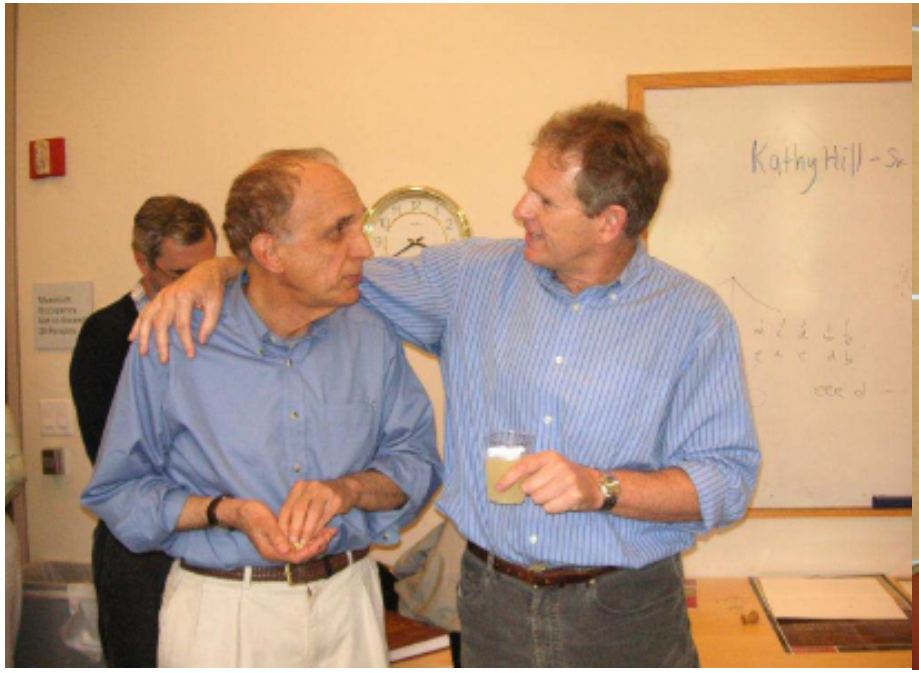
Validación Cruzada (cross-validation) y Remuestreo (bootstrapping)

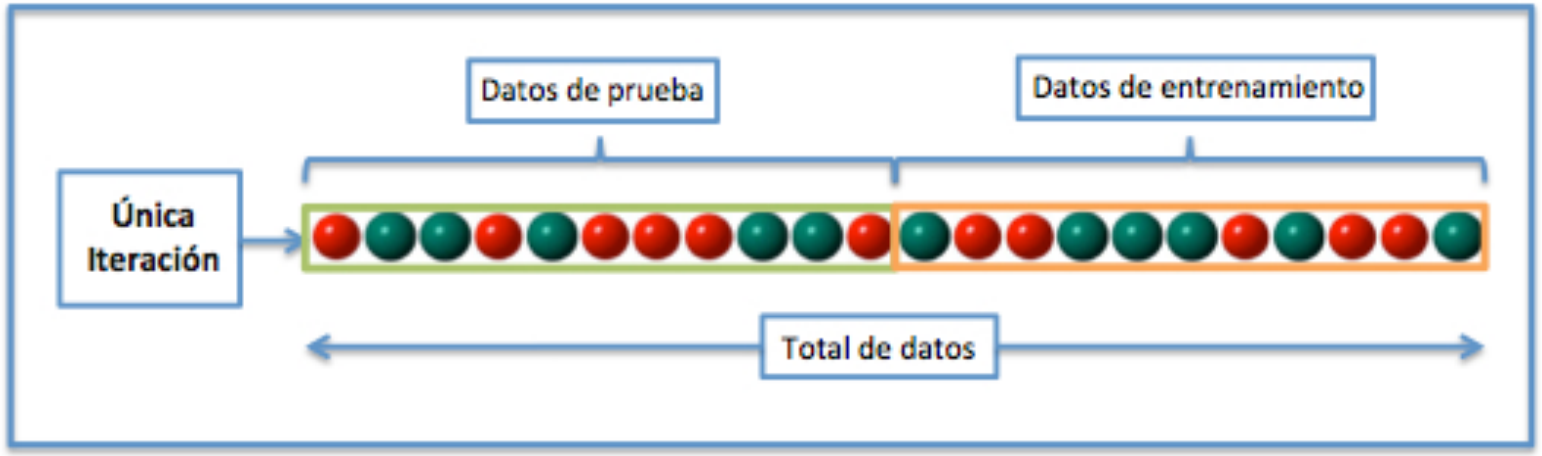
Padres de cross-validation y el bootstrapping
Bradley Efron y *Rob Tibshirani*

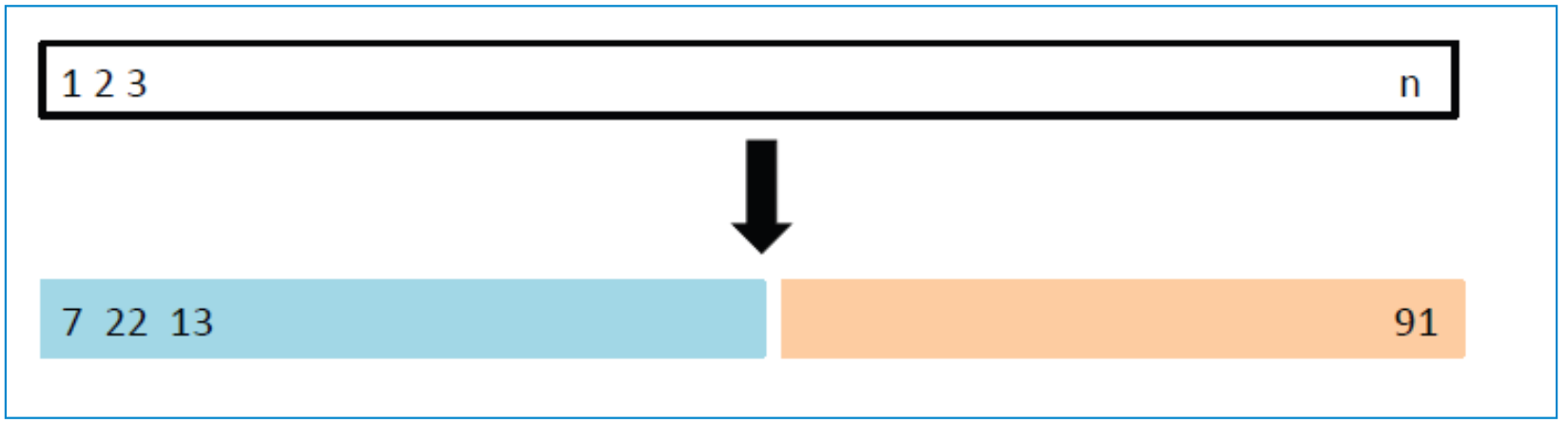


Bradley Efron y Trevor Hastie

Rob Tibshirani

Enfoque: “tabla de aprendizaje y tabla de testing” (the validation test approach)





