CIENCIA DE DATOS PARA DECISIONES DE NEGOCIO

Diapositivas 02 · Missing Data: MCAR, MAR y MNAR

Interpretar por qué faltan los datos antes de eliminarlos, imputarlos o convertirlos en señal.

Módulo 02 · Calidad de datos y valores faltantes

Material elaborado por el profesor Sergio Gevatschnaider

IDEA CLAVE

Missing data no significa lo mismo en todos los casos

El valor faltante puede ser azar, estructura del proceso o señal del propio valor.

El porcentaje de faltantes no explica por qué faltan.

Dos bases con 25 % de ausencia pueden tener riesgos opuestos: una puede ser casi neutral y otra profundamente sesgada.



Ausencia estructural vs. omisión problemática

Antes de imputar hay que interpretar el significado del vacío.

Ausencia estructural

- No existe el atributo para esa observación.
- Ejemplo: una casa sin pileta no tiene calidad de pileta.
- Debe codificarse como información semántica.
- Puede ser una señal predictiva válida.

Omisión problemática

- El atributo existe pero no fue registrado.
- Ejemplo: ingreso no declarado en una solicitud.
- Puede concentrarse por segmento o valor.
- Requiere hipótesis y sensibilidad.

COMPARACIÓN

MCAR, MAR y MNAR

Tres hipótesis de trabajo, tres riesgos distintos.

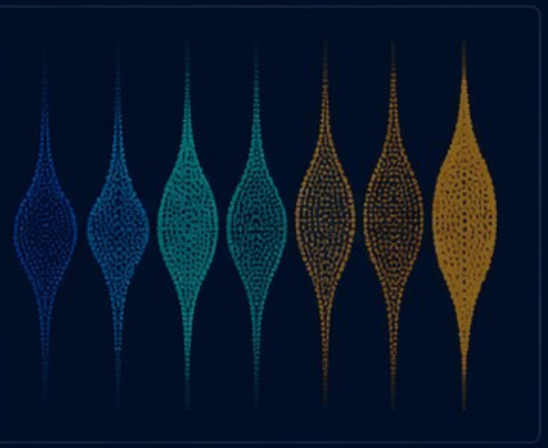
Mecanismo	Lectura	Riesgo
MCAR	Falta completamente al azar.	Pérdida de precisión, poco sesgo sistemático.
MAR	Falta según variables observadas.	Sesgo si se ignora el grupo condicionante.
MNAR	Falta según el valor no observado.	Sesgo severo difícil de corregir sin supuestos.
Estructural	El atributo no aplica.	Error si se imputa como dato perdido.
Operativo	Falla de captura o integración.	Puede indicar proceso roto.
Estratégico	No respuesta deliberada o sensible.	Puede revelar asimetría de información.

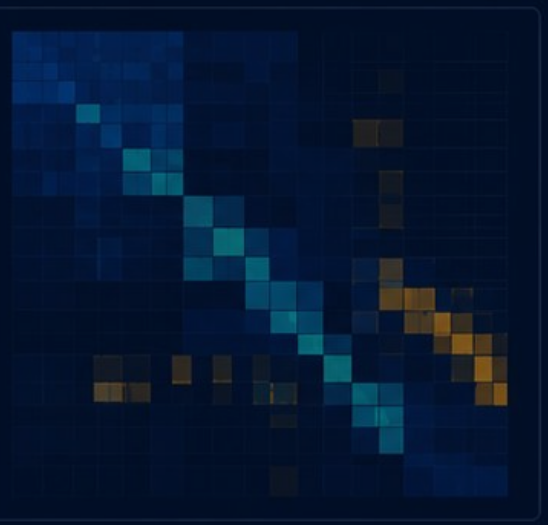
CONCEPTO

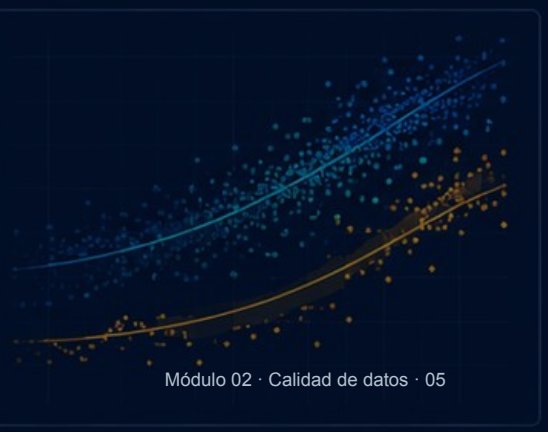
Diagnóstico visual del patrón

La matriz de ausencia permite ver estructura antes de calcular.

- MCAR suele parecer disperso, sin concentración clara por grupo o rango.
- MAR aparece concentrado en segmentos observables: canal, región, perfil, producto.
- MNAR puede alinearse con extremos: valores altos, bajos, riesgosos o sensibles.
- La visualización no prueba el mecanismo, pero ayuda a formular hipótesis.







Ejemplo empresarial: crédito

Los ingresos no faltan igual para todos los solicitantes.

Lectura ingenua

- Eliminar registros incompletos.
- Calcular media observada.
- Entrenar un score con casos completos.
- Concluir que la base “mejoró”.

Lectura profesional

- Medir faltantes por tipo de empleo.
- Comparar aprobación por segmento.
- Testear imputación por grupo.
- Reportar impacto en riesgo y equidad.

SECUENCIA

Cómo formular una hipótesis de ausencia

La hipótesis debe escribirse antes del tratamiento.

