



ESCUELA DE INGENIERÍA
FACULTAD DE INGENIERÍA

EDUCACIÓN
PROFESIONAL

Diplomado en Big Data y Ciencia de Datos
Ciencia de Datos y sus Aplicaciones

Clase 03: Detección de Anomalías

Roberto González

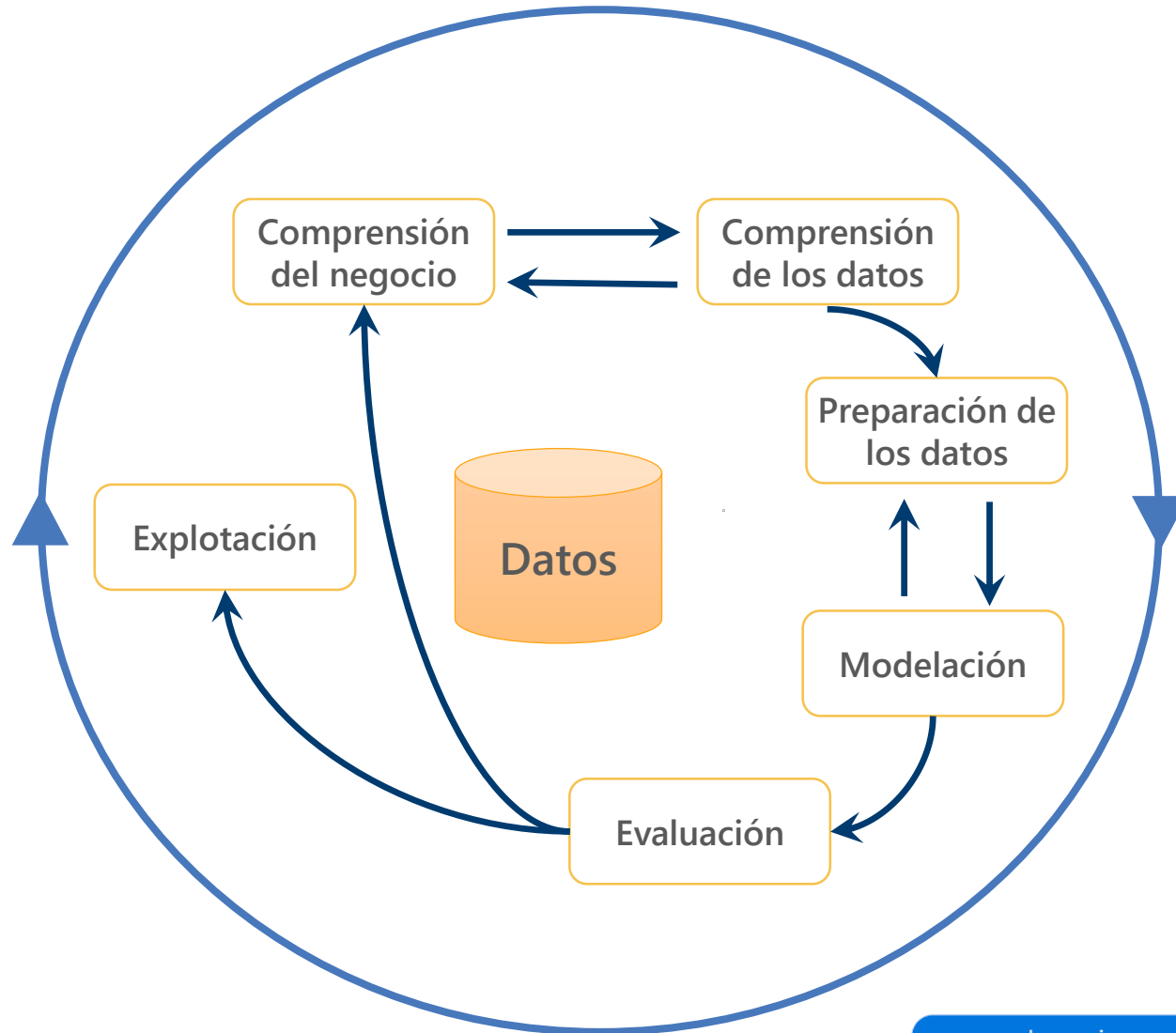


regonzar@uc.cl





CRISP-DM



● Recording

Hardest Part of ML isn't ML, it's Data

"Hidden Technical Debt in Machine Learning Systems," Google NIPS 2015

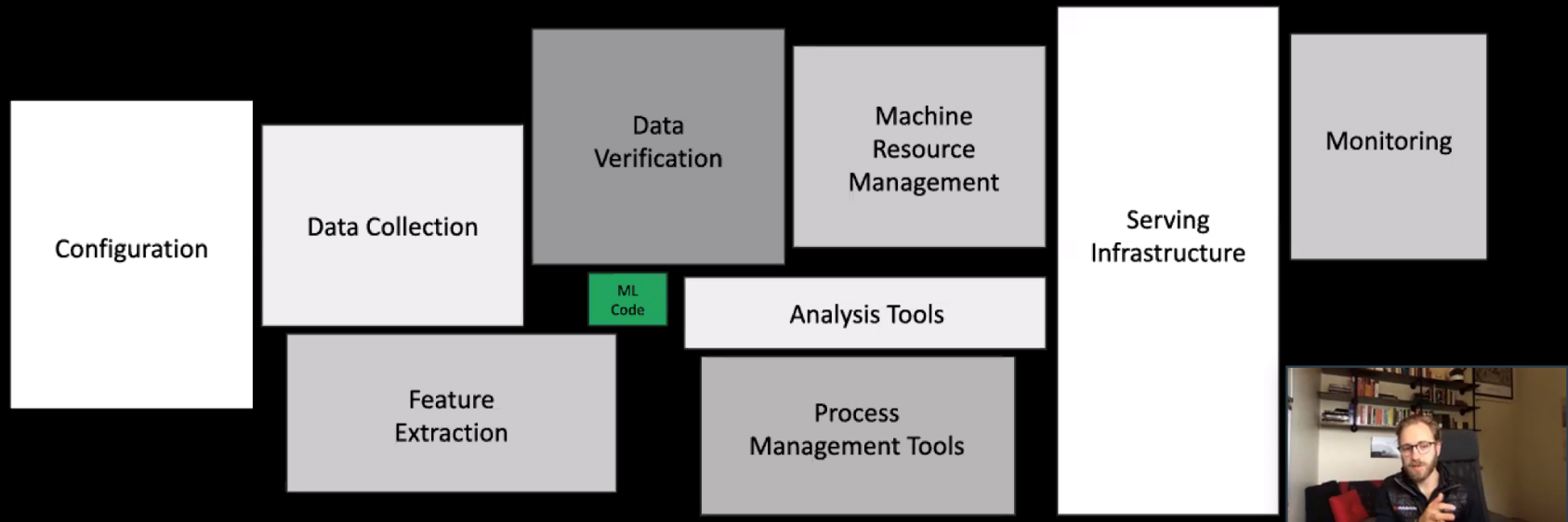



Figure 1: Only a small fraction of real-world ML systems is composed of the ML code, as shown by the small green box in the middle. The required surrounding infrastructure is vast and complex.

 databricks

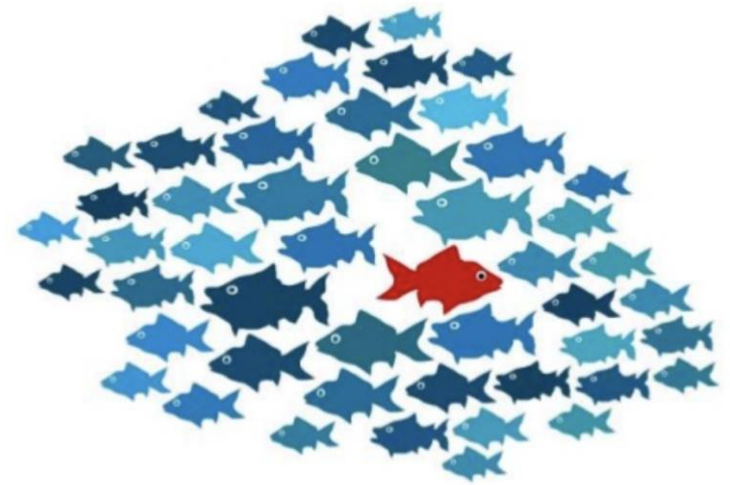


Clase 03: Anomalías

DETECCIÓN DE ANOMALÍAS

¿Qué son las anomalías?

- Las anomalías o valores atípicos son puntos de datos que parecen desviarse notablemente de los resultados esperados.
- Detección de anomalías es el proceso de encontrar patrones en los datos que no se ajustan a un comportamiento esperado previo.