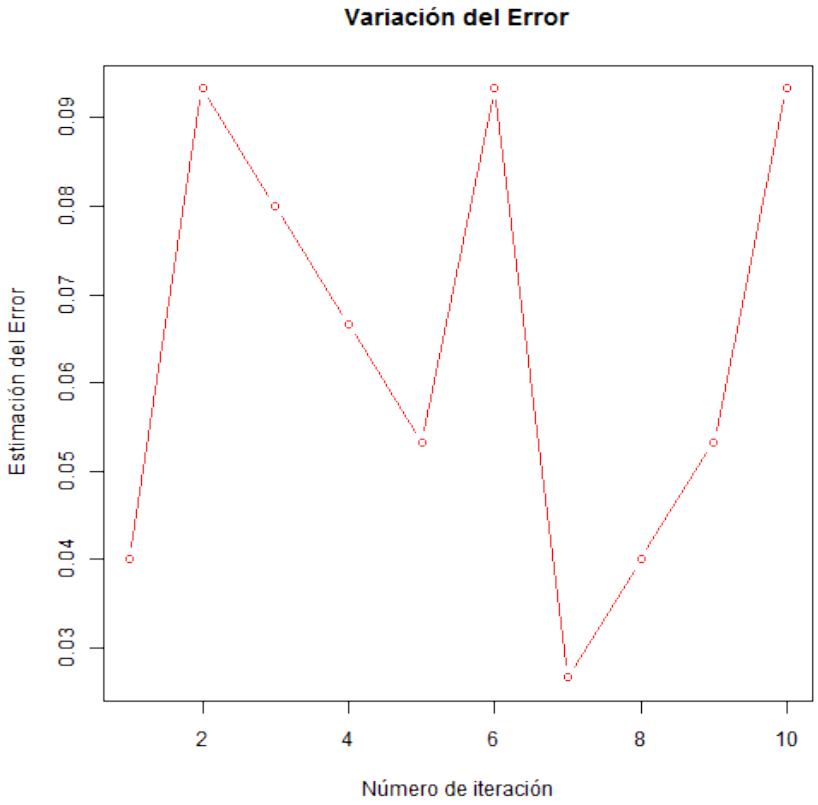
Enfoque: “tabla de aprendizaje y tabla de testing” (the validation test approach)

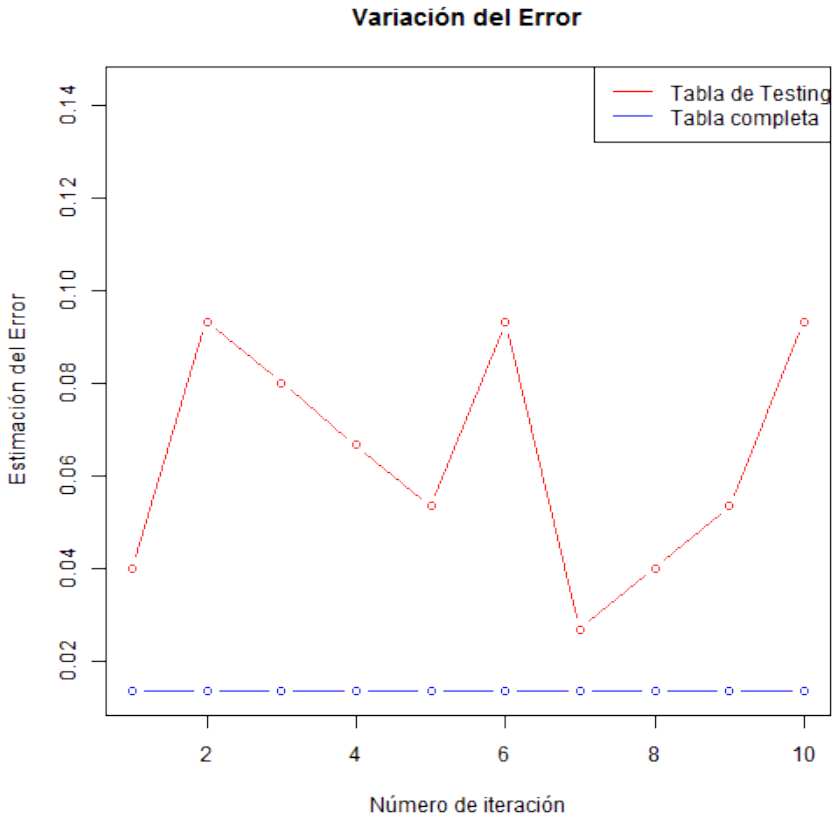
Tiene dos grandes problemas:

1. La estimación del error tiende a ser muy variable dependiendo de cuáles datos quedan en la tabla de aprendizaje y cuáles en la tabla de testing.

2. Se tiende a sobrestimar la estimación del error, es decir, es mucho mayor el error en la tabla de testing que en toda la tabla de datos.

