



INTRODUCCIÓN

APRENDIZAJE AUTOMÁTICO - CEIOT - FIUBA

Dr. Ing. Facundo Adrián Lucianna

INTRODUCCIÓN

- Materia de 8 clases teórico-prácticas
- Clases con diapositivas y desarrollo en notebooks
- Estructuras de las clases:
 - 10 minutos repaso de clase anterior
 - 3 bloques de 50 minutos de clases teórico-prácticas
 - 2 recreos de 10 minutos
 - Trabajos prácticos entre las clases.

INTRODUCCIÓN

Repositorio de la materia:

- https://github.com/facundolucianna/Apre_Aut_CEIoT

Consultas

- Servidor en Discord: <https://discord.gg/hXzXv4wRF>

Correo

- Facundo Adrian Lucianna: facundolucianna@gmail.com

EVALUACIÓN

- Trabajos prácticos por cada tema con entrega a 14 días posteriores de cuando fue presentado. La entrega es por correo electrónico o formulario de Google, ya sea el envío de la notebook o el link a repositorio.
- Los trabajos prácticos se aprueban con 6. Tiene una instancia de corrección una semana después de la fecha final de entrega y se puede corregir cualquier Trabajo práctico independientemente de si está aprobado o no.
- Se puede desaprobar hasta 2 TPs sin corrección.
- La nota final es la **mediana** de las notas de los TPs.
- La entrega consiste en un Jupyter Notebook o Google Colab mediante correo electrónico o repositorio en Github o Gitlab.
- El trabajo no solo consiste en resolver los ejercicios, sino una correcta presentación. Visualización es una de las partes mas importantes. No olvidar usar referencias.
- Criterio de evaluación: https://github.com/facundolucianna/Apre_Aut_CElot/blob/main/CriteriosAprobacion.md

HERRAMIENTAS

Lenguaje de Programación

- Python >=3.8
- Herramienta pip para instalar librerías de código y dependencias

Librerías

- Numpy, Pandas, SciPy, Statsmodels
- Matplotlib, Seaborn
- Scikit-Learn, Tensorflow

Consola Interactiva de Python

- IPython
- Jupiter Notebook

Herramientas

- GitHub para repositorios

IDE Recomendados

- Visual Studio Code
- PyCharm Community Edition



The Python logo consists of two interlocking snakes, one blue and one yellow, forming a stylized 'P' shape. To the right of the logo, the word "python" is written in a large, lowercase, sans-serif font. A small "TM" symbol is located at the top right of the "python" text.

```
101
102
103
104
105
106
107
108
109
110
111
112
113
114
115
116
117
118
119
120
121
122
123
124
125
126
127
128
129
130
131
132
133
134
135
136
137
138
139
140
141
142
143
144
145
146
147
148
149
150
151
152
153
154
155
156
157
158
159
160
161
162
163
164
165
166
167
168
169
170
171
172
173
174
175
176
177
178
179
180
181
182
183
184
185
186
187
188
189
190
191
192
193
194
195
196
197
198
199
200
201
202
203
204
205
206
207
208
209
210
211
212
213
214
215
216
217
218
219
220
221
222
223
224
225
226
227
228
229
230
231
232
233
234
235
236
237
238
239
240
241
242
243
244
245
246
247
248
249
250
251
252
253
254
255
256
257
258
259
260
261
262
263
264
265
266
267
268
269
270
271
272
273
274
275
276
277
278
279
280
281
282
283
284
285
286
287
288
289
290
291
292
293
294
295
296
297
298
299
300
301
302
303
304
305
306
307
308
309
310
311
312
313
314
315
316
317
318
319
320
321
322
323
324
325
326
327
328
329
330
331
332
333
334
335
336
337
338
339
340
341
342
343
344
345
346
347
348
349
350
351
352
353
354
355
356
357
358
359
360
361
362
363
364
365
366
367
368
369
370
371
372
373
374
375
376
377
378
379
380
381
382
383
384
385
386
387
388
389
390
391
392
393
394
395
396
397
398
399
400
401
402
403
404
405
406
407
408
409
410
411
412
413
414
415
416
417
418
419
420
421
422
423
424
425
426
427
428
429
430
431
432
433
434
435
436
437
438
439
440
441
442
443
444
445
446
447
448
449
450
451
452
453
454
455
456
457
458
459
460
461
462
463
464
465
466
467
468
469
470
471
472
473
474
475
476
477
478
479
480
481
482
483
484
485
486
487
488
489
490
491
492
493
494
495
496
497
498
499
500
501
502
503
504
505
506
507
508
509
510
511
512
513
514
515
516
517
518
519
520
521
522
523
524
525
526
527
528
529
530
531
532
533
534
535
536
537
538
539
540
541
542
543
544
545
546
547
548
549
550
551
552
553
554
555
556
557
558
559
560
561
562
563
564
565
566
567
568
569
570
571
572
573
574
575
576
577
578
579
580
581
582
583
584
585
586
587
588
589
589
590
591
592
593
594
595
596
597
598
599
600
601
602
603
604
605
606
607
608
609
610
611
612
613
614
615
616
617
618
619
620
621
622
623
624
625
626
627
628
629
630
631
632
633
634
635
636
637
638
639
640
641
642
643
644
645
646
647
648
649
650
651
652
653
654
655
656
657
658
659
660
661
662
663
664
665
666
667
668
669
670
671
672
673
674
675
676
677
678
679
680
681
682
683
684
685
686
687
688
689
689
690
691
692
693
694
695
696
697
698
699
700
701
702
703
704
705
706
707
708
709
709
710
711
712
713
714
715
716
717
718
719
719
720
721
722
723
724
725
726
727
728
729
729
730
731
732
733
734
735
736
737
738
739
739
740
741
742
743
744
745
746
747
748
749
749
750
751
752
753
754
755
756
757
758
759
759
760
761
762
763
764
765
766
767
768
769
769
770
771
772
773
774
775
776
777
778
779
779
780
781
782
783
784
785
786
787
788
789
789
790
791
792
793
794
795
796
797
798
799
800
801
802
803
804
805
806
807
808
809
809
810
811
812
813
814
815
816
817
818
819
819
820
821
822
823
824
825
826
827
828
829
829
830
831
832
833
834
835
836
837
838
839
839
840
841
842
843
844
845
846
847
848
849
849
850
851
852
853
854
855
856
857
858
859
859
860
861
862
863
864
865
866
867
868
869
869
870
871
872
873
874
875
876
877
878
879
879
880
881
882
883
884
885
886
887
888
889
889
890
891
892
893
894
895
896
897
898
899
900
901
902
903
904
905
906
907
908
909
910
911
912
913
914
915
916
917
918
919
919
920
921
922
923
924
925
926
927
928
929
929
930
931
932
933
934
935
936
937
938
939
939
940
941
942
943
944
945
946
947
948
949
949
950
951
952
953
954
955
956
957
958
959
959
960
961
962
963
964
965
966
967
968
969
969
970
971
972
973
974
975
976
977
978
979
979
980
981
982
983
984
985
986
987
988
989
989
990
991
992
993
994
995
996
997
998
999
```

PROGRAMA

CLASE A CLASE

01

Introducción a Machine Learning. Rol del Data Scientist. Tipos de Datos. Análisis de Datos. Tipos de aprendizaje. Supervisado y No Supervisado. Clasificación de Algoritmos.

02

Aprendizaje supervisado. Conceptos de Regresión. Regresión lineal simple y múltiple. Regresión polinómica. Métodos de evaluación de regresiones. Variables dummies. Construcción de modelos.

03

Aprendizaje supervisado. Conceptos de Clasificación. Regresión logística. K-NN. Métodos de evaluación de clasificadores.

04

Arboles de decisión. Arboles de clasificación y arboles regresión. Random forest. Métodos de selección de modelos. Cross-validation. Exactitud y precisión.

PROGRAMA

CLASE A CLASE

05

Support Vector Machines. Clasificación y regresión usando SVM. Métodos heurísticos de búsqueda de hiperparámetros.

06

Redes neuronales simples. Red feed-forward y Backpropagation.

07

Aprendizaje No Supervisado. Clustering. K-Means. Suma de Cuadrados Intracluster. Implementación.

BIBLIOGRAFIA

- Practical Statistics for Data Scientists: 50+ Essential Concepts Using R and Python - Peter Bruce (Ed. O'Reilly)
- The Elements of Statistical Learning - Trevor Hastie (Ed. Springer)
- An Introduction to Statistical Learning - Gareth James (Ed. Springer)
- Pattern Recognition And Machine Learning - Christopher Bishop (Ed. Springer)
- Deep Learning - Ian Goodfellow <https://www.deeplearningbook.org/>

The background features a minimalist design with abstract, wavy shapes in shades of purple and blue. These shapes are layered and overlap, creating a sense of depth and movement against a solid dark blue background.