Detección de anomalías

La detección de anomalías se emplea cada vez más en la industria

- Presencia en datos capturados por sensores (IoT)
- Plataformas de redes sociales
- Sistemas de producción y distribución energía
- Dispositivos médicos
- Bancos
- Ciberseguridad y detección de hackers

Casos de uso en la industria

TELECOM  Detect roaming abuse, revenue fraud, service disruptions	BANKING  Flag abnormally high purchases/deposits, detect cyber intrusions	FINANCE & INSURANCE  Detect and prevent out of pattern or fraudulent spend, travel expenses	HEALTHCARE  Detect fraud in claims and payments; events from RFID and mobiles	MANUFACTURING  Detect abnormal machine behavior to prevent cost overruns
TRANSPORTATION  Ensure external communications to the vehicle are not intrusion	SOCIAL MEDIA  Detect compromised accounts, bots that generate fake reviews	NETWORKING  Detect intrusion into networks, prevent theft of source code or IP	SMART HOUSE  Detect energy leakage, standardize smart sensor datasets	VIDEO SURVEILLANCE  Detect or track objects and persons of interest in monotonous footage

Algunos ejemplos

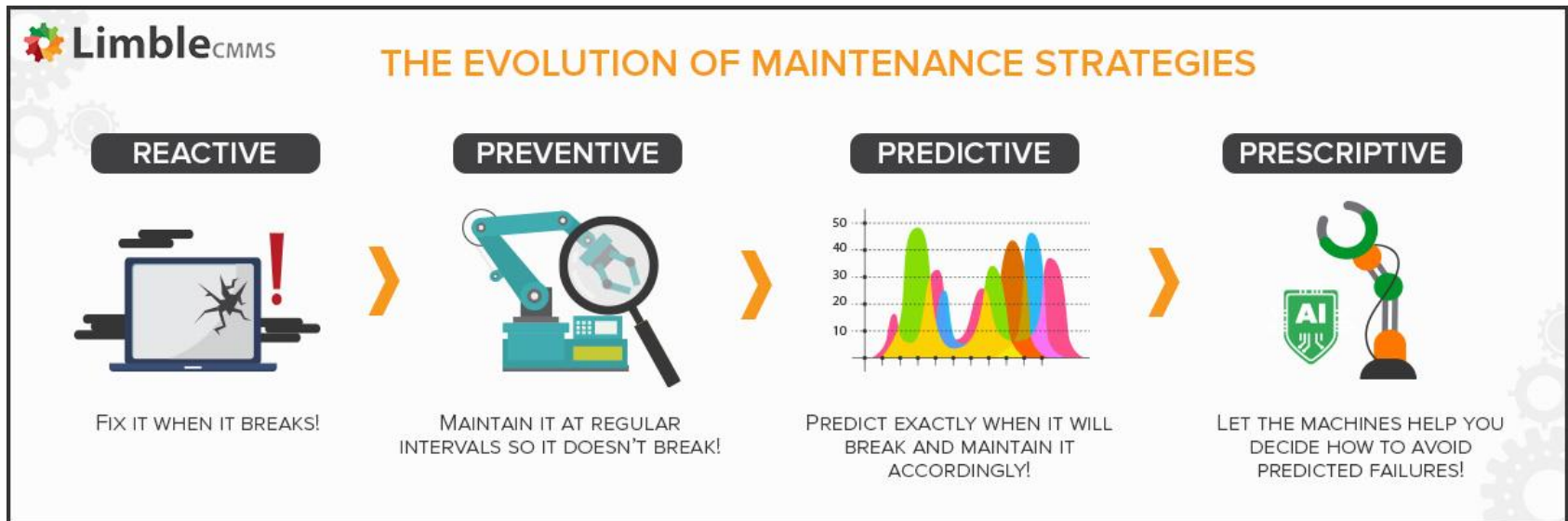
- Detección de fraudes
 - Fraudes con tarjetas de crédito
 - Llamadas maliciosas
- Mercado de acciones
 - Subida o caída abrupta en valor de acciones
 - Sistema Identifica comportamiento errático de humanos
- Comercio electrónico
 - Aumento en visitas
 - Error en los precios



Mantenimiento Preventiva de fallas

Las industrias de las telecomunicaciones y manufactura están constantemente recolectando datos de sus operaciones. Cuentan con máquinas equipadas con múltiples sensores.

Usar modelos para identificar fallas de manera temprana y hacer mantenciones a la maquinaria.





¿Qué hay de nuevo?

Los métodos clásicos se basan en el uso de reglas que dependen del negocio. Poca flexibilidad y adaptabilidad.

Acercamiento moderno basado en Data Science y ML

- Más eficiente
- Integración con datos en tiempo real
- Mejorar detección usando múltiples canales
- Aprender y detectar variaciones
- Adaptabilidad a múltiples dominios



Metodologías

Existen 3 tipos de metodologías usadas para la detección de anomalías

- Análisis gráfico
- Análisis estadístico
- Análisis basado en machine learning



Análisis gráfico

Existen múltiples gráficos que pueden ser usados para detectar anomalías

- Gráfico de caja o boxplot
- Gráfico de puntos o scatter plot
- Gráfico de percentiles ajustado

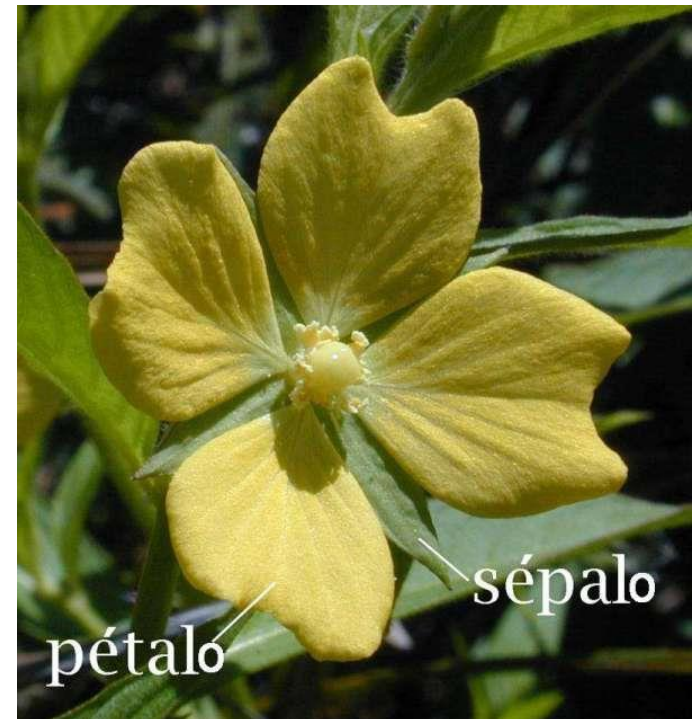
Dataset Iris

El dataset Iris fue recolectado por el estadístico y biólogo Ronald Fisher. El conjunto de datos contiene 50 muestras de cada una de tres especies de flores Iris

- Iris setosa
- Iris virginica
- Iris versicolor

Se midieron cuatro rasgos

- Largo de sépalo
- Ancho de sépalo
- Largo pétalos
- Ancho de pétalo



Dataset Iris

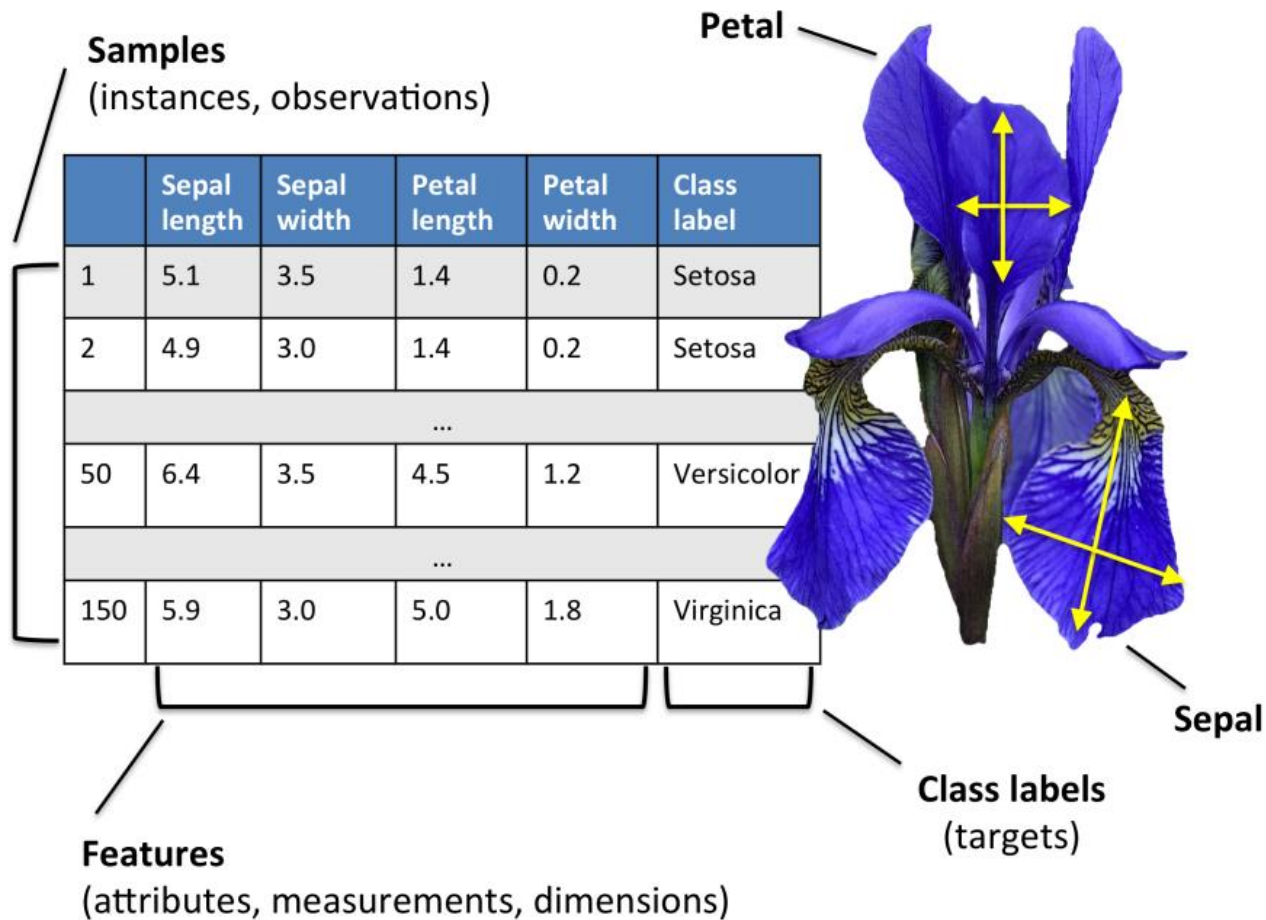
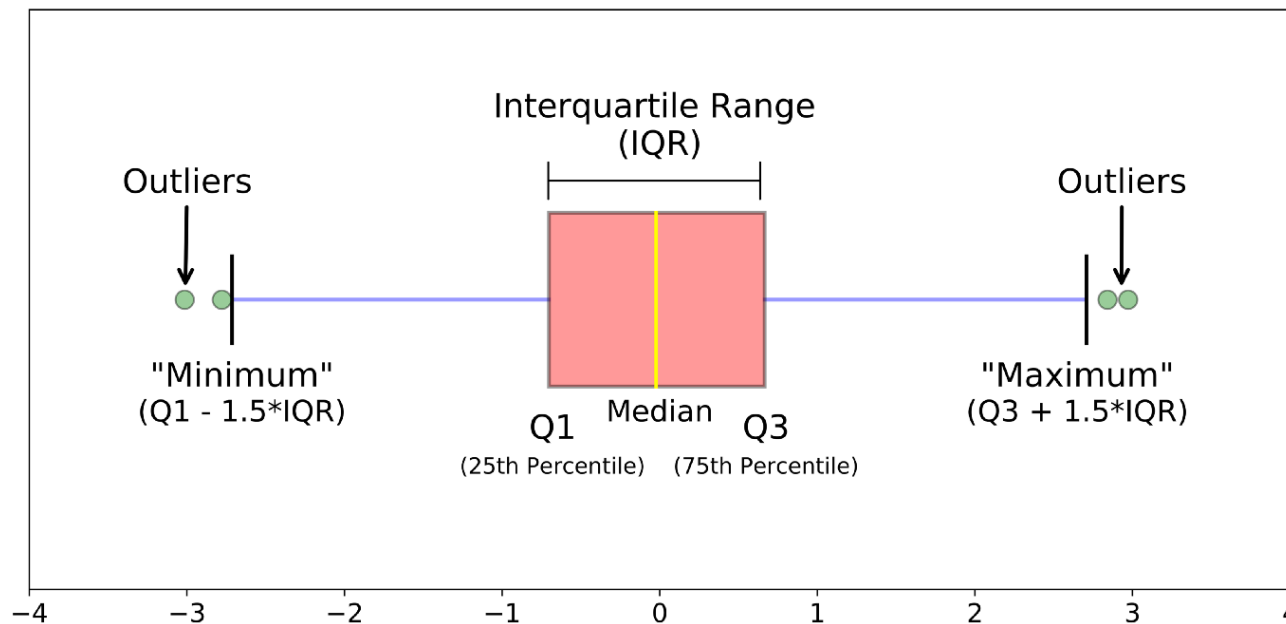