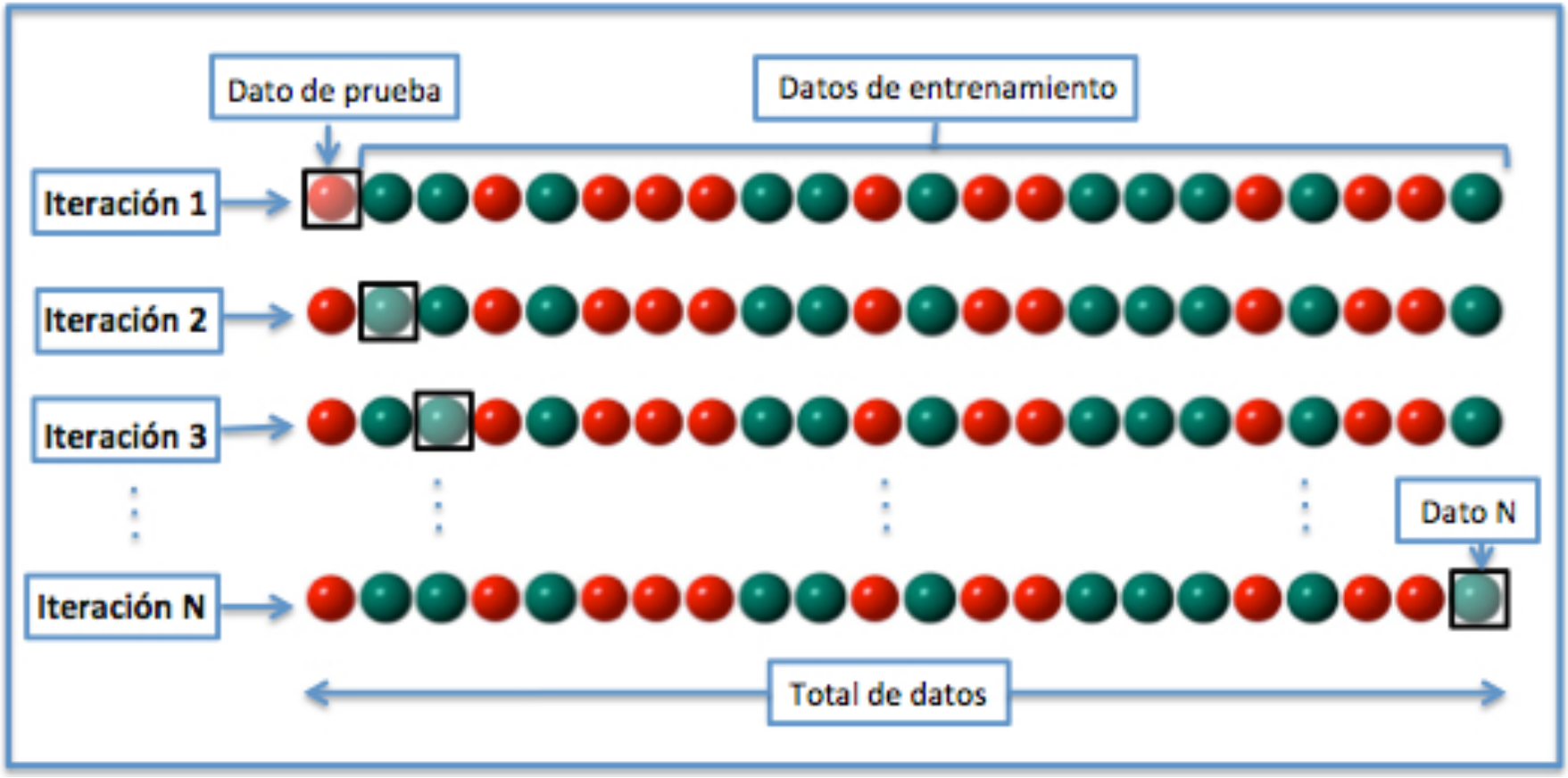
Problemas en : “tabla de aprendizaje y tabla de testing” (the validation test approach)



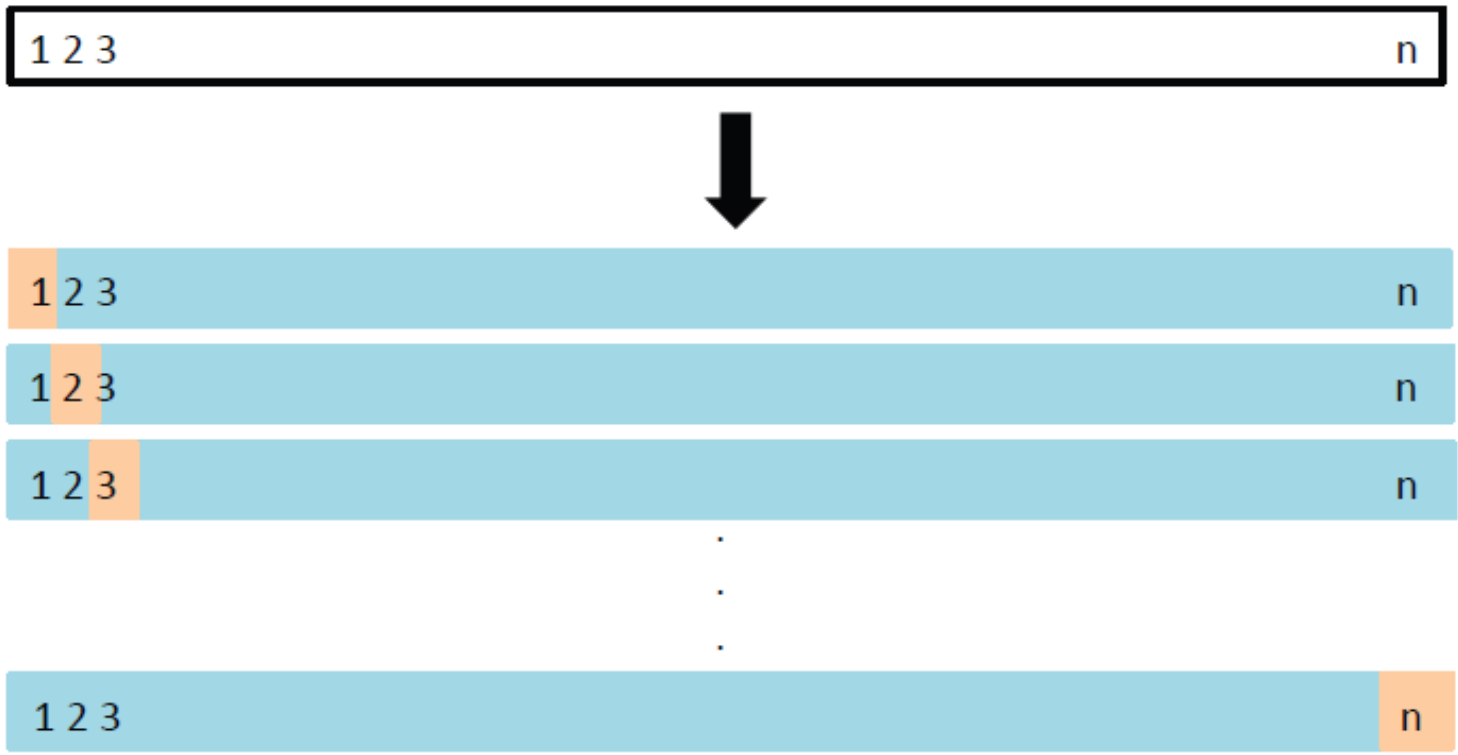


(Ver: Validacion_Cruzada.html)

Validación cruzada dejando uno fuera (Leave-one-out cross-validation LOOCV)



Validación cruzada dejando uno fuera (Leave-one-out cross-validation LOOCV)



Los *n* datos pueden ser tomados en orden.