MACHINE LEARNING

MACHINE LEARNING

- Masificación del uso de internet
- Surgimiento de las redes sociales
- Crecimiento exponencial de dispositivos móviles
- Interfaces de usuario mas simples e intuitivas

Cada día se crean
(2.5 Exabytes)

El 90% de los datos del mundo
de hoy se generaron en los
últimos 2 años

MACHINE LEARNING

UNA PRIMERA DEFINICIÓN

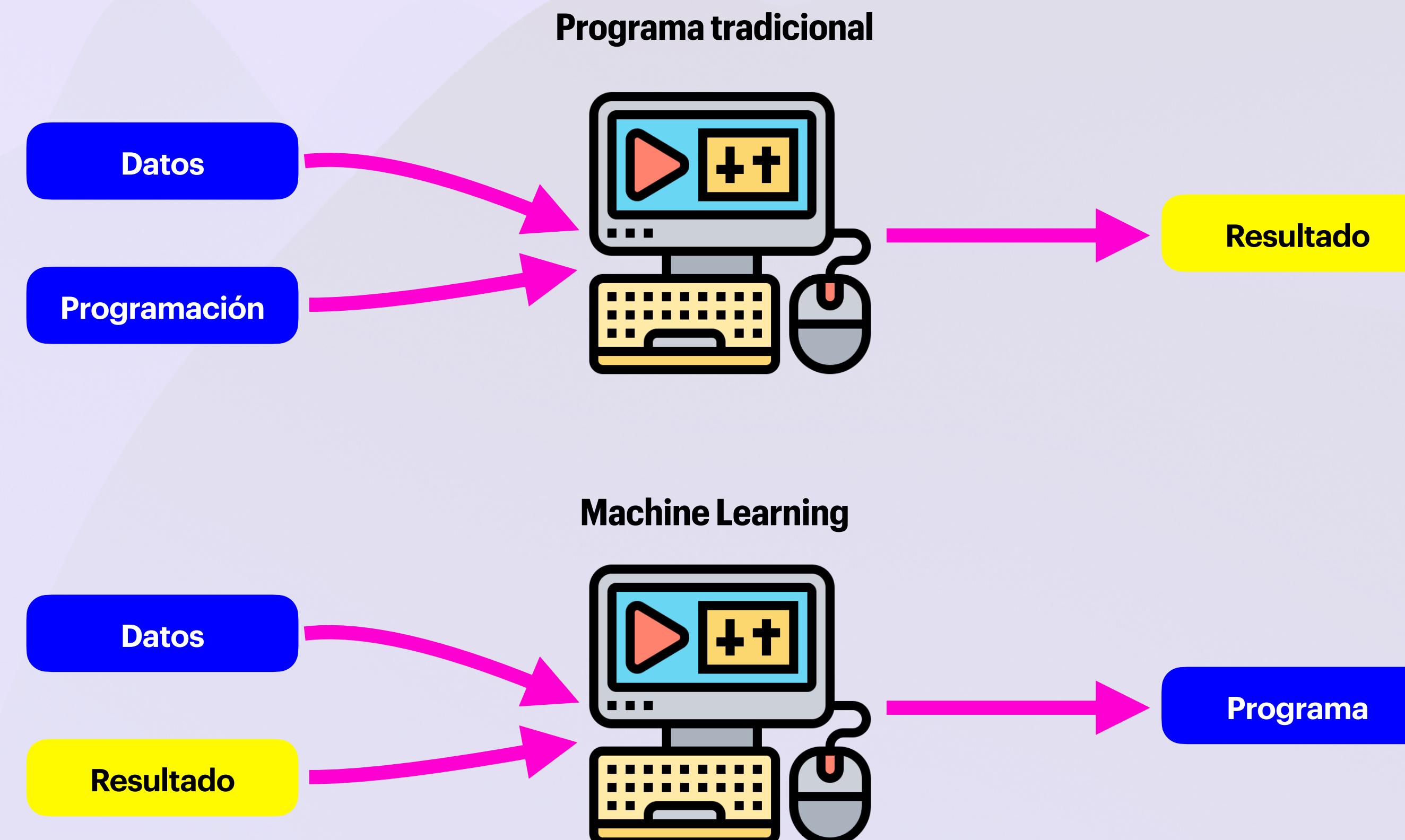
Es la **ciencia** que permite que las computadoras aprendan y actúen como lo hacen los humanos, mejorando su aprendizaje a lo largo del tiempo de **una forma autónoma**, alimentándose con **datos e información en forma de observaciones e interacciones** con el **mundo real**.

- Recibe un conjunto de datos y aprende por sí mismo.
- Reconoce patrones entre los datos y hace una predicción
- No requiere que una persona programe instrucciones



MACHINE LEARNING

VS. PROGRAMACIÓN TRADICIONAL



BIG DATA

UNA PRIMERA DEFINICIÓN

Volumen masivo de datos, tanto estructurados como no-estructurados, los cuales son demasiado grandes y difíciles de procesar con las bases de datos y el software tradicionales.



BIG DATA

LAS TRES V

Volumen

Gran cantidad de volúmenes de datos no estructurados. Para algunas organizaciones, esto puede suponer decenas de terabytes de datos. Para otras, incluso cientos de petabytes.

Velocidad:

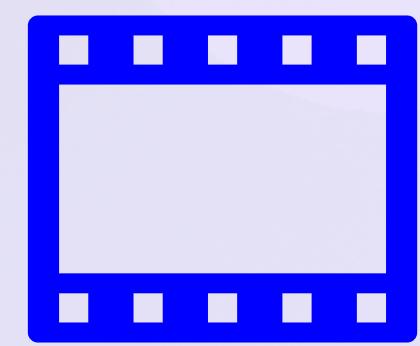
La velocidad es el ritmo al que se reciben los datos y (posiblemente) al que se aplica alguna acción. Muchas aplicaciones responden a “tiempo real”.

Variedad

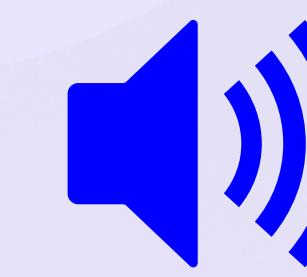
Gran cantidad de tipos de datos, principalmente no estructurado o semiestructurado, como el texto, audio o vídeo.

BIG DATA

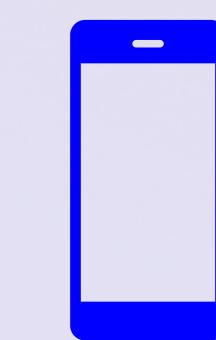
DATA LAKE



Media: Videos y audio



Datos estructurados



IoT y sensores



Logs

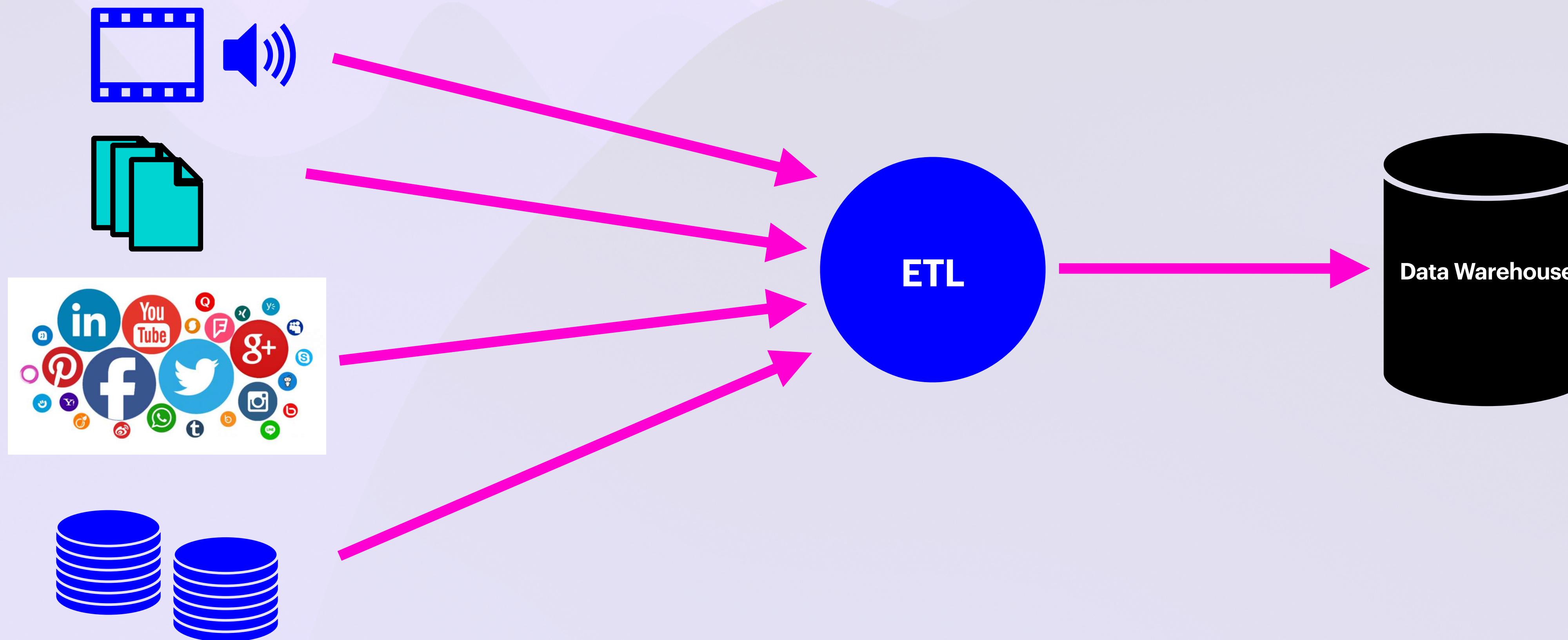


Bases de datos



BIG DATA

DATA WAREHOUSE

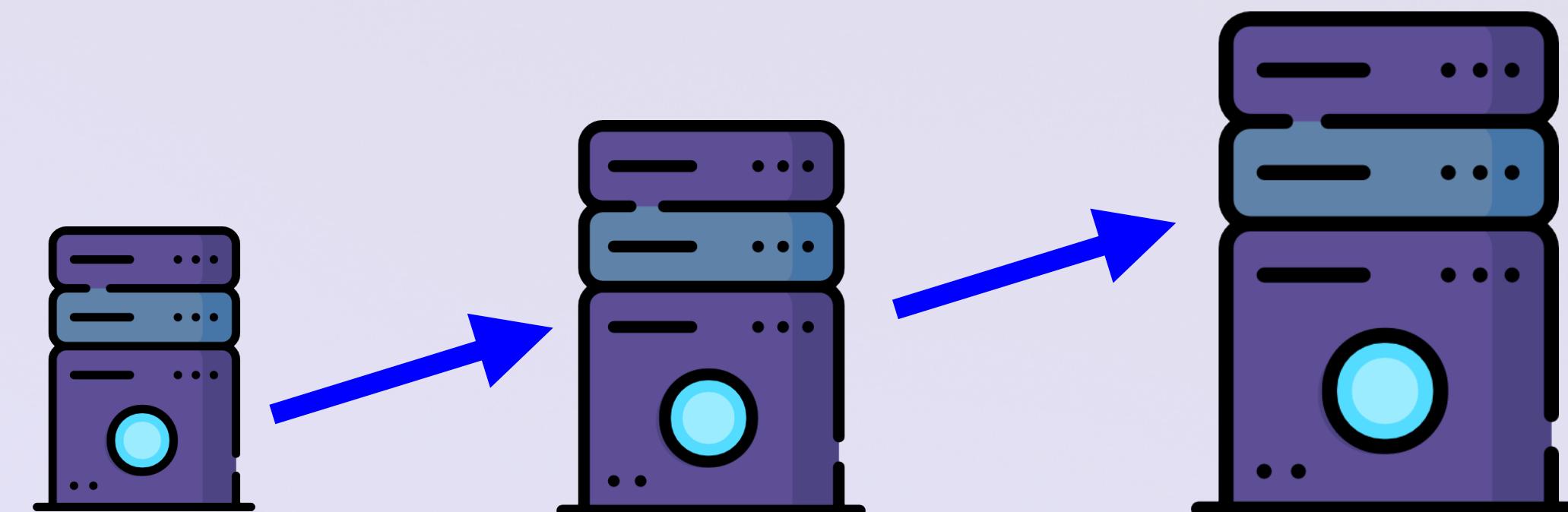


PROCESAMIENTO EN BIG DATA

ESCALAMIENTO

Escalamiento vertical

- Escalamiento dentro de un mismo servidor.
- Implica incrementar la capacidad de un Servidor agregando más recursos de CPU, memoria y de almacenamiento.

