SINCRONA

Machine Learning para el Modelamiento y Gestión de Sistemas Complejos

Sesión 6:

Machine Learning

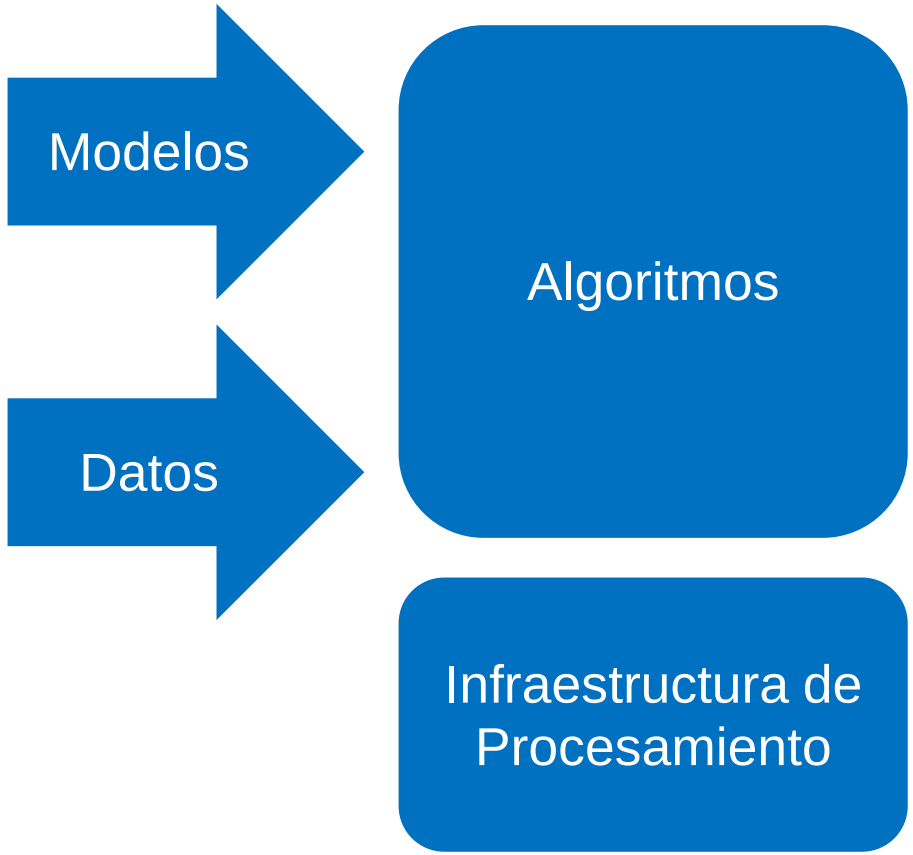
José Carlos Machicao

40 años
GRADE
Grupo de Análisis para el Desarrollo

S6

Modelamiento: Machine Learning

Componentes de los modelos predictivos



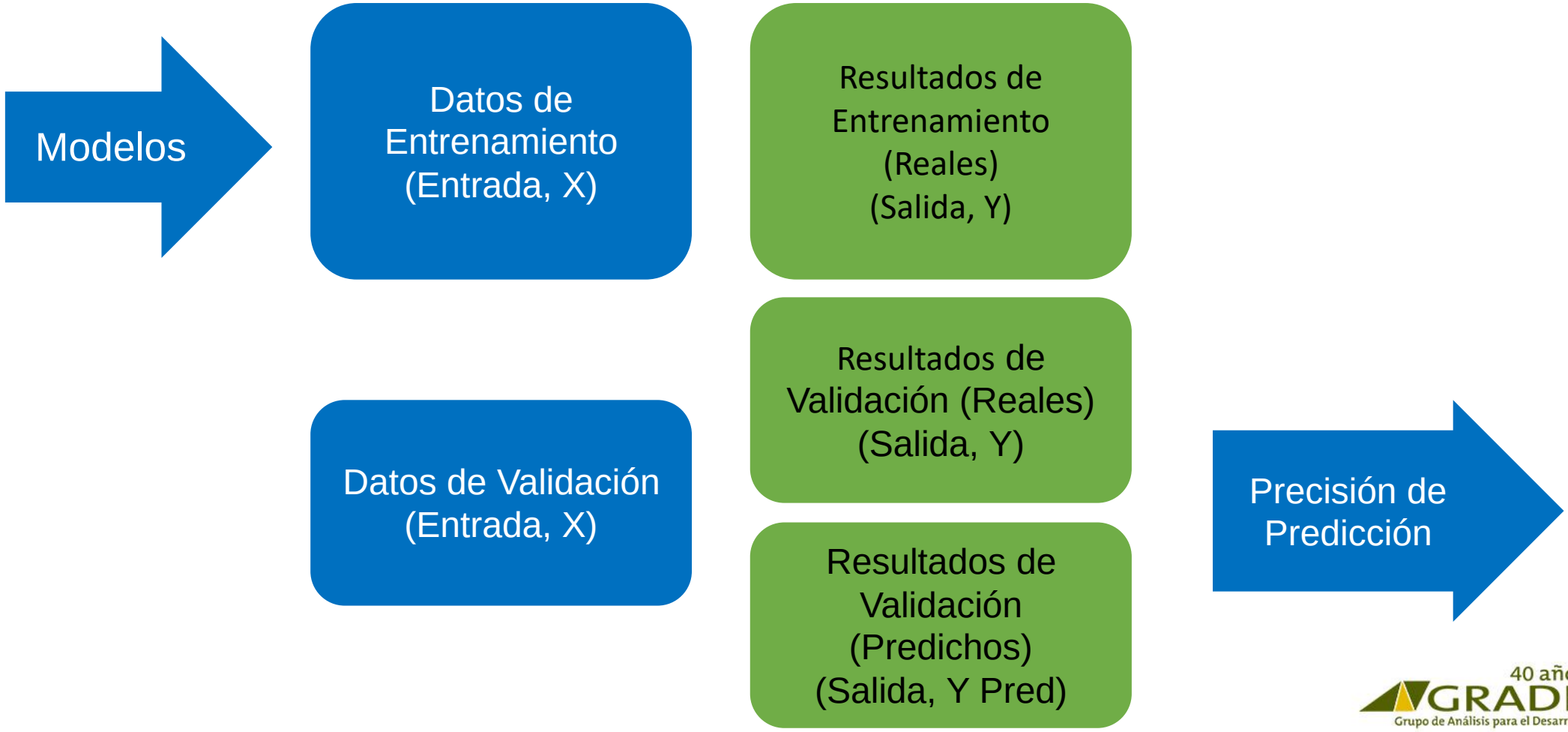
Información

Algoritmos

Infraestructura de Procesamiento

Capacidad de modelamiento algorítmico

Componentes de los modelos predictivos



Clasificación y Regresión

01	Clasificación	• Se tiene un número finito (pequeño) de clases • Los atributos de los datos contribuyen a la pertenencia a una clase
