Centro de Transformación Digital

# Aprendizaje No Supervisado

## Conceptos Básicos

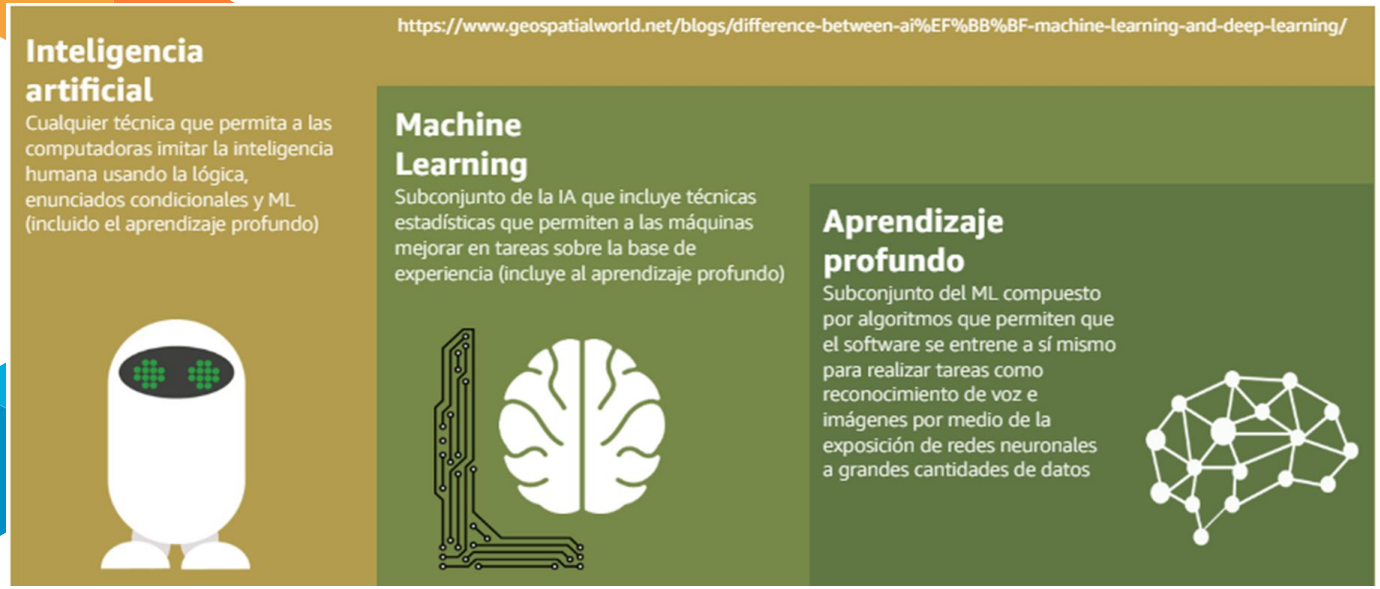
Universidad del Desarrollo
Facultad de Ingeniería
Centro de Transformación Digital

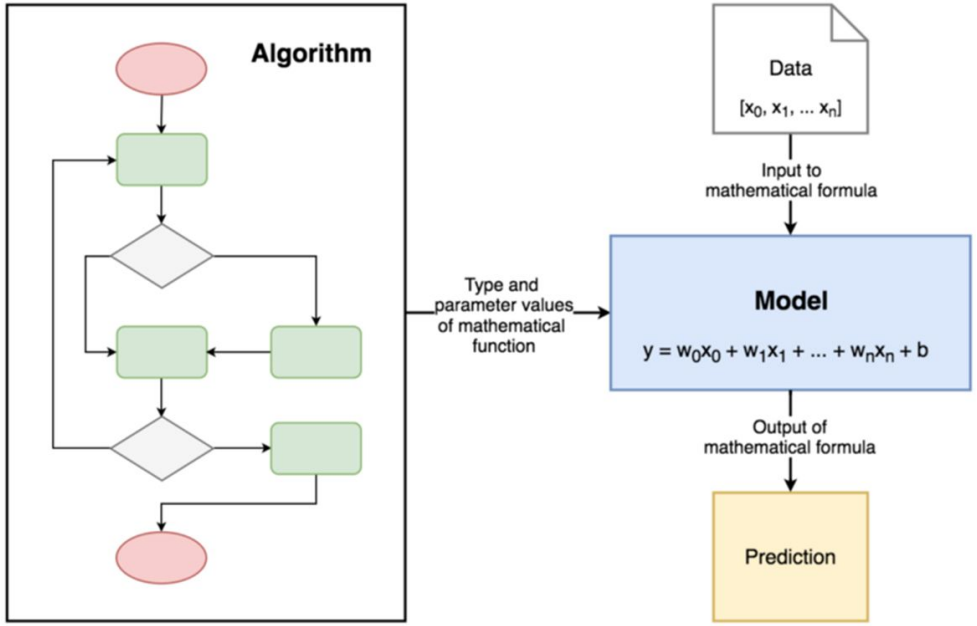
Proyecto apoyado por
CORFO
Gobierno de Chile

# Índice de la clase

- Conceptos Iniciales
- Machine Learning
- Tipos de Aprendizaje
- Aplicaciones
- Algoritmos de Machine Learning
- Etapas de un Proyecto de Machine Learning
- Aplicaciones Reales

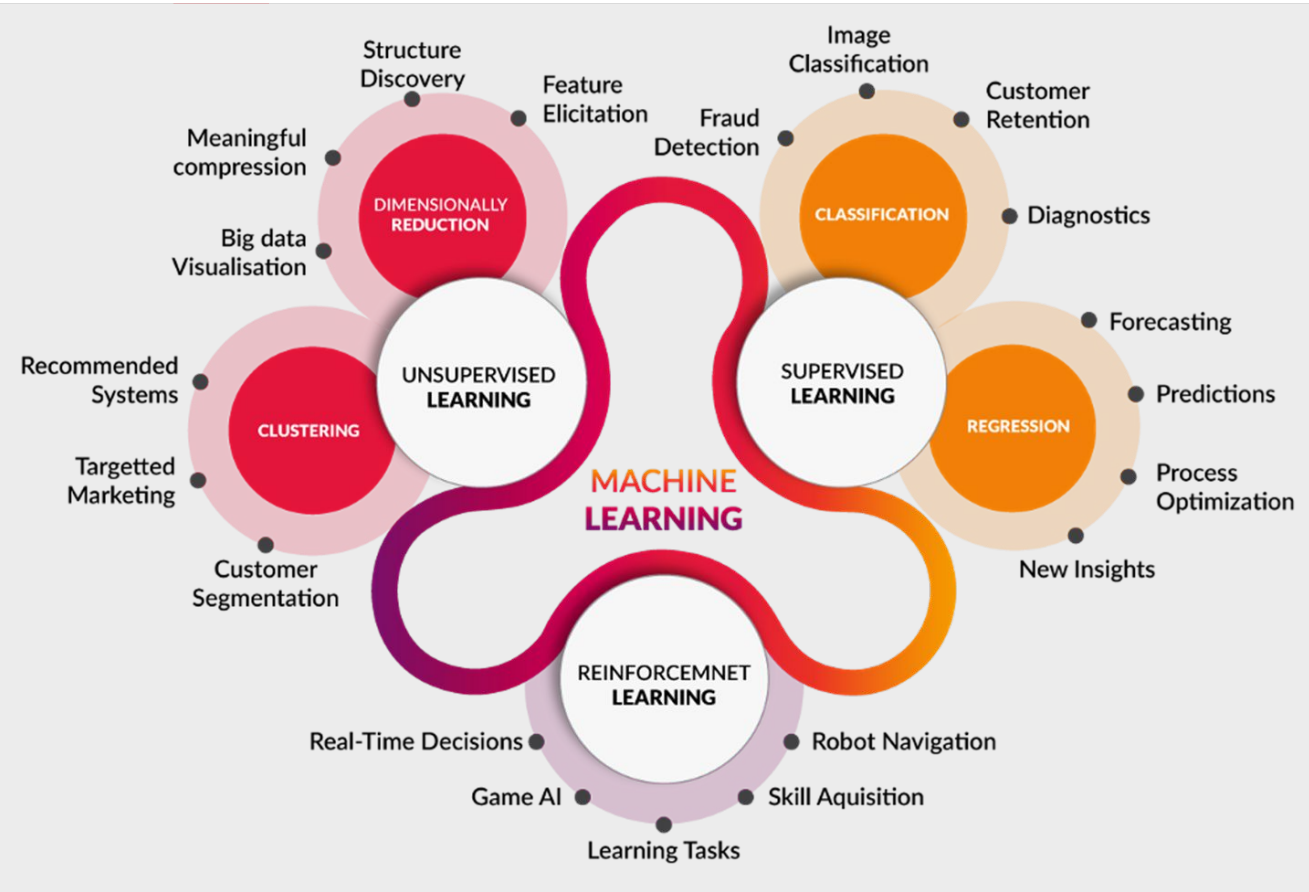
# ¿IA, ML, Deep Learning?




- La base de todo modelo son los **datos**.
- Los modelos aprenden de los datos **patrones** que si fuéramos a programarlos serían cientos o miles de líneas de código.
- El objetivo es que el modelo pueda **generalizar** para dar predicciones sobre datos con los cuales no fue entrenado

# Machine Learning



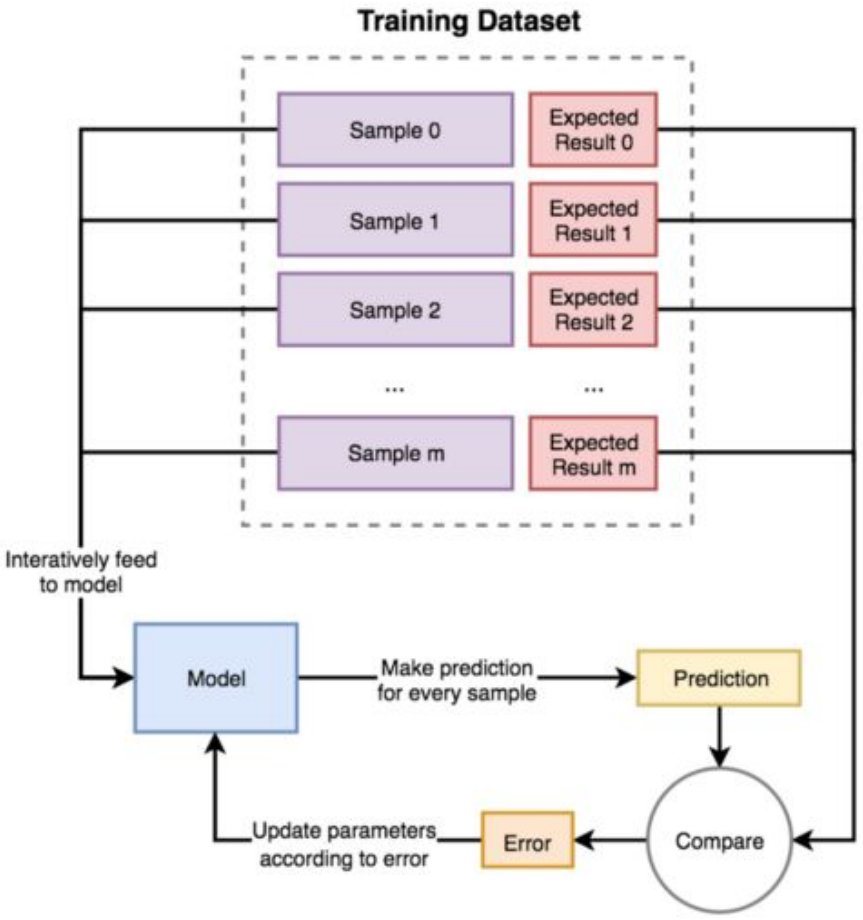

Dentro del ML encontramos 2 enfoques principales, el **aprendizaje supervisado** y **el aprendizaje no supervisado.**

- El **aprendizaje supervisado** consiste en predecir los valores de un conjunto de datos de salida, a partir de un conjunto de datos de entrada. Se le llama supervisado porque conforme el modelo predice las salidas para datos de prueba, se calcula el error entre lo que predijo el algoritmo y el valor real. El objetivo es minimizar el error, ajustando la función de densidad de probabilidad que relaciona las entradas con las salidas.

