End-to-end ML

Guillem Sitges i Puy

Annex III

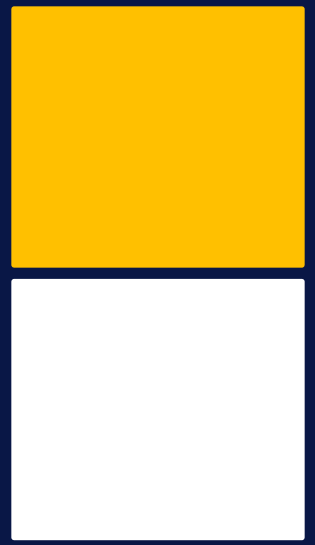
Métodos de rebalanceo

Datasets desbalanceados

A partir de <20% puede haber problemas.
Se notan más cuanto menos registros tenemos.

Por ejemplo, es peor tener 5k registros de la clase minoritatia y que representen un 20% del dataset total que tener 50k registros de la clase minoritaria que representen un 15% del dataset total.

Nuestro objetivo no es llegar al 50%-50% sinó a un balance que solucione lo sufieciente el problema.



Detección del problema

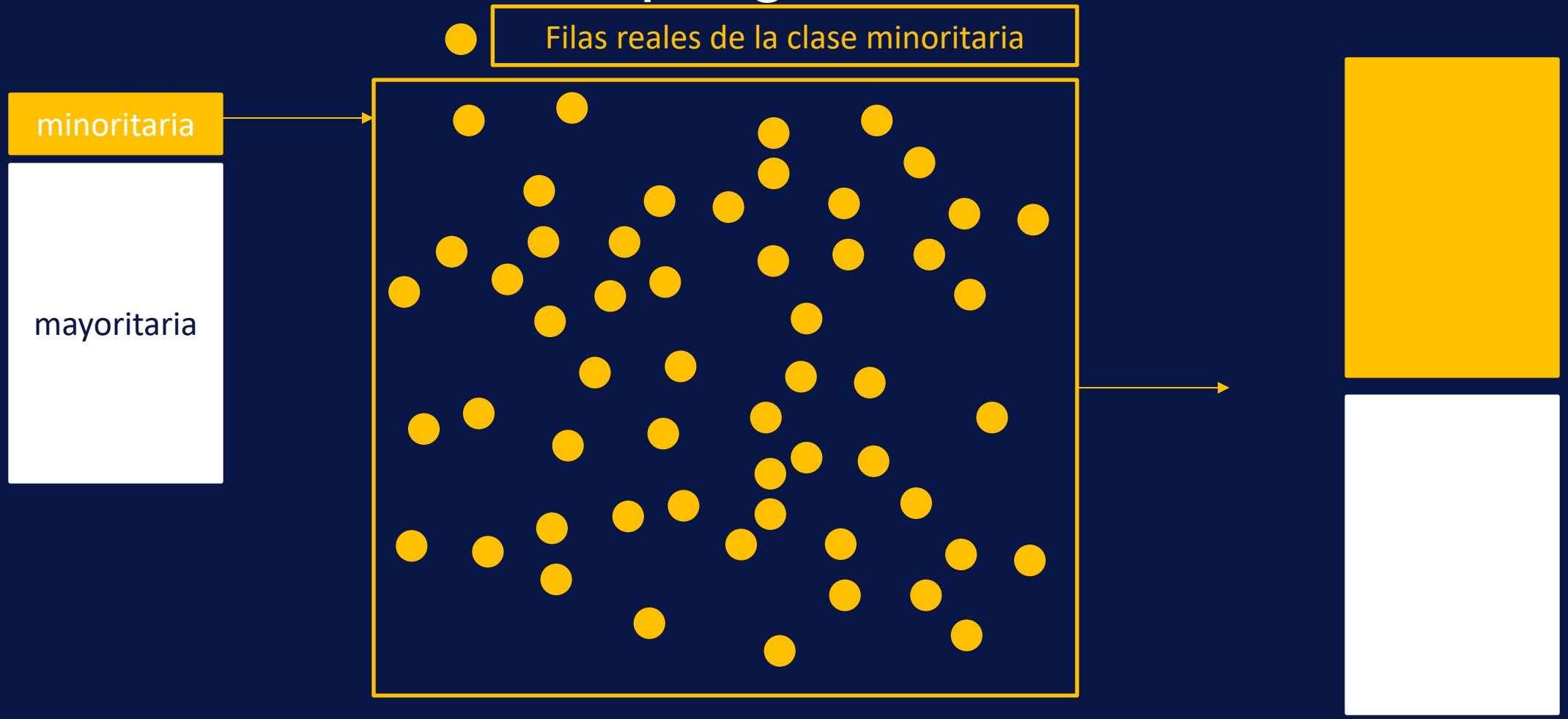
En clasificación:

Mi modelo predice menos ejemplos de los que realmente existen (recall muy baja en caso que la clase minoritaria sean los 1)

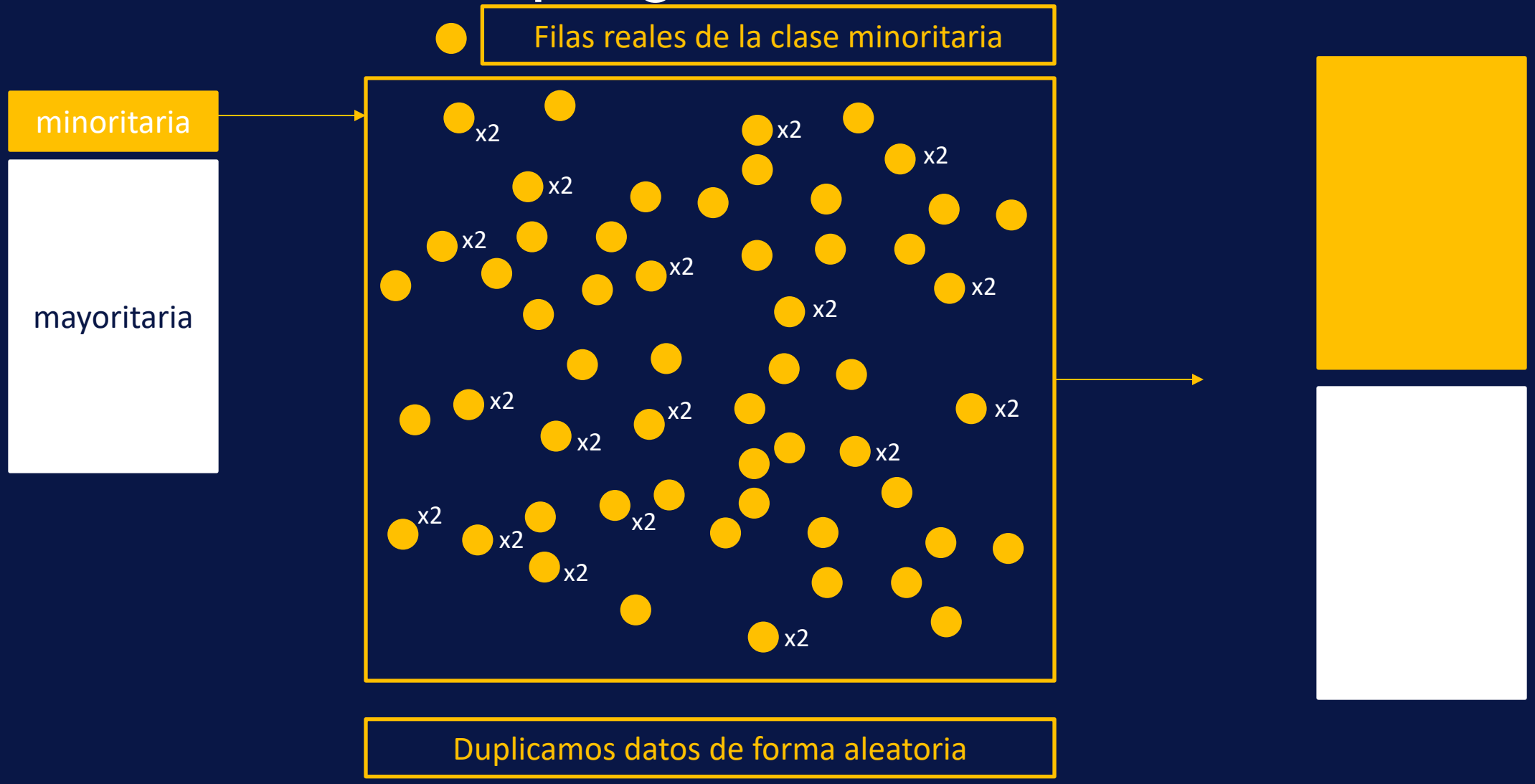
En regressión:

Mi modelo no es capaz de predecir valores de la distribución que aparecen poco.
Por ejemplo, en un target sesgado a la izquierda, casi nunca detecta valores altos

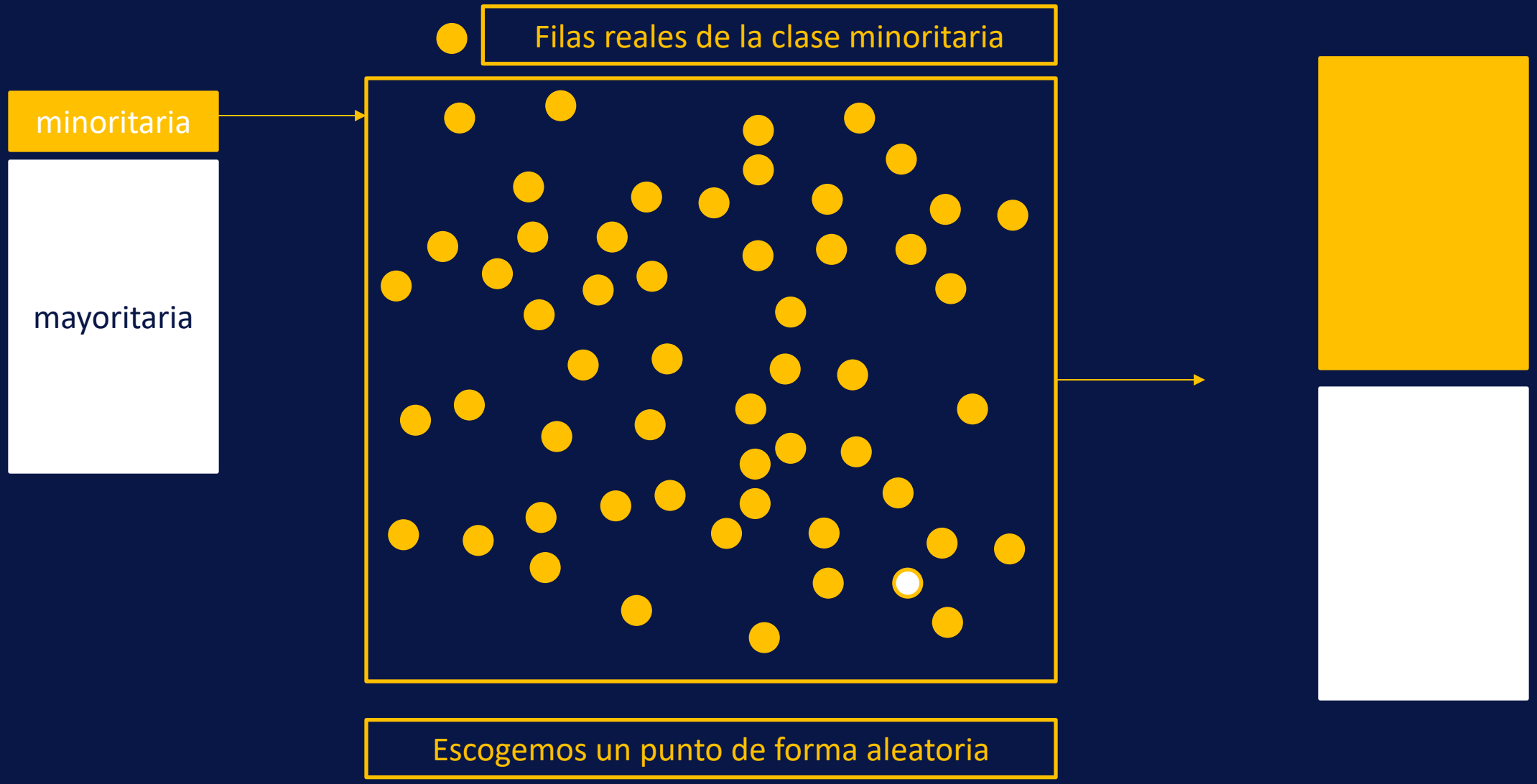
Técnicas de oversampling



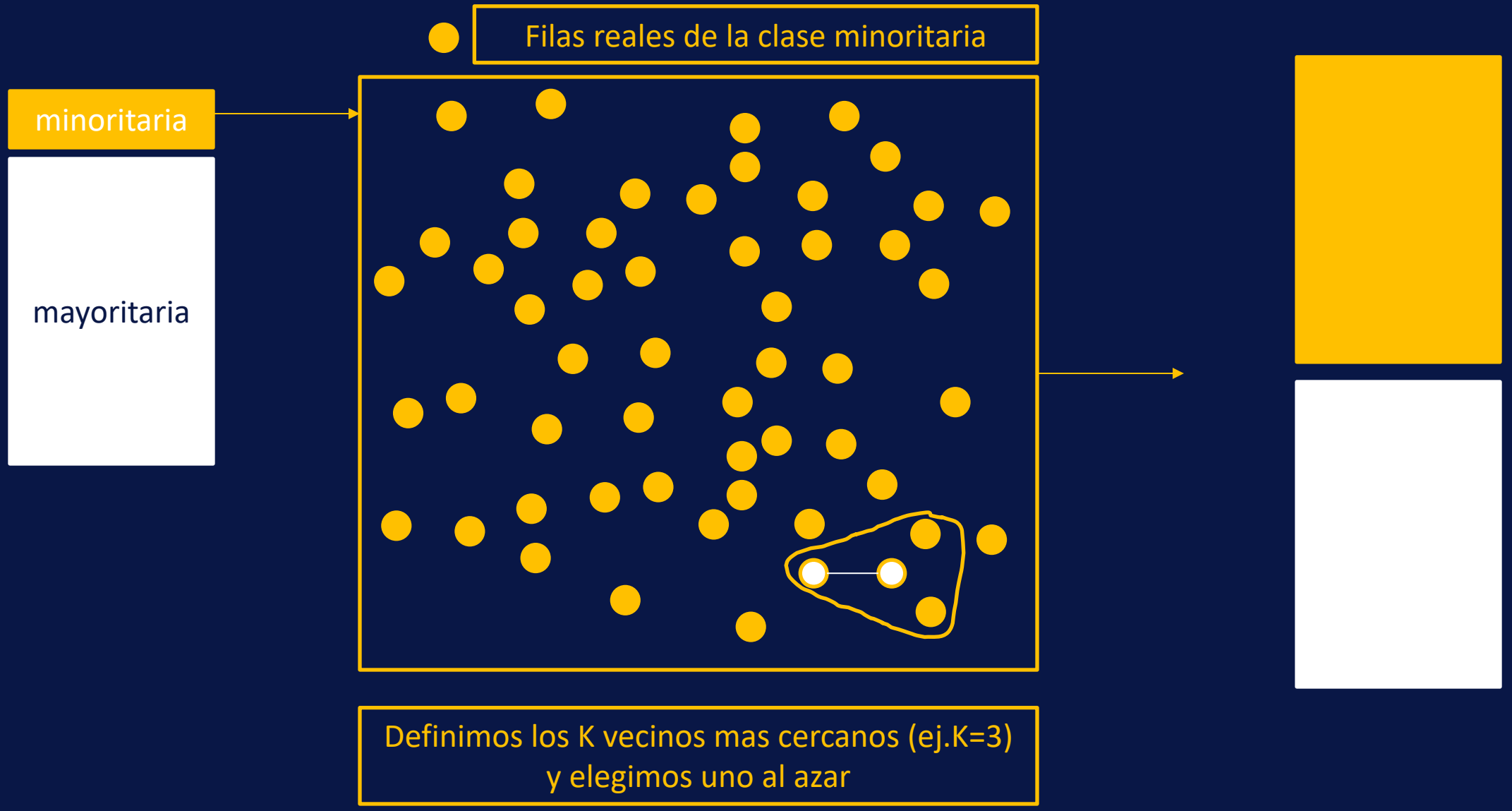
Random Oversampling