02	Regresión	• Se tiene un valor continuo (número continuo) como variable de salida • Los atributos contribuyen al cálculo del número de la variable de salida

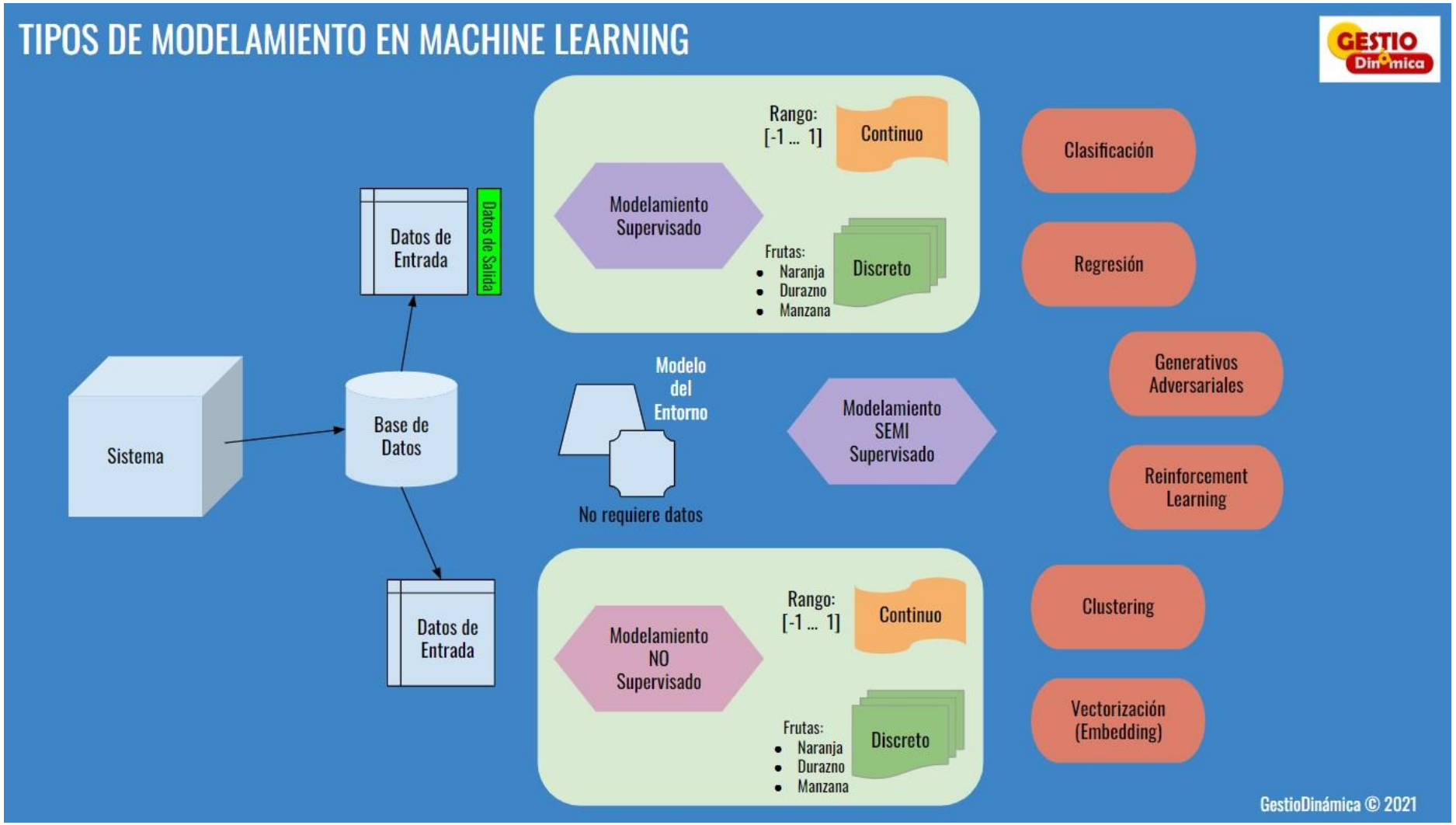
Supervisado y No Supervisado

01	Supervisado	• Existe una clasificación previa elaborada por seres humanos u otro algoritmo • Los atributos de la data están vinculados a las etiquetas
02	No Supervisado	• No existen etiquetas previamente definidas • Los atributos proponen una configuración auto-organizada sin un aprendizaje previo

