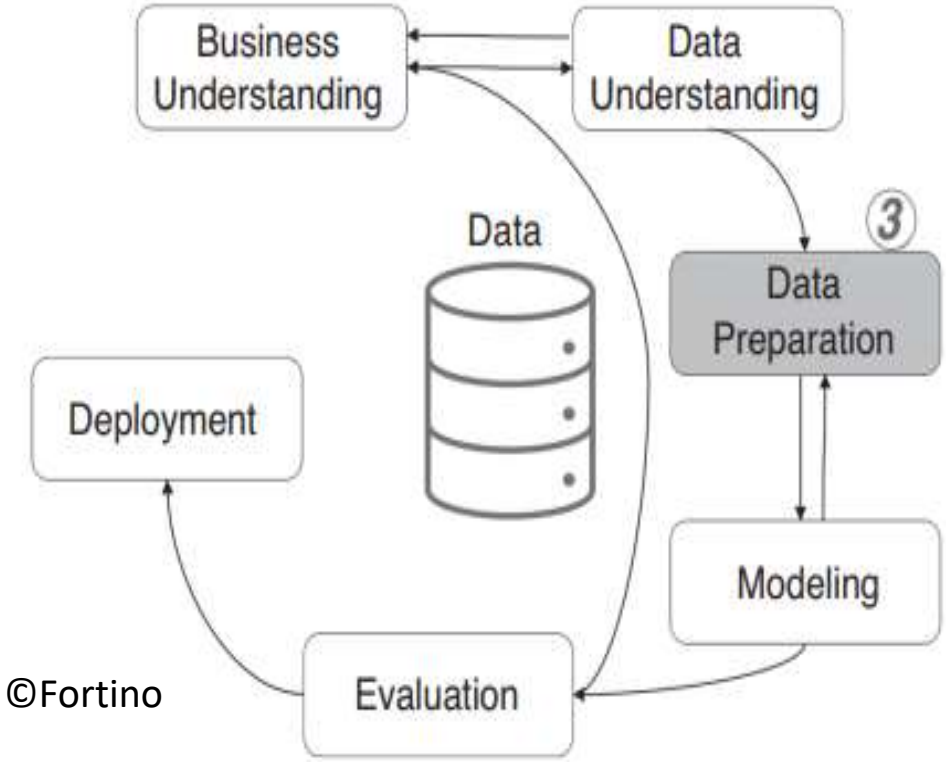
Preparación de datos

Prof. Dr. Jorge Zavaleta
zavaleta.jorge@gmail.com

Preparación de datos

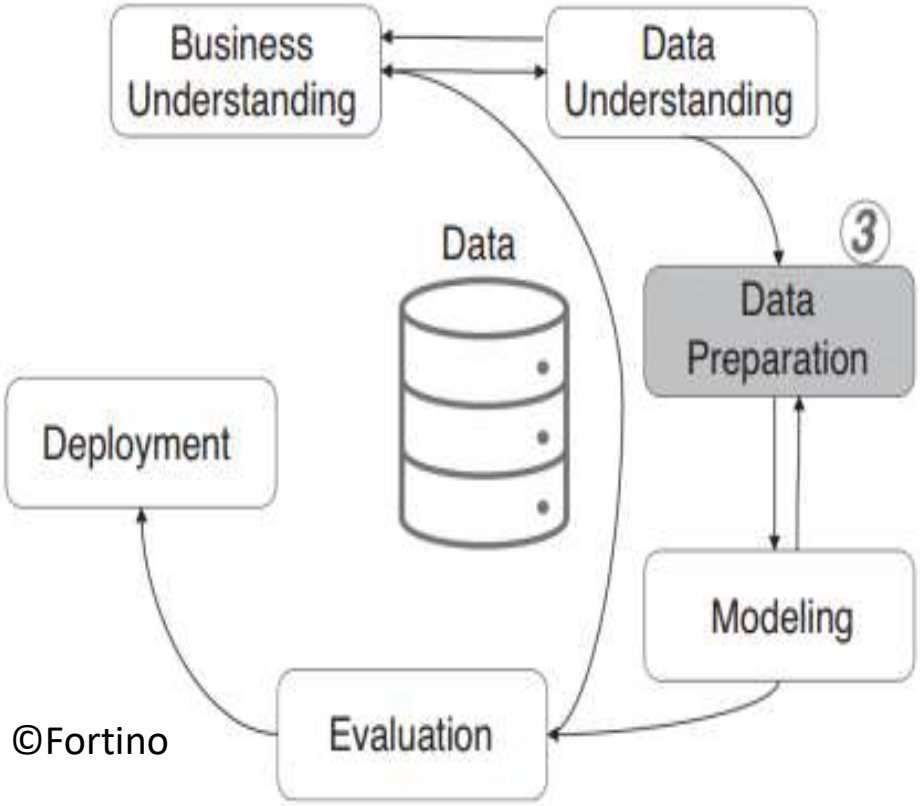


CRISP-DM

- La fase de **preparación de datos** cubre todas las actividades necesarias para construir el conjunto de datos final a partir de los datos iniciales sin procesar.
- Es probable que las tareas de preparación de datos se realicen varias veces y no en ningún orden prescrito.