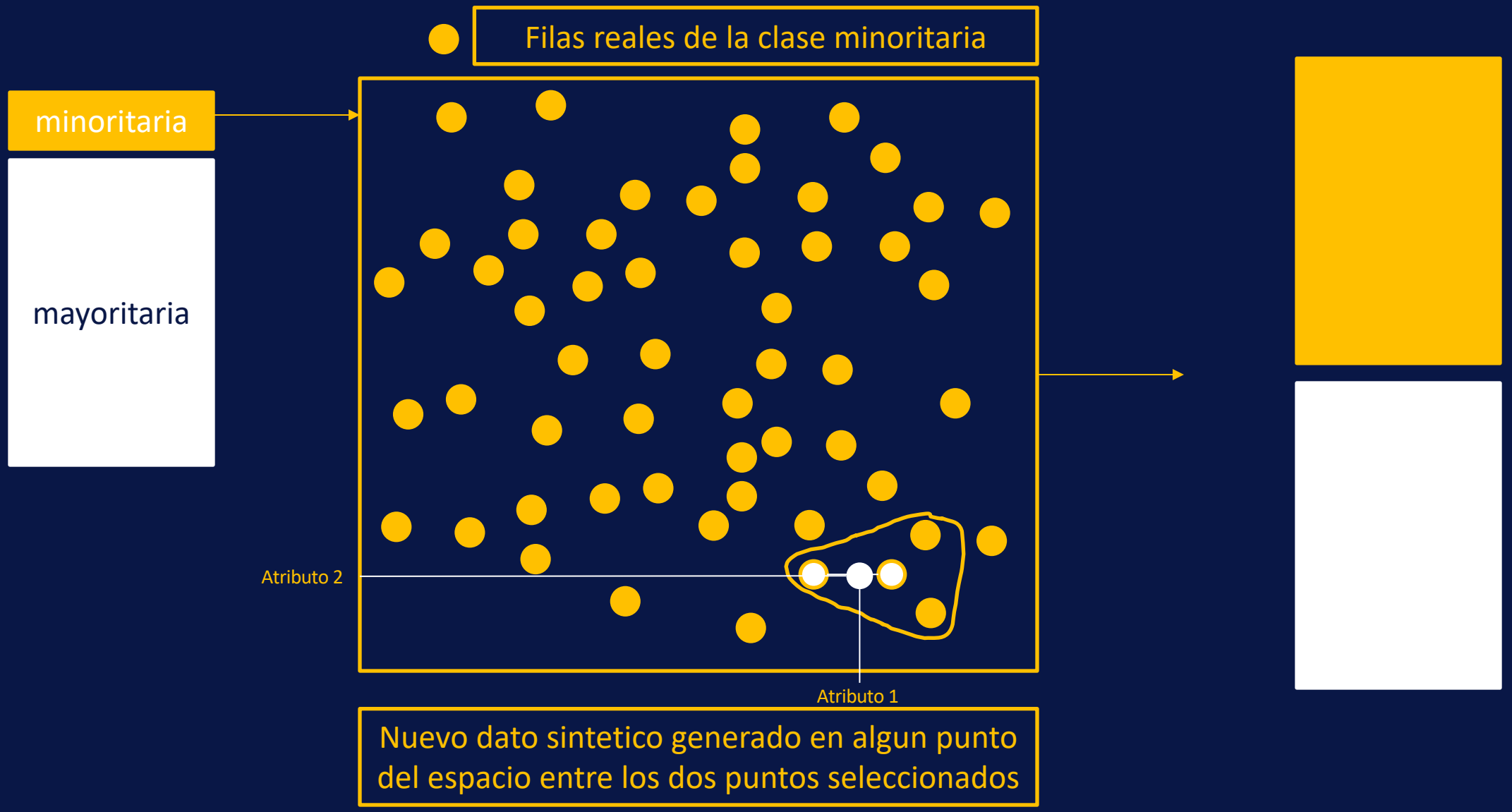
SMOTE


Filas reales de la clase minoritaria
minoritaria
mayoritaria
Escogemos un punto de forma aleatoria