Validación cruzada dejando uno fuera (Leave-one-out cross-validation LOOCV)

- La validación cruzada dejando uno fuera o Leave-one-out cross-validation (LOOCV) implica separar los datos de forma que para cada iteración tengamos un solo dato de prueba y todo el resto de los datos para entrenamiento.
- El error se calcula como el promedio de los errores cometidos:

Si $\mathrm{MSE}_i = (y_i - \hat{y}_i)^2$ donde $\hat{y}_i$ es la predicción para y_i, entonces se define el error por:

$$\mathrm{CV}_{(n)} = \frac{1}{n} \sum_{i=1}^{n} \mathrm{MSE}_i$$

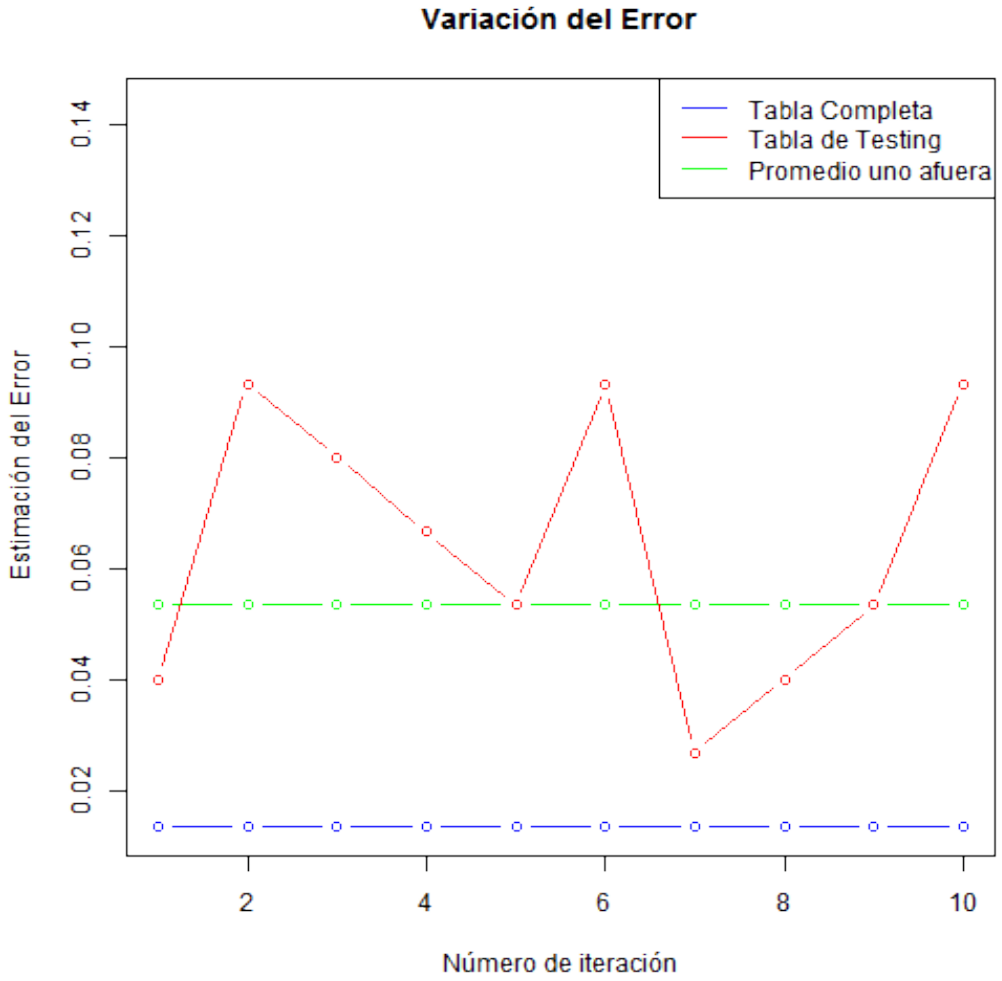
Validación cruzada dejando uno fuera (Leave-one-out cross-validation LOOCV)

Tiene dos ventajas:

1. La estimación del error NO es muy variable dependiendo de cuáles datos quedan en la tabla de aprendizaje y cuáles en la tabla de testing, es decir, la estimación del error es mucho más estable.

2. NO se tiende a sobrestimar el error, es decir, como pasa en el Enfoque: “tabla de aprendizaje y tabla de testing” donde es mucho mayor el error en la tabla de testing que en toda la tabla de datos.

Ventajas: Validación cruzada dejando uno fuera (Leave-one-out cross-validation LOOCV)



(Ver: Validacion_Cruzada.html)

Validación cruzada dejando uno fuera (Leave-one-out cross-validation LOOCV)