Aspectos de Experimentación en Modelamiento

Preprocesamiento de datos:	• Limpiar y preparar los datos. • Manejo de valores perdidos, escalado de características y selección de características relevantes
Selección de modelo:	• Elegir un algoritmo de aprendizaje automático apropiado • Alinearse con la tarea en función de las características de los datos y los objetivos del proyecto.
Ajuste de hiper-parámetros:	• Optimización del rendimiento de un modelo de aprendizaje automático • Selección de los mejores valores para los hiperparámetros del modelo.
Métricas de evaluación:	• Elegir las métricas adecuadas para evaluar el rendimiento de un modelo de aprendizaje automático: exactitud, precisión y recuperación.
Interpretabilidad del modelo:	• Comprender cómo un modelo de aprendizaje automático hace predicciones para mejorar su precisión o identificar sesgos en los datos.

Los tipos de modelos para machine learning



40 años GRADE
Grupo de Análisis para el Desarrollo

Riesgos de la Experimentación en Modelamiento

Sobreajuste (Over-fitting)	• un modelo es demasiado complejo y tiene demasiados parámetros, lo que hace que se ajuste muy bien a los datos de entrenamiento pero tenga un rendimiento deficiente en datos nuevos e invisibles.
Infraajuste (Under-fitting)	• un modelo es demasiado simple y no puede capturar los patrones subyacentes en los datos, lo que lleva a un rendimiento deficiente tanto en los conjuntos de entrenamiento como de prueba.
Sesgo	• si los datos utilizados para entrenar el modelo no son representativos de la población del mundo real, lo que lleva a predicciones que son sistemáticamente incorrectas para ciertos grupos.
Consideraciones éticas	• los modelos de aprendizaje automático pueden perpetuar y amplificar los sesgos existentes en los datos • genera preocupaciones éticas sobre su uso en ciertos contextos • tomar medidas para mitigarlos en el diseño experimental

40 años GRADE
Grupo de Análisis para el Desarrollo

Colinearidad en Atributos

	Diámetro (cm)	Color	Peso (gr)
M1	5	Rojo	10
M2	4	Rojo	8
M3	3	Verde	6
M4	5	Marrón	10
M5	4	Verde	8

d * 2

10
8
6
10
8

40 años GRADE
Grupo de Análisis para el Desarrollo

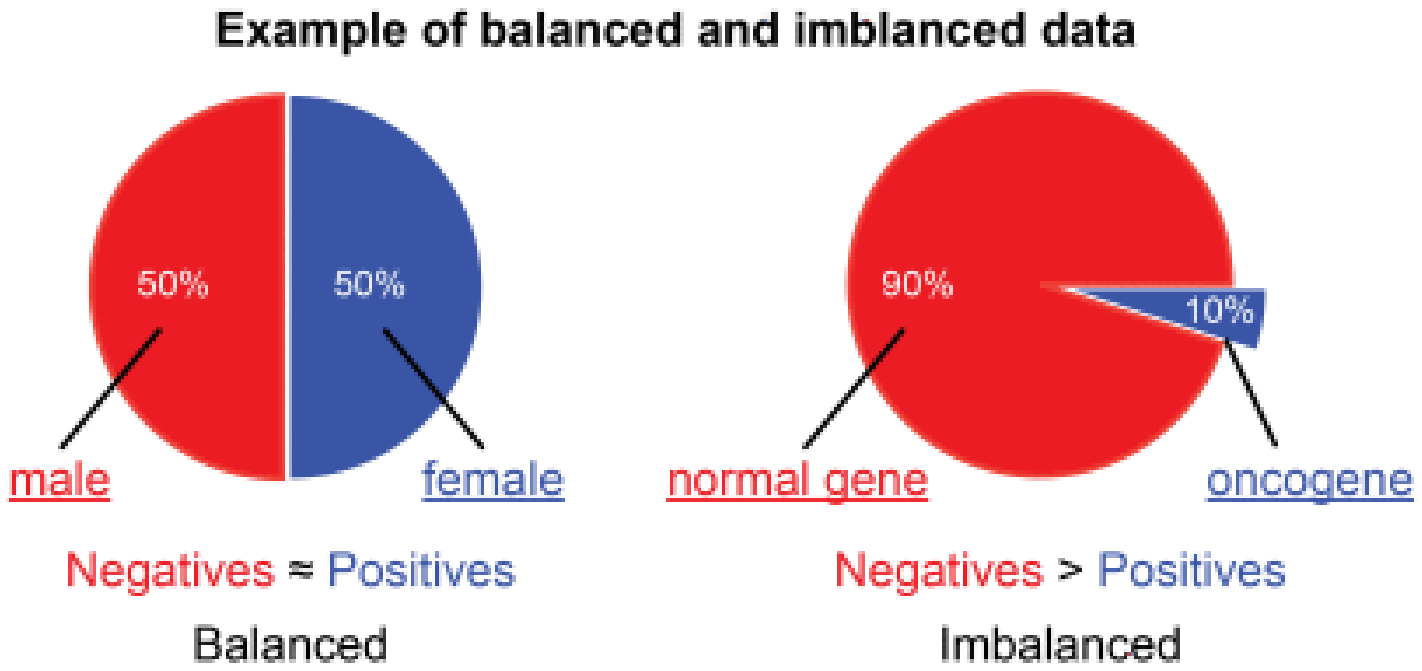
Expansión de Data (Data Augmentation)