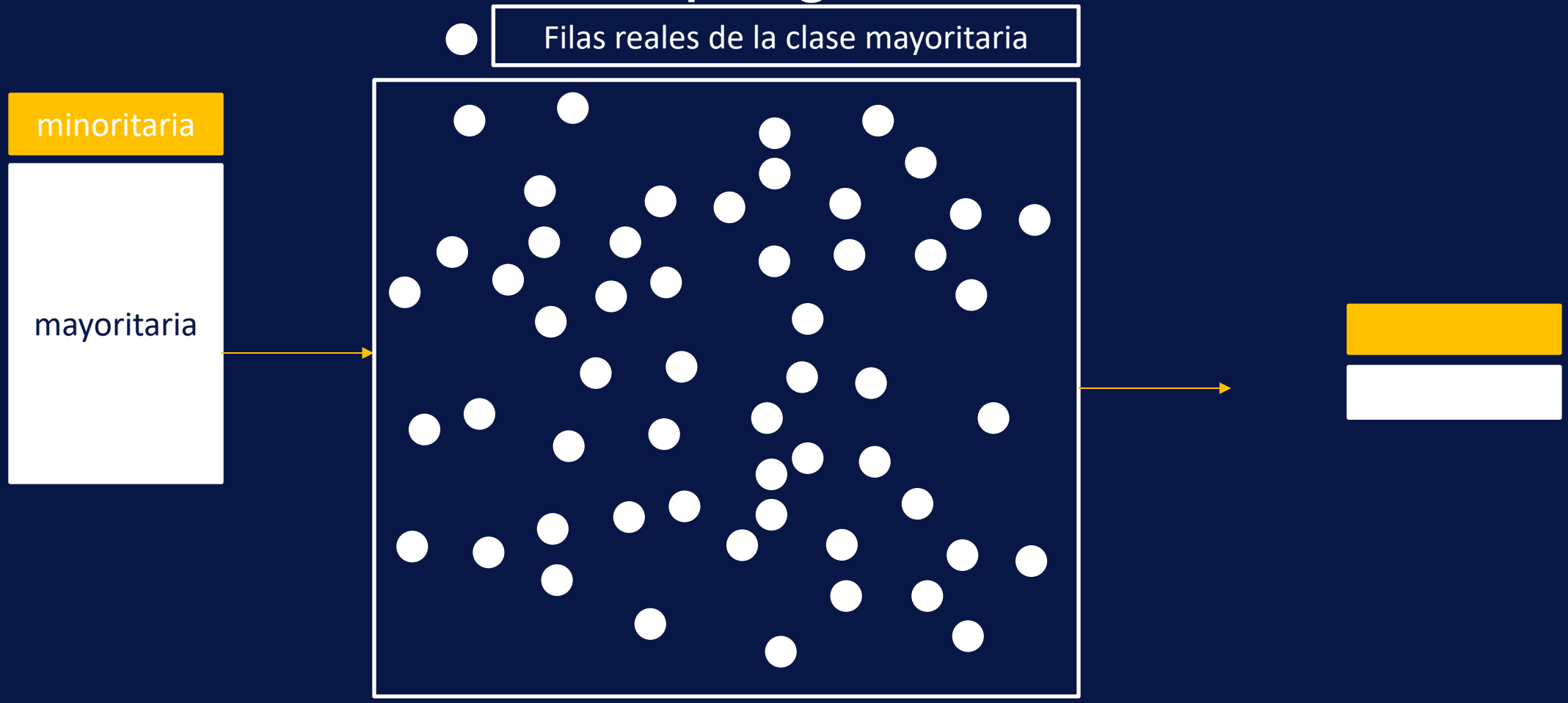
SMOTE


Filas reales de la clase minoritaria
minoritaria
mayoritaria
Definimos los K vecinos mas cercanos (ej.K=3)
y elegimos uno al azar