- Las tareas incluyen **selección de tablas**, registros y atributos, así como transformación y **limpieza de datos** para herramientas de modelado.

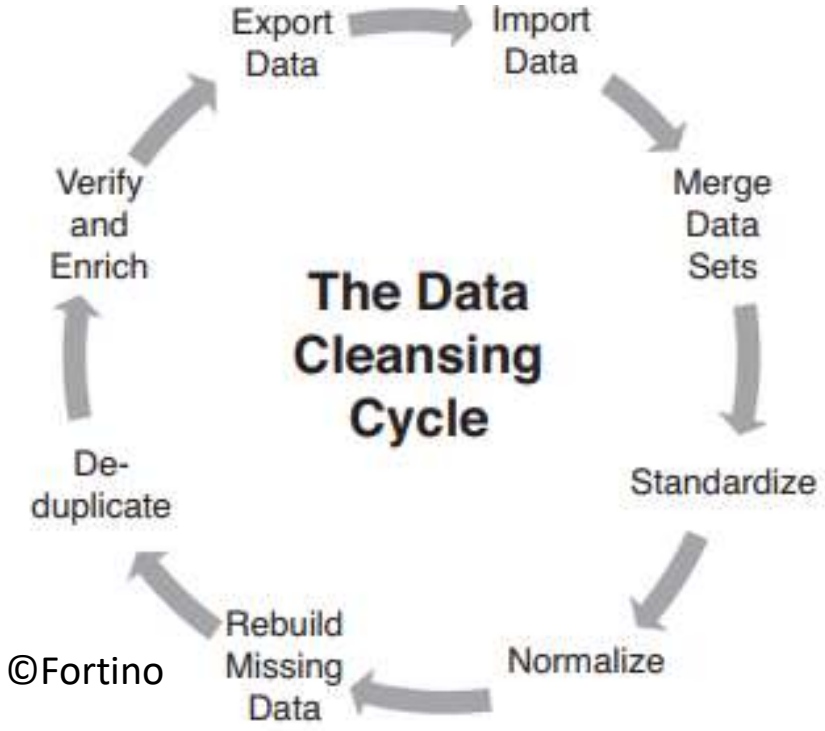
Preparación de datos

- A **preparação de dados** é uma etapa crucial no processo de mineração de dados, muitas vezes consumindo mais tempo do que as próprias técnicas de mineração.
- É o processo de transformar dados brutos em um formato limpo, consistente e adequado para análise.
- Sem uma preparação adequada, os resultados da mineração de dados podem ser distorcidos e inúteis.