- Cuando la data es escasa
- Pero además se tienen parámetros cuya variación ligera no afectan la integridad de los datos

- En estos casos se puede aplicar la técnica de “expansión de data” (data augmentation)

- Permite incrementar el número de observaciones

Desbalance de Categorías



Data Ausente (Missing Data)

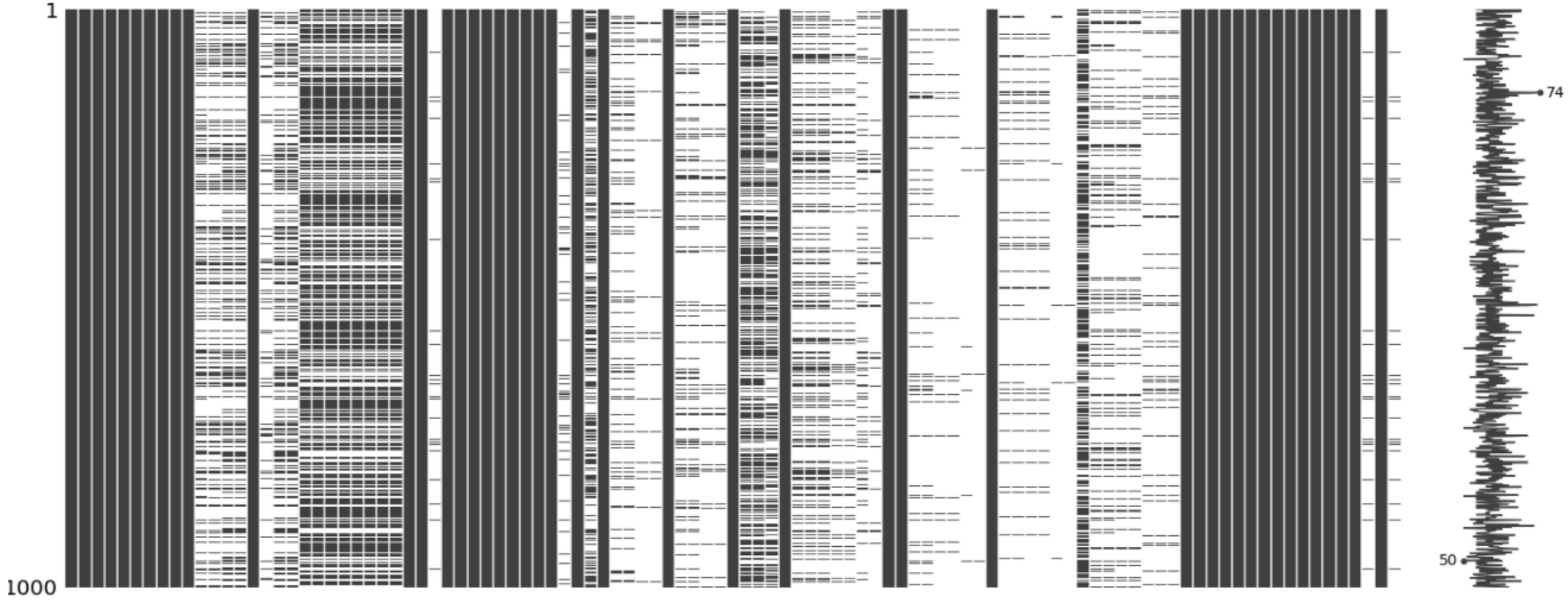


Diagrama de data ausente (missing data) para una base de datos de acceso financiero en Perú. Los espacios oscuros muestran donde está presente un dato, los espacios blancos donde está ausente. Columnas verticales, observaciones horizontales.