Gráfico de caja

Una forma estandarizada de mostrar la variación de datos basada en el resumen de cinco números, que incluye mínimo, primer cuartil, mediana, tercer cuartil y máximo



Ejemplo en R

```
### boxplot, scatter plot, aqplot, chi-square plot, symbolplot  
### Boxplot  
data <- data.frame(iris)  
head(data)  
boxplot(data$Petal.Length~data$Species,  
        main="Boxplot of Petal.Length and Species")  
Sepal.width <- data[,2]  
boxplot(Sepal.width,main="Boxplot of Sepal.width and Species")  
boxplot.stats(Sepal.width)$out
```

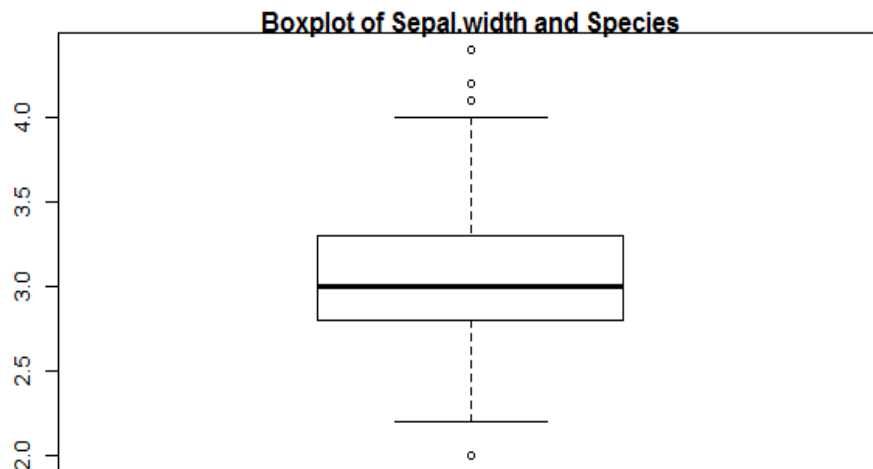
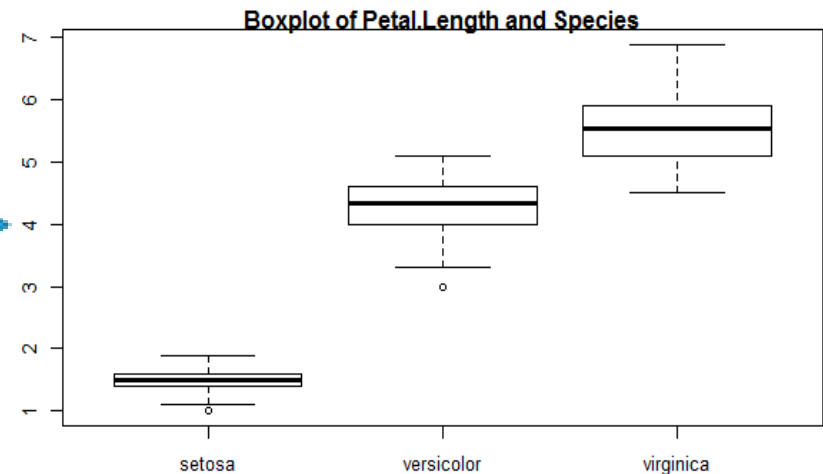


Gráfico de Violín

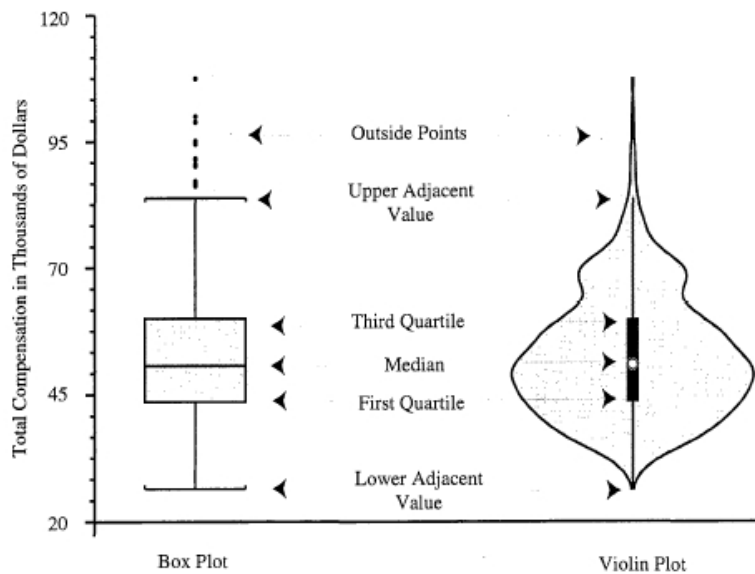


Figure 1. Common Components of Box Plot and Violin Plot. Total compensation for all academic ranks.

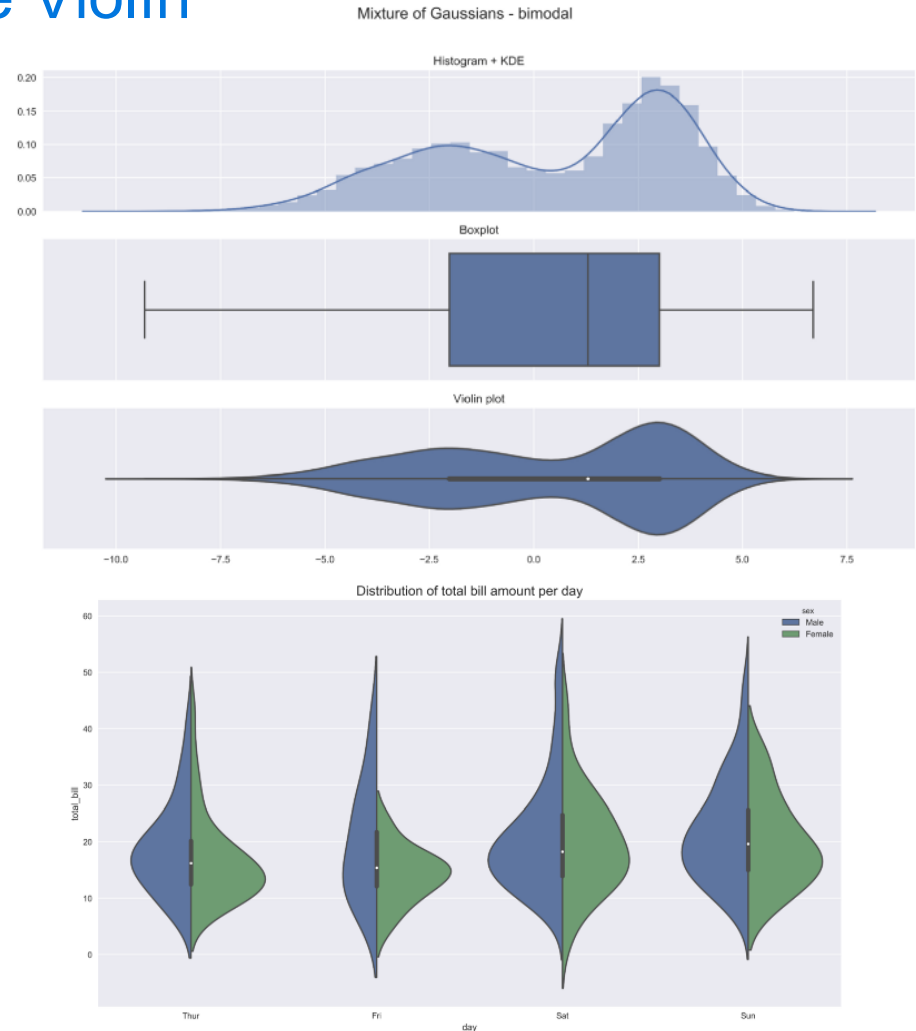
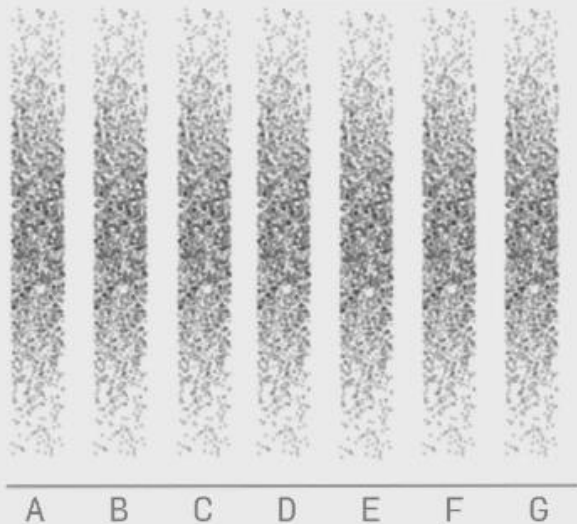
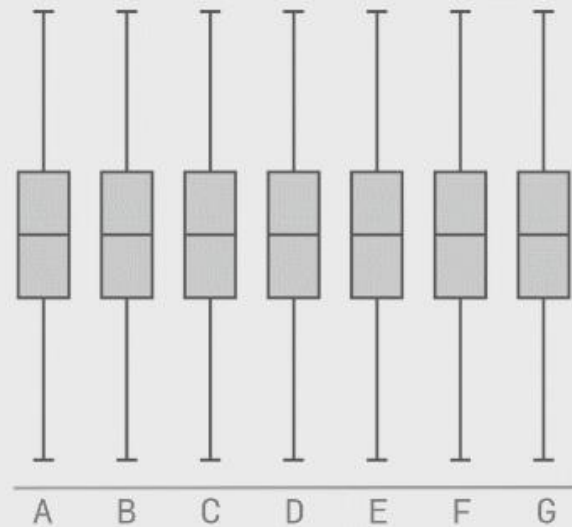