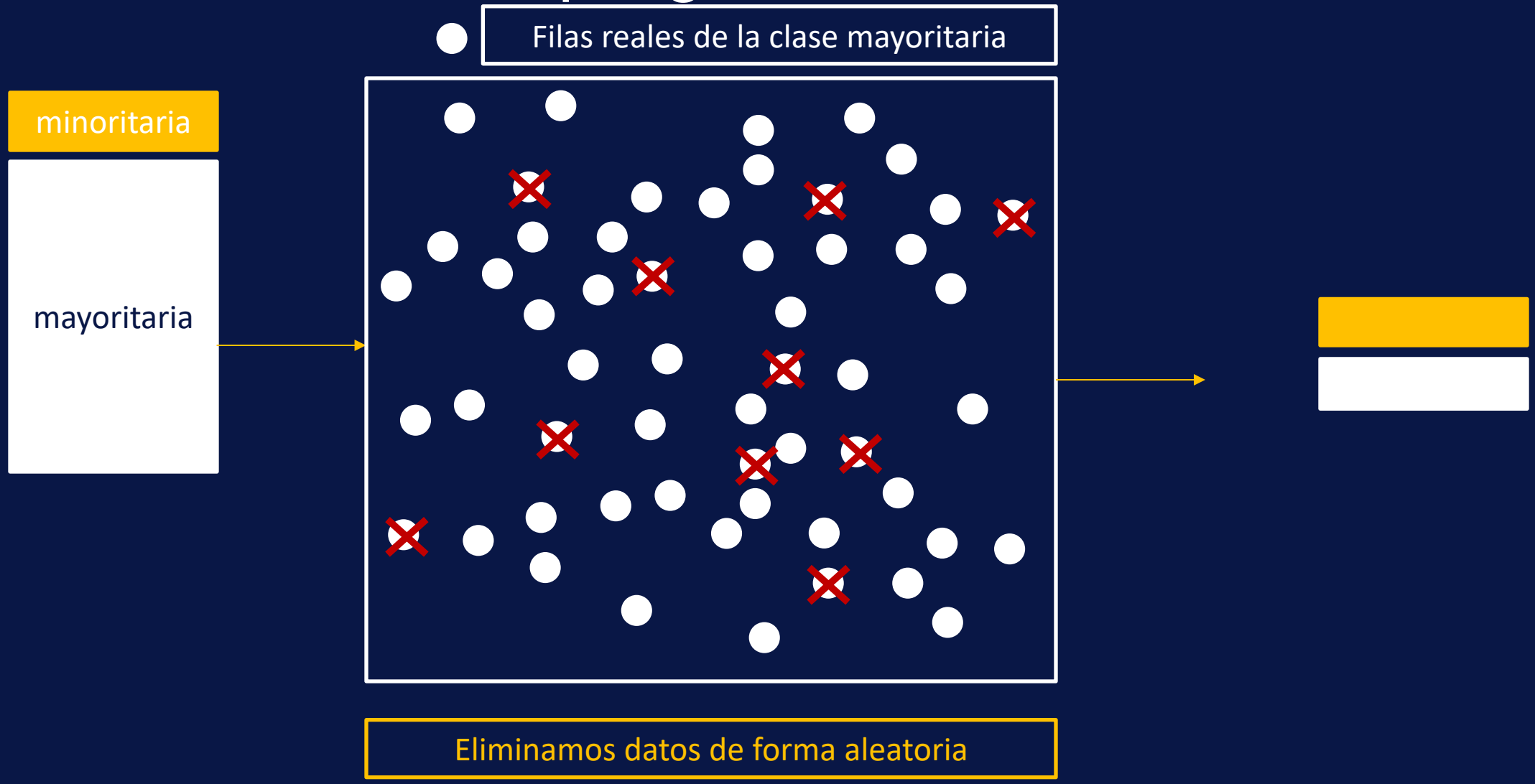
SMOTE


Filas reales de la clase minoritaria
minoritaria
mayoritaria
Atributo 2
Atributo 1
Nuevo dato sintetico generado en algun punto del espacio entre los dos puntos seleccionados