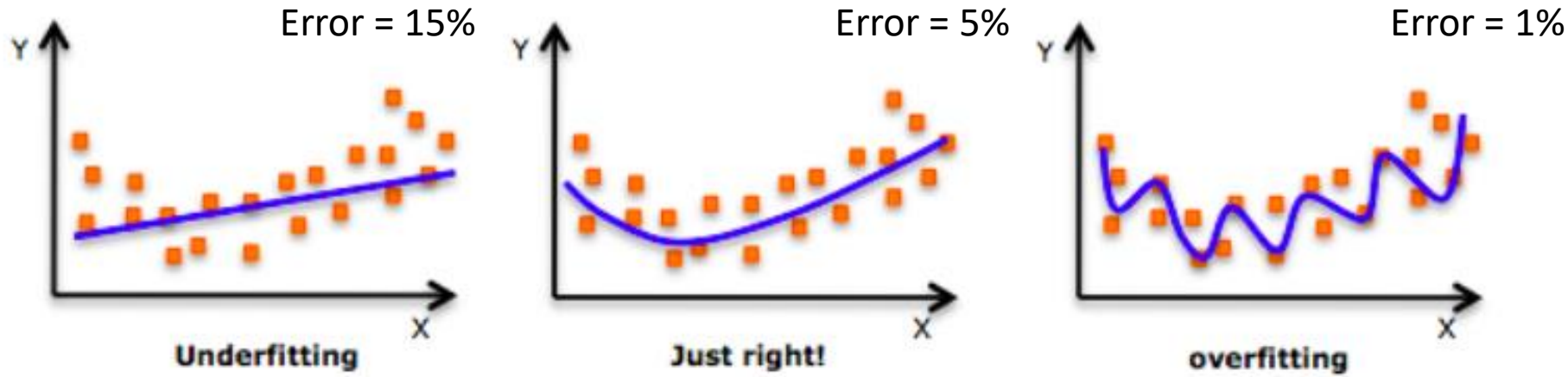
Anonimización y De-Identificación

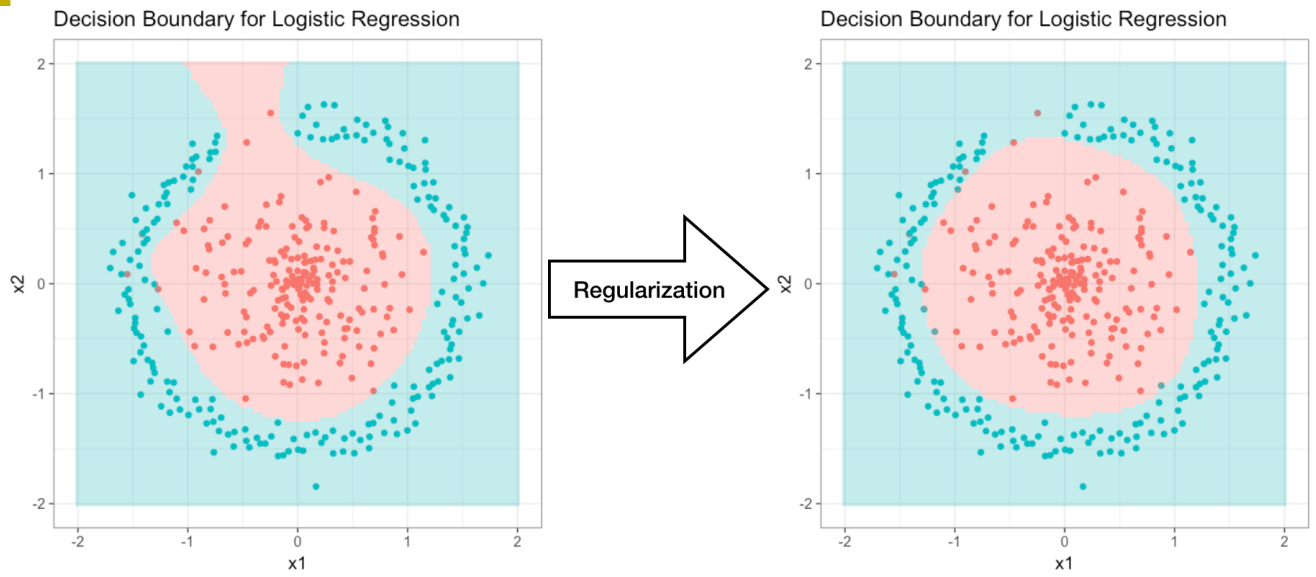
•*Anonymization*: The act of permanently and completely removing personal identifiers from data, such as converting personally identifiable information into aggregated data. Anonymized data is data that can no longer be associated with an individual in any manner. Once this data is stripped of personally identifying elements, those elements can never be re-associated with the data or the underlying individual.

https://www.educause.edu/

- **Anonimización**:
  - Se quitan todas las etiquetas de la observación que mencionan o pueden mencionar algo referido a su identidad
  - Se puede incluir agregación de la data
  - Cuando se garantiza que no haya absolutamente nada en la data que pueda conducir a la identidad de la observación
- **De-identificación**:
  - Sólo cuando se quitan todas las etiquetas de la observación que mencionan o pueden mencionar algo referido a su identidad