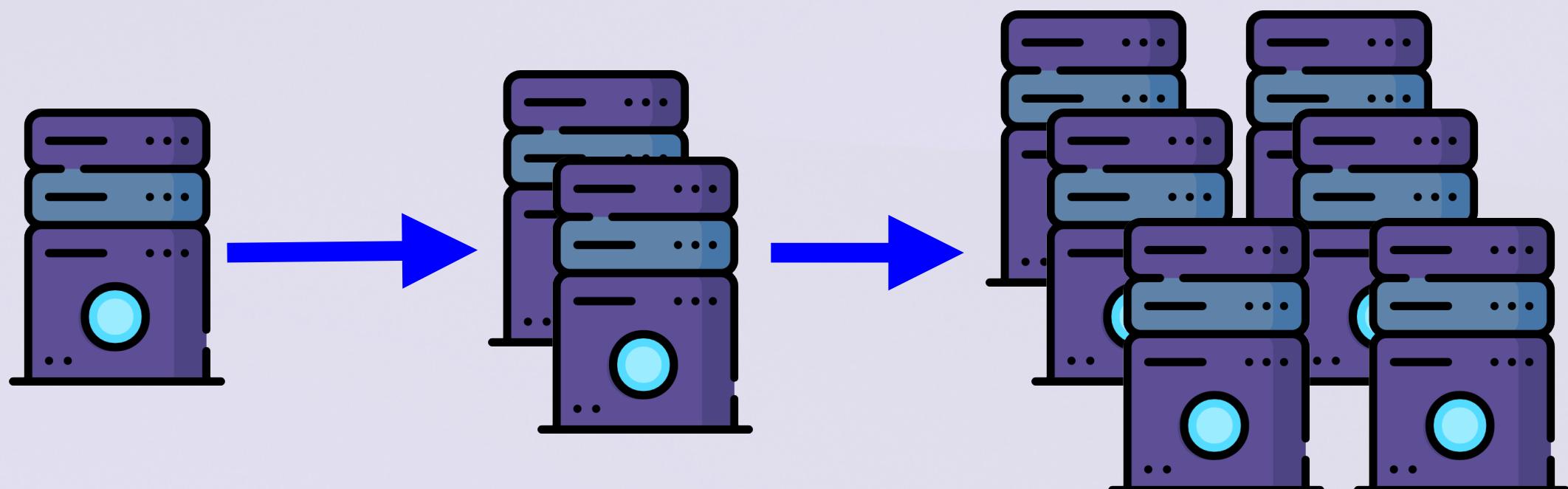


PROCESAMIENTO EN BIG DATA

ESCALAMIENTO

Escalamiento horizontal

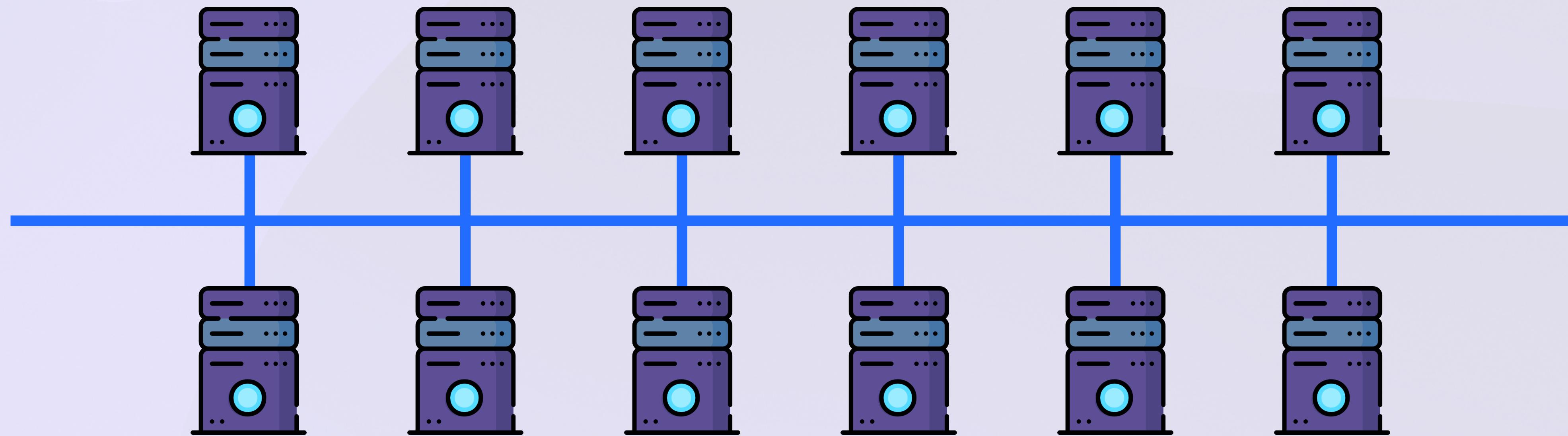
- Escalamiento en varios servidores
- Cluster de servidores
- Replicación de datos
- Particionamiento de datos
- Procesamiento paralelo



PROCESAMIENTO EN BIG DATA

CLUSTER

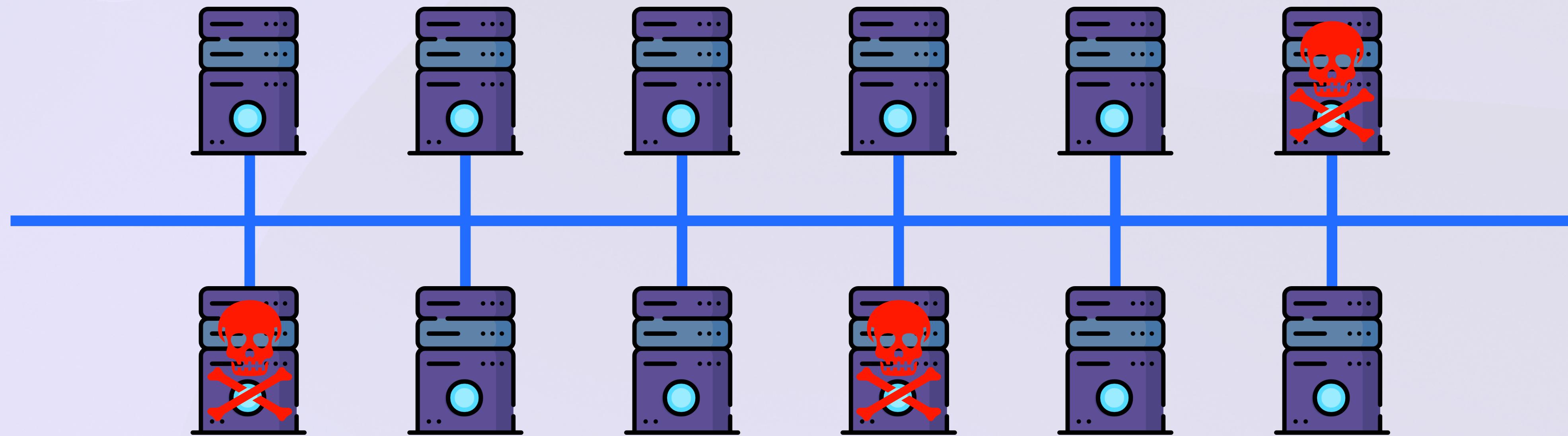
Grupo de servidores independientes interconectados a través de una red dedicada que trabajan como un único recurso de procesamiento.