Gráfico de Violin

Raw Data



Box-plot of the Data



Violin-plot of the Data

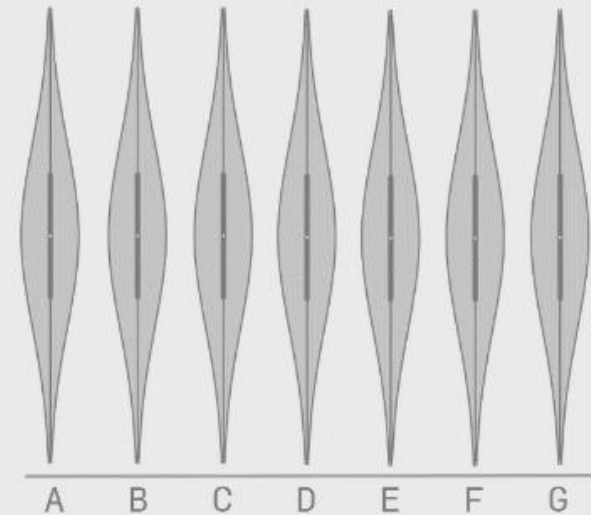
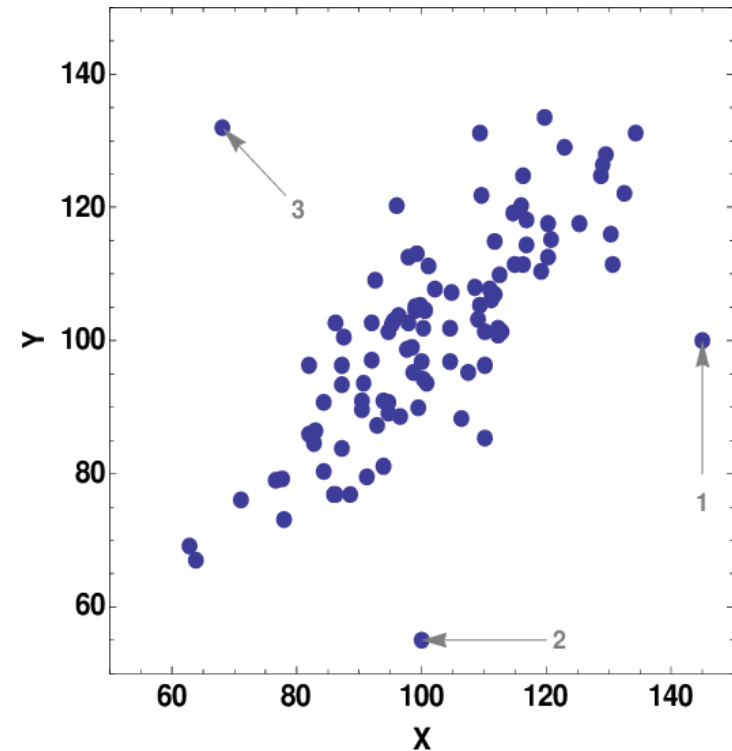


Gráfico de puntos

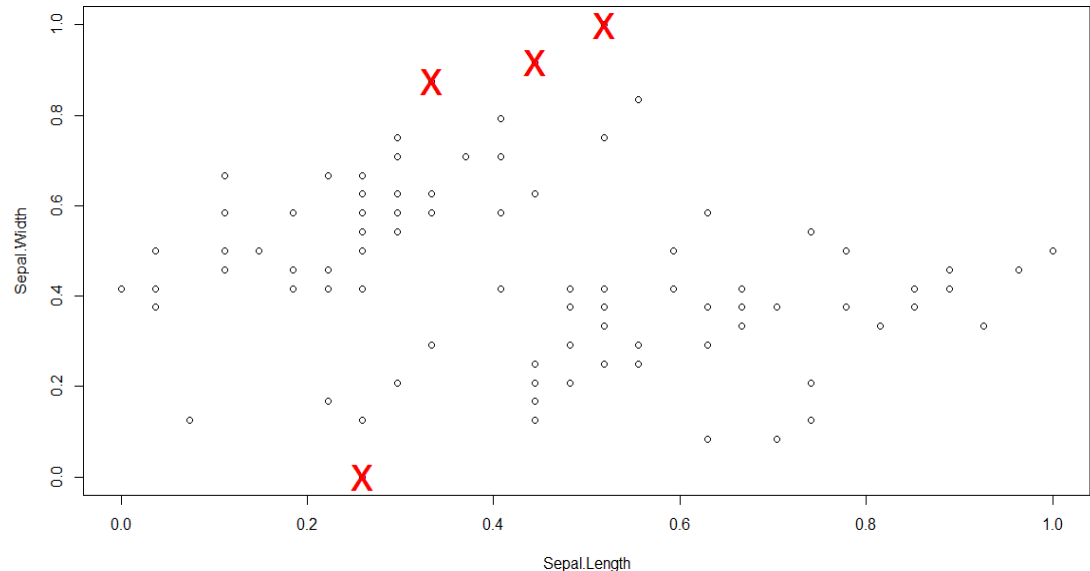
- Los gráficos de puntos se usan para mostrar pares de datos. Analizar si existe o no correlación. Típicamente dos variables numéricas.
- Un valor atípico se define como un punto que no parece encajar con el resto de los datos.



Ejemplo en R

```
### Scatterplot
data <- data[1:100,c(1,2)]
data <- data.Normalization(data,type="n4",normalization="column")
sepal.Length <- data[,1]
sepal.width <- data[,2]
(Sep.out1 <- which(Sepal.Length %in% boxplot.stats(Sepal.Length)$out))
(Sep.out2 <- which(Sepal.width %in% boxplot.stats(Sepal.width)$out))

### outliers in either Sepal.Length or Sepal.width
(outlier.list <- union(Sep.out1,Sep.out2 ))
plot(data)
points(data[outlier.list,], col="red", pch="x", cex=3)
```



Análisis estadístico

Existen múltiples métodos de análisis estadístico que pueden ser usados para detectar anomalías

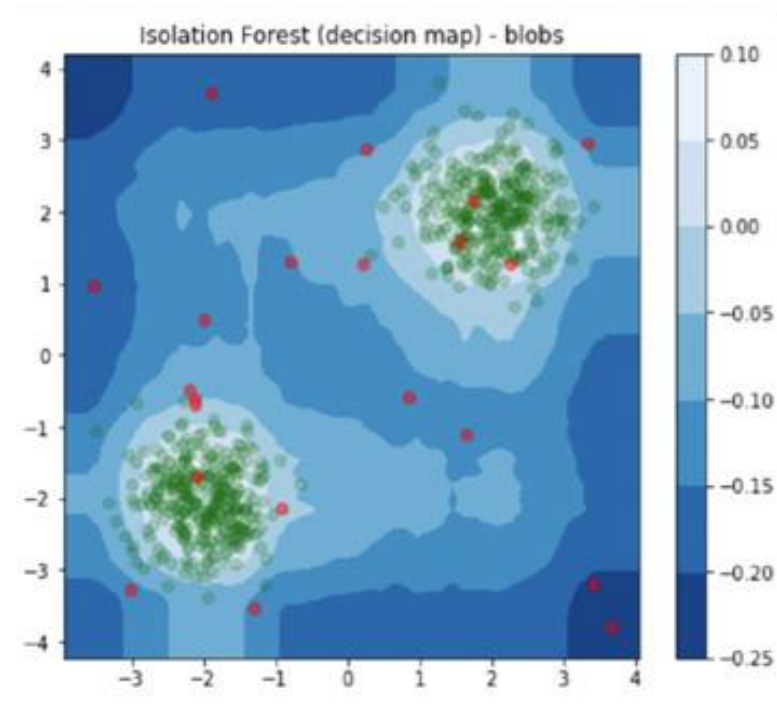
- Test de hipótesis
 - Test de Grubbs
- Uso de scores
 - Distribución normal
 - T-student
 - IQR ($Q3 - Q1$)