1. Identificar variable faltante y decisión afectada.
2. Medir ausencia total y por segmentos relevantes.
3. Comparar distribución de variables observadas entre completos e incompletos.
4. Proponer mecanismo probable: MCAR, MAR, MNAR o estructural.
5. Definir qué evidencia refutaría esa hipótesis.

COMPARACIÓN

Qué hacer según el mecanismo

La decisión de tratamiento debe ser coherente con la hipótesis.

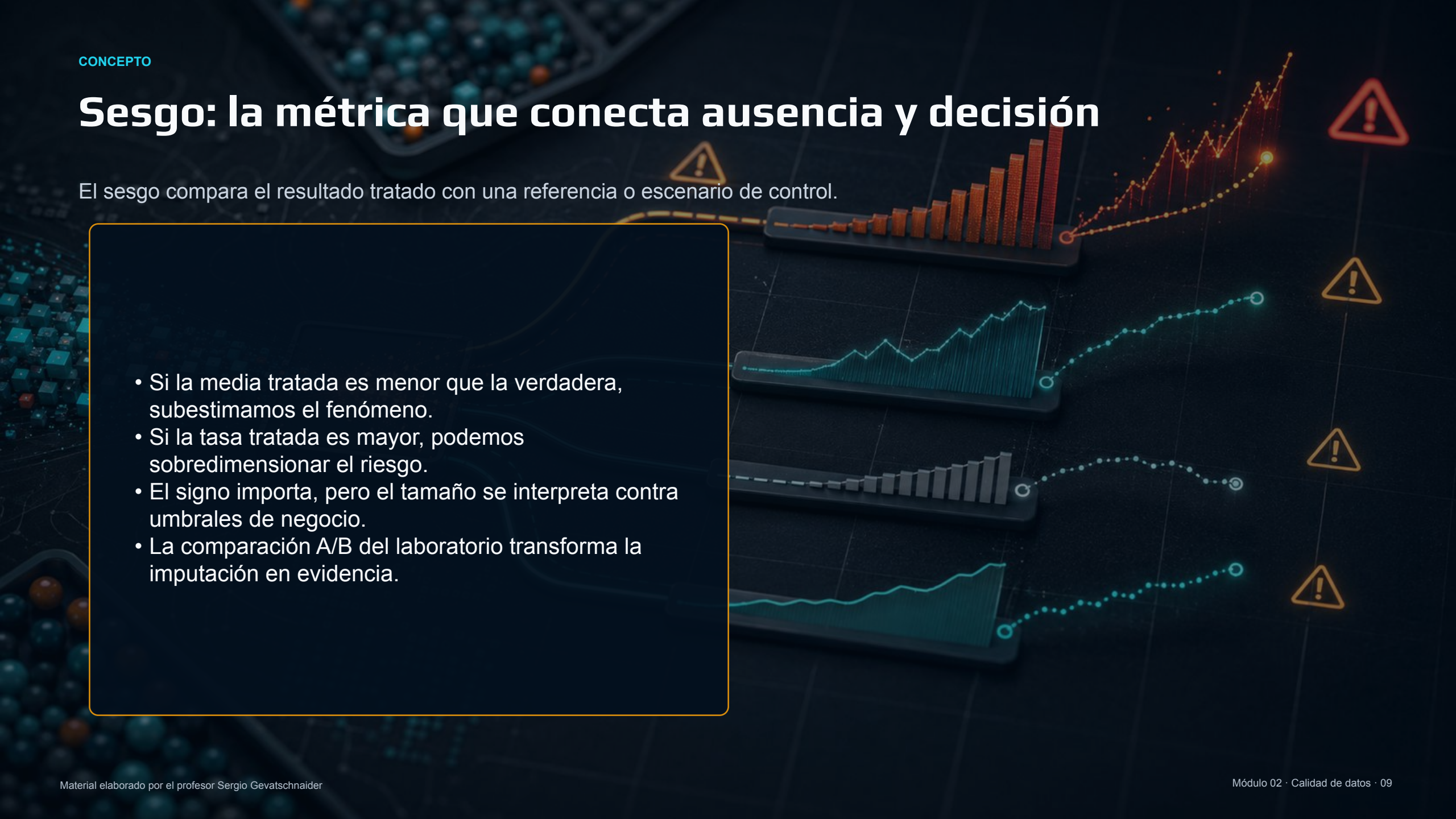
Hipótesis	Tratamiento posible	Límite
MCAR	Eliminar o imputar simple con sensibilidad.	Pierde tamaño muestral.
MAR	Imputar por grupo o modelo usando variables observadas.	Depende de variables explicativas completas.
MNAR	Escenarios, cotas, recolección adicional.	No se resuelve solo con técnica.
Estructural	Crear categoría/indicador de no aplica.	No confundir con error.
Operativo	Corregir proceso de captura.	Puede requerir rediseño del sistema.
Sensible	Diseñar captura ética y segura.	El valor ausente puede ser señal.

CONCEPTO

Sesgo: la métrica que conecta ausencia y decisión

El sesgo compara el resultado tratado con una referencia o escenario de control.

- Si la media tratada es menor que la verdadera, subestimamos el fenómeno.
- Si la tasa tratada es mayor, podemos sobredimensionar el riesgo.
- El signo importa, pero el tamaño se interpreta contra umbrales de negocio.
- La comparación A/B del laboratorio transforma la imputación en evidencia.



IDEA CLAVE

Principio de cierre

El mecanismo no se “descubre” automáticamente: se argumenta con evidencia.

Missing data es una hipótesis sobre el proceso que produjo los datos.

La recomendación debe separar lo observado, lo supuesto y lo que falta recolectar para reducir incertidumbre.