PROCESAMIENTO EN BIG DATA

CLUSTER

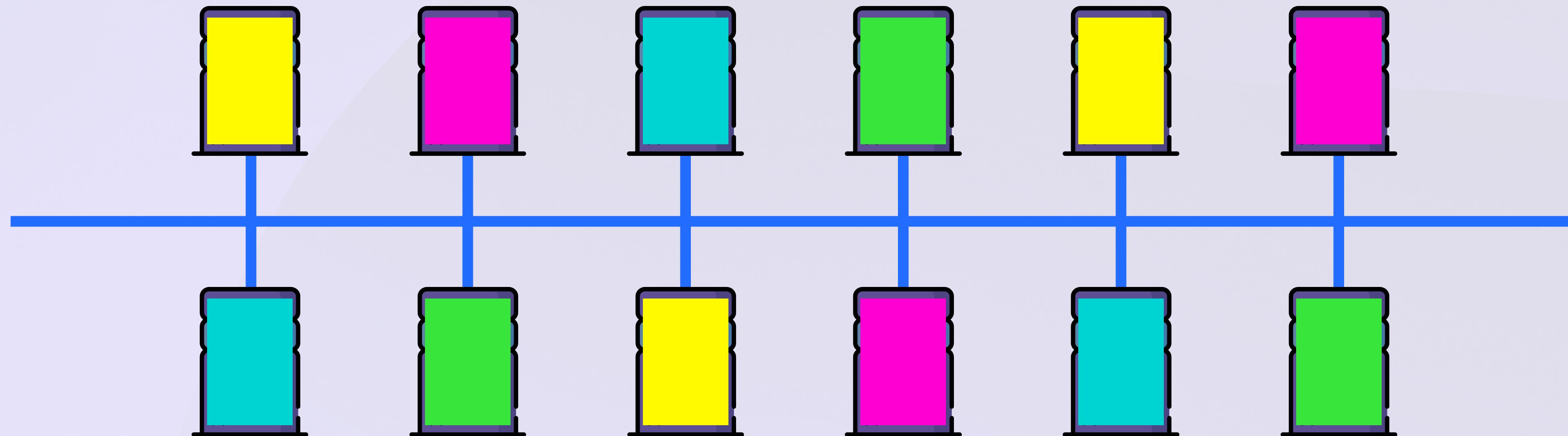
Bien configurados, poseen alta disponibilidad y tolerancia a fallos



PROCESAMIENTO EN BIG DATA

PARTICIONAMIENTO DE DATOS

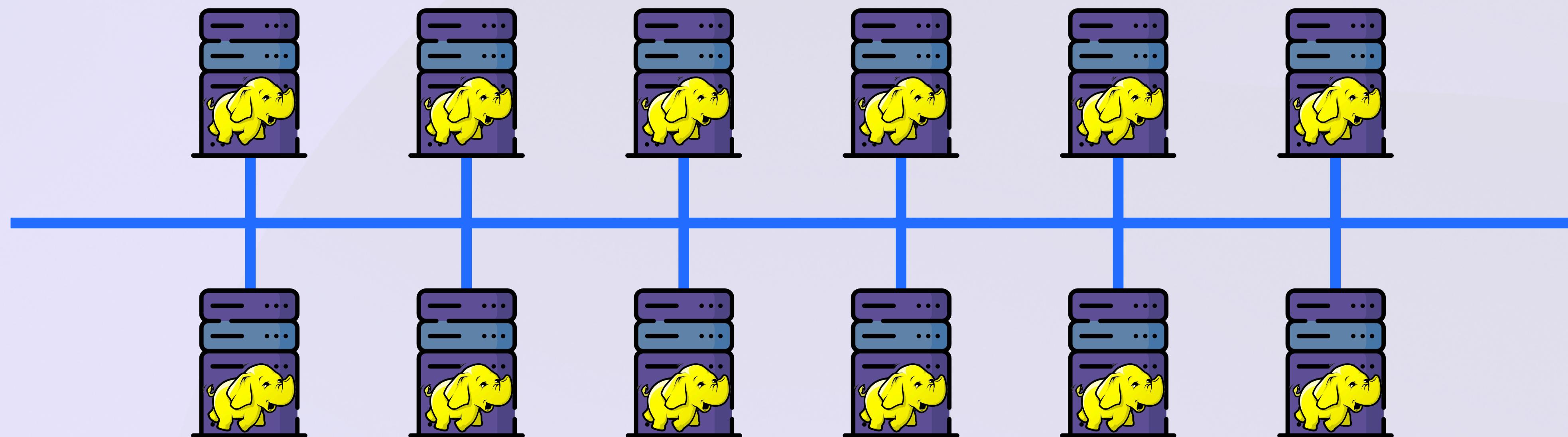
Un solo servidor no soporta almacenar la totalidad de los datos. Se deben particionar los datos en múltiples **servidores del cluster**. Además los datos se encuentran replicados.



PROCESAMIENTO EN BIG DATA

PROCESAMIENTO PARALELO

Varios servidores procesan un mismo programa de forma simultánea para resolver un determinado problema.



TERMINOLOGIA BÁSICA

En Machine Learning generalmente se utilizan arrays 2D y notaciones vectoriales para referirnos a los datos, de la siguiente forma:

- Cada fila del array es una **muestra, observación** o dato puntual
- Cada columna es una **características (feature o atributo)**, de la **observación** mencionada en el punto anterior.
- En el caso más general habrá una columna, que llamaremos **objetivo, label, etiqueta o respuesta**, y que será el valor que se pretende predecir.

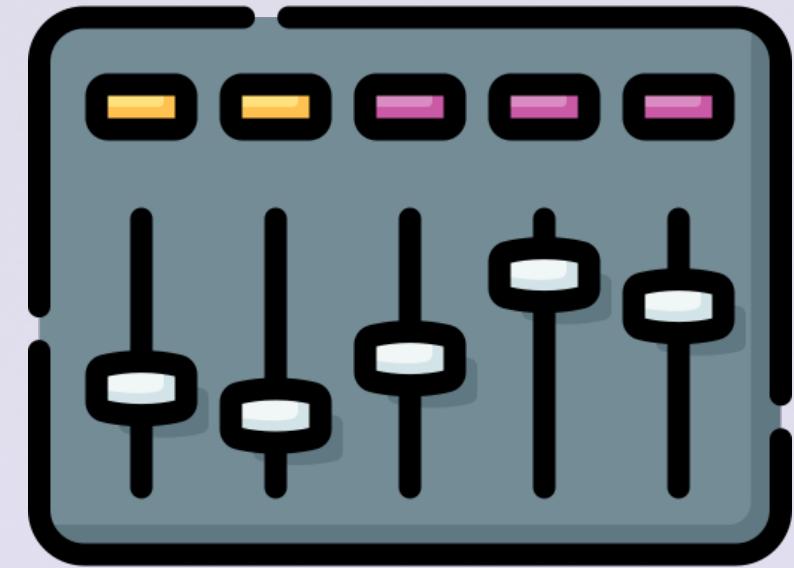
TERMINOLOGIA BÁSICA

Observation →

Features					Label
Position	Experience	Skill	Country	City	Salary (\$)
Developer	0	1	USA	New York	103100
Developer	1	1	USA	New York	104900
Developer	2	1	USA	New York	106800
Developer	3	1	USA	New York	108700
Developer	4	1	USA	New York	110400
Developer	5	1	USA	New York	112300
Developer	6	1	USA	New York	116100
Developer	7	1	USA	New York	117800

TERMINOLOGIA BÁSICA

- Los algoritmos de Machine Learning tienen parámetros “internos” que no dependen de los datos. Estos parámetros se llaman **hiperparámetros**. Por ejemplo, una red neuronal tiene como hiperparametros la función de activación o la constante de entrenamiento.



- Llamamos **generalización** a la capacidad del modelo de hacer predicciones nuevas utilizando datos nuevos.

TIPOS DE APRENDIZAJE

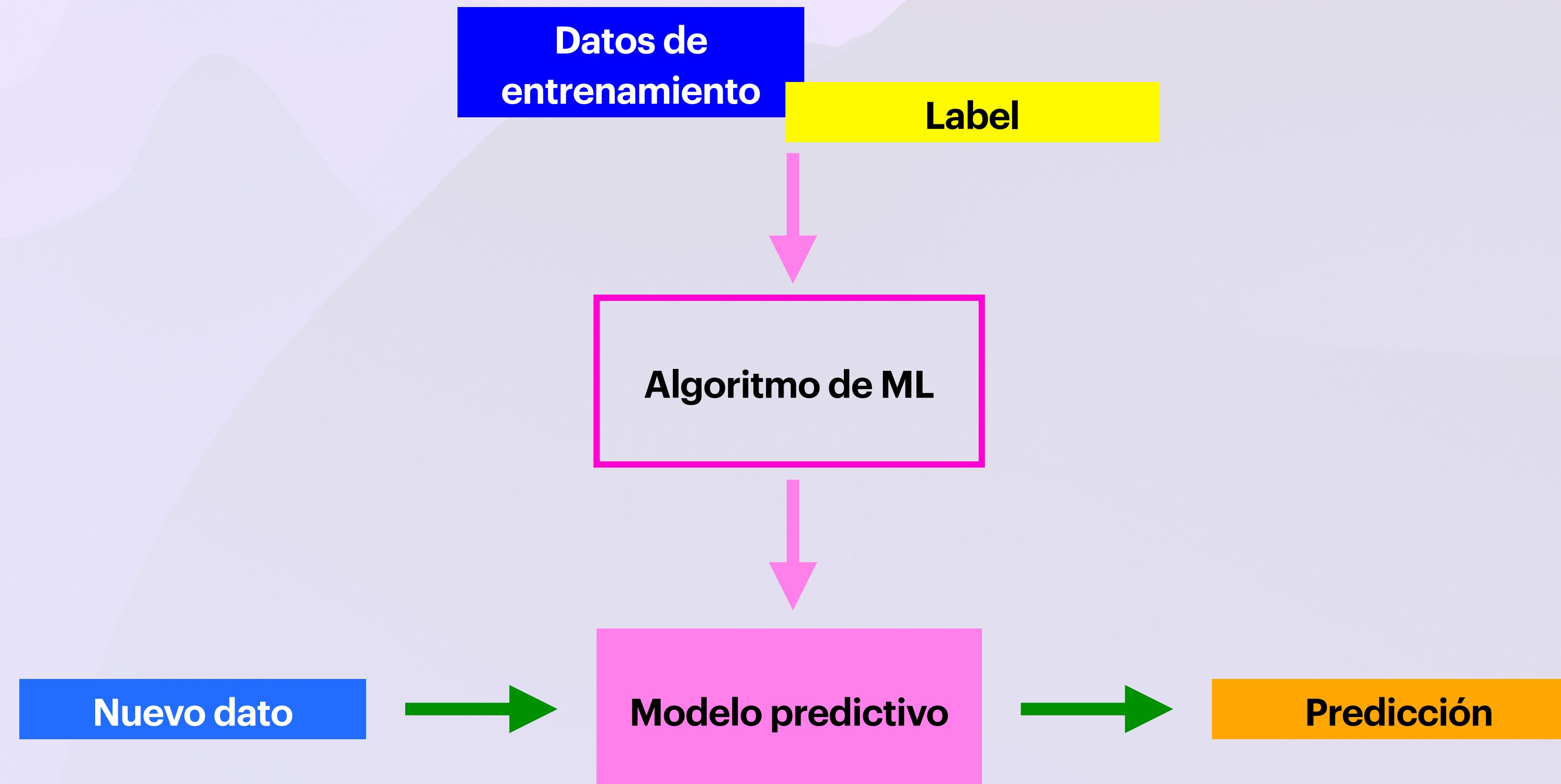
- **Aprendizaje supervisado:** Se refiere a un tipo de modelos de Machine Learning que se entrena con un conjunto de ejemplos en los que los resultados de salida son conocidos.
- **Aprendizaje no supervisado:** El objetivo será la extracción de información significativa, sin la referencia de variables de salida conocidas, y mediante la exploración de la estructura de dichos datos sin etiquetar.
- **Aprendizaje profundo:** Es un subcampo de Machine Learning, que usa una estructura jerárquica de redes neuronales artificiales, que se construyen de una forma similar a la estructura neuronal del cerebro humano, con los nodos de neuronas conectadas.