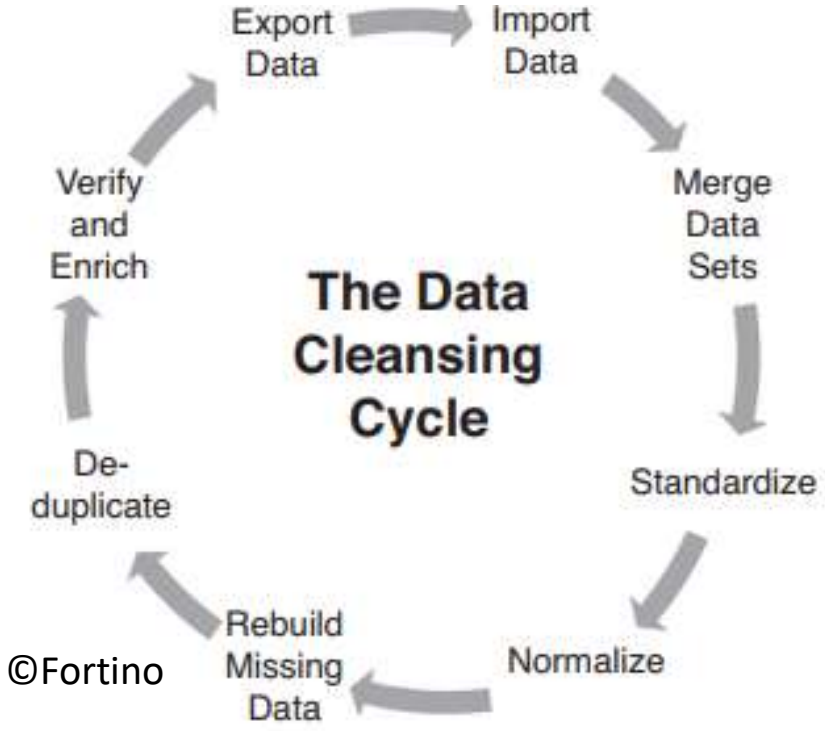
Ciclo de limpieza de datos

- La Figura ilustra el **ciclo de limpieza de datos**, con varias actividades para preparar los datos para el análisis.
- Las actividades incluyen **importación**, **fusión**, **estandarización**, **normalización**, **reconstrucción de datos faltantes**, **eliminación de duplicados** y **verificación/enriquecimiento** del dataset.

