DATA SCIENCE
& BIG DATA
RESEARCH LAB

Workshop 2021

Predicción en streaming mediante vecinos cercanos

Laura Melgar García

lmelgar@upo.es

Índice

1. Data streaming
2. Algoritmo vecinos cercanos tradicional (nearest-neighbors: NN)
3. Algoritmo StreamWNN
4. Aplicaciones

1. Data streaming

Flujo de datos masivos continuos

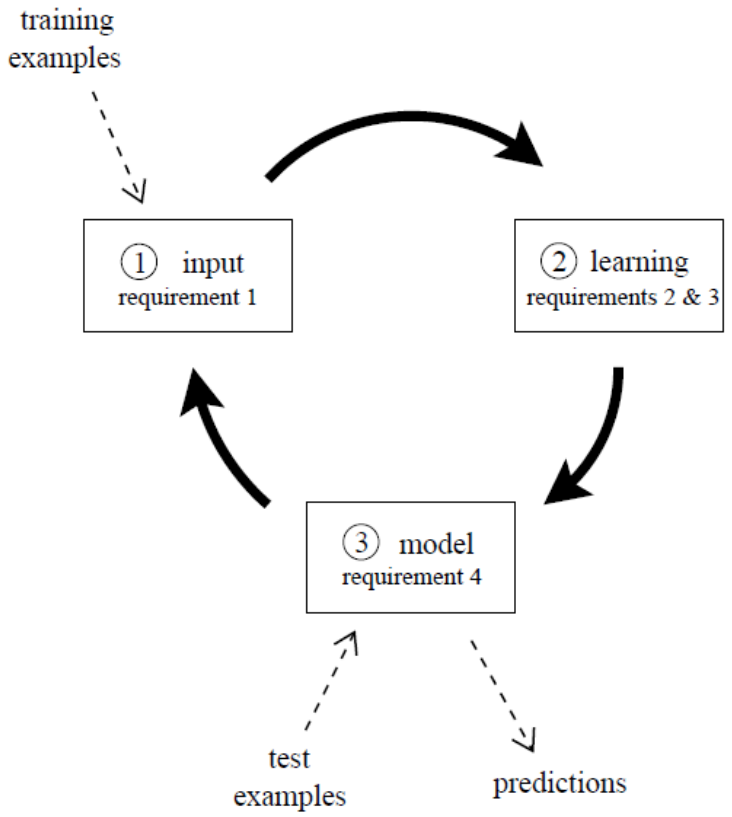
Batch/Tradicionales	Streaming
Recalcular el modelo	Incorporar al modelo