APRENDIZAJE SUPERVISADO

- Los modelos aprenden de los resultados conocidos y realizan ajustes en sus parámetros interiores para adaptarse a los datos de entrada.
- Una vez que el modelo es entrenado adecuadamente, y los parámetros internos son coherentes con los datos de entrada y los resultados de los datos de entrenamiento, **el modelo podrá realizar predicciones adecuadas ante nuevos datos**



APRENDIZAJE SUPERVISADO



APRENDIZAJE SUPERVISADO

CLASIFICACIÓN

Clasificación es una subcategoría de aprendizaje supervisado en la que el objetivo es predecir clases categóricas (valores discretos, no ordenados, pertenencia a grupos).

- Detección de SPAM... clasificación binaria (es spam o no es spam).
- Clasificación multi-clase: Múltiple clases, como por ejemplo clasificación del nivel socioeconómico de una persona (alta, media y baja).

APRENDIZAJE SUPERVISADO

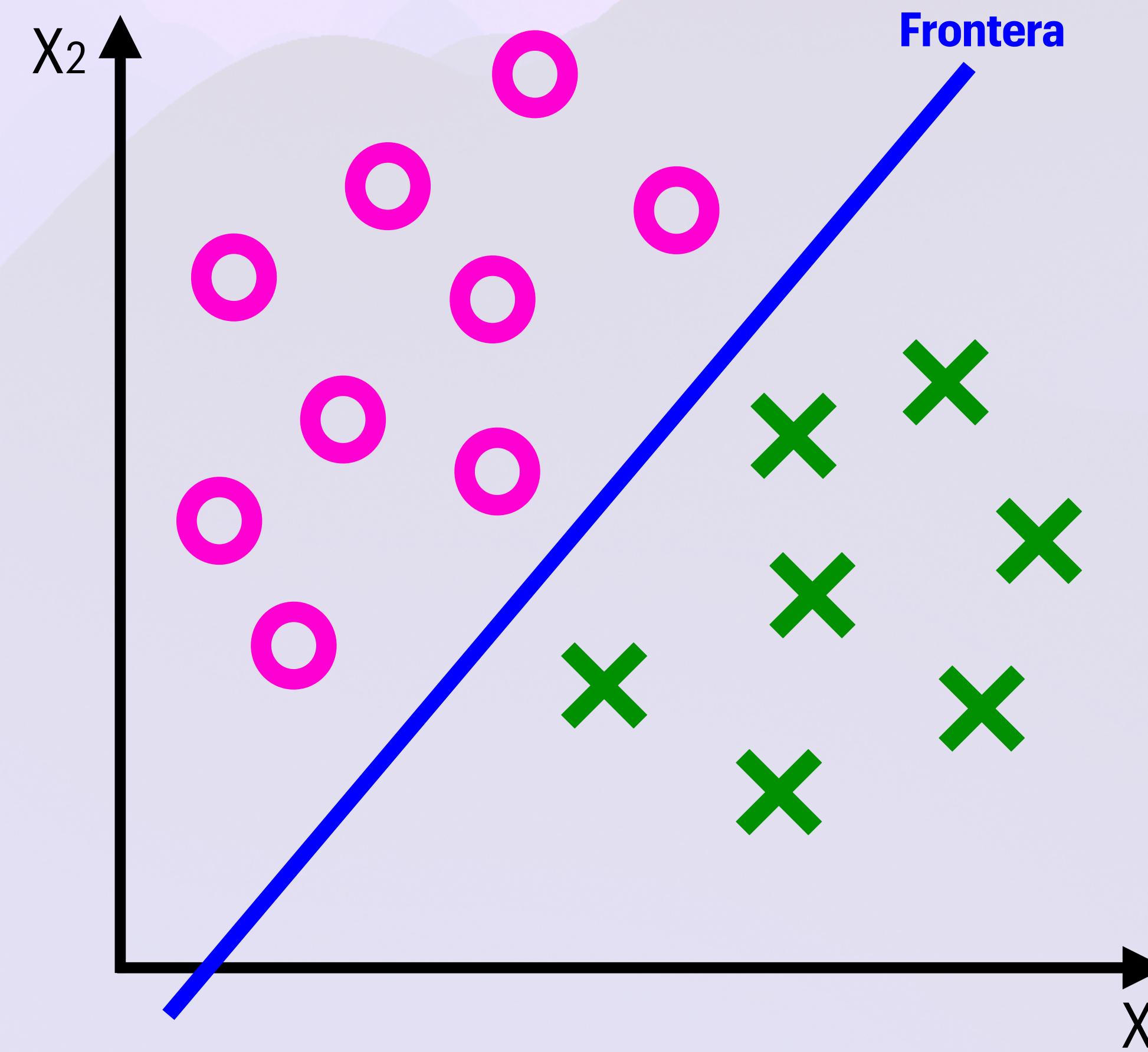
CLASIFICACIÓN

Un ejemplo de clasificación binaria: hay **dos clases** de objetos, círculos y cruces, y dos **características** de los objetos, X_1 y X_2

- Se entrena un modelo que dado las dos características puede predecir la clase (circulo o cruz).
- Lo que hace el modelo es crear una frontera que le permite separar a las dos clases. Si tenemos un nuevo dato, en función de que lado de la frontera queda, se clasificará de una forma u otra.