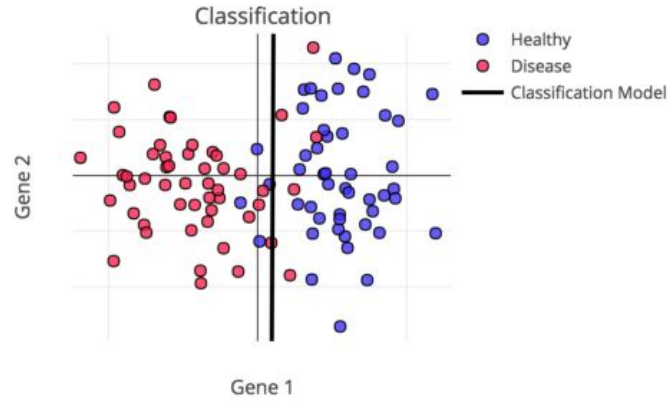
- En el **aprendizaje no supervisado**, solo se tienen conjuntos de datos de entrada, sin conocer su relación con variables de salida, por lo que no tiene como verificar fácilmente si el desempeño del modelo es el adecuado. En esta categoría, los modelos de ML realizan principalmente agrupaciones o clasificaciones de los datos, según las variables de entrada y su diferenciación. Son modelos con métricas no muy precisas y que en ocasiones dependen de ciertas heurísticas para validar sus resultados.

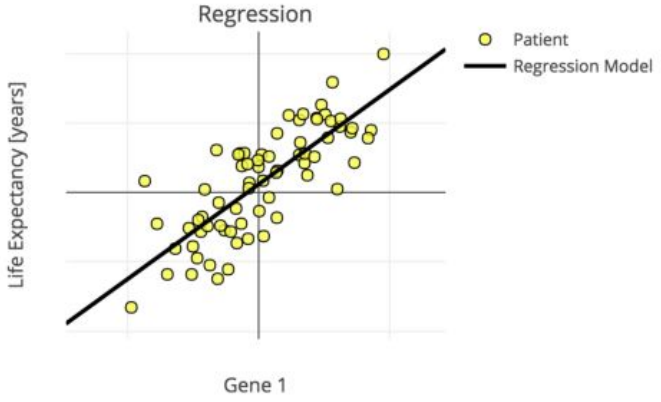
*El **aprendizaje por refuerzo** es una categoría de algoritmos de aprendizaje automático en la cual un agente interactúa con un entorno y aprende a tomar decisiones para maximizar una recompensa acumulada a lo largo del tiempo. El agente toma una serie de acciones en el entorno y, en función de esas acciones, el entorno responde con una recompensa o una penalización. El objetivo del agente es aprender una política óptima que determine qué acción tomar en cada estado para maximizar la recompensa total a lo largo de una secuencia de acciones.

# Modelos Supervisados

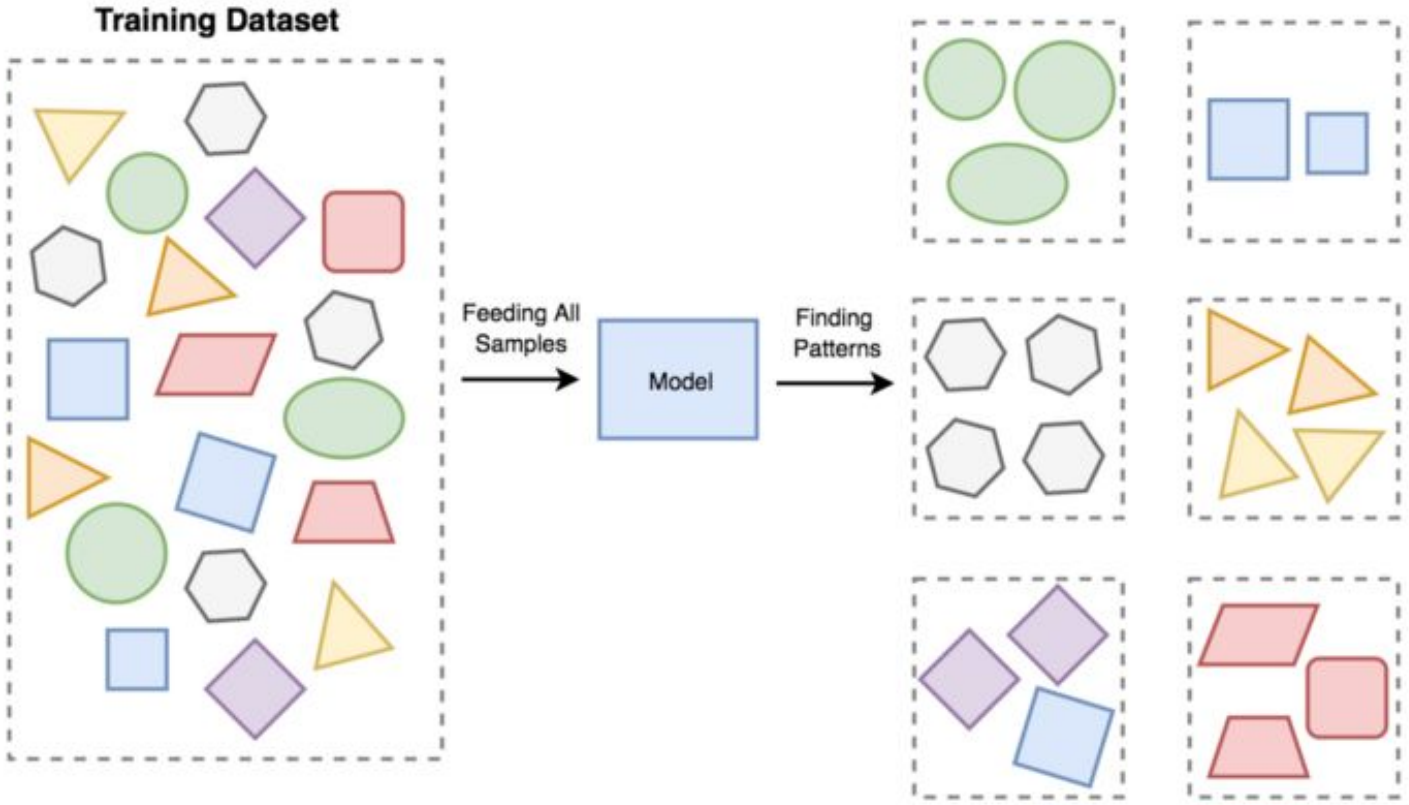



- Se requieren datos ya marcados con el objetivo (**Ground Truth**)
- Se considera un loop de **feedback** sobre el set de entrenamiento y un proceso para ajustar el modelo
- Los modelos intentan encontrar los parámetros que les permitirán performar de manera correcta en **datos desconocidos**.
- Clasificación y Regresion



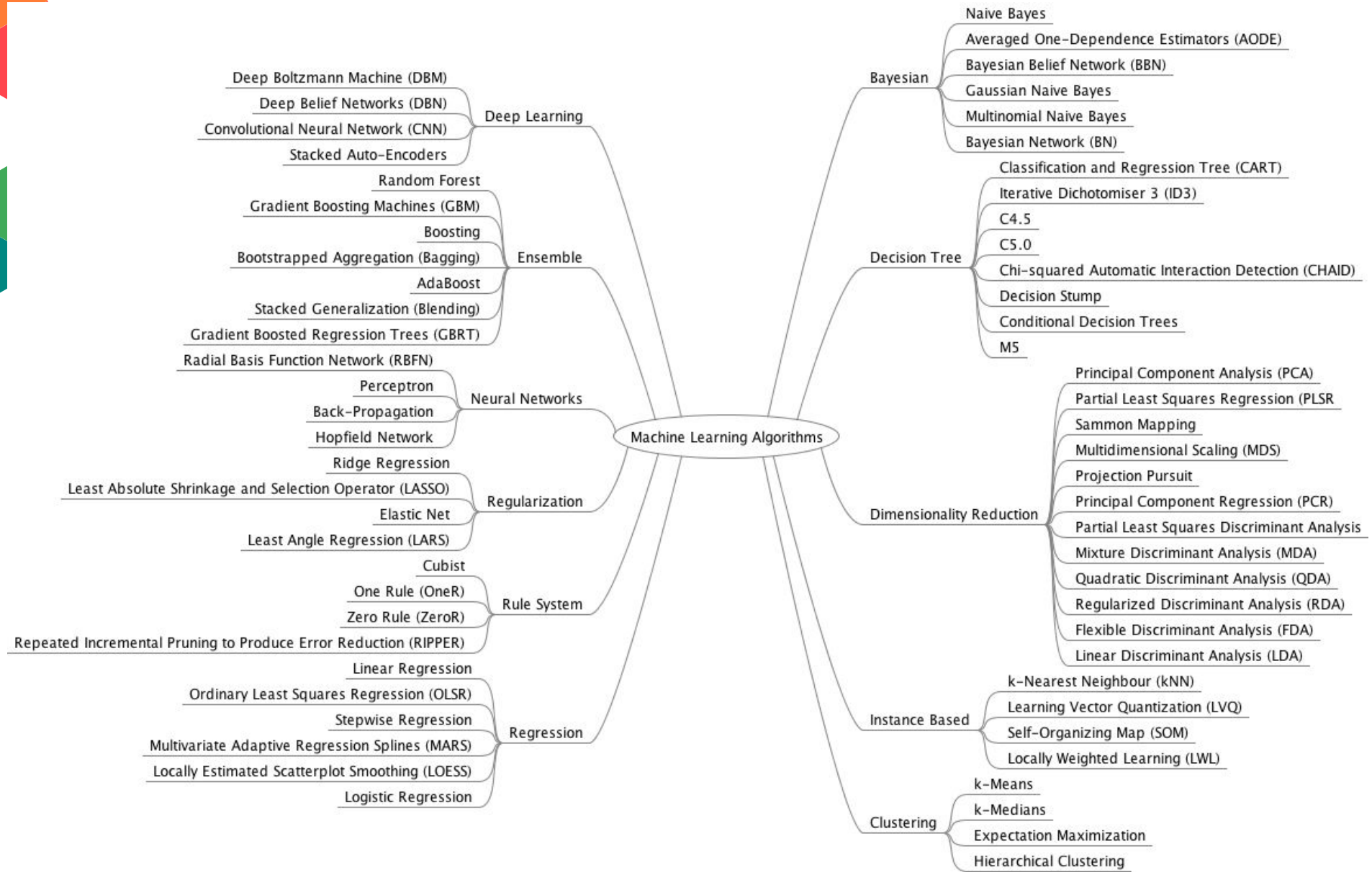



# No Supervisados



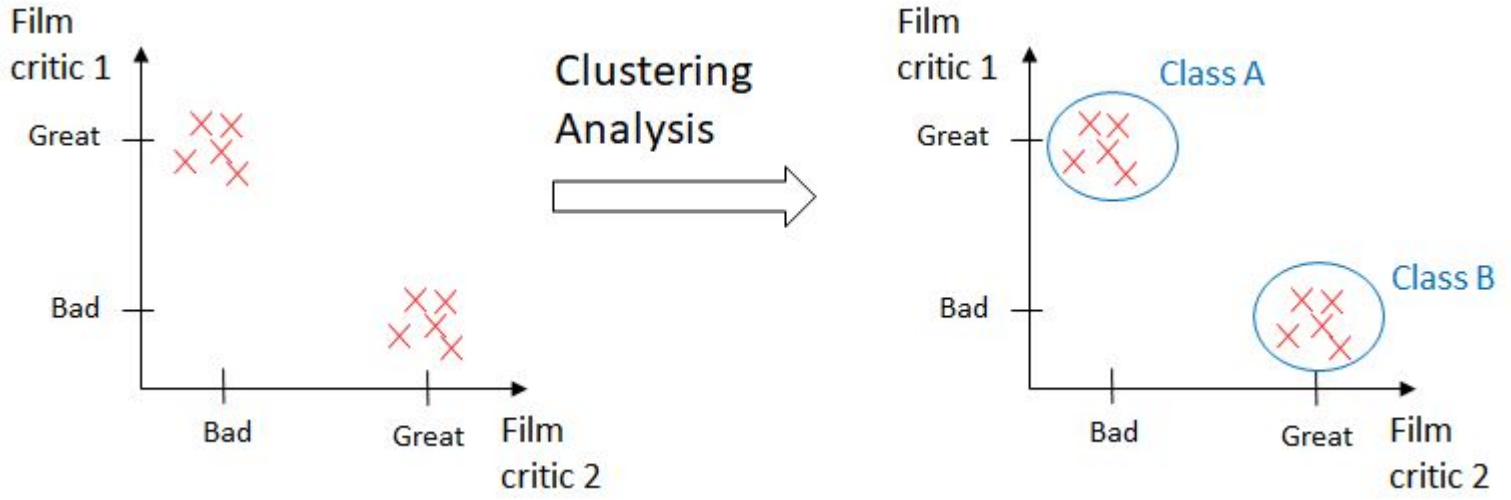

- **No requieren** datos marcados con objetivo.
- Se dice que “no hay respuestas correctas”.
- En base a los datos se aprenden los patrones para **generar agrupaciones**.
- Clustering, reducción de dimensiones y reglas de asociación.