- Tiempo real
- Gran volumen

1. Data streaming

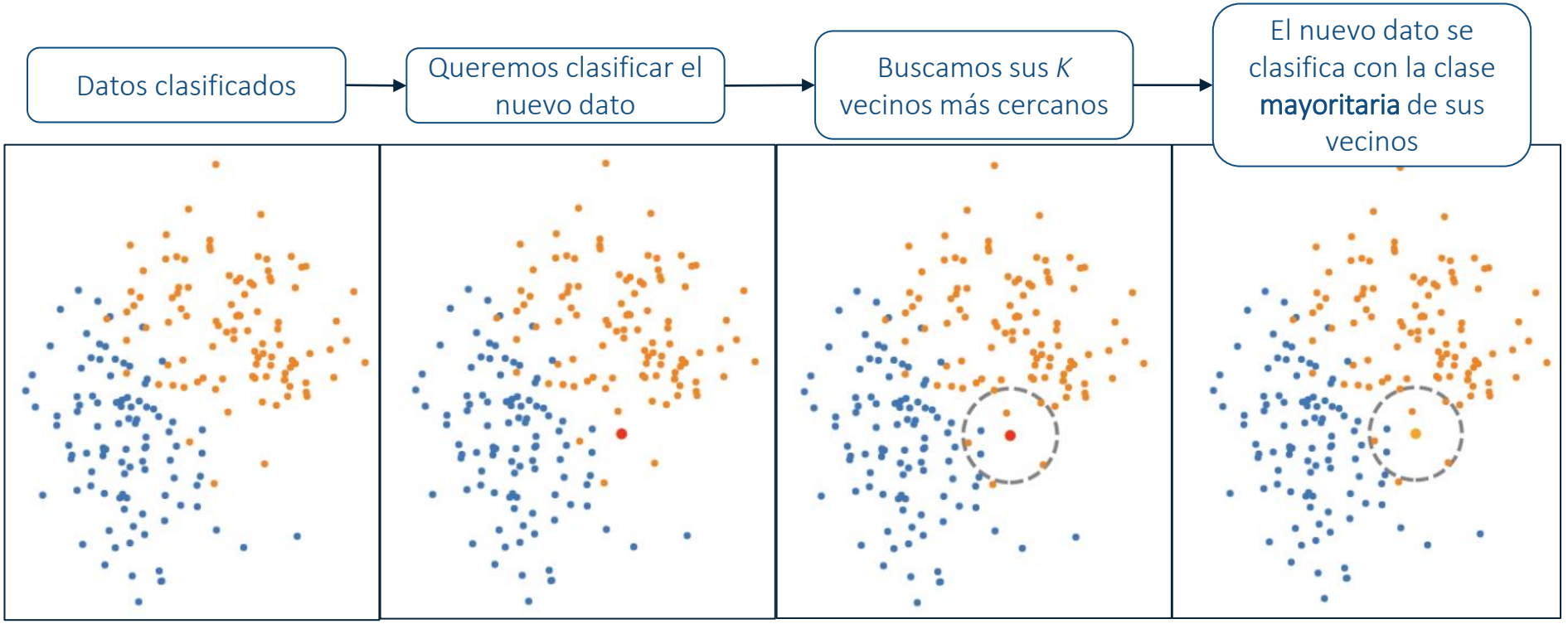
Requisitos:

1. Uno a uno, solo una vez y orden de llegada.
2. Memoria limitada.
3. Tiempo limitado.
4. Preparado.

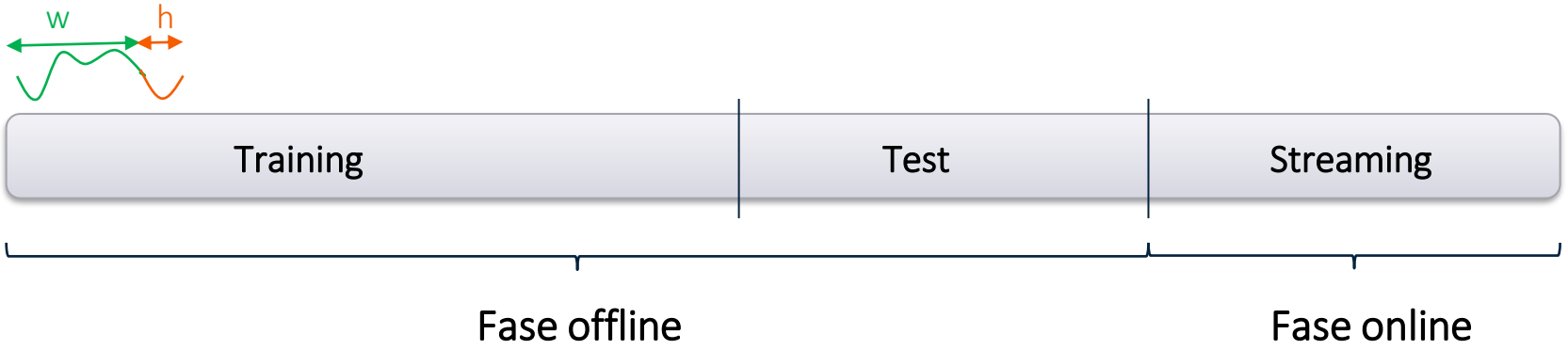


2. Algoritmo vecinos cercanos tradicional

Para clasificación



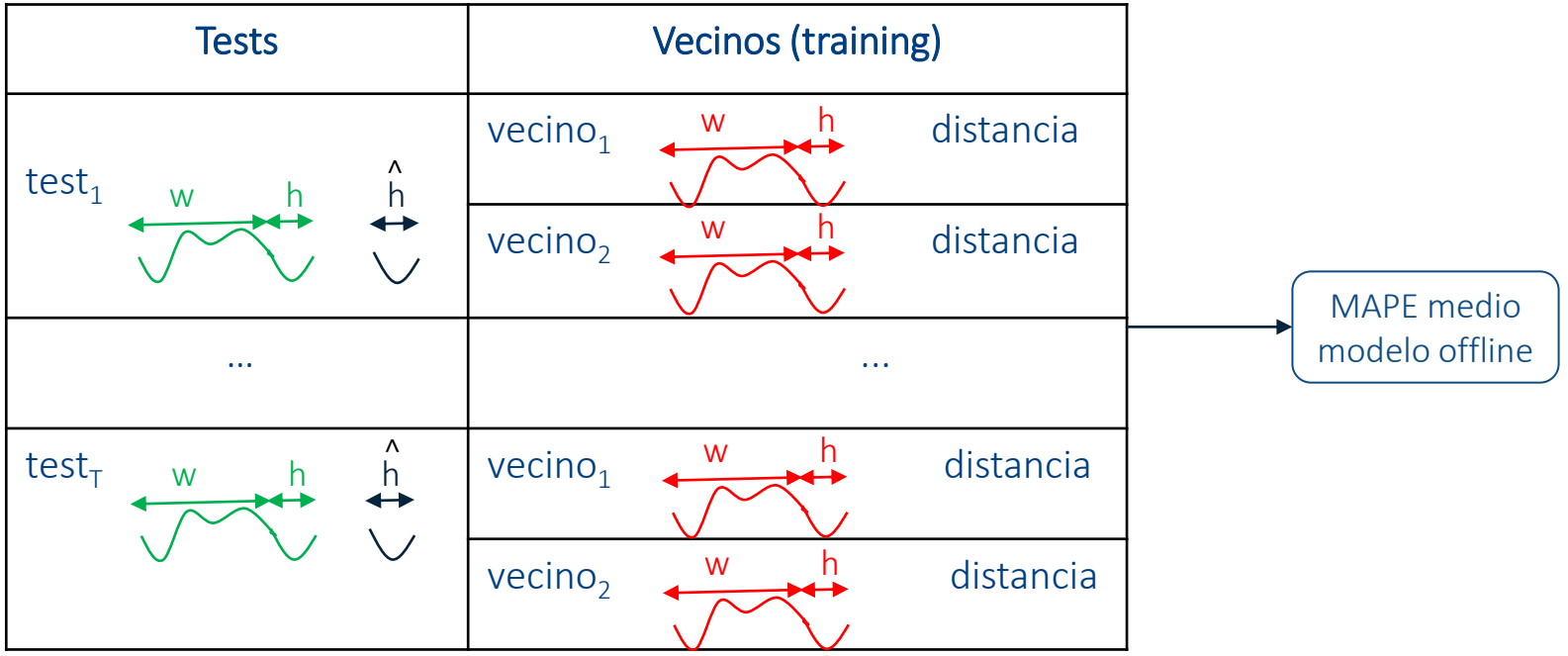
División del dataset



- Construcción modelo base

- Predicción en tiempo real
- Actualización del modelo

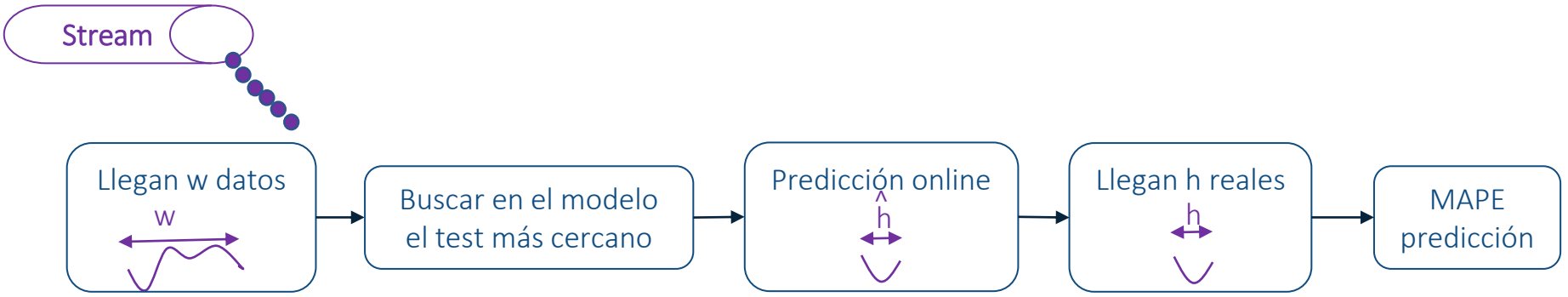
A) Fase offline



3. Algoritmo StreamWNN

B) Fase online

- Comienza el flujo de datos en tiempo real
- Se cumplen los requisitos del data streaming



