



ESCUELA DE INGENIERÍA  
FACULTAD DE INGENIERÍA

EDUCACIÓN  
PROFESIONAL

# Diplomado en Big Data y Ciencia de Datos

## Curso: *Ciencia de Datos y sus Aplicaciones*

Educación Profesional  
Escuela de Ingeniería UC

✉ regonzar@uc.cl

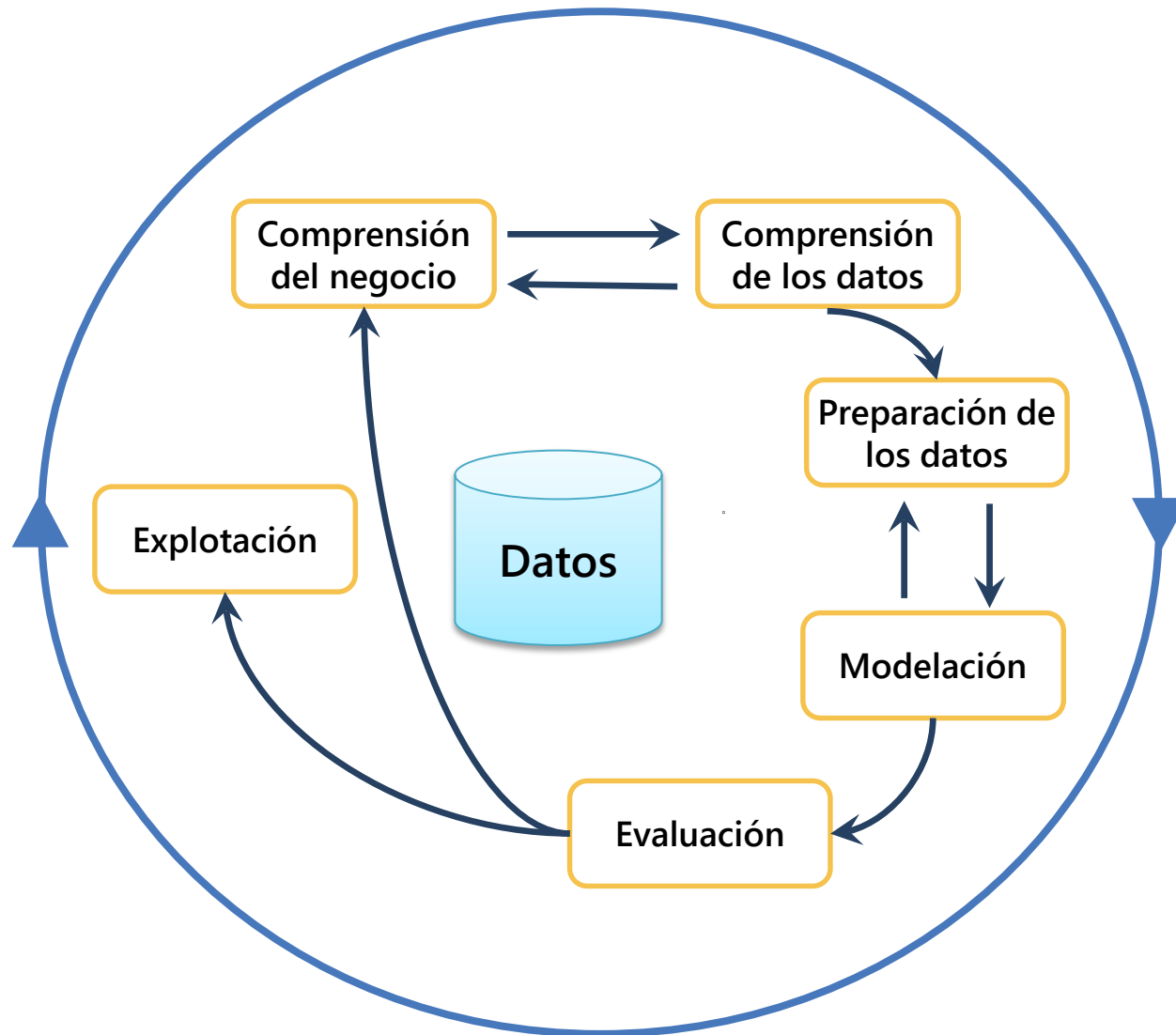
✉ rmunoz@uc.cl

✉ jcaiceo@uc.cl

Roberto González, Roberto Muñoz, Jaime Caiceo



# CRISP-DM



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

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