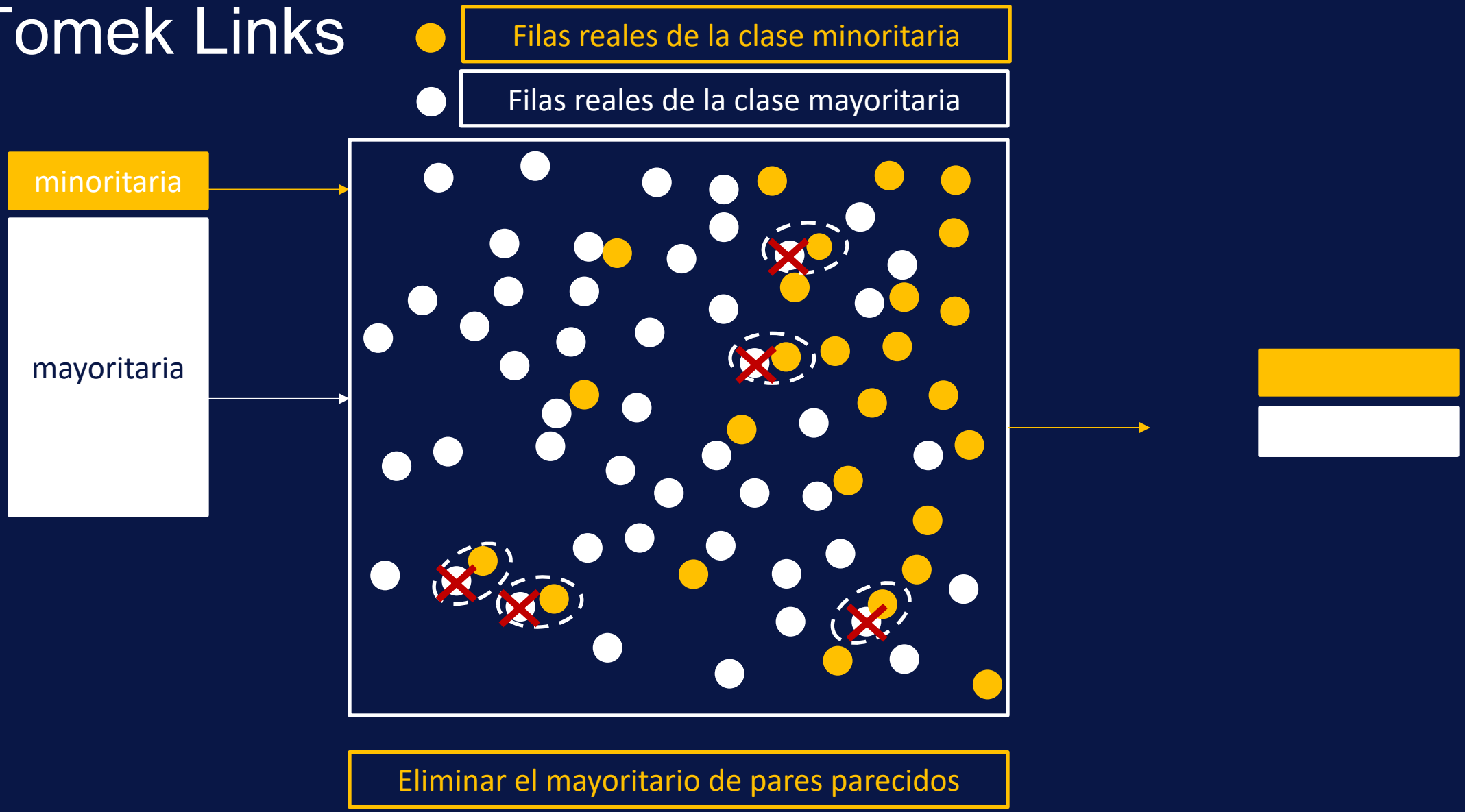
Técnicas de undersampling



Random undersampling



Tomek Links


Filas reales de la clase minoritaria
Filas reales de la clase mayoritaria
minoritaria
mayoritaria
Eliminar el mayoritario de pares parecidos