3 tipos de actualizaciones: interna, externa e interna+externa

3. Algoritmo StreamWNN

Fórmula predicción offline:

$$\hat{h}(l) = \frac{1}{\sum_{j=1}^{K} \alpha_j} \sum_{j=1}^{K} \alpha_j h_{vecino_j}(l) \qquad 1 \leq l \leq |h| \qquad ; \qquad \text{donde} \qquad \alpha_j = \frac{1}{distancia_j^2}$$

Fórmula predicción online:

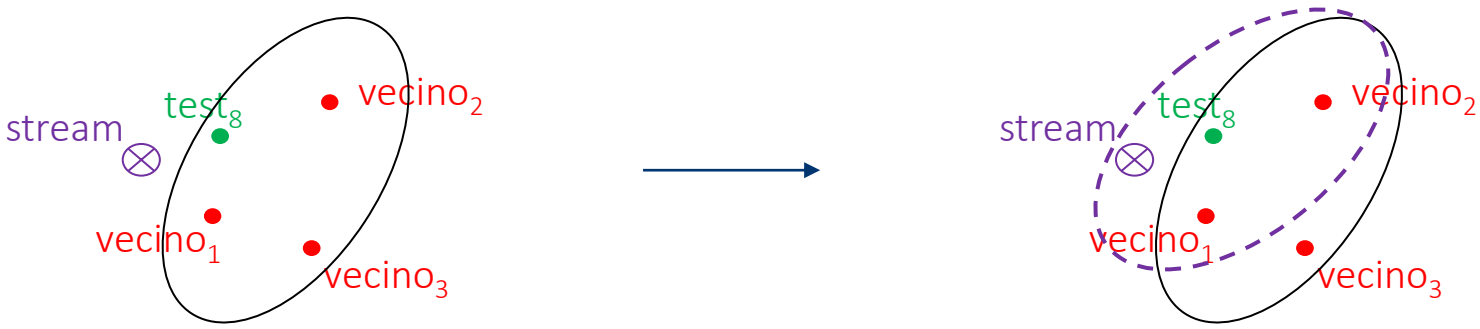
$$\hat{h}(l) = \frac{1}{\sum_{j=1}^{K} \alpha_j + \alpha_{test}} \left(\sum_{j=1}^{K} \alpha_j h_{vecino_j}(l) + \alpha_{test} h_{test}(l) \right) \qquad 1 \leq l \leq |h| \qquad ;$$