# Algoritmos en Machine Learning



# Clustering

En términos básicos, el objetivo de la clusterización es encontrar diferentes grupos dentro de los elementos de los datos. Para ello, los algoritmos de agrupamiento encuentran la estructura en los datos de manera que los elementos del mismo clúster sean más similares entre sí que con los elementos de clústeres diferentes.



# Clustering

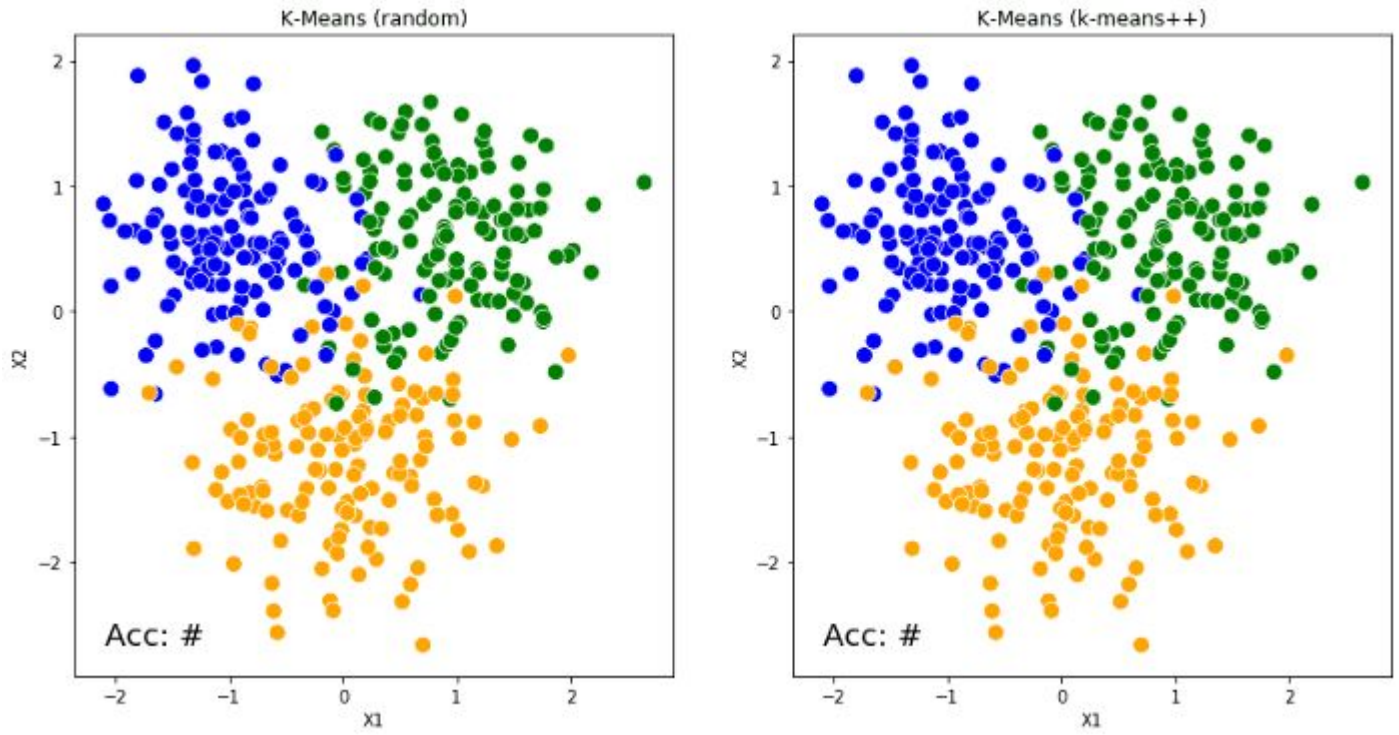
Estos algoritmos de aprendizaje no supervisado tienen una gama increíblemente amplia de aplicaciones y son muy útiles para resolver problemas del mundo real como la detección de anomalías, los sistemas de recomendación, la agrupación de documentos o la búsqueda de clientes con intereses comunes basados en sus compras.

Algunos de los algoritmos de agrupación más comunes son:

- **K-Means**
- **Clusterización Jerárquica**
- **Density Based Scan Clustering (DBSCAN)**
- **Modelo de Agrupamiento Gaussiano**

# Algoritmo K-means

El algoritmo K-means tiene como objetivo encontrar y agrupar en clases los puntos de datos que tienen una alta similitud entre ellos. En los términos del algoritmo, esta similitud se entiende como lo opuesto de la distancia entre puntos de datos. Cuanto más cerca estén los puntos de datos, más similares y con más probabilidades de pertenecer al mismo clúster serán.



# Pasos del Algoritmo K-means

1. Elegir k, el número de clusters que queremos que nos encuentren.
2. Luego, el algoritmo seleccionará aleatoriamente los centroides de cada grupo.
3. Se asignará cada punto de datos al centroide más cercano (utilizando la distancia euclídea).
4. Se calculará la inercia del conglomerado.
5. Los nuevos centroides se calcularán como la media de los puntos que pertenecen al centroide del paso anterior. En otras palabras, calculando el error cuadrático mínimo de los puntos de datos al centro de cada cluster, moviendo el centro hacia ese punto.
6. Volver al paso 3.

## Hiperparámetros K-means

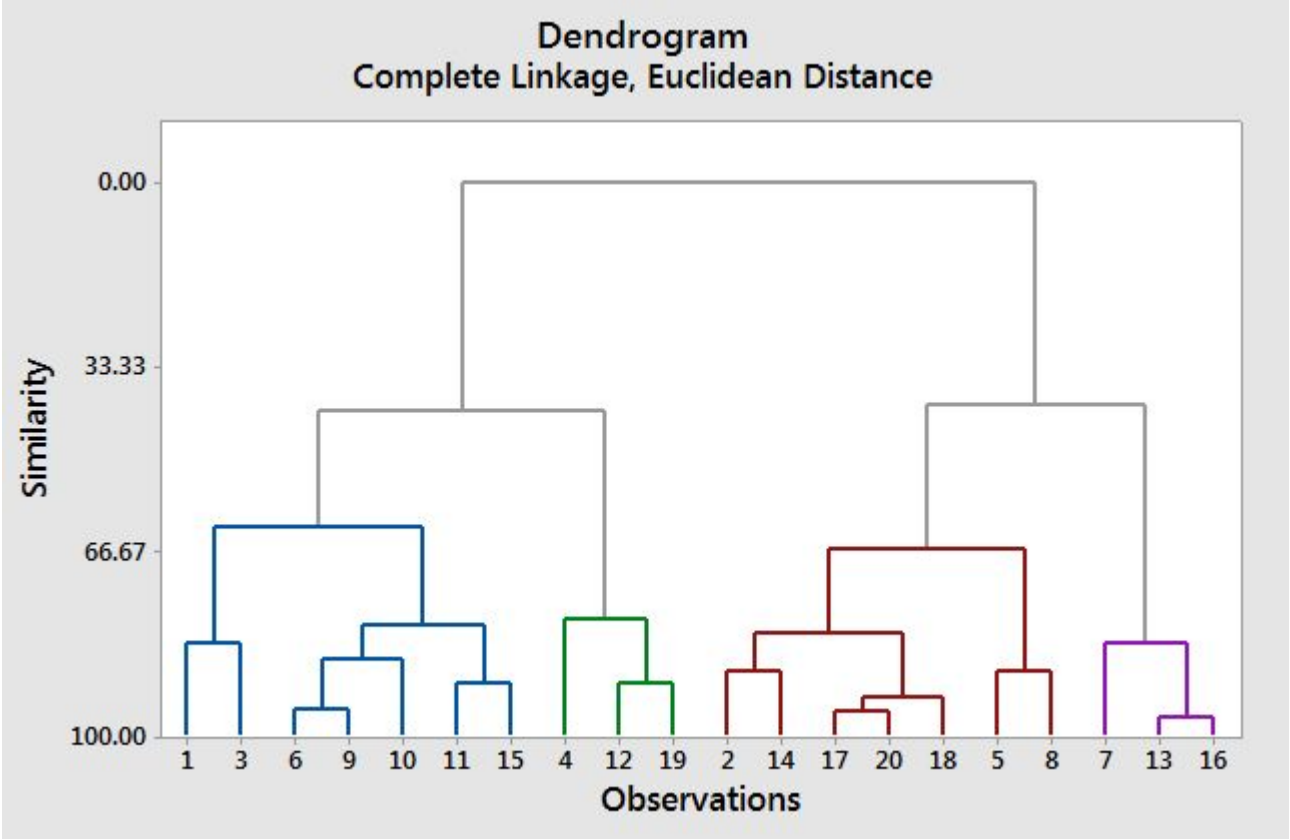
- Número de grupos: El número de clusters y centros de generación.
- Máximas iteraciones: del algoritmo para una sola ejecución.
- Número inicial: El número de veces que el algoritmo se ejecutará con diferentes semillas de centroide. El resultado final será el mejor rendimiento del número definido de corridas consecutivas, en términos de inercia.

## Limitaciones K-means

Aunque K-means es un gran algoritmo de agrupación, es más útil cuando sabemos de antemano el número exacto de grupos y cuando estamos tratando con distribuciones esféricas.

# Clustering Jerárquico

La agrupación jerárquica es una alternativa a los algoritmos de agrupación basados en prototipos. La principal ventaja de la agrupación jerárquica es que no es necesario especificar el número de agrupaciones, **la encontrará por sí misma**. Además, permite el trazado de dendogramas. Los dendogramas son visualizaciones de una agrupación jerárquica binaria.



# Tipos de Agrupación Jerárquica

Existen dos enfoques para este tipo de agrupación: aglomerativo y divisivo.

- **Divisivo**: este método es de tipo descendente, i.e., todas las observaciones comienzan en un único grupo. Luego, dividirá el grupo iterativamente en otros más pequeños hasta que cada uno de ellos contenga sólo una muestra.
- **Aglomerativo**: este método es de tipo ascendente, i.e., cada observación comienza en su propio grupo. Luego los pares de grupos se van fusionando conforme se avanza en la jerarquía, hasta que sólo haya un grupo.

# Ventajas

- Las representaciones jerárquicas resultantes pueden ser muy informativas.
- Los dendogramas proporcionan una forma interesante e informativa de visualización.
- Son especialmente potentes cuando el conjunto de datos contiene relaciones jerárquicas reales.

# Desventajas

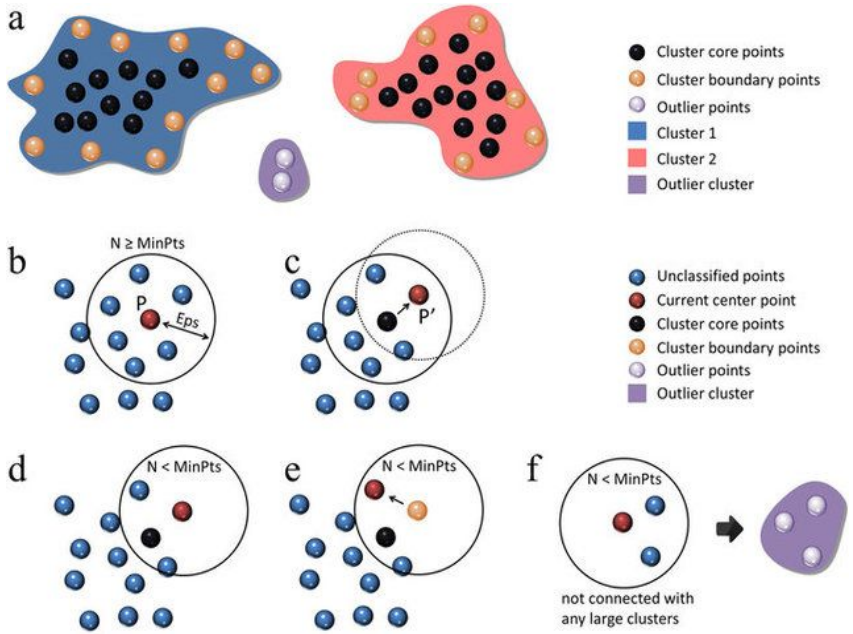
- Son muy sensibles a los valores atípicos y, en su presencia, el rendimiento del modelo disminuye significativamente.
- Son muy exigentes, desde el punto de vista informático y computacional (caso genera: $O(n^3)$

# Agrupación Espacial de Aplicaciones Basadas en la Densidad con Ruido (DBSCAN)

La Agrupación Espacial Basada en Densidad de Aplicaciones con Ruido, o DBSCAN (Density-Based Spatial Clustering of Applications with Noise), es otro algoritmo de agrupación especialmente útil para identificar correctamente el ruido en los datos.