Regularización del Modelo

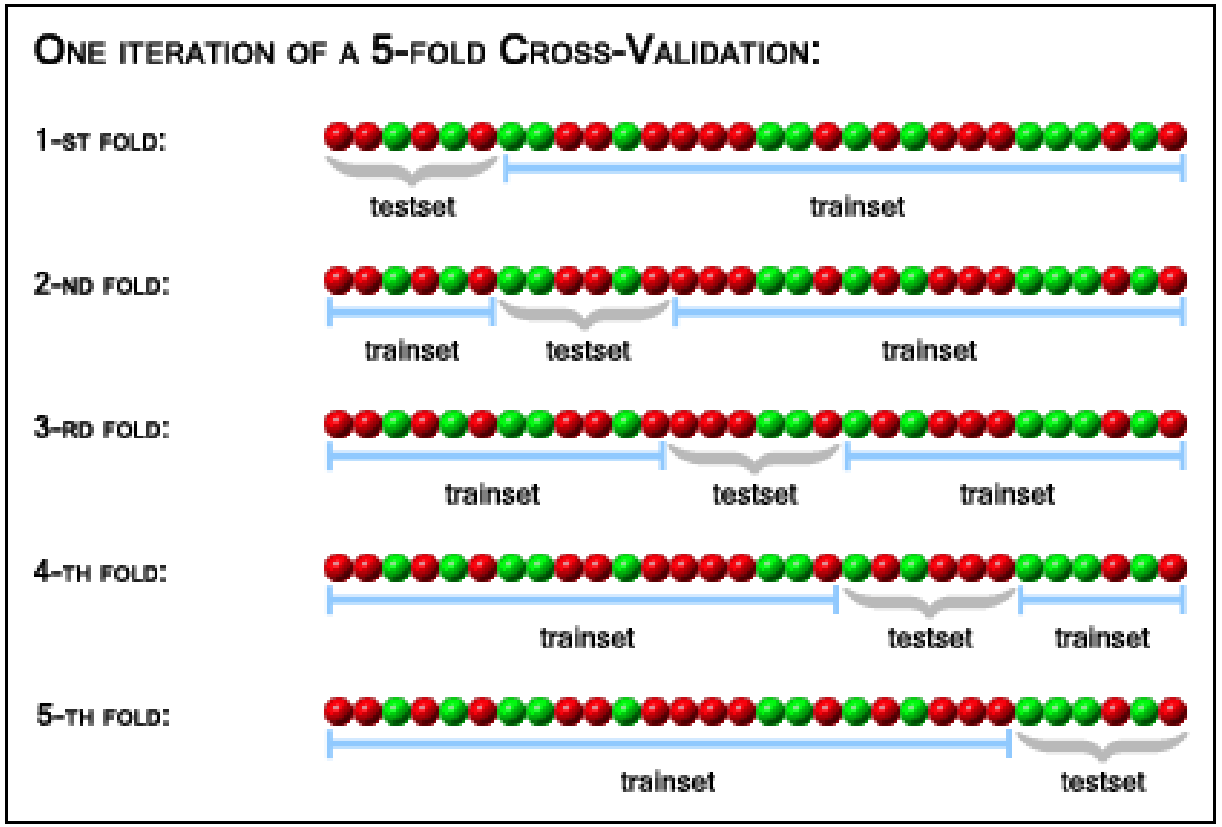




- ¿Siempre se busca exactitud total? NO
- La exactitud del modelo con la data de entrenamiento tiene que ser suficiente para dejar flexibilidad al modelo para funcionar correctamente con data nueva
- Se requiere
  - Data de entrenamiento
  - Data de prueba
  - Data de validación

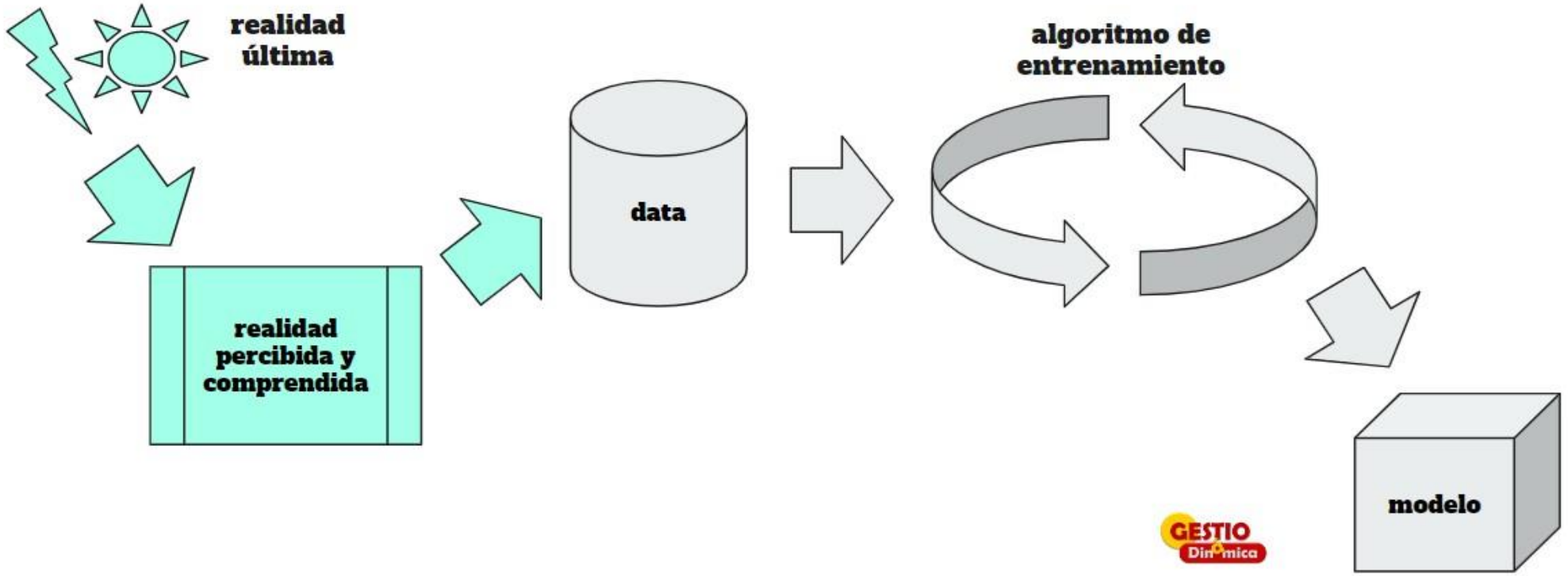
40 años GRADE Grupo de Análisis para el Desarrollo