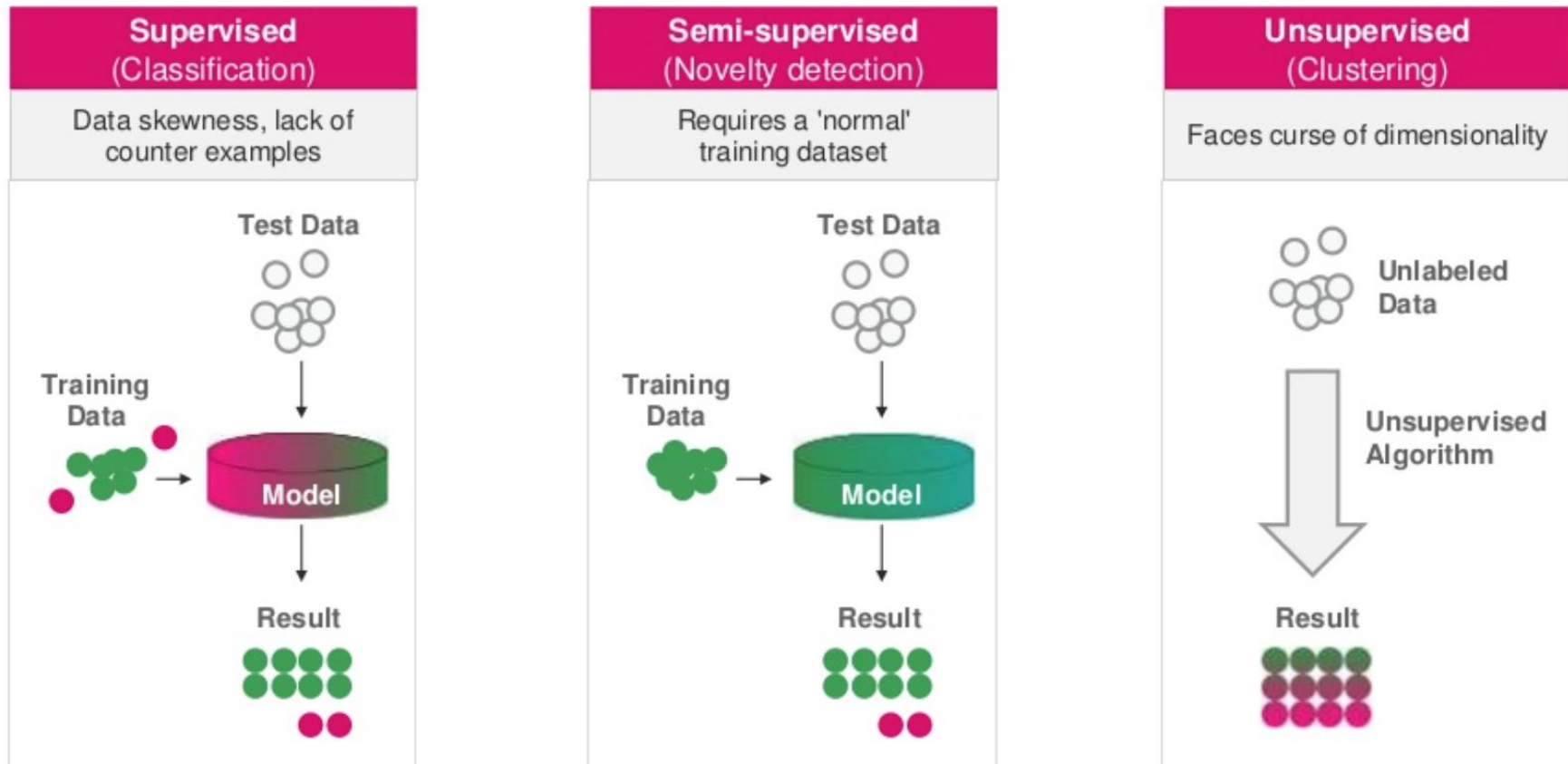
Análisis basado en ML

Existen múltiples métodos de ML





- Regresión lineal
- Random Forest
- Isolation Forest
- Clustering
- One-class SVM

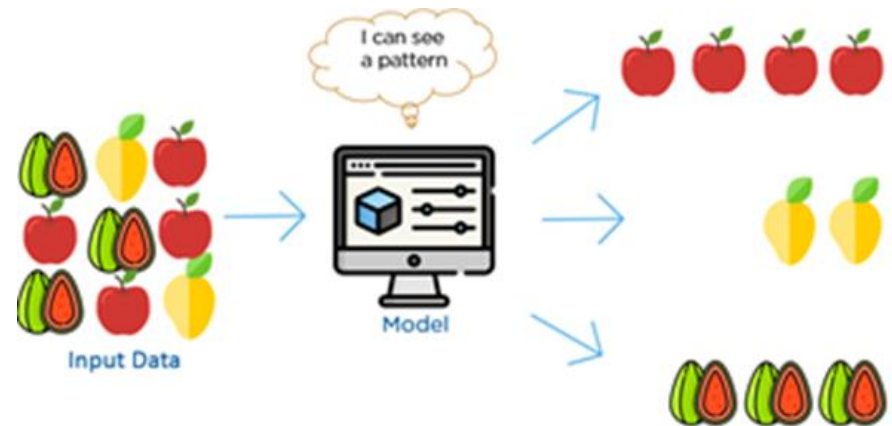


Análisis basado en ML

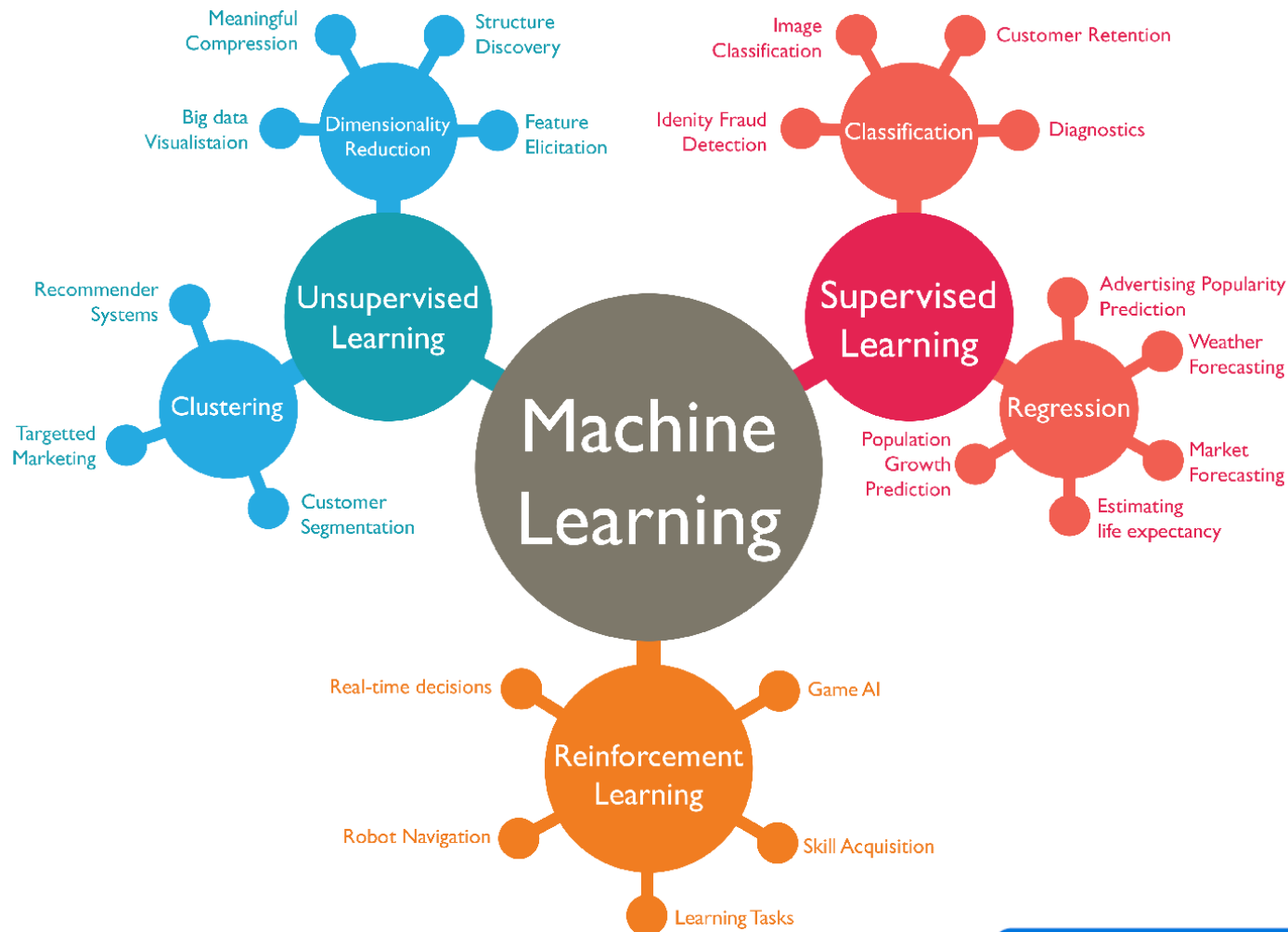


Análisis basado en ML

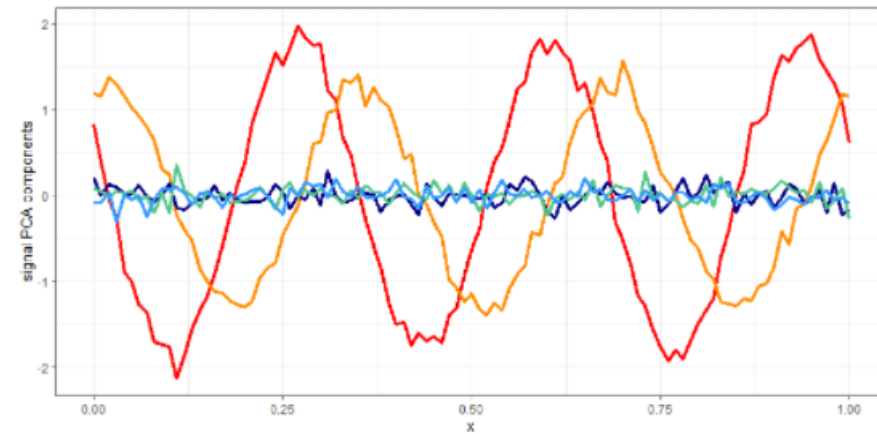
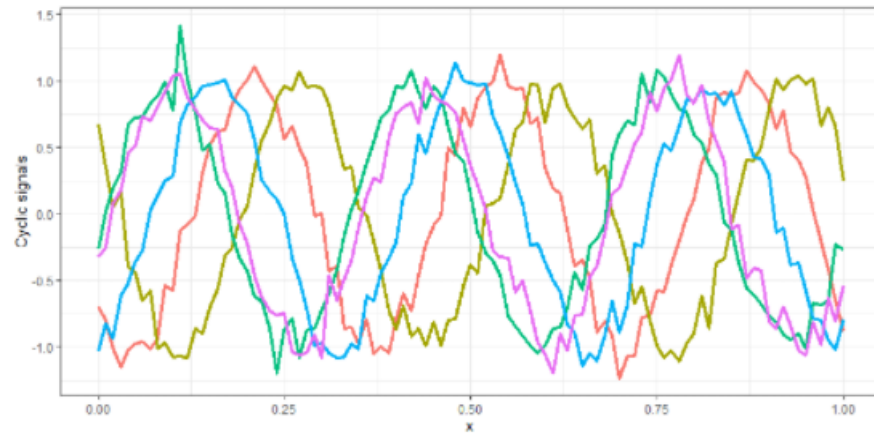
Input	Label	Prediction
	CAT	
	NOT CAT	
	CAT	
		?