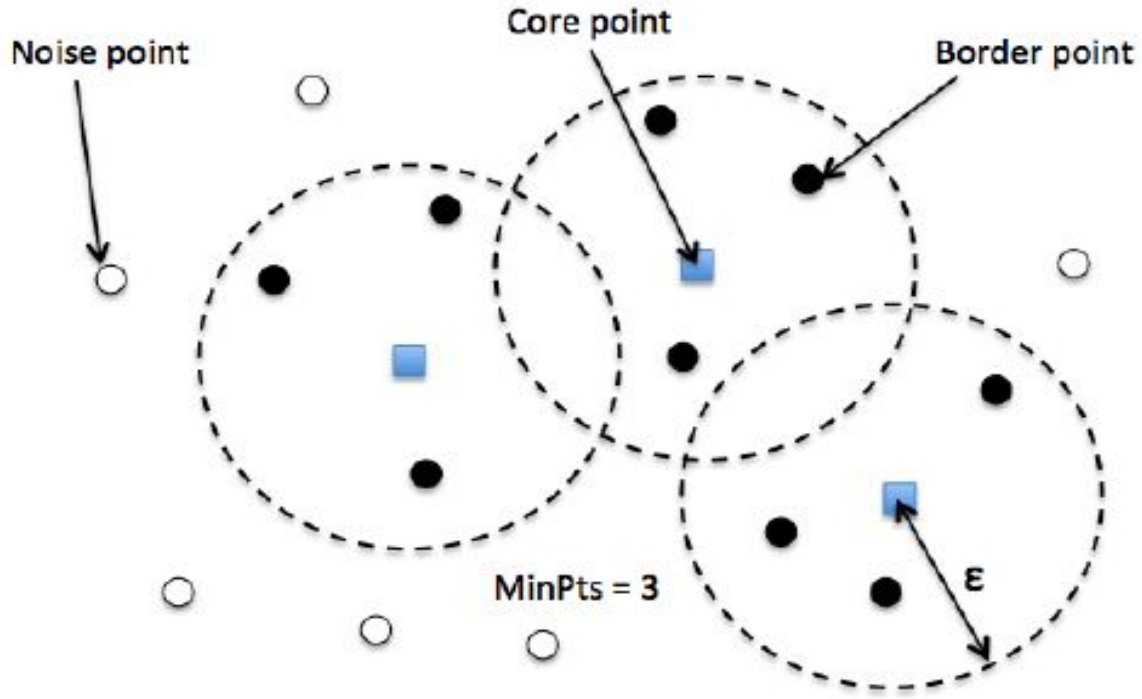
Se basa en un número de puntos con un radio especificado ε y hay una etiqueta especial asignada a cada punto de datos. El proceso de asignación de esta etiqueta es el siguiente:

- Es un número especificado (MinPts) de puntos vecinos. Se asignará un punto central si existe este número de puntos de los MinPts que caen en el radio ε
- Un punto fronterizo caerá en el radio de ε de un punto central, pero tendrá menos vecinos que el número de MinPts.
- Todos los demás puntos serán puntos de ruido.



# Algoritmo DBSCAN

1. Identificar un punto central y hacer un grupo para cada uno, o para cada grupo conectado de puntos centrales (si es que establecen que el criterio es el punto central).
2. Identificar y asignar puntos fronterizos a sus respectivos puntos centrales.



## Ventajas

- No es necesario especificar el número de grupos.
- Existe una gran flexibilidad en las formas y tamaños que pueden adoptar los grupos.
- Permite encontrar clusters no separables linealmente.
- Es muy útil para identificar y tratar con datos de ruido y valores atípicos.

## Desventajas

- Se enfrenta a dificultades a la hora de tratar los puntos de borde que son alcanzables por dos grupos.
- No encuentra bien racimos de densidades variables.

# Modelos de Mezcla Gaussiana (MMG)

Los Modelos de Mezcla Gaussiana son modelos probabilísticos que asumen que todas las muestras son generadas a partir de una mezcla de un número finito de distribución gaussiana con parámetros desconocidos.

Pertenece al grupo de algoritmos de agrupamiento blando en el que cada punto de datos estará presente en cada grupo existente en el conjunto de datos, pero con diferentes niveles de pertenencia a cada grupo. Esta membresía se asigna como la probabilidad de pertenecer a un determinado grupo, que oscila entre 0 y 1.

# Algoritmo MMG

Es un algoritmo de maximización de expectativas cuyo proceso podría resumirse de la siguiente manera:

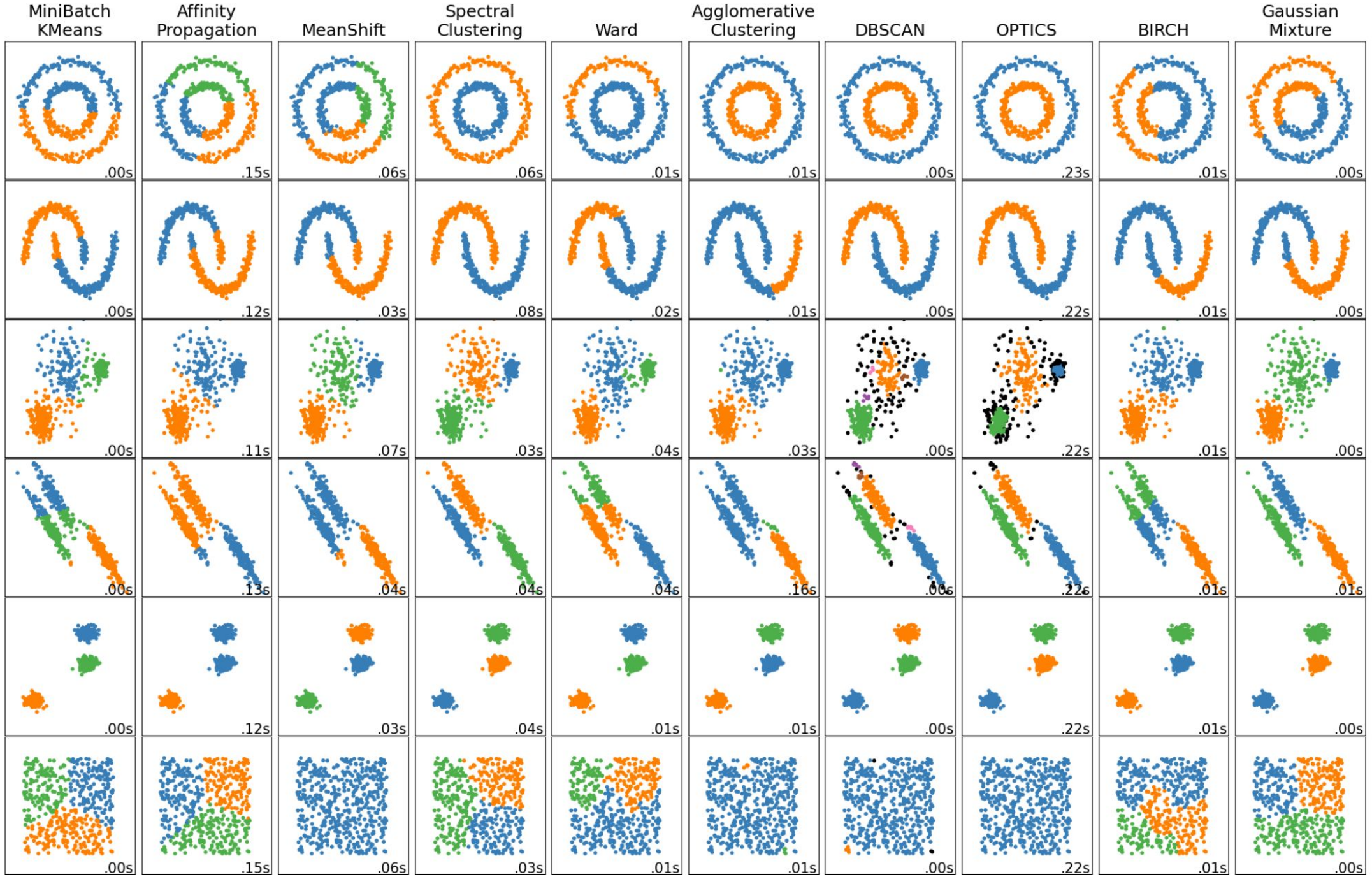
1. Inicializar las distribuciones de K Gaussian. Lo hace con los valores μ (media) y σ (desviación estándar). Se pueden tomar del conjunto de datos (método ingenuo) o aplicando K-means.
2. Grupo 'blando' de datos: es la fase de'Expectativa' en la que todos los puntos de datos se asignarán a cada grupo con su respectivo nivel de membresía.
3. Reestimar a los gaussianos: es la fase de 'Maximización' en la que se comprueban las expectativas y se utilizan para calcular nuevos parámetros para los gaussianos: nuevos μ y σ.
4. Evaluar la probabilidad de registro de los datos para verificar la convergencia. Cuanto mayor sea la similitud con el registro, más probable es que la mezcla del modelo que creamos se ajuste a nuestro conjunto de datos. Por lo tanto, esta es la función de maximizar.
5. Repetir desde el paso 2 hasta la convergencia.

## Ventajas

- Es un método de agrupación en grupos blandos, que asigna etiquetas de pertenencia a varios grupos. Esta característica lo convierte en el algoritmo más rápido para aprender modelos de mezclas.
- Existe una gran flexibilidad en el número y la forma de los grupos.

## Desventajas

- Es muy sensible a los valores iniciales que condicionarán en gran medida su rendimiento.
- El MMG puede converger a un mínimo local, lo que constituiría una solución no óptima.
- Al tener puntos insuficientes por mezcla, el algoritmo diverge y encuentra soluciones con infinitas probabilidades a menos que regularicemos artificialmente las covarianzas entre los puntos de datos.