donde $\alpha_j = \frac{1}{distancia_j^2}$

B) Fase online

Actualización interna del modelo

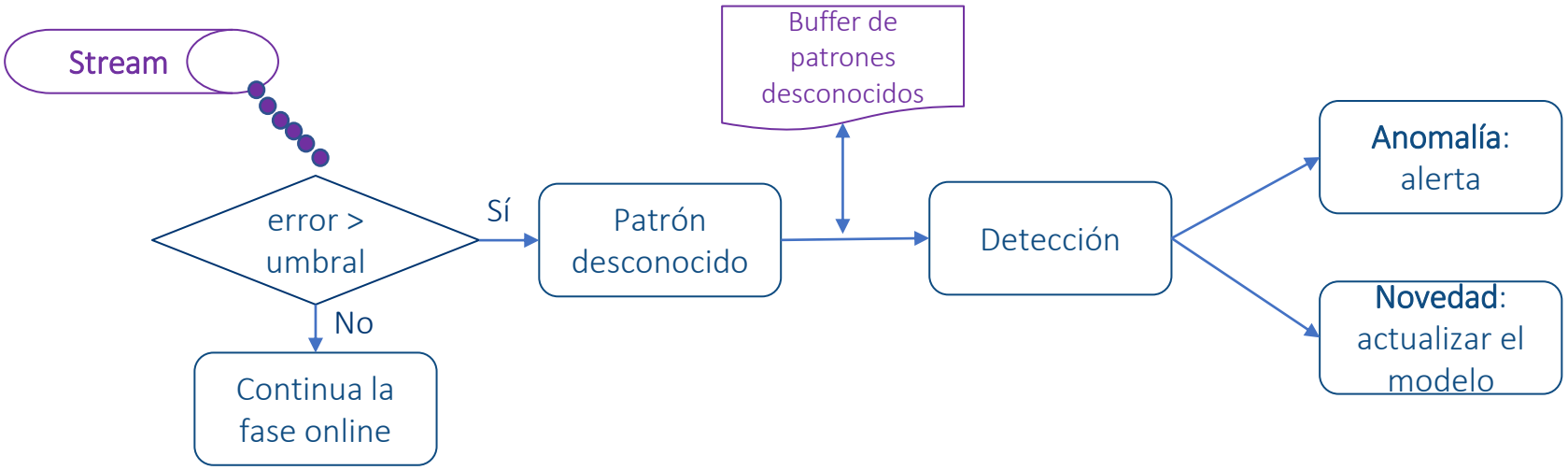
- Actualización de los vecinos con streams
- Buffer de streams si distancia del *K* vecino del modelo es mayor
- Actualización temporal