Validación Cruzada

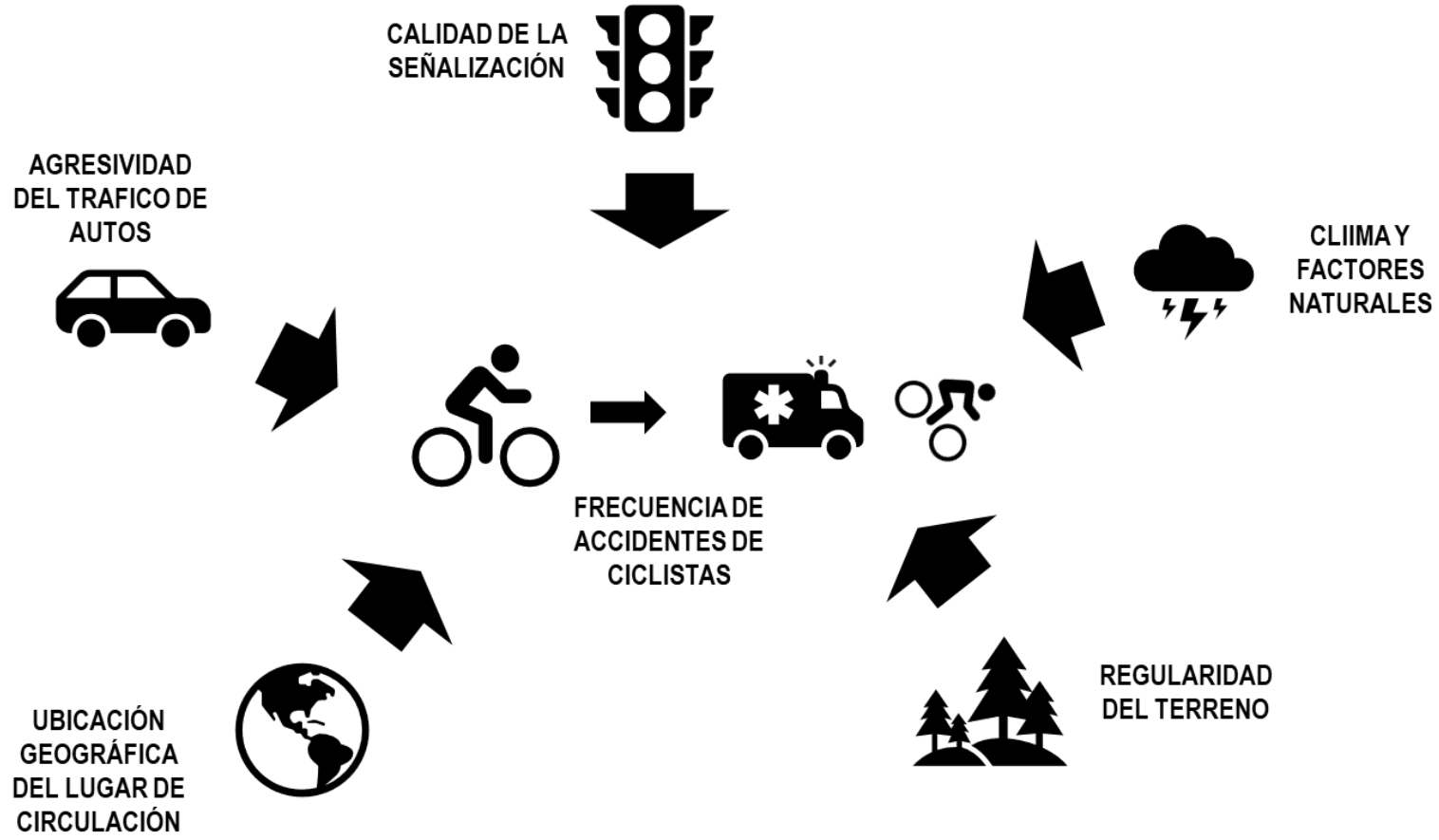


- La validación cruzada (crossvalidation) es la distribución diversificada de datos de entrenamiento con datos de validación.
- Lo que asegura es que no haya un sesgo debido a la primera forma en que se distribuyan los datos train/test.
- El objetivo es buscar cuál es la **exactitud** más baja que podría encontrar en todas las configuraciones.