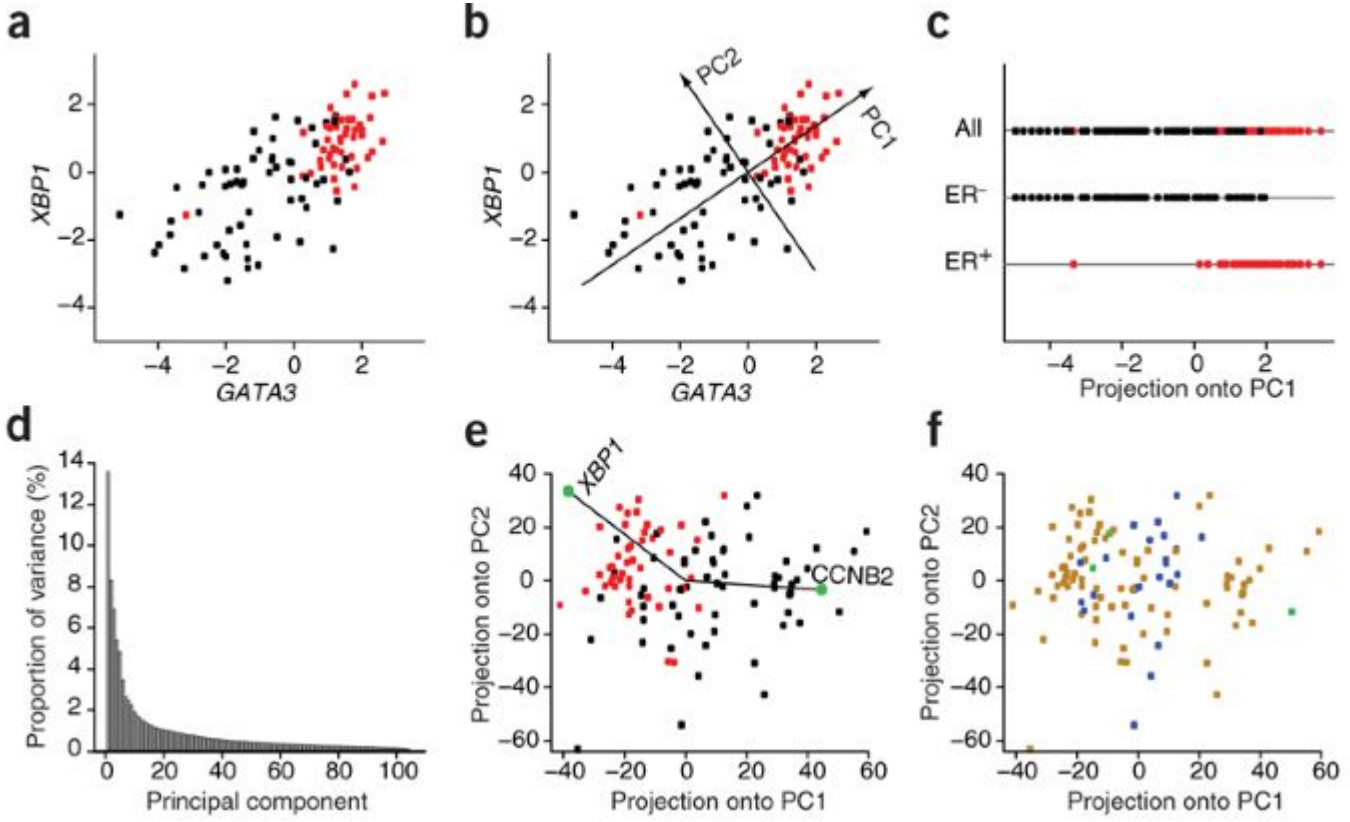

MiniBatch KMeans
Affinity Propagation
MeanShift
Spectral Clustering
Ward
Agglomerative Clustering
DBSCAN
OPTICS
BIRCH
Gaussian Mixture
.00s .15s .06s .06s .01s .01s .00s .23s .01s .00s
.00s .12s .03s .08s .02s .01s .00s .22s .01s .00s
.00s .11s .07s .03s .04s .03s .00s .22s .01s .00s
.00s .13s .04s .04s .04s .16s .00s .22s .01s .01s
.00s .12s .03s .04s .01s .01s .00s .22s .01s .00s
.00s .15s .06s .03s .01s .01s .00s .22s .01s .00s

# Resumen de Algoritmos

| Method name | Parameters | Scalability | Usecase | Geometry (metric used) |
|---|---|---|---|---|
| K-Means | number of clusters | Very large `n_samples`, medium `n_clusters` with MiniBatch code | General-purpose, even cluster size, flat geometry, not too many clusters | Distances between points |
| Affinity propagation | damping, sample preference | Not scalable with n_samples | Many clusters, uneven cluster size, non-flat geometry | Graph distance (e.g. nearest-neighbor graph) |
| Mean-shift | bandwidth | Not scalable with `n_samples` | Many clusters, uneven cluster size, non-flat geometry | Distances between points |
| Spectral clustering | number of clusters | Medium `n_samples`, small `n_clusters` | Few clusters, even cluster size, non-flat geometry | Graph distance (e.g. nearest-neighbor graph) |
| Ward hierarchical clustering | number of clusters or distance threshold | Large `n_samples` and `n_clusters` | Many clusters, possibly connectivity constraints | Distances between points |
| Agglomerative clustering | number of clusters or distance threshold, linkage type, distance | Large `n_samples` and `n_clusters` | Many clusters, possibly connectivity constraints, non Euclidean distances | Any pairwise distance |
| DBSCAN | neighborhood size | Very large `n_samples`, medium `n_clusters` | Non-flat geometry, uneven cluster sizes | Distances between nearest points |
| OPTICS | minimum cluster membership | Very large `n_samples`, large `n_clusters` | Non-flat geometry, uneven cluster sizes, variable cluster density | Distances between points |
| Gaussian mixtures | many | Not scalable | Flat geometry, good for density estimation | Mahalanobis distances to centers |
| Birch | branching factor, threshold, optional global clusterer. | Large `n_clusters` and `n_samples` | Large dataset, outlier removal, data reduction. | Euclidean distance between points |

# Principal Component Analysis (PCA)

Principal Component Analysis es una técnica de Extracción de Características donde combinamos las entradas de una manera específica y podemos eliminar algunas de las variables “menos importantes” manteniendo la parte más importante de todas las variables. Como valor añadido, luego de aplicar PCA conseguiremos que todas las nuevas variables sean independientes unas de otras.



# Algoritmo PCA

1. Estandarizar los datos de entrada (ó Normalización de las Variables)
2. Obtener los autovectores y autovalores de la matriz de covarianza
3. Ordenar los autovalores de mayor a menor y elegir los “k” autovectores que se correspondan con los autovectores “k” más grandes (donde “k” es el número de dimensiones del nuevo subespacio de características).
4. Construir la matriz de proyección W con los “k” autovectores seleccionados.
5. Transformamos el dataset original “X estandarizado” vía W para obtener las nuevas características k-dimensionales.

## Ventajas

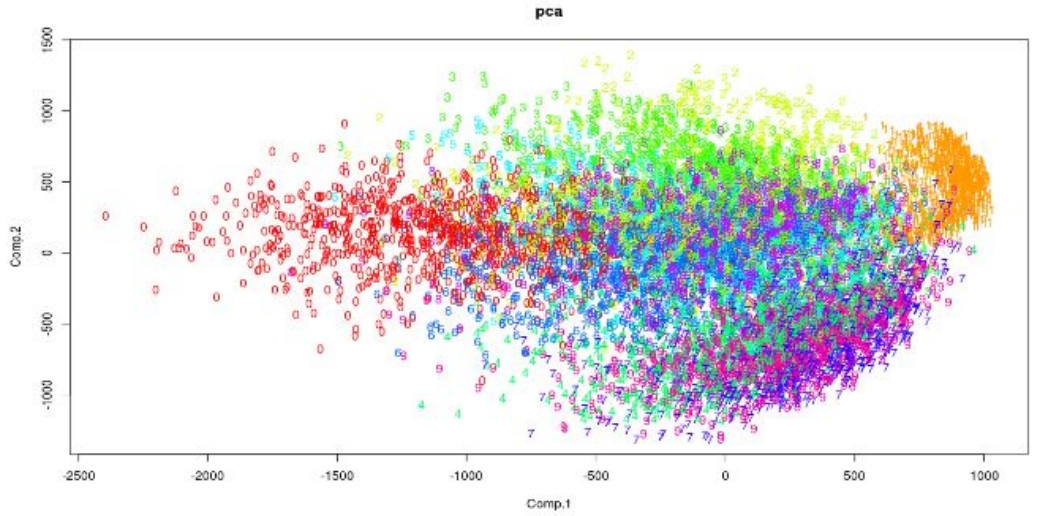
- Nos entrega una medida de como cada variable se asocia con las otras (matriz de covarianza)
- La dirección en las que nuestros datos están dispersos (autovectores)
- La relativa importancia de esas distintas direcciones (autovalores)
- PCA combina nuestros predictores y nos permite deshacernos de los autovectores de menor importancia relativa.

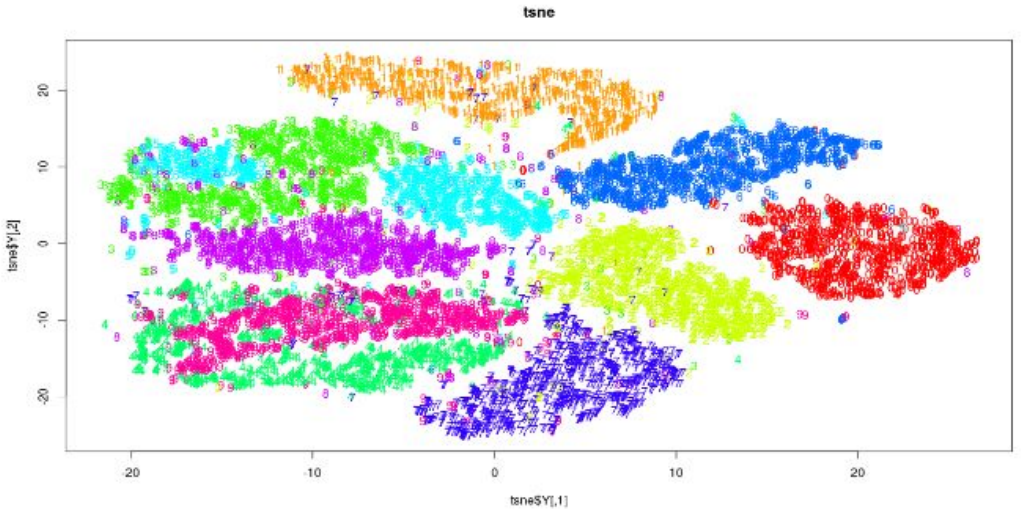
## Desventajas

- El algoritmo de PCA es muy influenciado por los outliers en los datos. Por esta razón, surgieron variantes de PCA para minimizar esta debilidad. Entre otros se encuentran: RandomizedPCA, SparcePCA y KernelPCA.

# T-distributed Stochastic Neighbor Embedding (t-SNE)

Es un algoritmo diseñado para la visualización de conjuntos de datos de alta dimensionalidad. Si el número de dimensiones es muy alto, se recomienda utilizar un método de reducción de dimensionalidad previo (como PCA) para reducir el conjunto de datos a un número de dimensiones razonable, lo que reducirá el ruido y aligerará la ejecución de t-SNE.



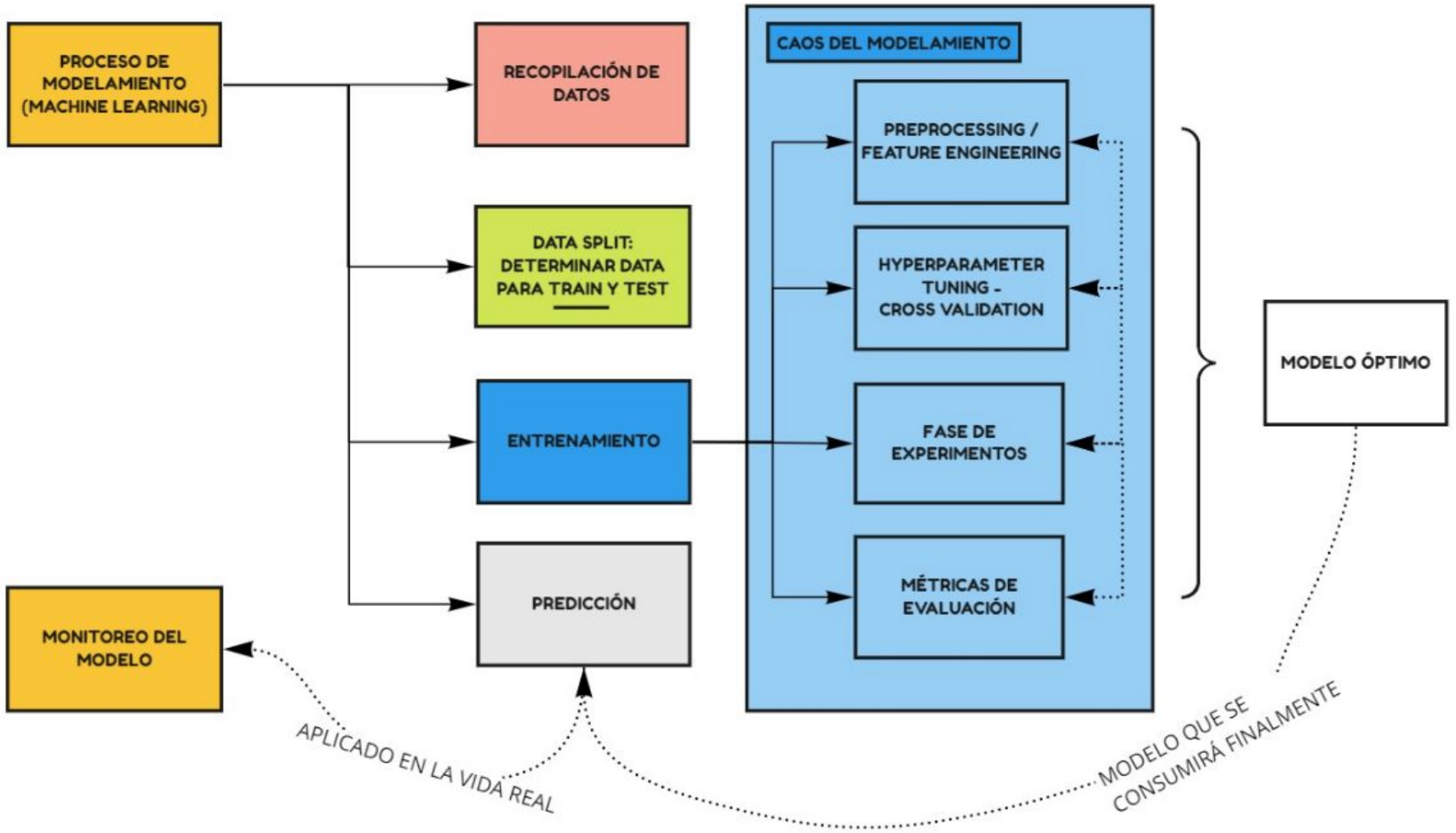



# Algoritmo t-SNE

t-SNE se ejecuta en dos pasos:

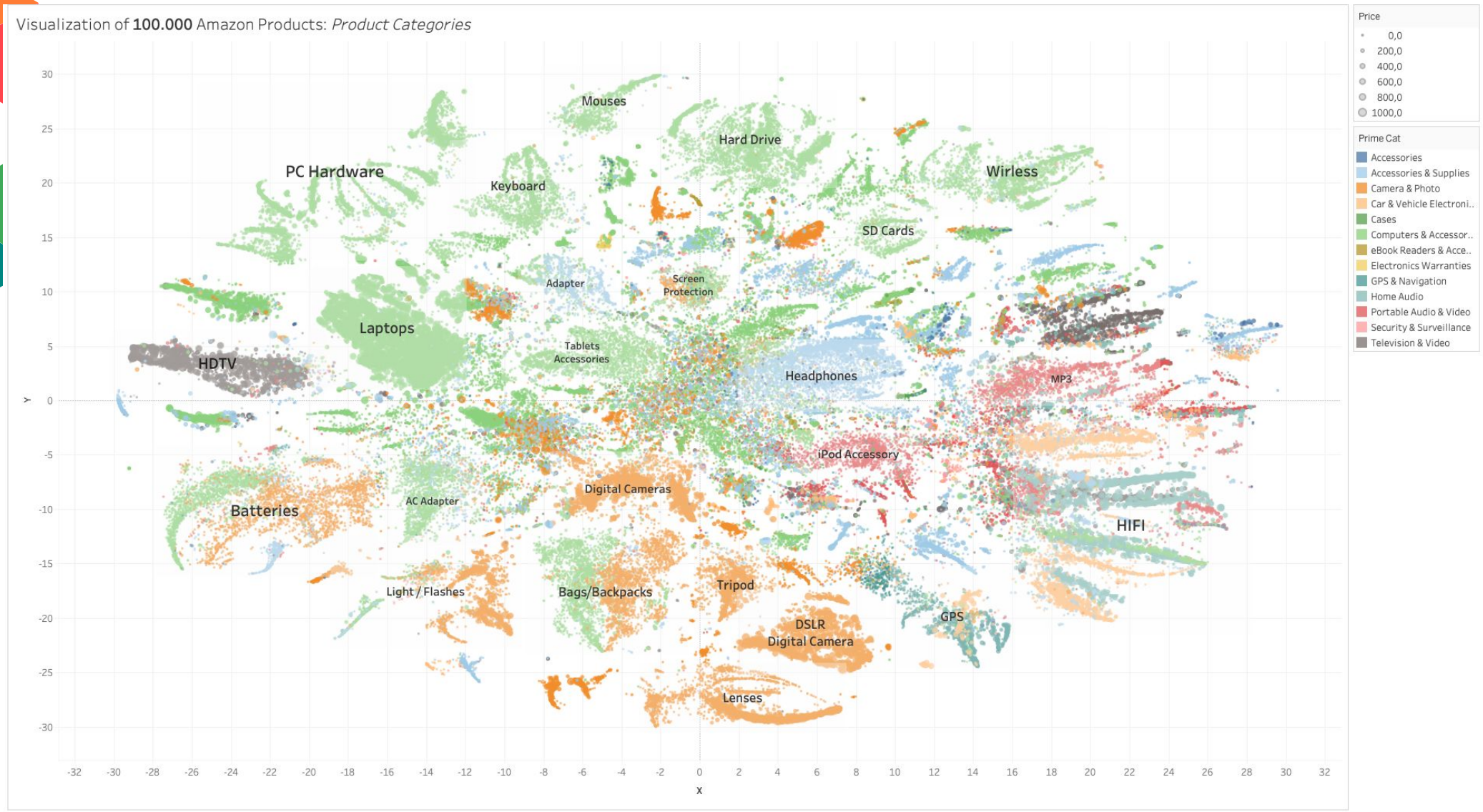
1. Construye una distribución de probabilidad sobre parejas de muestras en el espacio original, de forma tal que las muestras semejantes reciben alta probabilidad de ser escogidas, mientras que las muestras muy diferentes reciben baja probabilidad de ser escogidas. El concepto de "semejanza" se basa en la distancia entre puntos y densidad en las proximidades de un punto. Tal y como lo describen los autores:
2. t-SNE lleva los puntos del espacio de alta dimensionalidad al espacio de baja dimensionalidad de forma aleatoria, define una distribución de probabilidad semejante a la vista en el espacio destino (el espacio de baja dimensionalidad), y minimiza la denominada divergencia Kullback-Leibler entre las dos distribuciones con respecto a las posiciones de los puntos en el mapa (la divergencia de Kullback-Leibler mide la similitud o diferencia entre dos funciones de distribución de probabilidad). Dicho con otras palabras: t-SNE intenta reproducir la distribución que existía en el espacio original en el espacio final.

# Flujo de un Problema de Machine Learning



# Aplicaciones Reales

## Visualización de 1000 productos de Amazon (t-SNE)



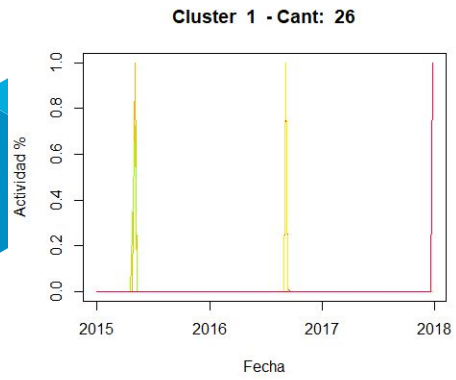

https://towardsdatascience.com/vis-amz-83dea6fcb059

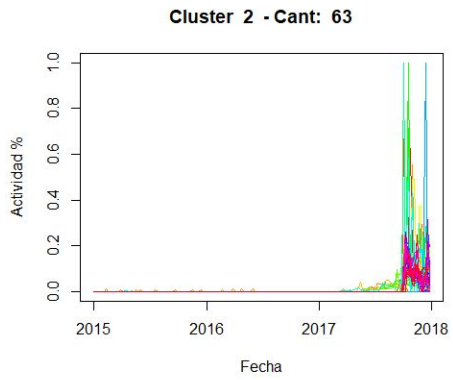
# Aplicaciones Reales

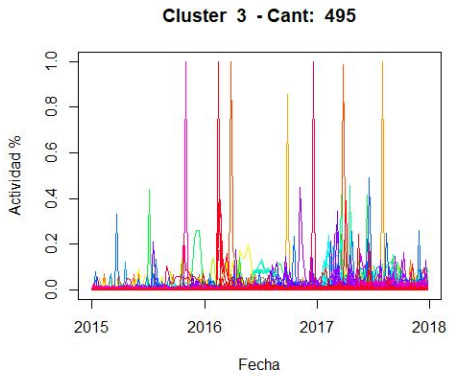
## Retail – Ejemplo importancia EDA

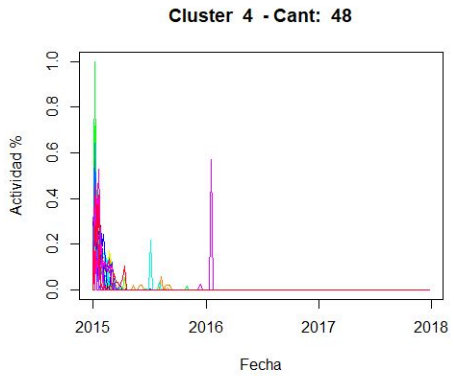
- Retailer con +700 productos: ¿Es posible generar un modelo de predicción de demanda?

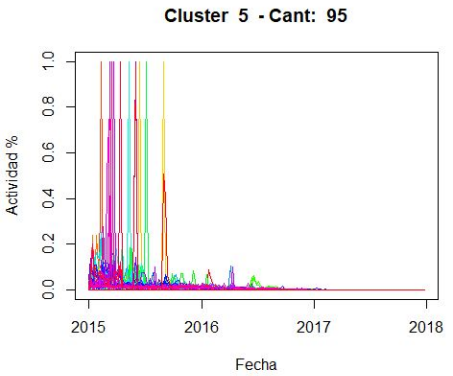
*Cluster de Tendencias*

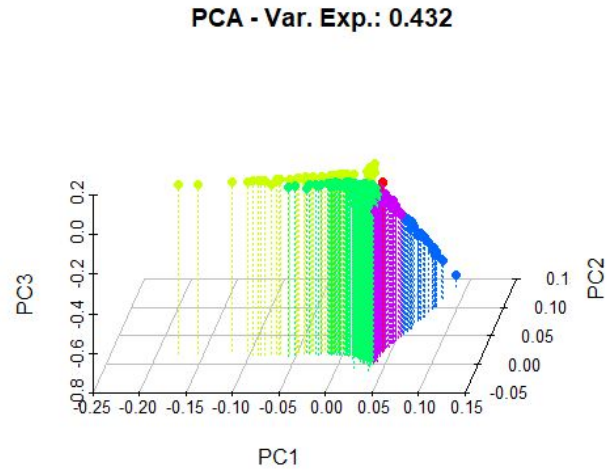


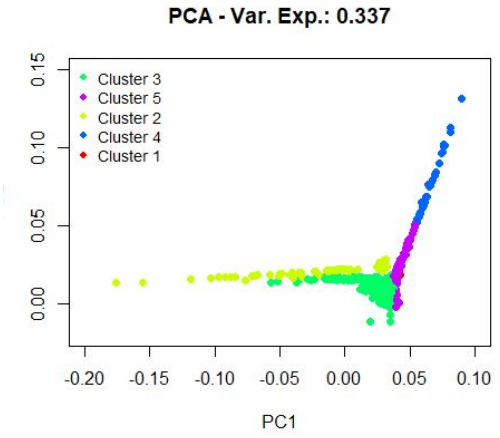



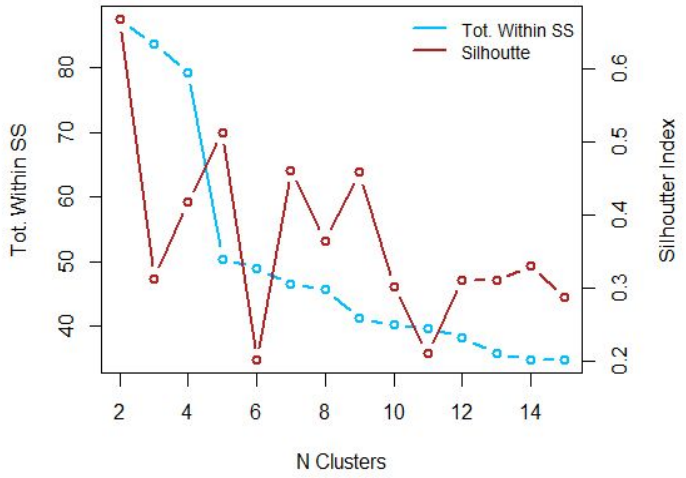



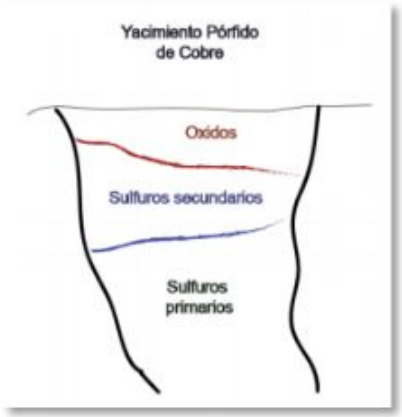



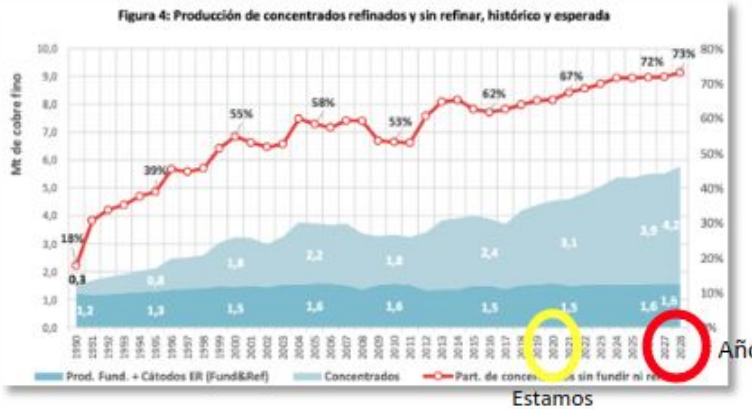
# Aplicaciones Reales

## RELEVANCIA DEL PROBLEMA: ¿Qué problema intenta resolver la tecnología?




En Chile los óxidos (Hidrometalurgia) explican un tercio de la producción de cobre (31% en 2015), y los sulfuros (Flotación) explican los dos tercios restantes (69% en 2015), y la mitad de las exportaciones.
El 50% que no se exporta como concentrado es procesado adicionalmente a través de Fundiciones y Refinerías (hornos y electrorrefinación), y exportado como cátodos.