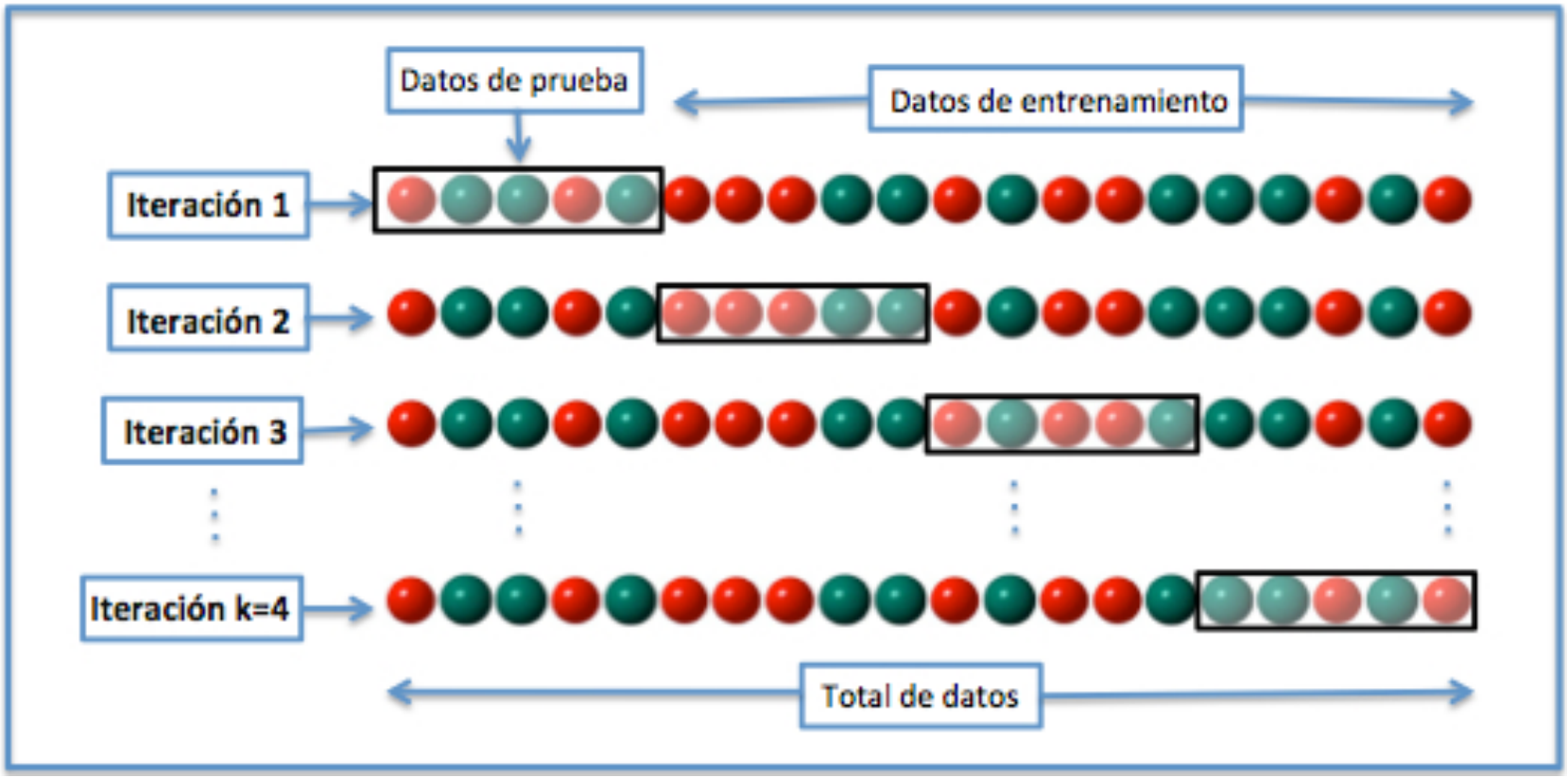
Tiene dos grandes desventajas:

1. La programación se vuelve mucho más complicada.

2. El tiempo de ejecución puede ser muy alto, pues se debe generar el modelo ***n*** veces.

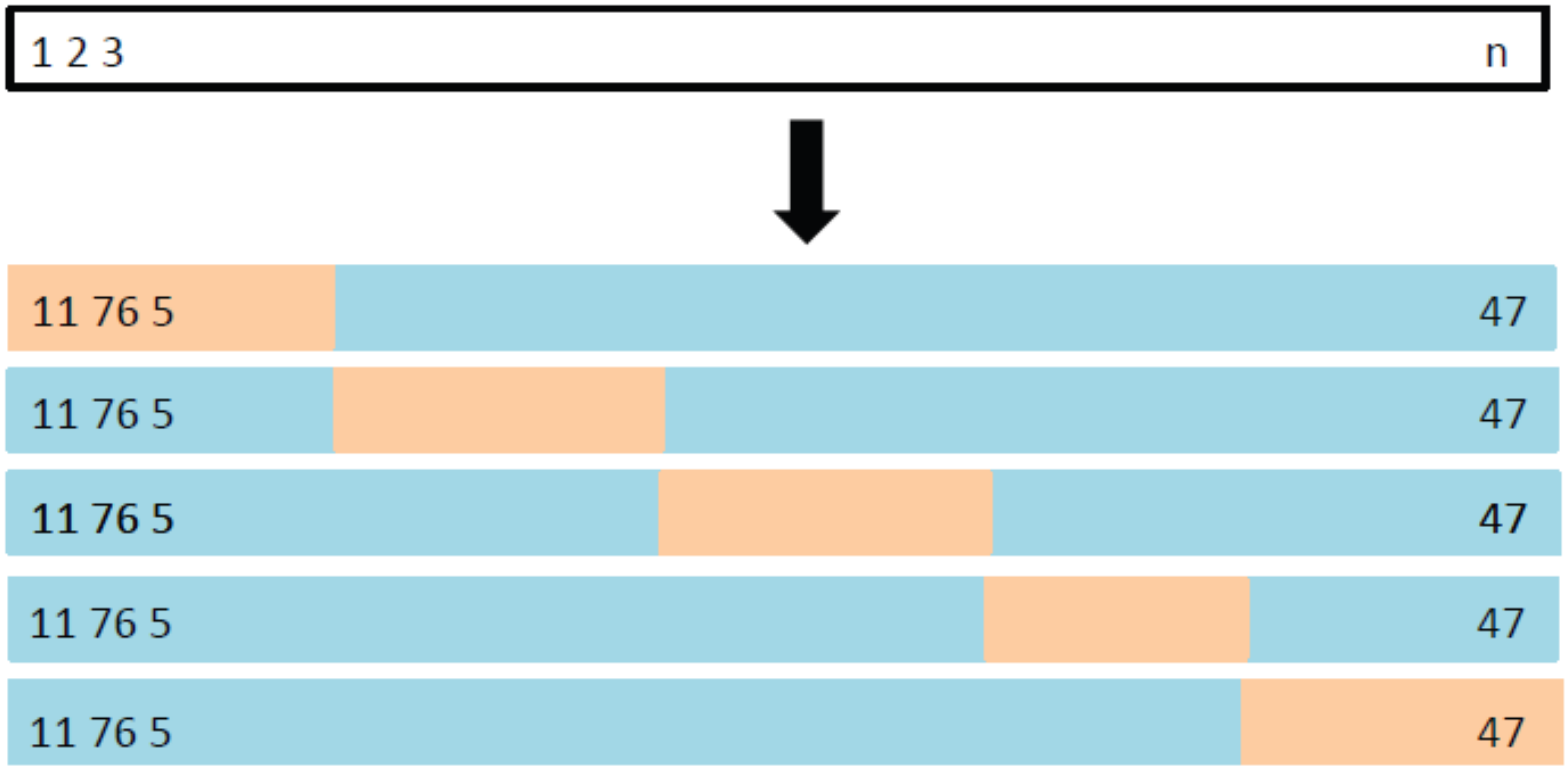
(Ver: Validacion_Cruzada.html)