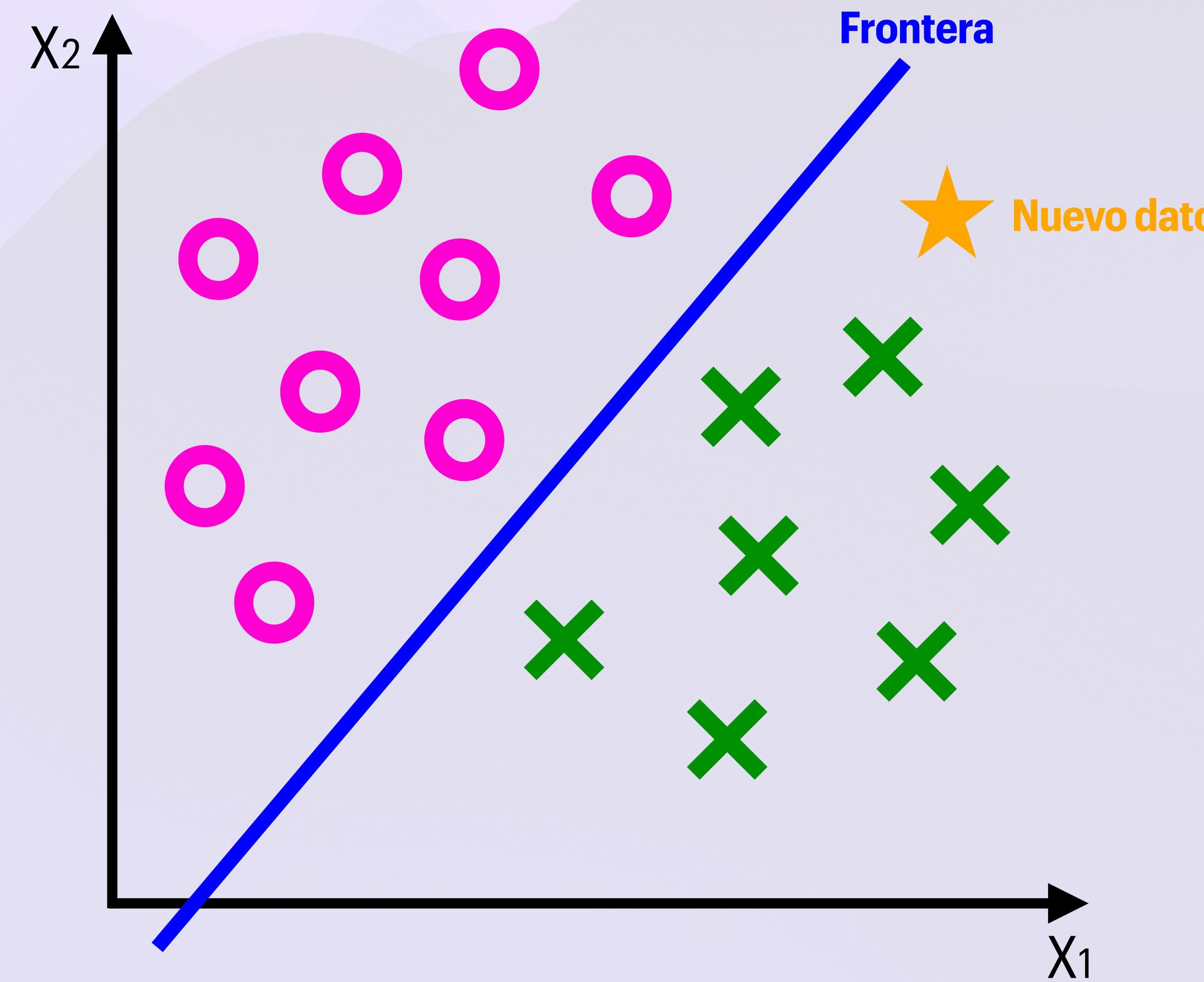
APRENDIZAJE SUPERVISADO

CLASIFICACIÓN



APRENDIZAJE SUPERVISADO

CLASIFICACIÓN



APRENDIZAJE SUPERVISADO

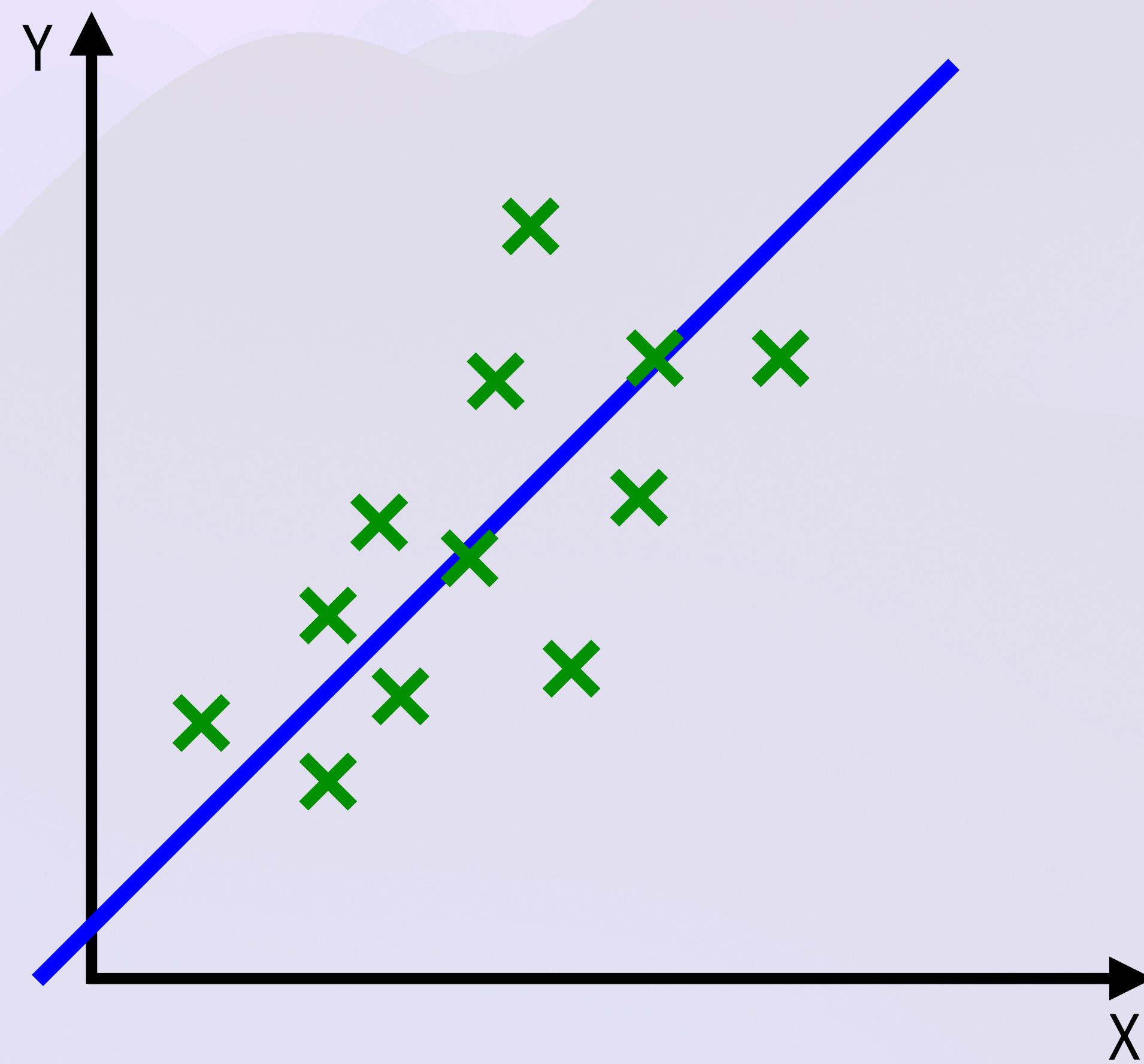
REGRESIÓN

En este tipo de aprendizaje tenemos un número de **variables predictoras** (explicativas) y una variable de **respuesta continua** (resultado), y se tratará de encontrar una relación entre dichas variables que nos proporciona un resultado continuo.

- La regresión lineal: dados X e Y, establecemos una línea recta que minimice la distancia (con el método de mínimos cuadrados) entre los puntos de muestra y la línea ajustada. Después, utilizaremos la fórmula obtenida de la regresión para predecir nuevos datos de salida.
- Un ejemplo típico es la regresión del precio de casas en ventas en una ciudad dado barrio, cantidad de habitaciones, etc.