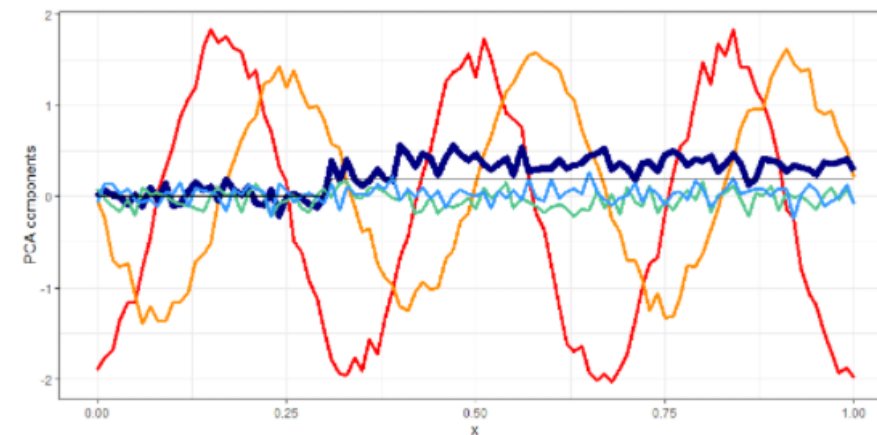
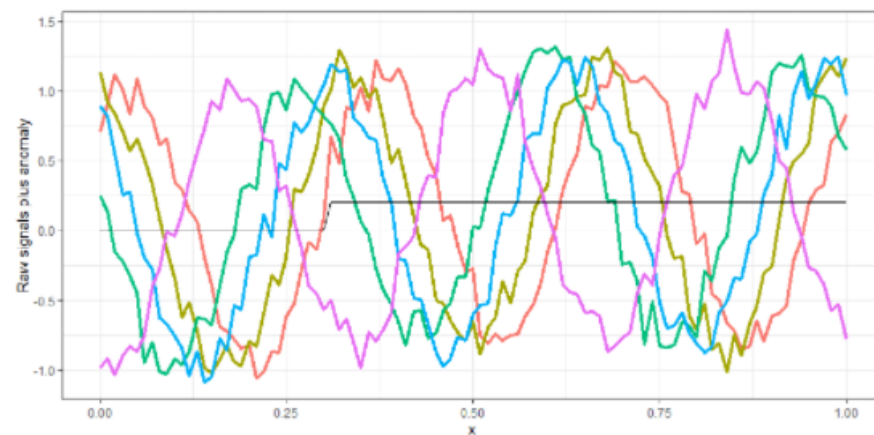
Análisis basado en ML



Otros Análisis - PCA



PCA isolates the common Fourier component among timeseries.



An anomaly introduced in the cyclic features is evident in the PCA components.

Volumen de datos vs Técnicas

Y LeCun

■ "Pure" Reinforcement Learning (cherry)

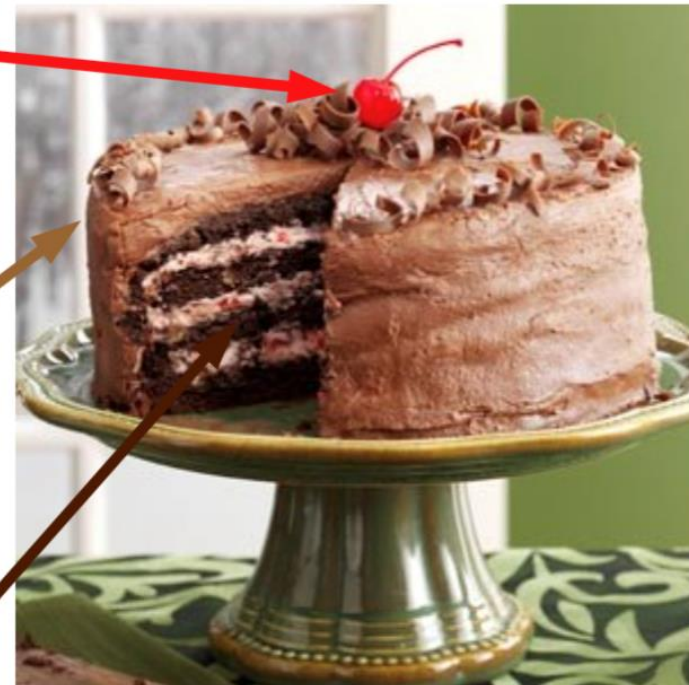
- ▶ The machine predicts a scalar reward given once in a while.
- ▶ **A few bits for some samples**

■ Supervised Learning (icing)

- ▶ The machine predicts a category or a few numbers for each input
- ▶ Predicting human-supplied data
- ▶ **10→10,000 bits per sample**

■ Unsupervised/Predictive Learning (cake)

- ▶ The machine predicts any part of its input for any observed part.
- ▶ Predicts future frames in videos
- ▶ **Millions of bits per sample**



■ (Yes, I know, this picture is slightly offensive to RL folks. But I'll make it up)



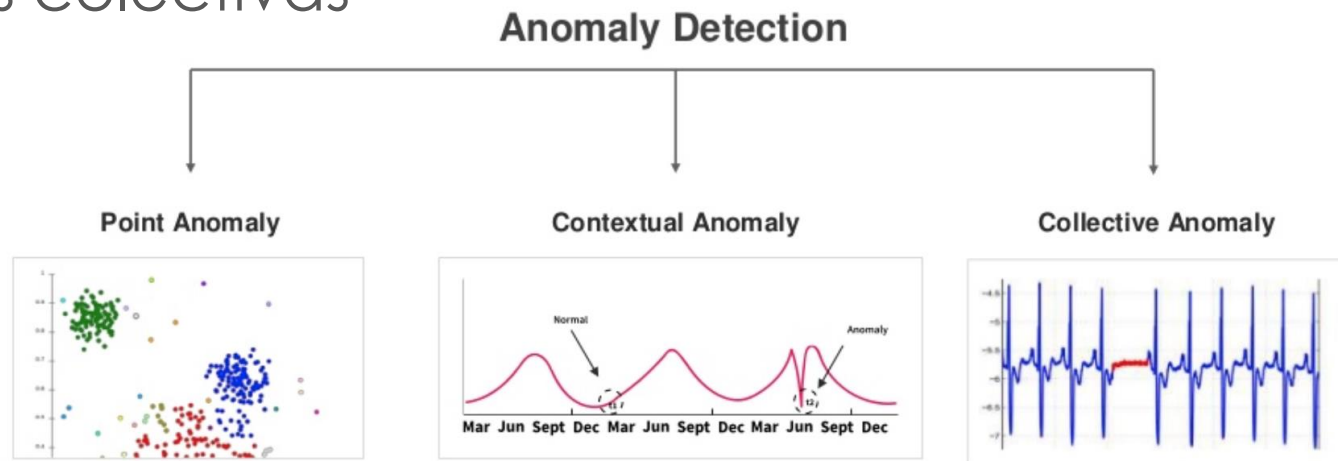
Clase 03: Anomalías

TIPOS DE ANOMALIAS

Tipos de anomalías

Existen 3 tipos de anomalías que pueden detectarse a partir del análisis de datos

- Anomalías puntuales
- Anomalías contextuales
- Anomalías colectivas

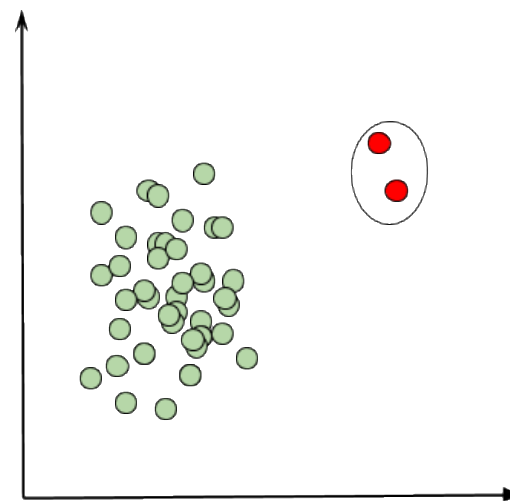
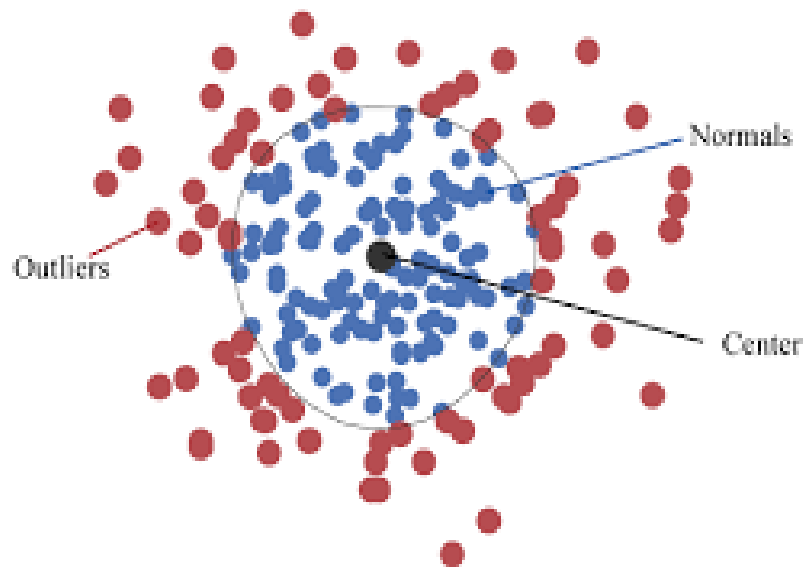




Anomalías puntuales

Las anomalías puntuales son simplemente instancias anómalas individuales dentro de un conjunto de datos más grande.