B) Fase online

Actualización externa del modelo, considerando novedades

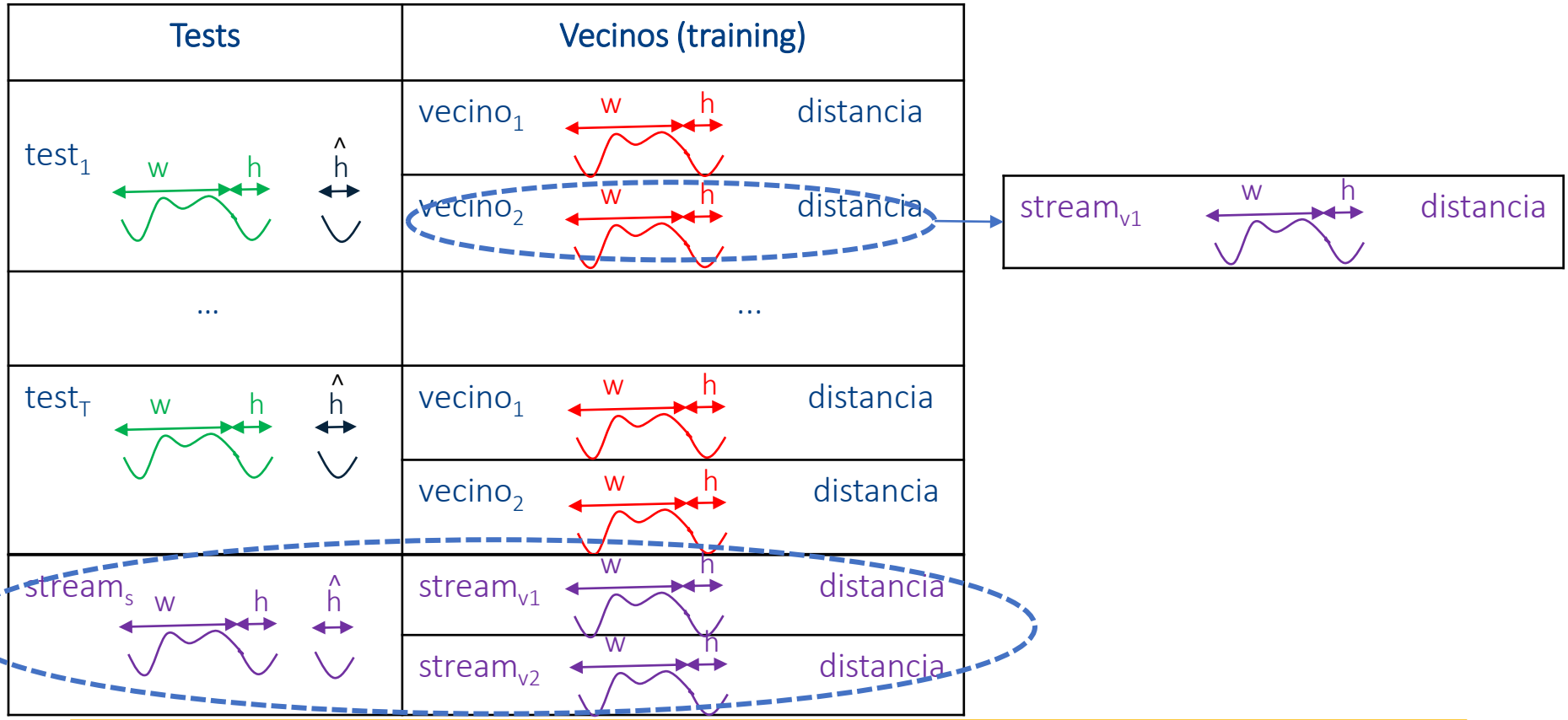
- Actualización añadiendo patrones
- Buffer de patrones desconocidos
- Actualización cuando se supera un umbral de error



B) Fase online

Actualización interna y externa del modelo

- Actualización de los vecinos (interna)
- Actualización añadiendo patrones (externa)



4. Aplicaciones

- Demanda eléctrica española
- Datos hospital: consumo eléctrico y consumo de gas

- Desarrollo StreamWNN multivariante
- Datos acústicos

Gracias

Laura Melgar García

lmelgar@upo.es

Extra

Implementación:

- Scala con Apache Spark
- HDFS
- Apache Kafka para el streaming

Extra

Bibliografía básica:

- **Data streams:**

  Libro “Industrial Applications of Machine Learning”, 2.6.1

- **Vecinos cercanos tradicional en big data:**

  “Big data time series forecasting based on nearest neighbours distributed computing in Spark”, Knowlegde-Based Systems, 2018.

- **Novedades y anomalías en data streaming:**

  “Novelty detection in data streams”, Springer Science, 2016.

Extra

Referencias a artículos propios – predicción en streaming basada en vecinos cercanos:

- "Nearest neighbors-based forecasting for electricity demand time series in streaming". CAEPIA, 2021. *(Submitted)*

Referencias a artículos propios – identificación de patrones en tres dimensiones en streaming:

- "A new big data triclustering approach for extracting three-dimensional patterns in precision agriculture". Neurocomputing, 2021. *(Submitted)*
- "Discovering three-dimensional patterns in real-time from data streams: an online triclustering approach". Information Science, 2021.
- "Generating a seismogenic source zone model for the Pyrenees: A GIS-assisted triclustering approach". Computers & Geosciences, 2021.
- "Discovering spatio-temporal patterns in precision agriculture based on triclustering". 15th International Conference on Soft Computing Models in Industrial and Environmental Applications (SOCO 2020), Advances in Intelligent Systems and Computing, Springer, Cham, 2020.
- "High-content screening images streaming analysis using the STriGen methodology". The 35th ACM/SIGAPP Symposium on Applied Computing (SAC 2020), Association of Computing Machinery, 2020.