Las operaciones se están encontrando cada vez con menos óxidos y más sulfuros, para lo cual las operaciones de concentrado tomarán mayor relevancia con el tiempo. **En 2016 la capacidad instalada de concentradoras fue de 631 MMTon, y se estima para el 2028 una capacidad instalada de 1.000 MMTon (Cochilco).**




Año 2028

Estamos acá

El aumento en la capacidad instalada de plantas concentradoras y la baja en la ley de mineral de entrada de los yacimientos, requiere de procesos que permitan **mejorar los niveles de recuperación de cobre, de manera de identificar y cuantificar patrones operacionales que condicionan la eficiencia de estos procesos.**

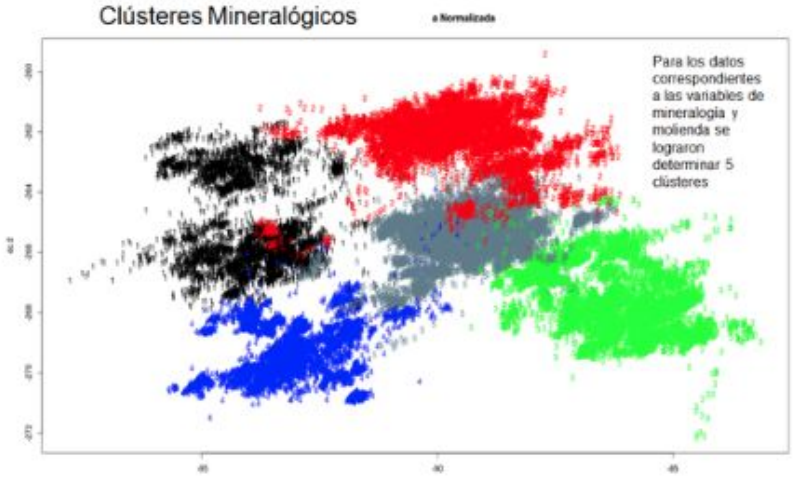
**Codelco División Ministro Hales** abrió las puertas de su proceso de concentración rougher para una prueba industrial en 2017, y se llegó a niveles adicionales de recuperación que varían entre **1,9% y 3,9%. En Dicha prueba se logró en un mes un estimado de 53 Ton adicionales de CuF, o unos US$250.000.**

**Optimización de planta concentradora mediante Software Inteligencia Artificial = Bajo CAPEX/OPEX + Alto ROI**

# Aplicaciones Reales

## SOLUCIÓN: ¿Cómo resuelven el problema?

- **Aplicando modelos de *machine learning* que permiten identificar y/o cuantificar patrones de entrada (características intrínsecas del mineral) y los patrones operacionales (configuración de la planta) que impactan en la eficiencia de la recuperación de Cu.**
  - **Caracterizando el mineral de entrada  clústeres mineralógicos**
  - **Identificando y seleccionando los mejores patrones de operación históricos para cada clúster mineralógico.**
  - **Recreando las mejores prácticas operacionales históricas que llevaron al mejor resultado posible de recuperación.**
  - **Generando recomendaciones de configuraciones operacionales, facilitando la toma de decisiones.**




| Copper recuperation (%) | Operational mode 1 | Operational mode 2 | Operational mode 3 | Operational mode 4 | Operational mode 5 | Operational mode 6 | Operational mode 7 | Operational mode 8 |
|---|---|---|---|---|---|---|---|---|
| Mineralogy – Gr. cluster 1 | 85,5% | 84,4% | 87,3% | 82,4% | 84,4% | 83,1% | 83,4% | 83,9% |
| Mineralogy – Gr. cluster 2 | 87,9% | 85,4% | 86,3% | 84,4% | 85,6% | 84,1% | 86,4% | 85,9% |
| Mineralogy – Gr. cluster 3 | 88,4% | 88,9% | 89,0% | 87,8% | 86,4% | 87,0% | 87,4% | 86,9% |
| Mineralogy – Gr. cluster 4 | 86,5% | 87,9% | 85,7% | 84,8% | 83,4% | 82,0% | 85,4% | 84,9% |
| Mineralogy – Gr. cluster 5 | 84,5% | 84,9% | 83,7% | 82,4% | 85,4% | 83,6% | 82,4% | 84,9% |

| FECHA RECOMENDACIÓN | 18-01-2017 15:00 |
|---|---|
| Clúster Mineralógico: | Muscovita-Pirofilita-Caolinita (7) |
| Clúster Modo de Operación en Flotación: | Modo de Operación 10 |
| Recuperación Promedio [%]: | 90,0 |
| Tratamiento [TPH] | 2300 - 2500 |
| Ley Alimentación [%] | 1.3 - 1.4 |
| Razón de Solubilidad | 8.5 - 9.5 |

- Recuperación de Cu Actual: 85.5%
- Recuperación de Cu Alcanzable Histórica por Similitud (Promedio): 90.0%
- Potencial de Mejora: 4.5%
- Rango Tratamiento Actual: 2300 - 2500 TPH
- Rango Ley Alimentación Actual: 1.3 - 1.4
- Rango Razón de Solubilidad Actual: 8.5 - 9.5

Cambios en Reactivos:

Colector 1: Bajar de 45 gr/ton a 40 gr/ton
Colector 2: Bajar de 25 gr/ton a 15 gr/ton
Espumante: Mantener en 21 gr/ton
NaHs: Bajar de 88 gr/ton a 83 gr/ton

| Indicador | Colector Primario | Colector Secundario | Espumante | NaHs | pH Molienda | pH Rougher 1 | pH Rougher 2 | pH Limpieza | Recuperación en Peso | Porcentaje Sólidos Rougher | Ley Concentrado Rougher | Recuperación Cu |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Valor Actual PI | 45,0 | 25,0 | 21,3 | 88,0 | 8,5 | 9,8 | 9,8 | 11,9 | 19,6 | 36,1 | 5,5 | 85,5 |
| Valor Medio | 38,9 | 15,5 | 12,6 | 82,9 | 9,6 | 10,3 | 10,5 | 12,0 | 18,3 | 32,1 | 5,4 | 90,0 |
| Desviación Estándar | 5,5 | 4,6 | 2,7 | 4,0 | 0,5 | 0,1 | 0,1 | 0,3 | 9,8 | 3,7 | 0,8 | 0,7 |
| Valor Mínimo | 33,5 | 10,9 | 10,0 | 78,8 | 9,1 | 10,4 | 10,3 | 11,7 | 8,5 | 28,4 | 4,6 | 89,3 |
| Valor Máximo | 44,4 | 20,1 | 15,3 | 86,9 | 10,0 | 10,6 | 10,6 | 12,4 | 28,1 | 35,8 | 6,2 | 90,7 |

- ✓ **Unificación de criterios para operar de mejor forma la planta, con los recursos ya disponibles.**
- ✓ **Los datos ya existen, el análisis efectivo de la información es difícil de realizar por metodologías convencionales.**
- ✓ **El resultado del conocimiento experto se rescata de la historia y se disponibiliza continuamente.**
- ✓ **El sistema aprende y se optimiza a medida que cuenta con más historia.**

# Aplicaciones Reales

## SOLUCIÓN: ¿Cómo resuelven el problema?

RO2P es altamente replicable y escalable tanto en minería como en otras industrias. La tecnología está diseñada y construida para ser fácilmente integrada en la operación, considerando las herramientas y los estándares tecnológicos con los que trabaja la minería, de manera de no requerir grandes ajustes para su pilotaje y puesta en producción.

Operación

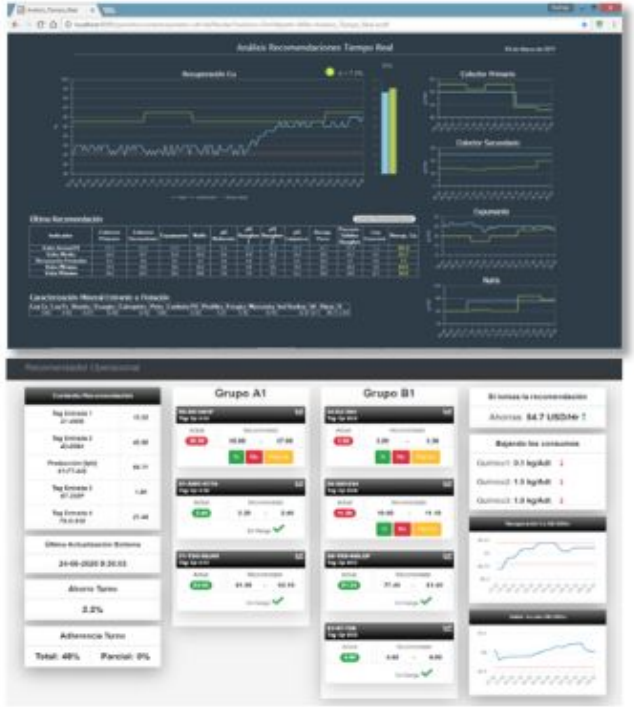
**1** Se recibe el plan de producción diario y/o información mineralógica, otras

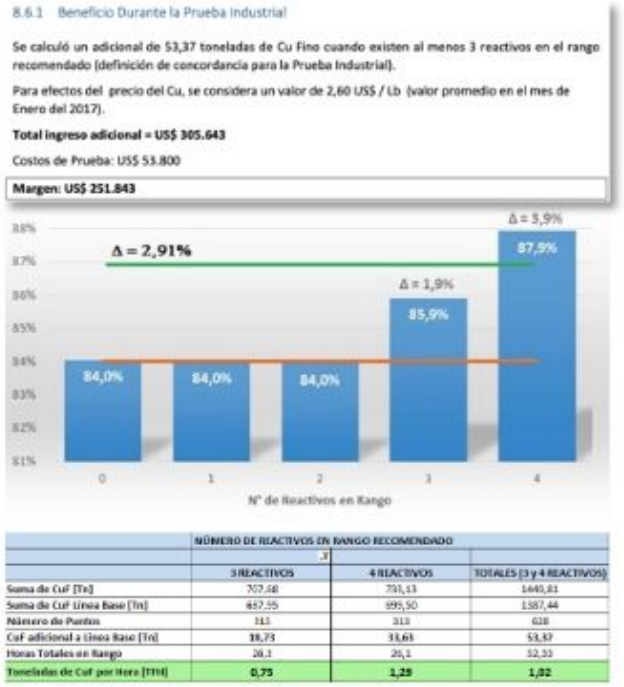
**2** Se procesa la información mineralógica en conjunto con la data de operación de la planta

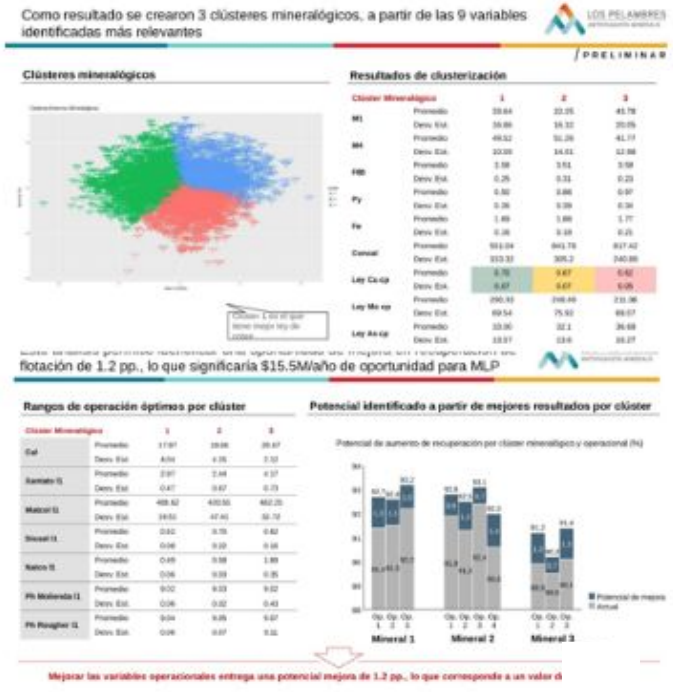
**3** Aplicación de modelos de caracterización del mineral entrando a la planta

**4** Se determinan los mejores modos operacionales para las variables de proceso

**5** Se genera recomendación si cumple con criterio de mejora



# Aplicaciones Reales

## Modelos para sobrevida cáncer de mama

Implementar una plataforma de software que permita el análisis y estudio de los tratamientos de paciente con cáncer de mama, dadas sus características individuales y aquellas asociadas a la enfermedad, de manera de generar los mejores resultados en términos de sobrevida.