Ejemplo: Una temperatura de 60 °C en un conjunto de datos sería una anomalía puntual, ya que sería la temperatura más alta jamás registrada en la Tierra



Monto de compra



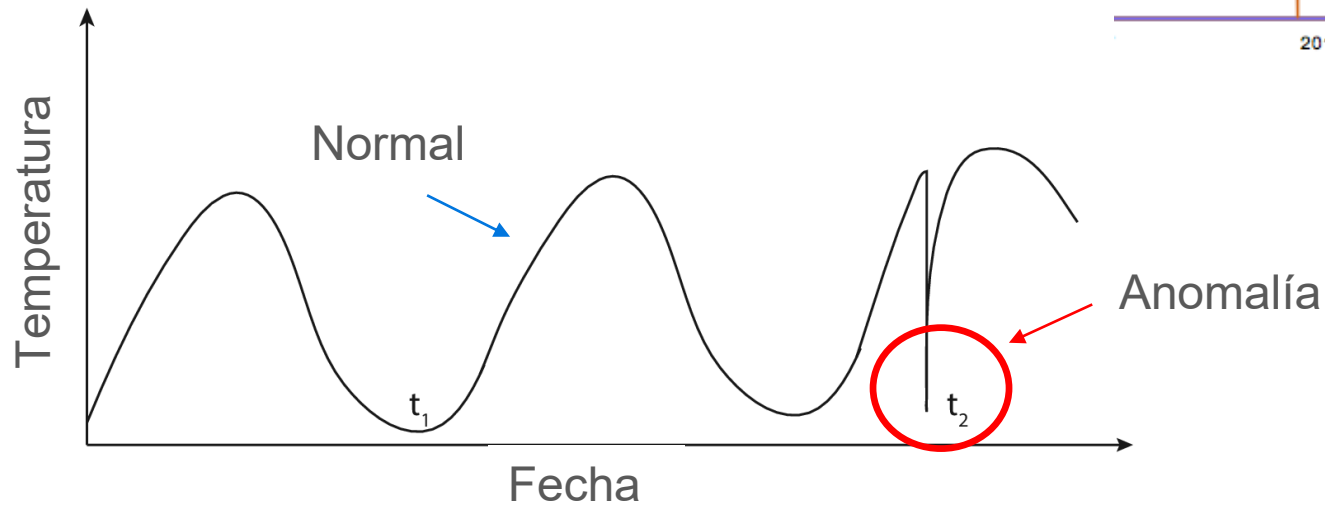
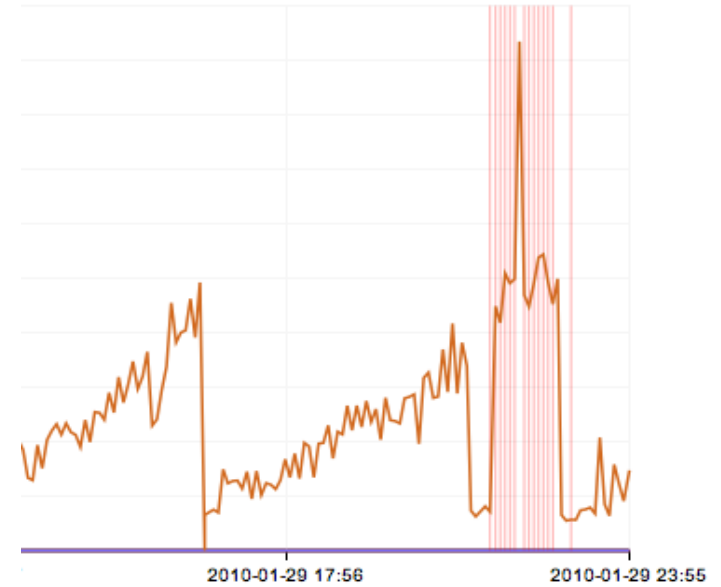
Anomalías contextuales

Estos son puntos que solo se consideran anómalos en cierto contexto. En datasets espaciales el contexto lo puede dar la latitud y longitud. En series de tiempo lo da el tiempo.

Ejemplo: Si bien se considera que 30 °C están dentro del rango de temperaturas posibles, dado el contexto de Julio en la ciudad de Santiago, este punto de datos es ciertamente una anomalía.



Eth1 ☒ Eth2 ☒ Eth3 ☒ Lo ☒ Anomalles ☒

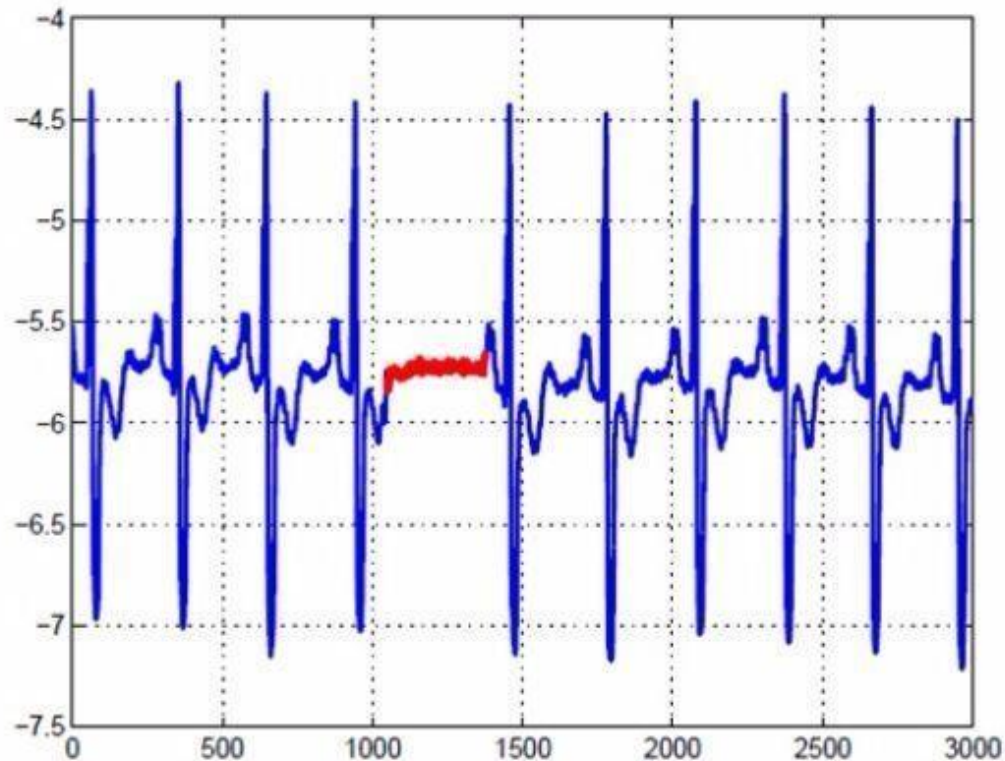




Anomalías colectivas

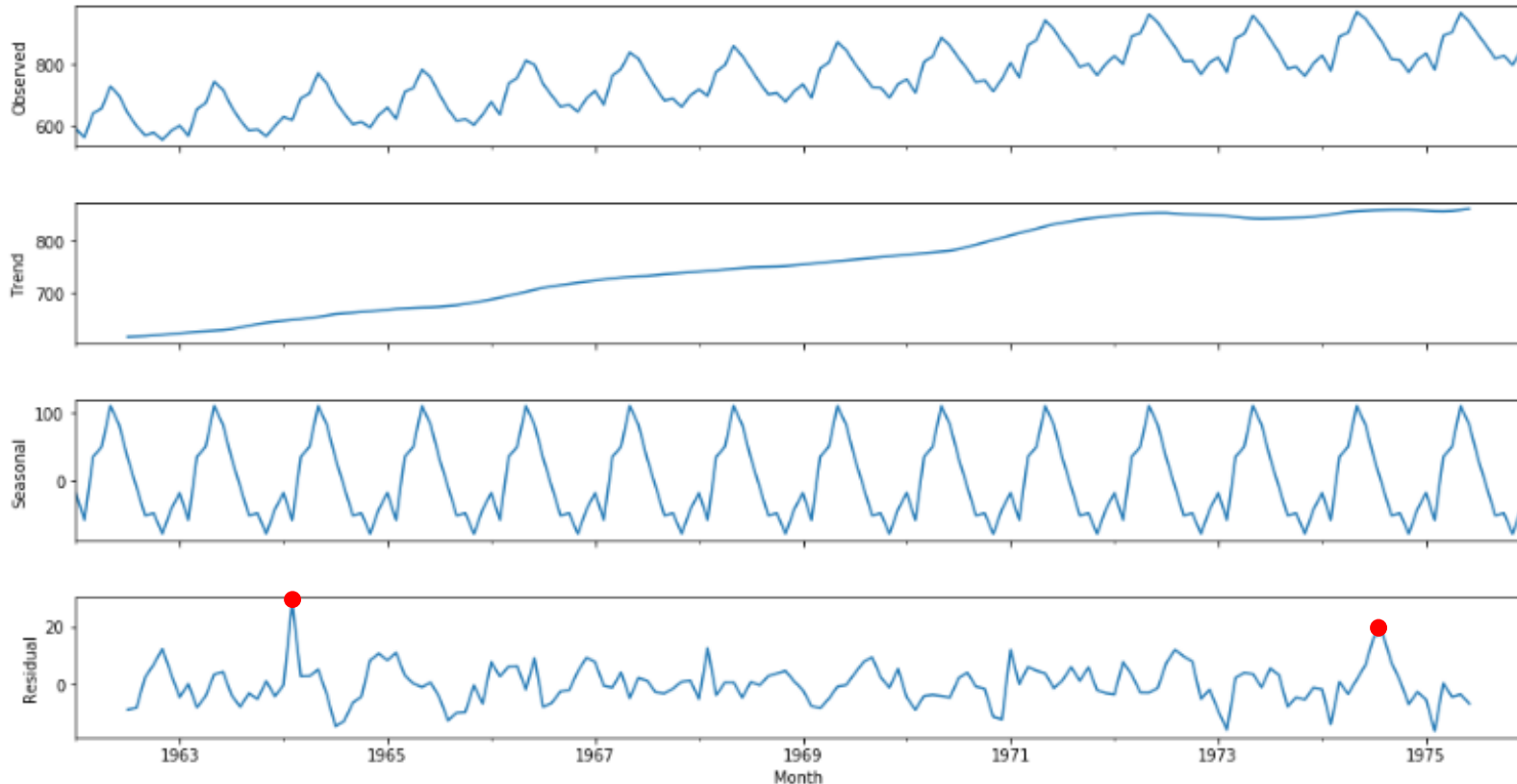
Cuando los conjuntos de datos relacionados o partes del mismo conjunto de datos en conjunto son anómalos con respecto al conjunto de datos completo.