La forma de modelar la realidad se ha hecho más compleja

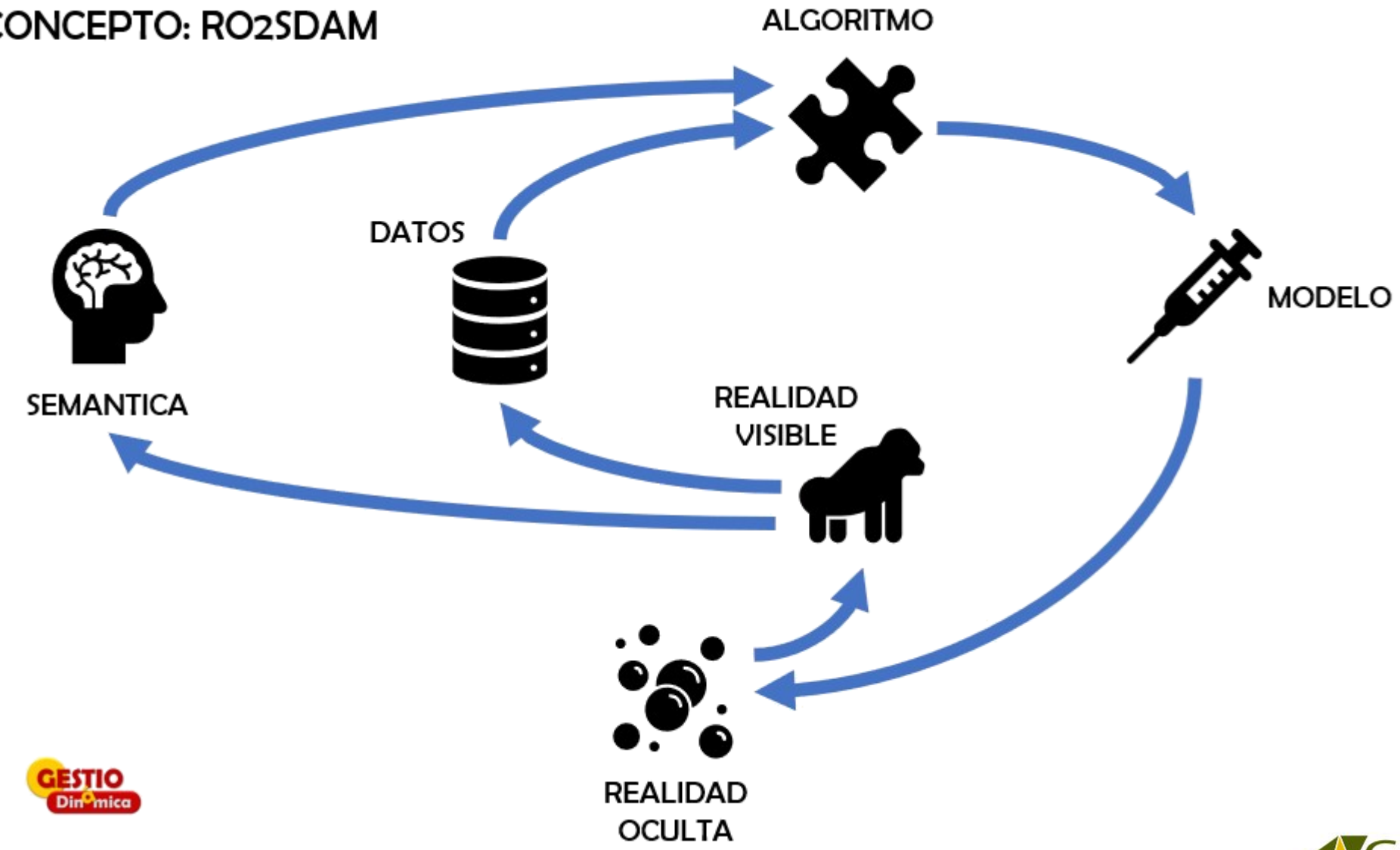


Modelamiento Conceptual Preliminar



40 años GRADE
Grupo de Análisis para el Desarrollo

CONCEPTO: RO2SDAM



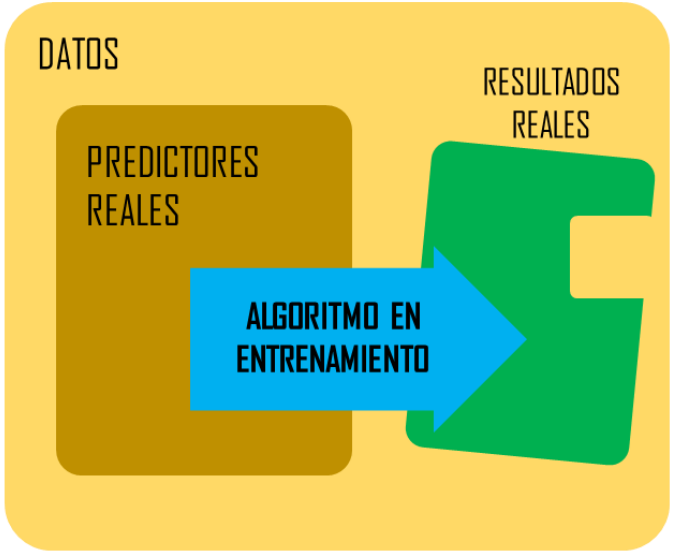
GESTIO Dinámica

40 años GRADE Grupo de Análisis para el Desarrollo

Machine Learning

Los datos reales se usan como base. Un algoritmo es entrenado vinculando las variables de entrada (predictores) para que coincidan con las variables de salida (resultados).

1



Estos resultado podrían no ser exactamente iguales a los reales, pero utilizan un razonamiento artificial.

2

La capacidad predictiva se muestra por la coincidencia entre los resultados reales de los datos nuevos y la predicción del algoritmo entrenado.

El algoritmo entrenado puede predecir resultados para predictores nuevos.

3

4

¡Muchas gracias!

40 años
GRADE
Grupo de Análisis para el Desarrollo