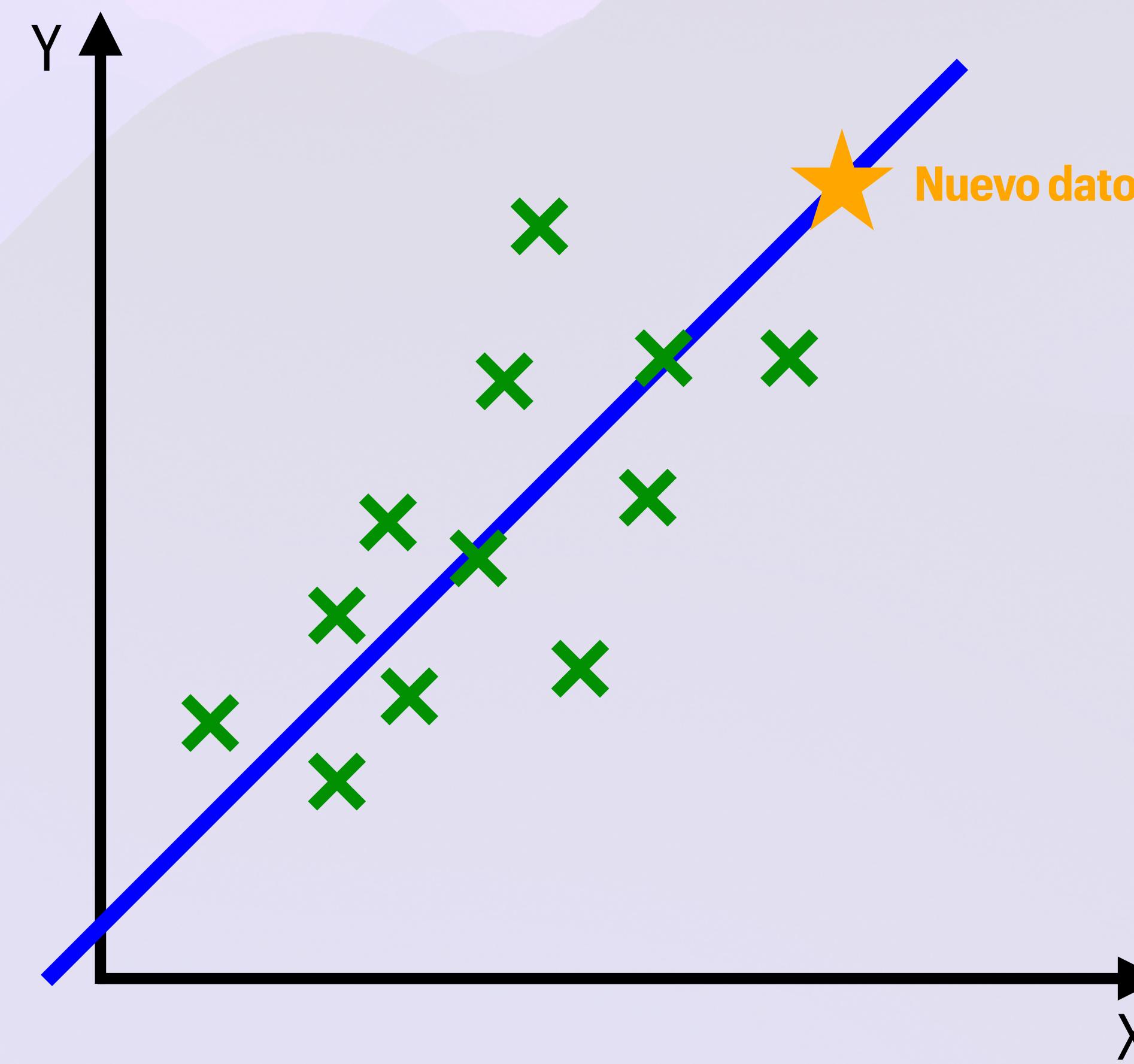
APRENDIZAJE SUPERVISADO

REGRESIÓN



APRENDIZAJE SUPERVISADO

REGRESIÓN



APRENDIZAJE NO SUPERVISADO

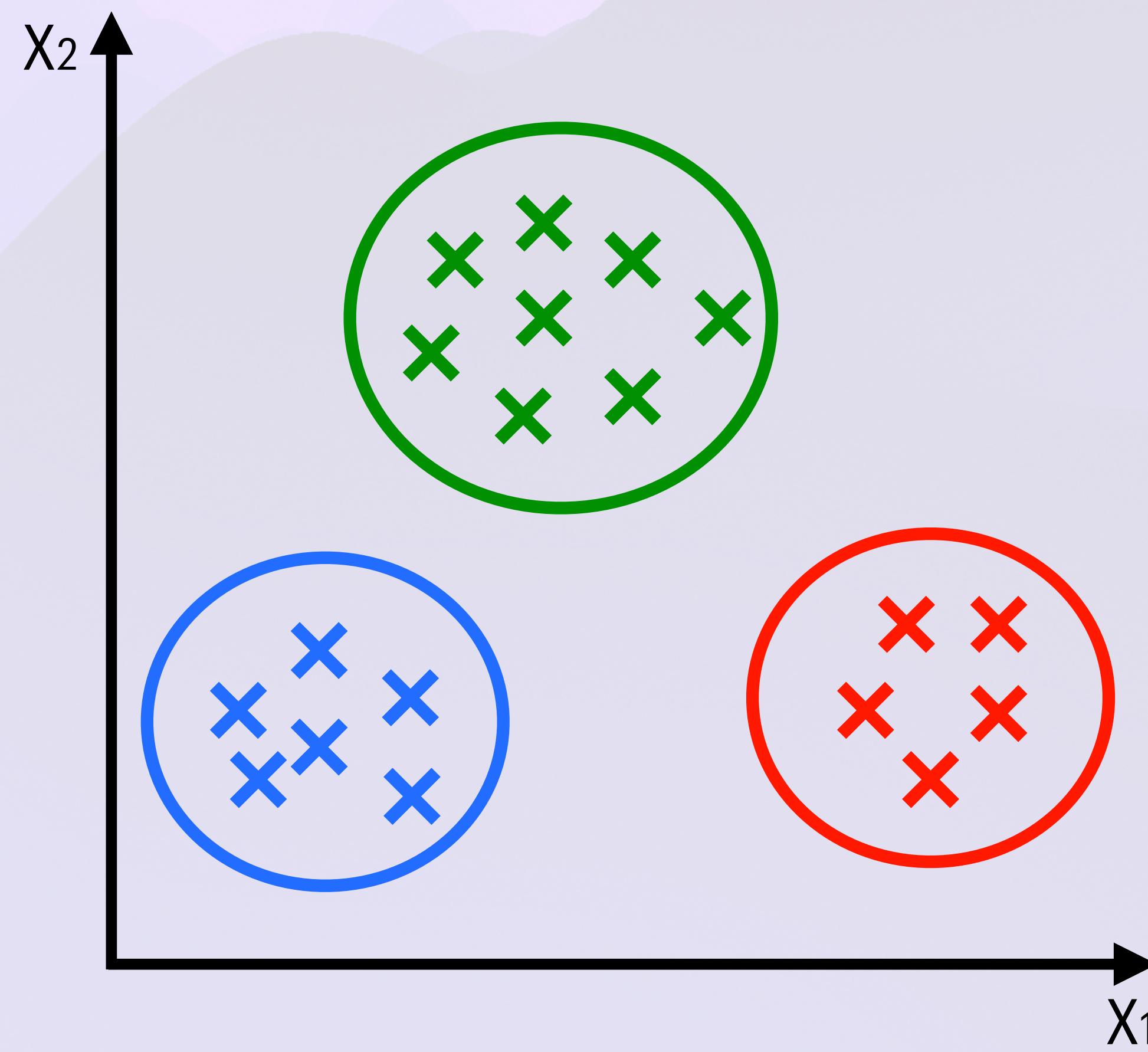
AGRUPAMIENTO O CLUSTERING

El agrupamiento es una técnica exploratoria de análisis de datos, que se usa para organizar información en grupos con significado sin tener conocimiento previo de su estructura.

- Cada grupo es un conjunto de objetos similares que se diferencia de los objetos de otros grupos.
- El objetivo es obtener un número de grupos de características similares.
- Un ejemplo de aplicación es establecer dado el comportamiento de muchas cadenas de comida, como se agrupan en el mercado en diferentes grupos para evaluar verdadero competidores.

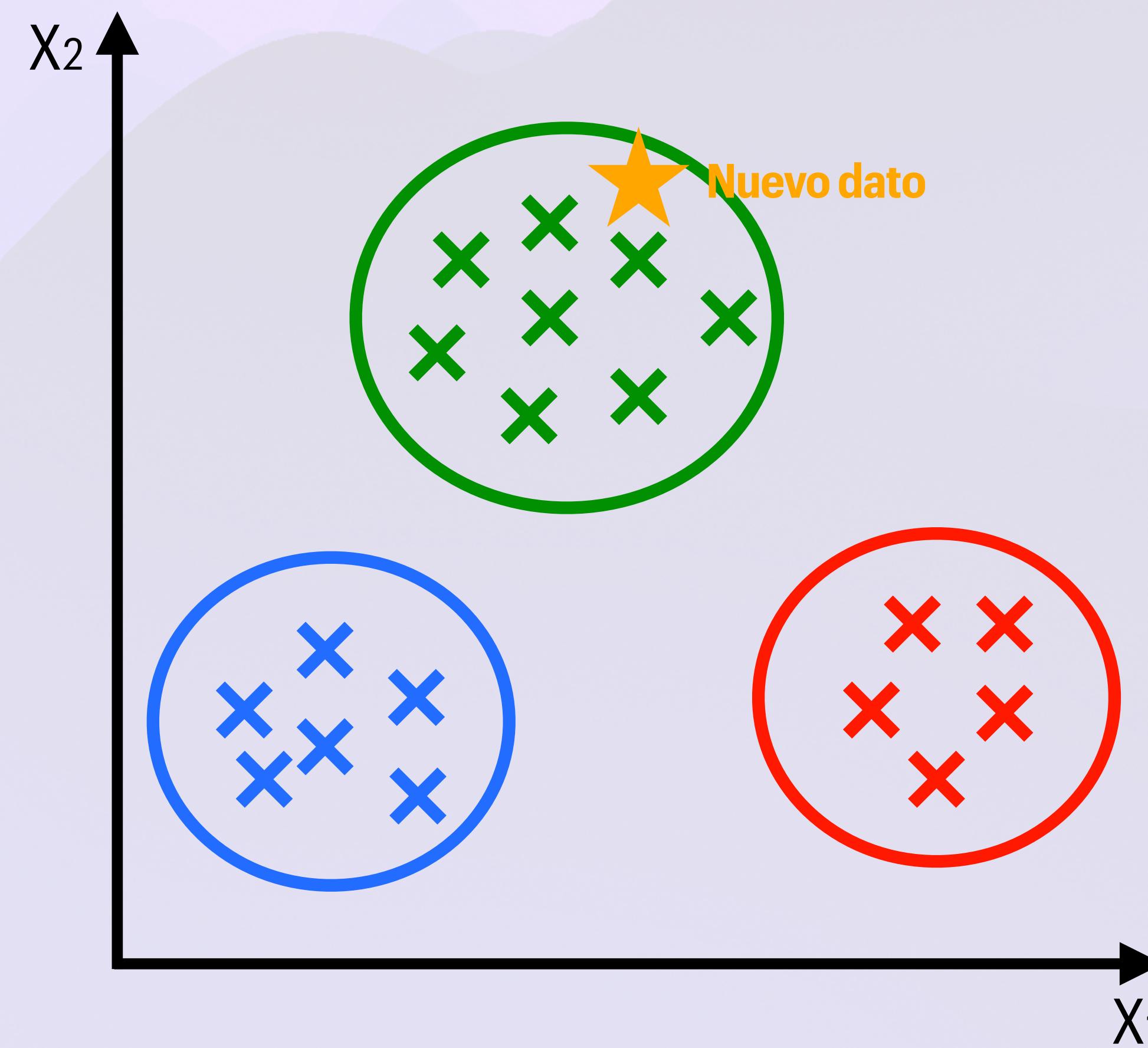
APRENDIZAJE NO SUPERVISADO

AGRUPAMIENTO O CLUSTERING



APRENDIZAJE NO SUPERVISADO

AGRUPAMIENTO O CLUSTERING

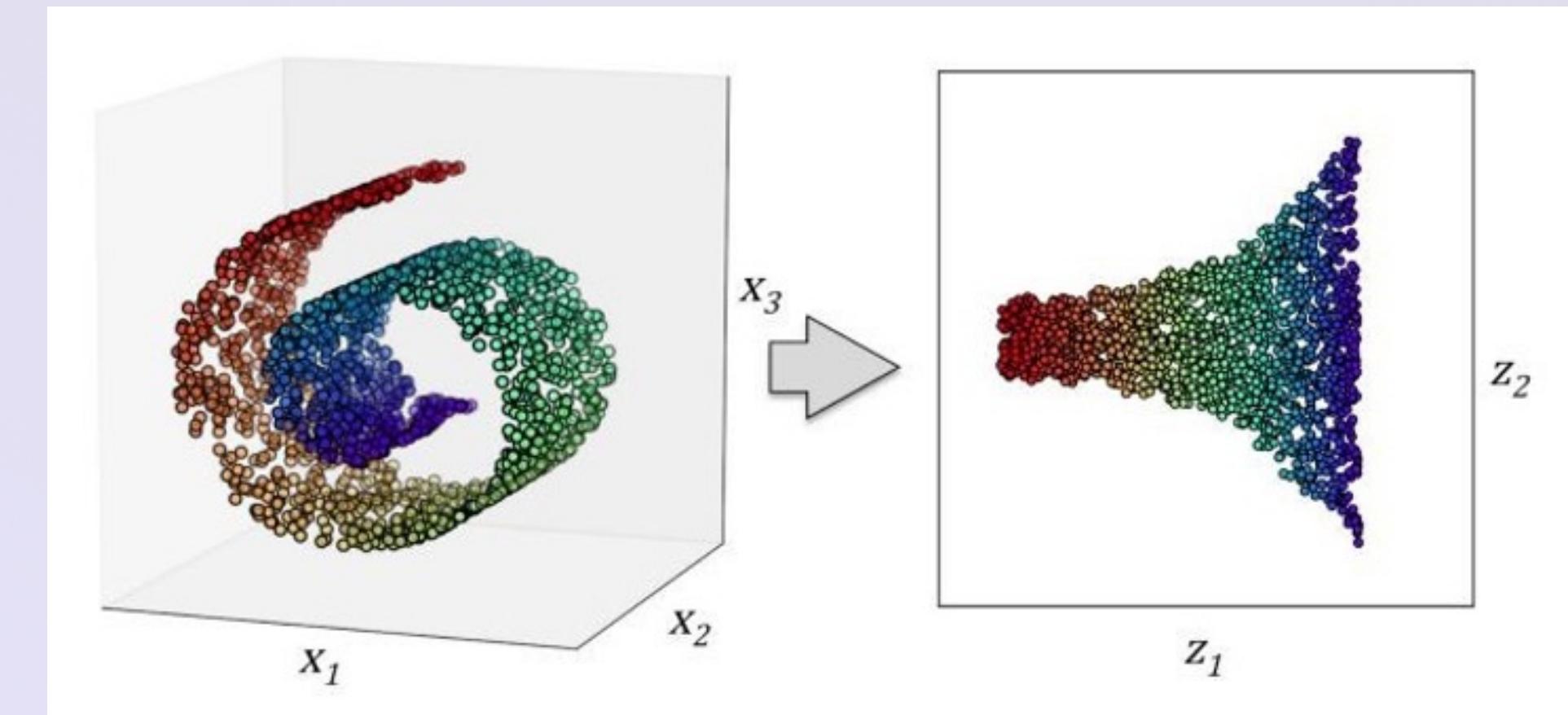


APRENDIZAJE NO SUPERVISADO

REDUCCIÓN DIMENSIONAL

La **reducción dimensional** funciona encontrando correlaciones entre las características, lo que implica que existe información redundante, ya que alguna característica puede explicarse parcialmente con otras (por ejemplo, puede existir dependencia lineal).