Ejemplo: Los datos de una tarjeta de crédito muestran que el cliente realiza una compra en los Miami, pero también muestran que al mismo tiempo se realiza un giro en un cajero automáticos en Santiago.



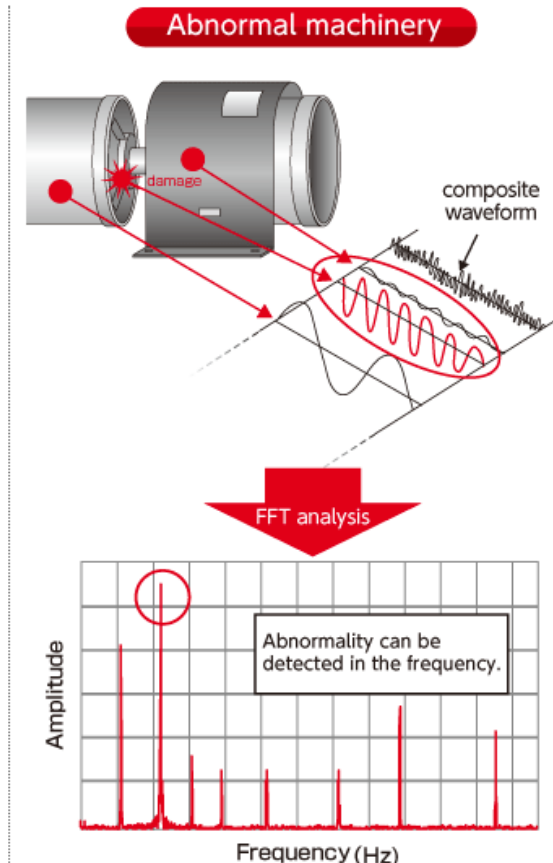
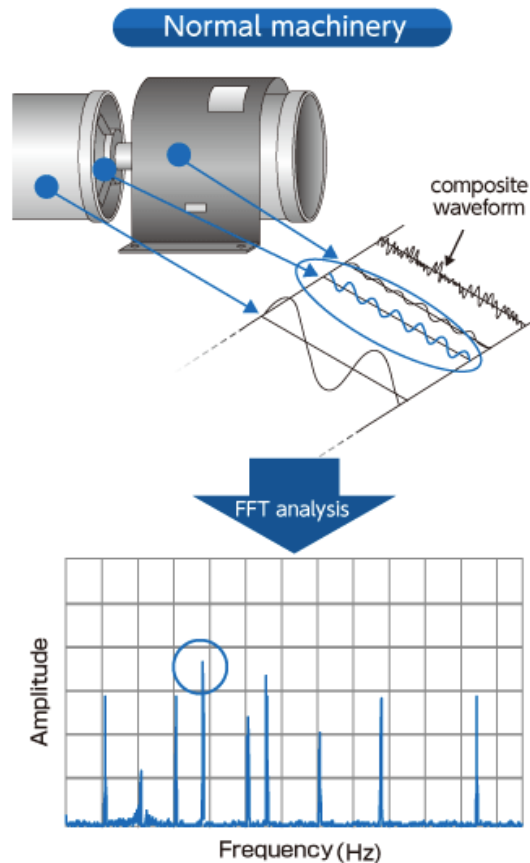
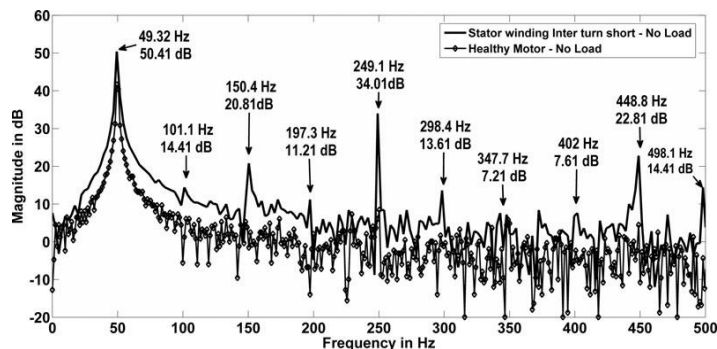
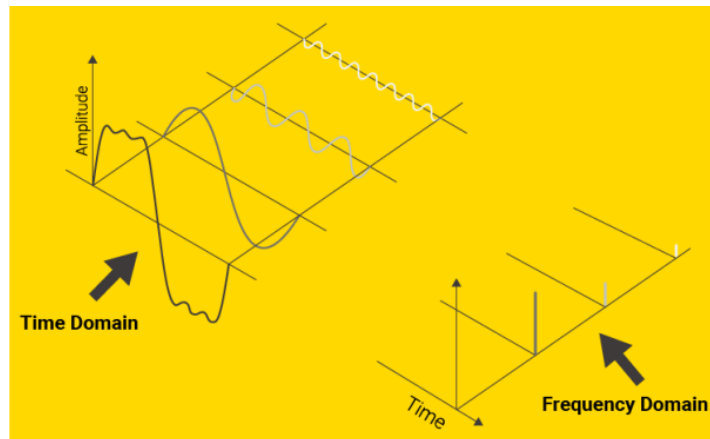
La parte roja de la señal es un valor atípico colectivo respecto al conjunto. Se mantiene en torno a un valor durante una duración significativamente más larga de lo normal.

Descomposición de Series



Para modelos de ML en general se puede entrenar la serie observada directamente, en donde el modelo debería aprender la estacionalidad, tendencia y varianza. Sin Embargo, con fines exploratorios y entendimiento de los datos conviene descomponer la serie de tiempo y en algunos casos entrenar modelos en donde se remueven tendencias o estacionalidad(ya entendidos y modelados) puede entregar mejores resultados.

Anomalías en Frecuencia



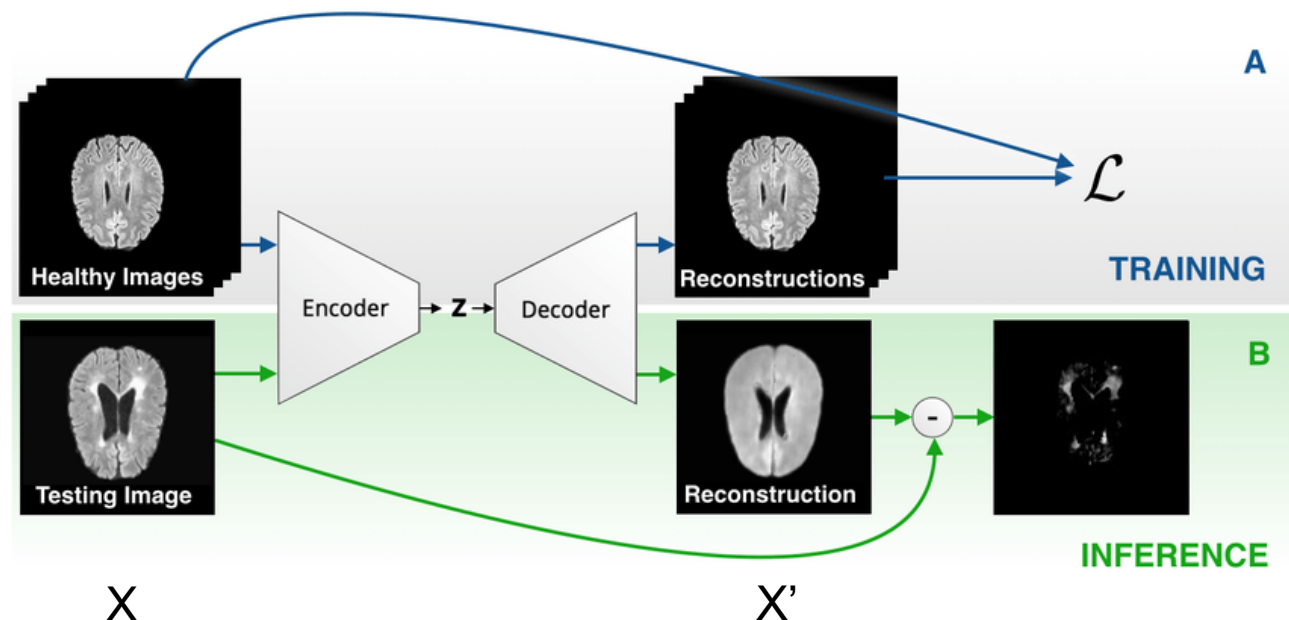
*generic example

Técnicas como la transformada de Fourier o PCA pueden facilitar la búsqueda de anomalías que se presentan en patrones asociados a series de tiempo.

Ejemplo: Patrón de vibración de un Motor o maquina rotatoria

Detección de anomalías usando ML: AutoEncoders

- Los autoencoders son un tipo de redes neuronales no supervisadas. Reciben un registro X (tabla, imagen, sonido), lo comprimen y generan la representación Z y finalmente buscan reconstruir el registro original y lo almacena en X' .





Clase 03: Anomalías

TALLER GRUPAL



Taller de Detección de anomalías

<https://colab.research.google.com/drive/1o11AhWuYY4WK2s7yAEdwyUvSHirNM8fL>



ESCUELA DE INGENIERÍA
FACULTAD DE INGENIERÍA

EDUCACIÓN
PROFESIONAL

Diplomado en Big Data y Ciencia de Datos *Ciencia de Datos y sus Aplicaciones*



www.educacionprofesional.ing.uc.cl