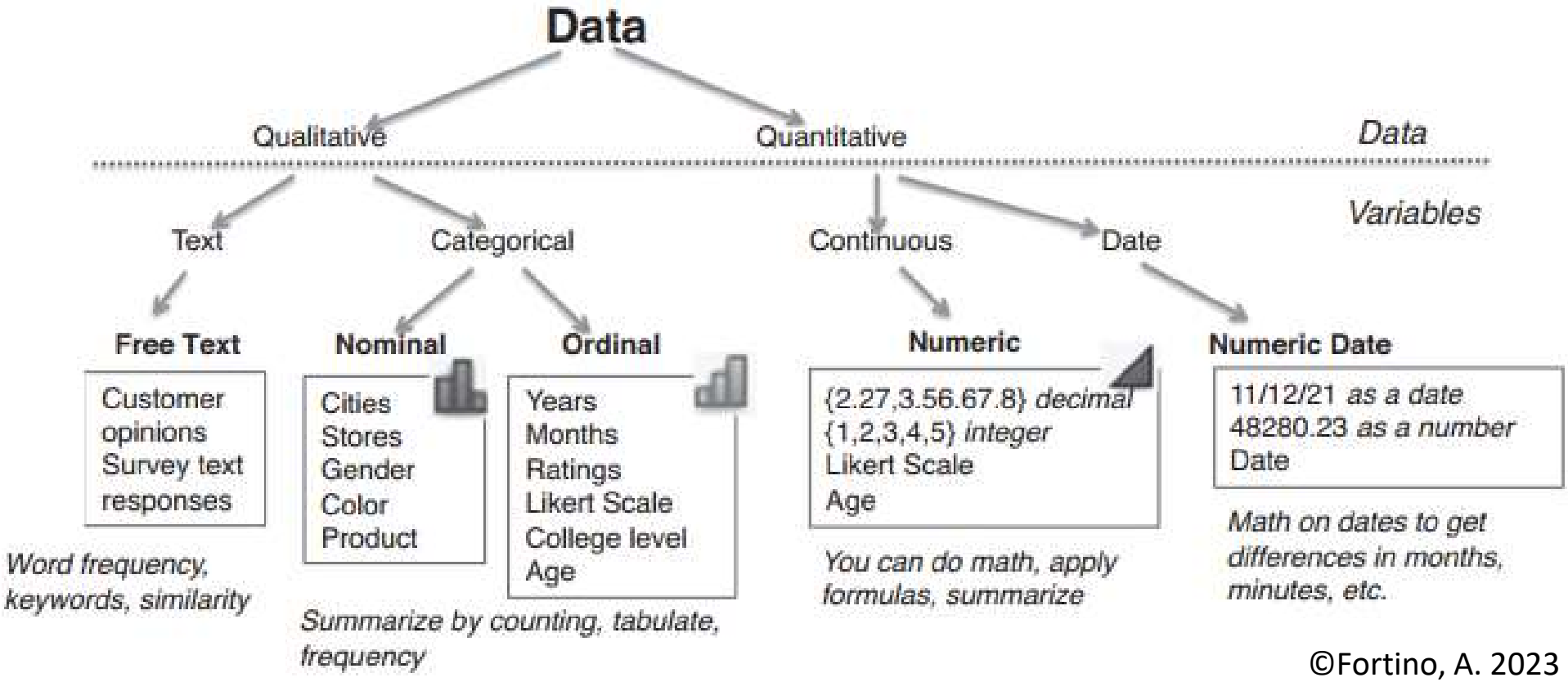


Formato de archivo plano

- El **objetivo de la limpieza** es transformar los datos en un formato de **archivo plano**.
- La primera fila de la tabla contiene nombres de las variables, cada fila tiene la misma naturaleza y no tiene líneas o columnas vacías.



Fuentes y formatos de datos



Importancia

- La **calidad de los datos** impacta directamente la **calidad** de los resultados de la minería.
- La **preparación de datos** garantiza que los datos sean confiables, consistentes y adecuados para el análisis.
- **Elimina inconsistencias**, corrige errores y maneja datos faltantes, lo que genera información más precisa y confiable.

PLanilha1: Bancos

	A	B
1	**Código**	**Banco**
2	1	Banco do Brasil S.A.

Planilha 2: Lançamentos

	A	B	C	D
1	**Código Lançamento**	**Descrição**	**Tipo**	**Sigla**
2	A1	Saque	Saída	S
3	A2	Pagamento Boleto	Saída	S
4	A3	Pagamento Cheque	Saída	S
5	A4	Pagamento DOC	Saída	S
6	A5	Pagamento Salário	Saída	S
7	B1	Depósito Dinheiro	Entrada	E
8	B2	Depósito Cheque	Entrada	E
9	B3	Entrada TED	Entrada	E
10	B4	Entrada DOC	Entrada	E
11	B5	Liquidação Boleto	Entrada	E
12				

Planilha 3: Fluxo

B	C	D
Código	**Lançamento**	**Valor**
492	B1	-4.810,00
79	A3	3.640,00
749	B1	-5.008,00
746	B3	-8.098,00
751	B4	-3.548,00
233	B5	-1.383,00
208	B3	-7.496,00

Etapas

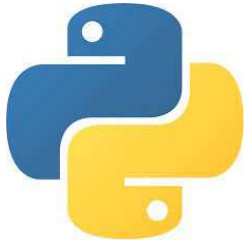
- **Recopilación de datos**: identificar y recopilar datos relevantes de diversas fuentes, como bases de datos, archivos, API y sensores.
- **Limpieza de datos**: corregir errores, inconsistencias y valores faltantes.
- **Integración de datos**: combinar datos de diferentes fuentes en un único conjunto de datos coherente.
- **Transformación de datos**: formatear datos en un formato adecuado para el análisis, como normalización y estandarización de valores.
- **Reducción de dimensionalidad**: Seleccionar los atributos más relevantes para el análisis y eliminar redundancias.

Técnicas

- **Detección y manejo de valores faltantes**: imputación de valores faltantes basándose en **métodos estadísticos** como la media, la mediana o la regresión.
- **Manejo de valores atípicos (outliers)**: identificar y eliminar valores atípicos que puedan distorsionar el análisis.
- **Detección y corrección de errores**: Identificar y corregir errores tipográficos, inconsistencias y valores no válidos.
- **Normalización de datos**: Estandarizar la escala de valores para facilitar la comparación entre diferentes atributos.
- **Transformación de datos**: Aplicar transformaciones matemáticas para mejorar la linealidad, normalidad u homogeneidad de los datos.

Herramientas

- Herramientas de Business Intelligence (BI) y análisis de datos, como Tableau, Power BI y QlikView.
- Herramientas de manipulación de datos como Python, R y SAS.
- Herramientas específicas para la preparación de datos, como OpenRefine, Trifacta y DataCleaner.



Consideraciones finales

- La preparación de datos es un **proceso iterativo** que se puede revisar a medida que avanza el análisis.
- La elección de **técnicas** y **herramientas** depende del tipo de datos, el objetivo del análisis y los recursos disponibles.
- Es importante **documentar el proceso de preparación de datos** para garantizar la **reproducibilidad** de los resultados.

Resumen

- La preparación de datos es una inversión fundamental para garantizar una extracción de datos exitosa.
- Al dedicar tiempo y atención a este paso crucial, puede obtener información más precisa y confiable a partir de sus datos, lo que generará mejores decisiones y resultados para su negocio.