Estas técnicas eliminan *ruido* de los datos (que puede también empeorar el comportamiento del modelo), y comprimen los datos en un sub-espacio más reducido, al tiempo que retienen la mayoría de la información relevante



APRENDIZAJE PROFUNDO

Esta arquitectura permite abordar el análisis de datos de forma no lineal.

La primera capa de la red neuronal toma datos en bruto como entrada, los procesa, extrae información y la transfiere a la siguiente capa como salida.

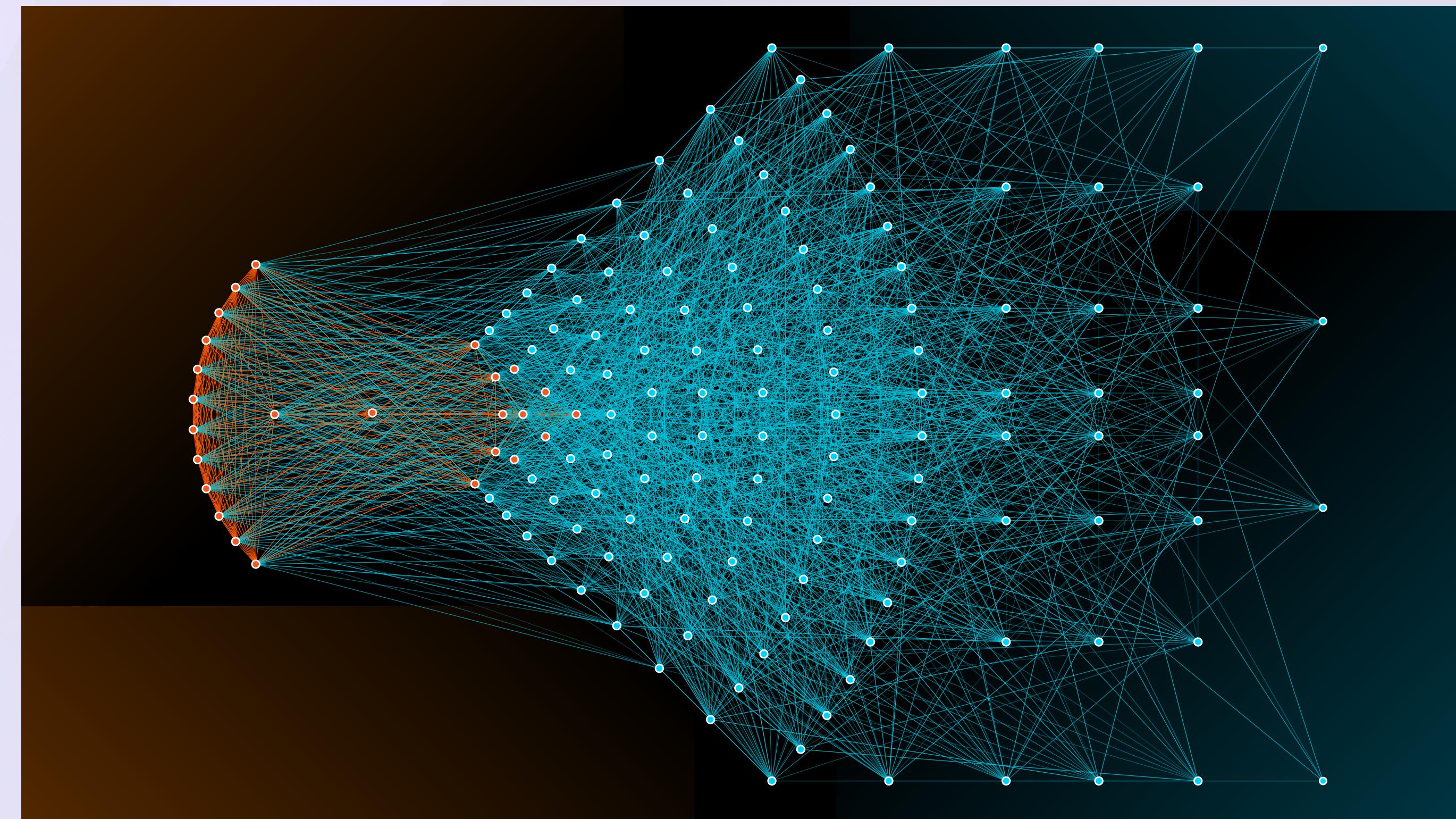
Este proceso se repite en las siguientes capas, cada capa procesa la información proporcionada por la capa anterior, y así sucesivamente hasta que los datos llegan a la capa final, que es donde se obtiene la predicción.

Esta predicción se compara con el resultado conocido, y así por análisis inverso el modelo es capaz de aprender los factores que conducen a salidas adecuadas.

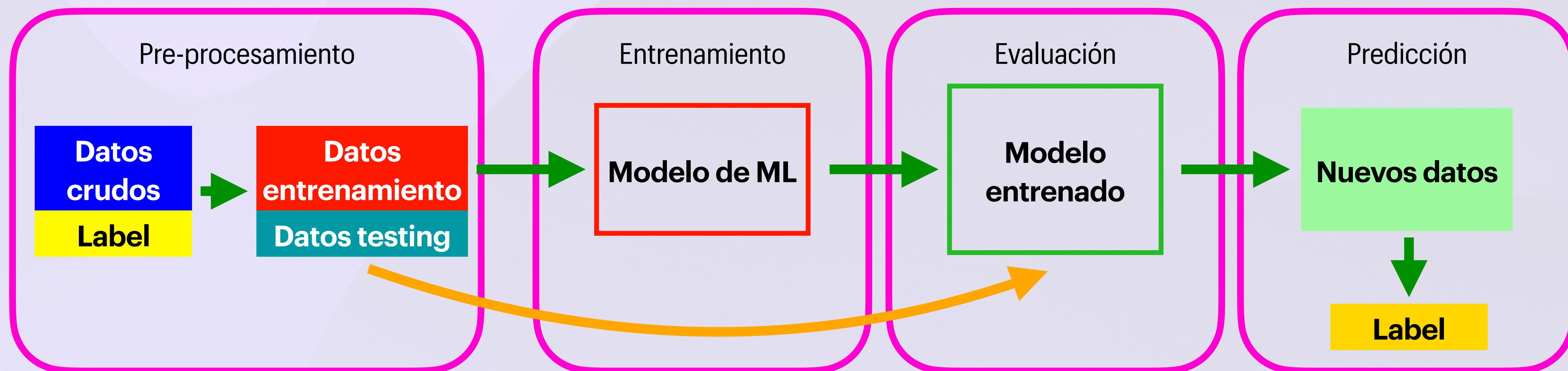
Es uno de los principales algoritmos utilizados en la creación de aplicaciones y programas para reconocimiento de imágenes.

APRENDIZAJE PROFUNDO

Para tener buenos resultados, necesitamos muchos datos y en general son "caras" de entrenar.



METODOLOGIA PARA CONSTRUIR ALGORITMOS DE ML



METODOLOGIA PARA CONSTRUIR ALGORITMOS DE ML

PRE-PROCESAMIENTO

- Es el paso mas importante. Recordar: Garbage in, garbage out.
- Usualmente los datos se presentan en formatos no óptimos (o incluso inadecuados) para ser procesados por el modelo.
- Muchos algoritmos requieren que las características estén en la misma escala para optimizar su rendimiento, lo que se realiza frecuentemente aplicando técnicas de normalización o estandarización en los datos.
- Podemos también encontrar en algunos casos que las características seleccionadas están correlacionadas, y por tanto son redundantes para extraer información con significado correcto de ellas. En este caso tendremos que usar técnicas de reducción dimensional para comprimir las características en subespacios con menores dimensiones.
- Hacer un correcto análisis estadístico es vital para lograr tener los datos para el algoritmo que queremos usar.

METODOLOGIA PARA CONSTRUIR ALGORITMOS DE ML

SELECCIÓN Y ENTRENAMIENTO DE MODELO

- Es esencial comparar los diferentes algoritmos de un grupo para entrenar y seleccionar el de mejor rendimiento. Para realizar esto, es necesario seleccionar una métrica para medir el rendimiento del modelo.
- Cuanto no tenemos nada para comparar inicialmente, elegimos un modelo sencillo que llamamos **baseline**, que nos servirá para comparar el desarrollo de nuevos modelos. No siempre debe ser un modelo.
- Para asegurarnos de que nuestro modelo funcionará adecuadamente con datos reales, podemos usar **validación cruzada (Cross Validation)** antes de utilizar el conjunto de datos de prueba para la evaluación final del modelo.
- En general, los parámetros por defecto de los algoritmos de Machine Learning proporcionados por las librerías no son los mejores para utilizar con nuestros datos, por lo que usaremos técnicas de optimización de **hiperparámetros**

METODOLOGIA PARA CONSTRUIR ALGORITMOS DE ML

EVALUANDO Y PREDICIENDO CON DATOS NUEVOS

- Una vez que hemos seleccionado y ajustado un modelo a nuestro conjunto de datos de entrenamiento, podemos usar los datos de prueba para estimar el rendimiento del modelo en los datos nuevos, por lo que podemos hacer una estimación del error de **generalización** del modelo, o evaluarlo utilizando alguna otra métrica.
- Si el modelo cumple nuestros objetivos, luego es ponerlo productivo en el contexto que se quiere utilizar. En esta materia no veremos este paso.