Validación cruzada usando K grupos (K-fold cross-validation)



K grupos → K iteraciones

Validación cruzada usando K grupos (K-fold cross-validation)



Los K grupos deben ser al azar y aproximadamente del mismo tamaño.

Validación cruzada usando K grupos (K-fold cross-validation)

- En la validación cruzada de K iteraciones o K-fold cross-validation los datos se dividen en K subconjuntos (folds). Uno de los subconjuntos se utiliza como datos de prueba y el resto (K-1) como datos de entrenamiento.

- El proceso de validación cruzada es repetido durante K iteraciones, con cada uno de los posibles subconjuntos de datos de prueba.

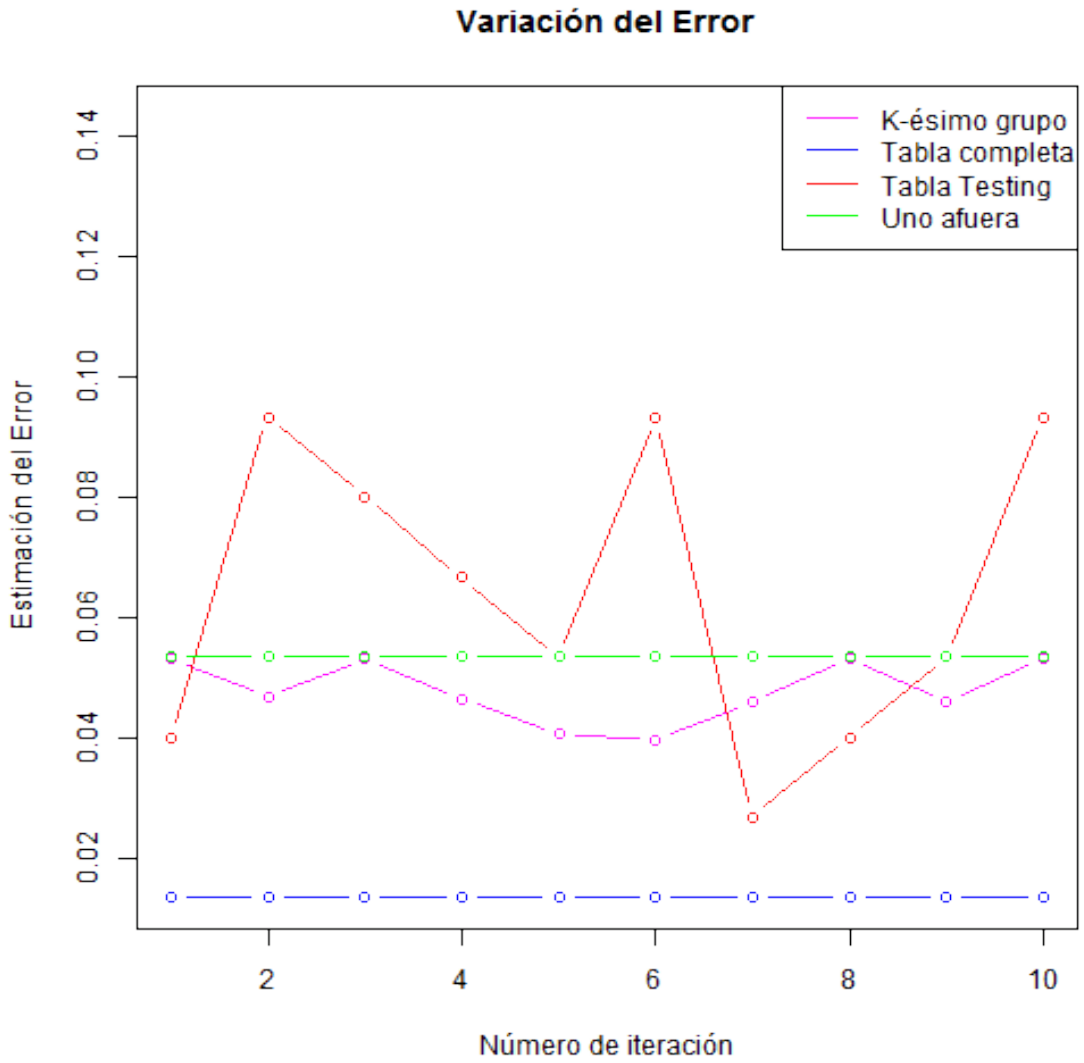
Validación cruzada usando K grupos (K-fold cross-validation)

- El error se calcula como la media aritmética de los errores de cada iteración para obtener un único resultado.

Si MSE_i denota el error en la iteración $i-$ésima, entonces:
El error de la Valización Cruzada se estima por:

$$\mathrm{CV}_{(k)} = \frac{1}{k}\sum_{i=1}^{k}\mathrm{MSE}_i$$

Validación cruzada usando K grupos (K-fold cross-validation)



(Ver: Validacion_Cruzada.html)

Remuestreo (Bootstrap)

"The bootstrap was introduced in 1979 by Bradley Efron"

- En Estadística y Minería de Datos, "bootstrapping" es usado para ***cuantificar la incertidumbre*** asociada con un estimador estadístico.

- "Bootstrap" es en general una herramienta para evaluar la precisión estadística.

- La idea básica es dado un conjunto datos de entrenamiento extraer de esta tabla aleatoriamente y con reemplazo nuevas tablas de datos, cada una de la cuales deberá tener el mismo tamaño que la tabla original.