- Aplicar modelos de analítica avanzada y machine learning para caracterizar a los pacientes y sus tratamientos.
- Aplicar modelos analíticos y estadísticos para estimar la sobrevida.
- Desarrollar una plataforma web que permita la administración y gestión del historial del paciente y que entregue los lineamientos que mejorar los resultados del tratamiento.
- Caracterizar a la población en estudio, entregando KPIs relevantes para el trabajo clínico, de docencia e investigación.

# Aplicaciones Reales

## K-MEDOIDS

- Permite construir modelos de clustering a partir de variables numéricas y variables categóricas.
- k-medoids escoge datapoints como centros y trabaja con una métrica arbitraria de distancias entre datapoints en vez de usar la distancia euclidiana → **Distancia de Gower.**
- Minimiza una suma de disimilaridades (entre pares de puntos) en vez de una suma de distancias euclidianas cuadradas (k-means).
- Distancia de Gower: puede ser usada para medir cuán diferentes son dos registros. Los registros pueden contener combinaciones de datos lógicos, numéricos, categóricos o de texto.

# Aplicaciones Reales

## BASE DE DATOS

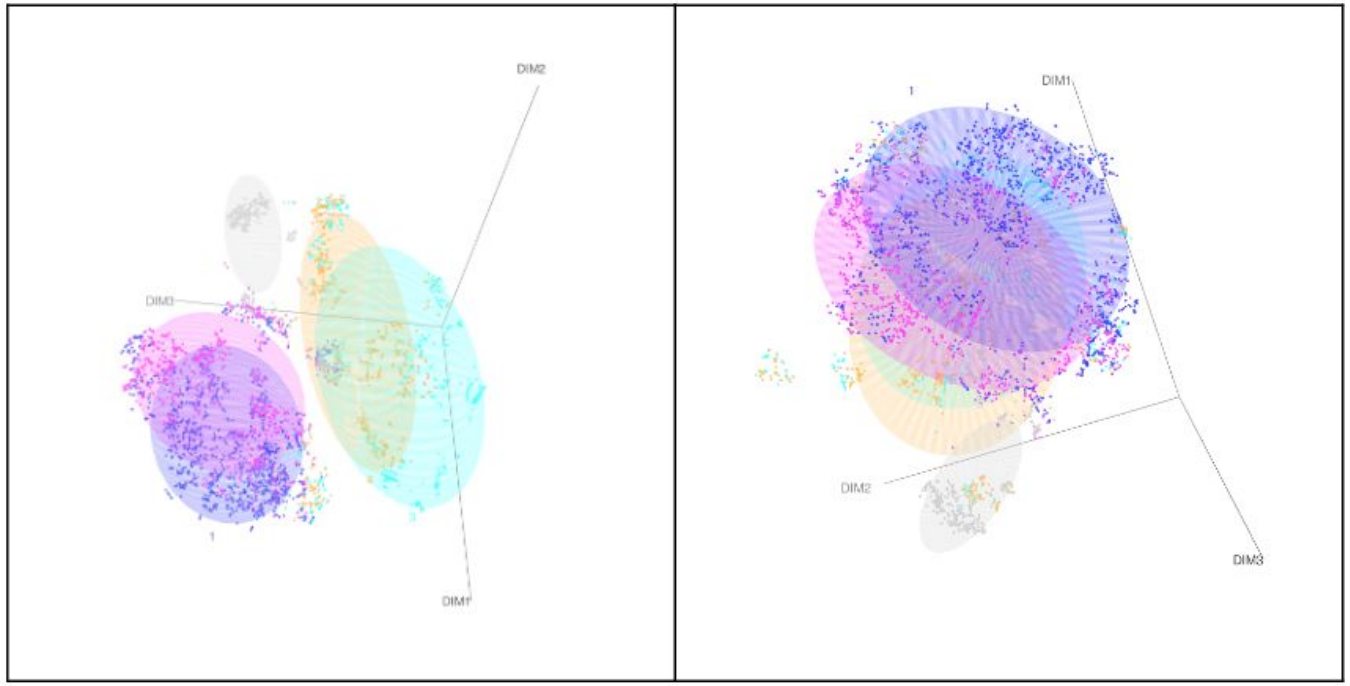
- Base de datos Registro de Cáncer
- Periodo 2010 - 2020
- Sub-muestra: 4384 pacientes, 38 variables

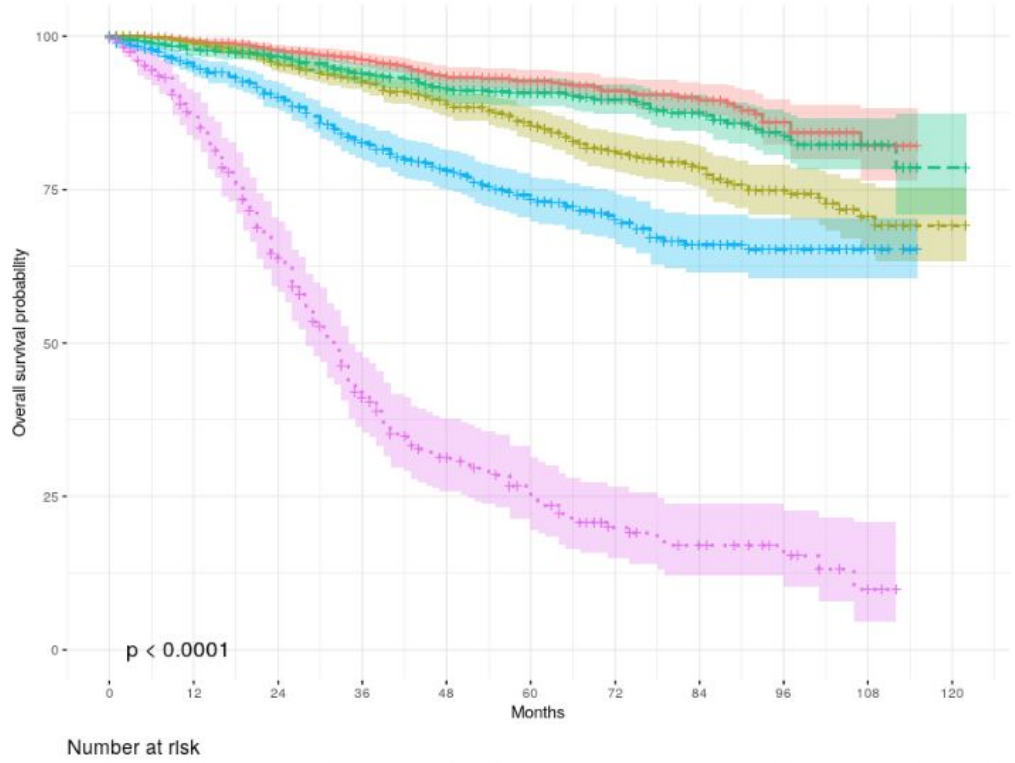
- Caracterización por grupo de pacientes:
  - Datos paciente (edad, imc, otros)
  - Antecedentes familiares
  - Comorbilidades
  - Características del cáncer
    - Generales (etapa, lado, ganglios, etc)
    - Específicos tumor

| N.° | Nombre Variable | N.° | Nombre Variable | N.° | Nombre Variable | N.° | Nombre Variable |
|---|---|---|---|---|---|---|---|
| 1 | Código | 11 | Partos | 21 | HTA | 31 | Sobrevida Libre Rec. Metastásica |
| 2 | Sexo | 12 | Abortos | 22 | DM2 | 32 | Sobrevida Global |
| 3 | Abortos | 13 | Edad Menarquia | 23 | Cáncer | 33 | Etapa |
| 4 | Edad Diagnóstico | 14 | Menopausia? | 24 | Hipotiroidismo | 34 | Subtipo |
| 5 | Peso | 15 | TRH | 25 | Recaída In Situ | 35 | Receptor Hormonal |
| 6 | Altura | 16 | Familia | 26 | Recidiva Invasora | 36 | HER2 |
| 7 | IMC | 17 | Antecedentes Mama u Ovario | 27 | Recidiva Ipsilateral | 37 | Lado Debut |
| 8 | Tiene Hijos? | 18 | Antecedentes Primer Grado | 28 | Recidiva Contralateral | 38 | Ganglios |
| 9 | Número de hijos | 19 | Primer Grado Mama U Ovario | 29 | Recidiva Metástasis | | |
| 10 | Gestaciones | 20 | Comorbilidades? | 30 | Sobrevida Libre Recurrencia | | |

# Aplicaciones Reales

## Resultados Sobrevida

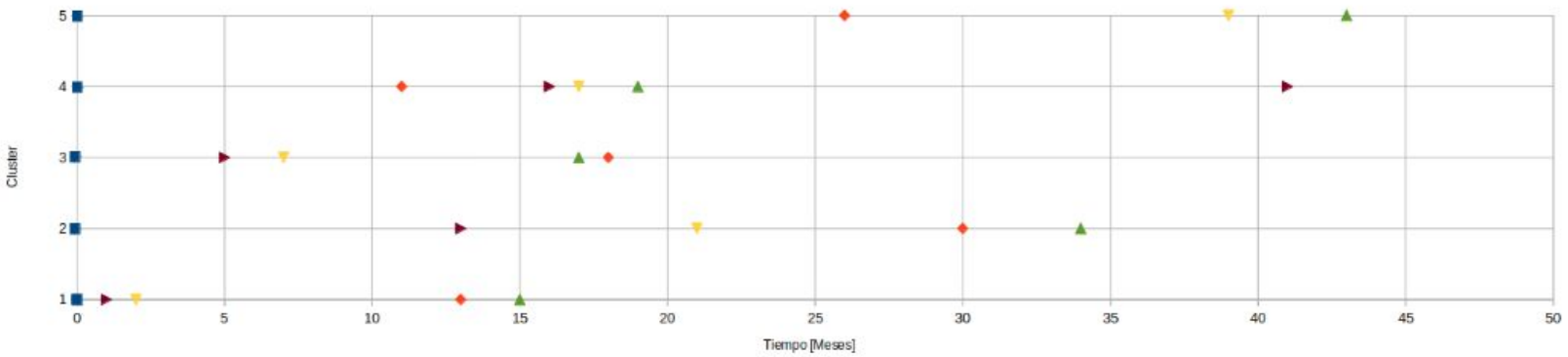



Number at risk

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 836 | 592 | 417 | 301 | 188 | 106 | 32 | 0 |
| 684 | 565 | 432 | 342 | 231 | 132 | 54 | 2 |
| 609 | 508 | 420 | 274 | 214 | 140 | 41 | 1 |
| 477 | 382 | 290 | 176 | 105 | 64 | 19 | 0 |
| 96 | 60 | 42 | 22 | 16 | 10 | 3 | 0 |
| 36 | 48 | 60 | 72 | 84 | 96 | 108 | 120 |

Months




Cluster 1 2 3 4 5

# Aplicaciones Reales


**Caracterización temporal** de cada agrupación de mediciones

Conozca en que hora, día y semana su operación es más eficiente

# Aplicaciones Reales


**Caracterización operacional** de cada agrupación de mediciones

Comprenda que parámetros operacionales definen cada comportamiento

# Muchas gracias

01001101 01110101 01100011 01101000 01100001 01110011 00100000 01100111 01110010 01100001 01100011 01101001 01100001 01110011

Centro de Transformación Digital

Universidad del Desarrollo
Facultad de Ingeniería

Proyecto apoyado por
CORFO
Gobierno de Chile