El enfoque del "Bootstrap"

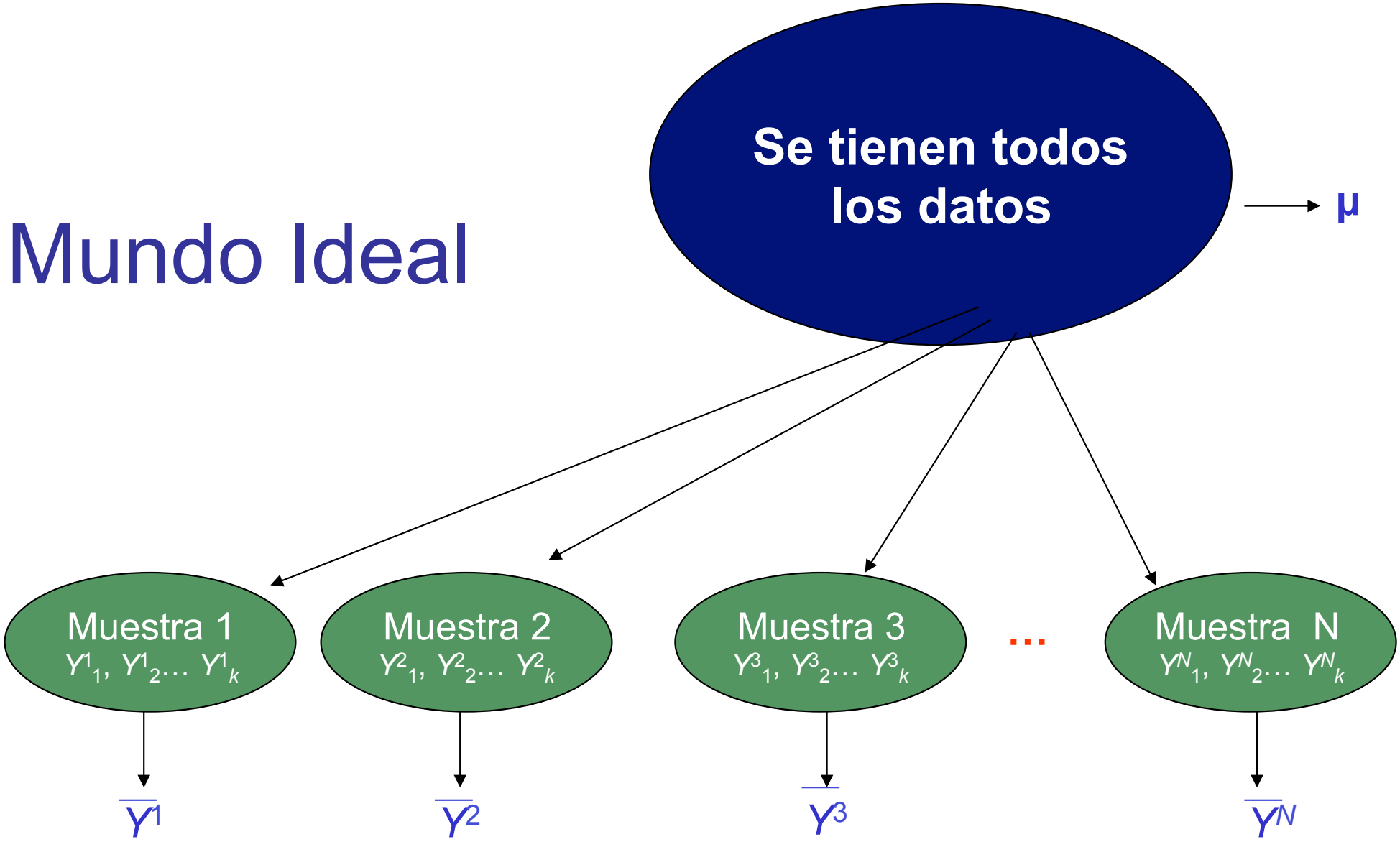
- En Bootstraping trabaja mediante el remuestreo *N* veces con reemplazo desde el conjunto de entrenamiento para formar nuevas tablas de datos (Bootstraps).

- Entonces modelo se estima en cada una de estas nuevas tablas (boostraps) y luego las predicciones se hacen para la tabla original de datos o conjunto de entrenamiento.

- Este proceso se repite muchas veces y se promedian los resultados.

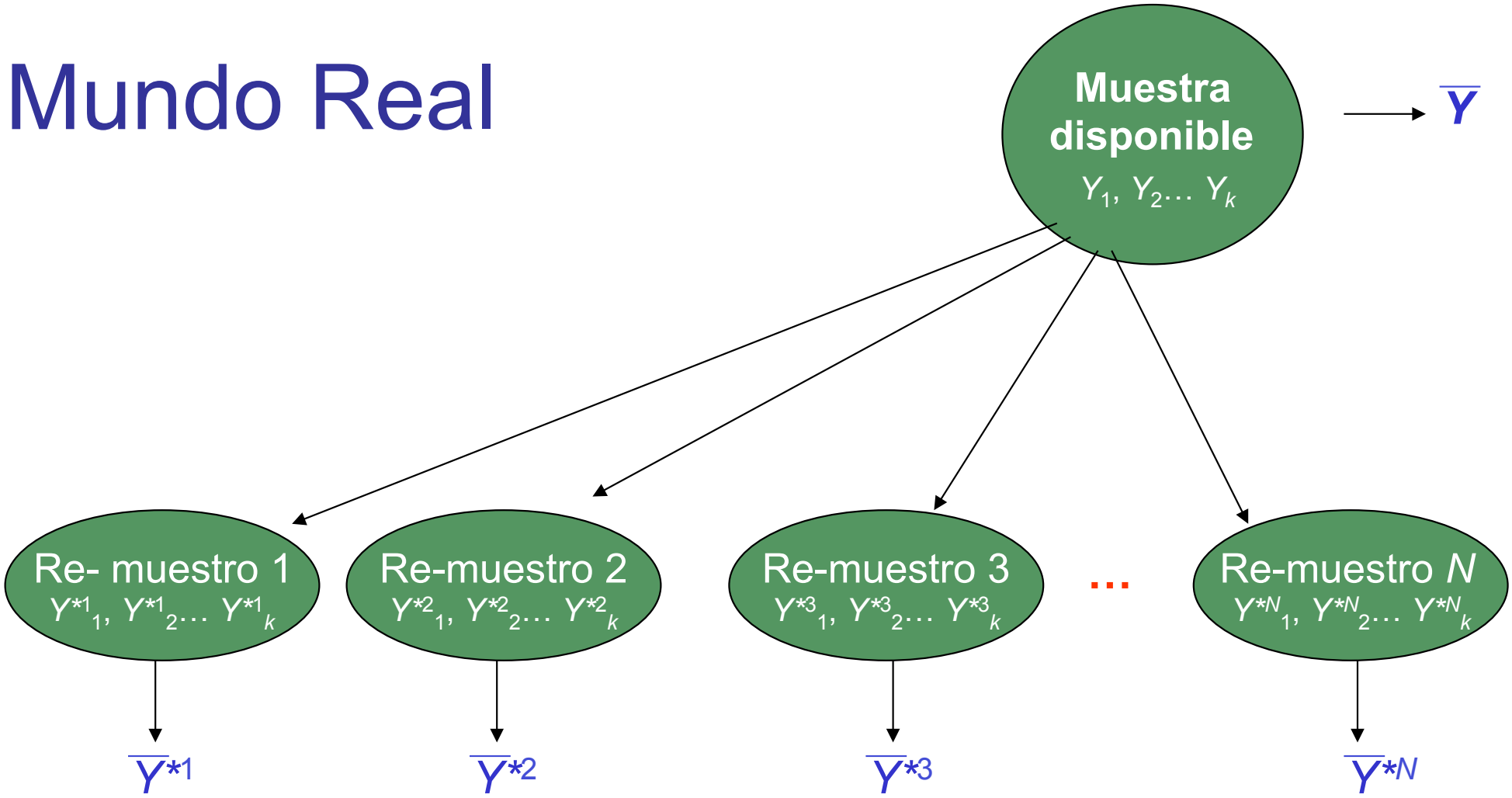
- Bootstrap es muy útil para estimar el error estándar en modelos predictivos y en algunas situaciones funciona mejor que la validación cruzada (cross-validation).

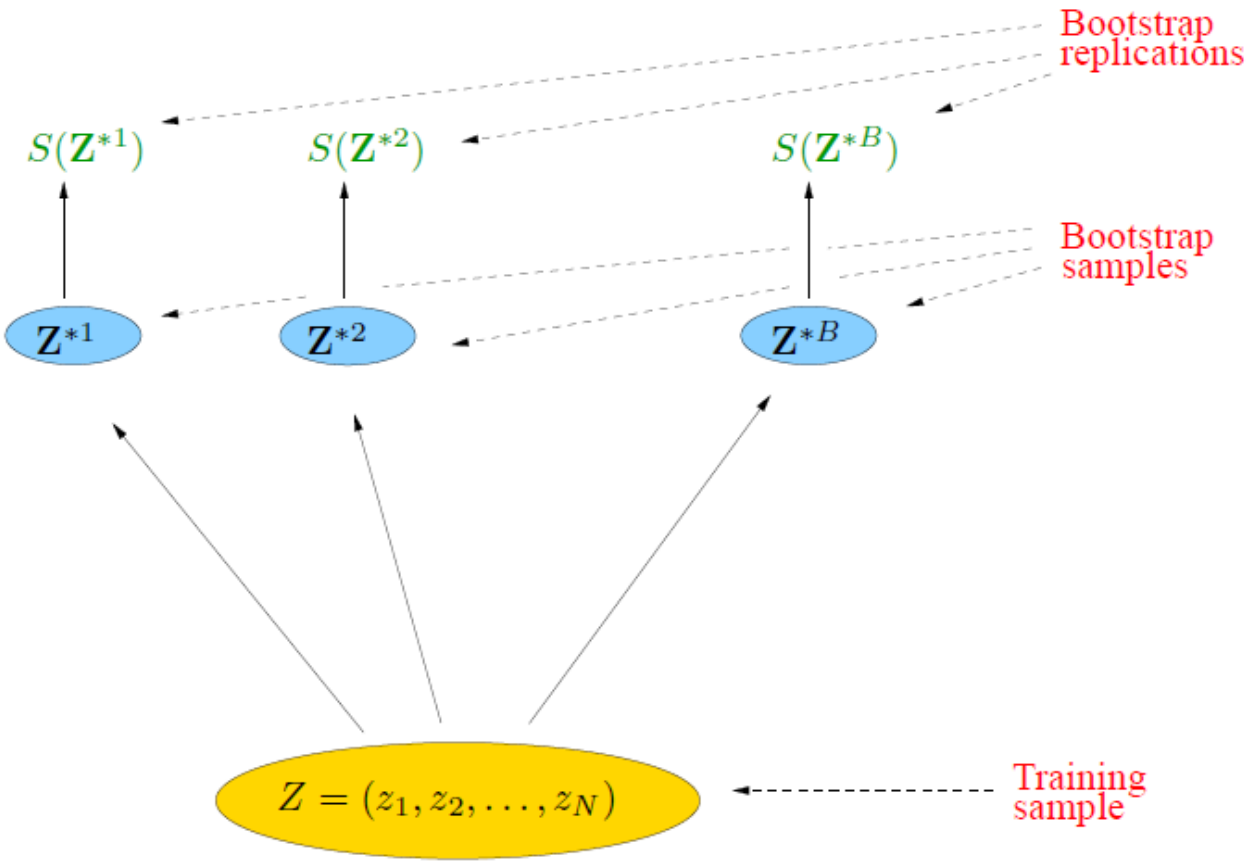
Mundo Ideal


Se tienen todos los datos
μ
Muestra 1
$Y^1_1, Y^1_2 \ldots Y^1_k$
Muestra 2
$Y^2_1, Y^2_2 \ldots Y^2_k$
Muestra 3
$Y^3_1, Y^3_2 \ldots Y^3_k$
...
Muestra N
$Y^N_1, Y^N_2 \ldots Y^N_k$
$\overline{Y}^1$
$\overline{Y}^2$
$\overline{Y}^3$
$\overline{Y}^N$

Proceso "Bootstrapping"

Mundo Real





Se desea calcular S(Z) entonces se generan B "bootstraps" Z^{*1}, Z^{*2},..., Z^{*B} (muestras con remplazo de tamaño N sobre Z) y se calcula S sobre esos "bootstraps", o sea $S(Z^{*1})$, $S(Z^{*2})$,...,$S(Z^{*B})$, para estimar la precisión estadística de S(Z).

Ejemplo de “Bootstraps”