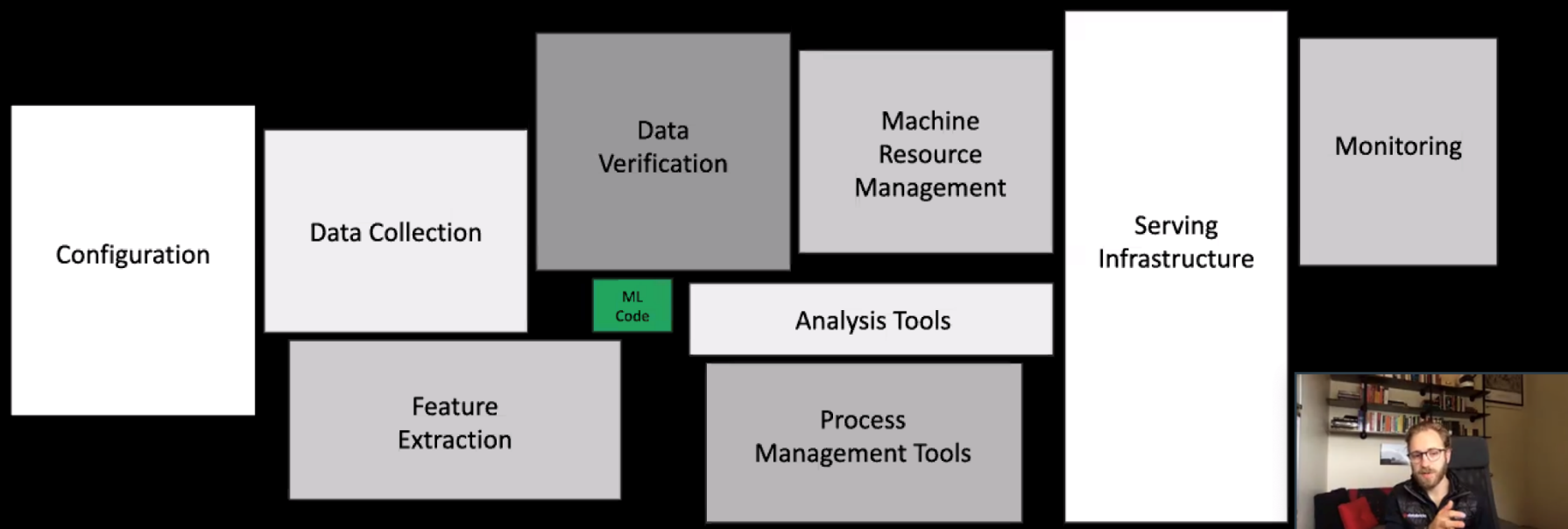


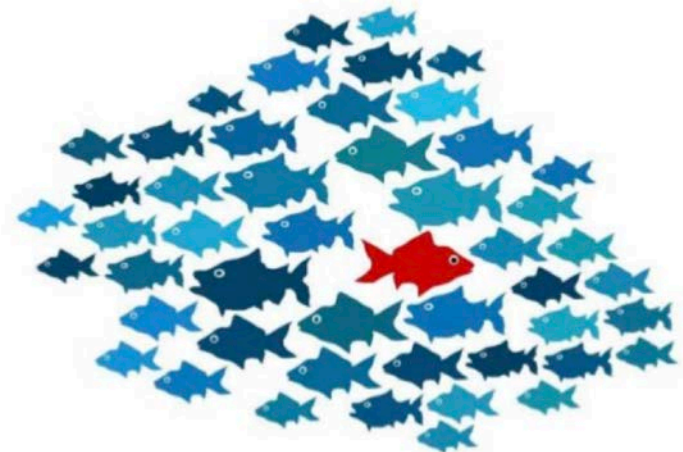
Figure 1: Only a small fraction of real-world ML systems is composed of the ML code, as shown by the small green box in the middle. The required surrounding infrastructure is vast and complex.

Clase 03: Anomalías

# DETECCIÓN DE ANOMALIAS

# ¿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

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





# ¿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

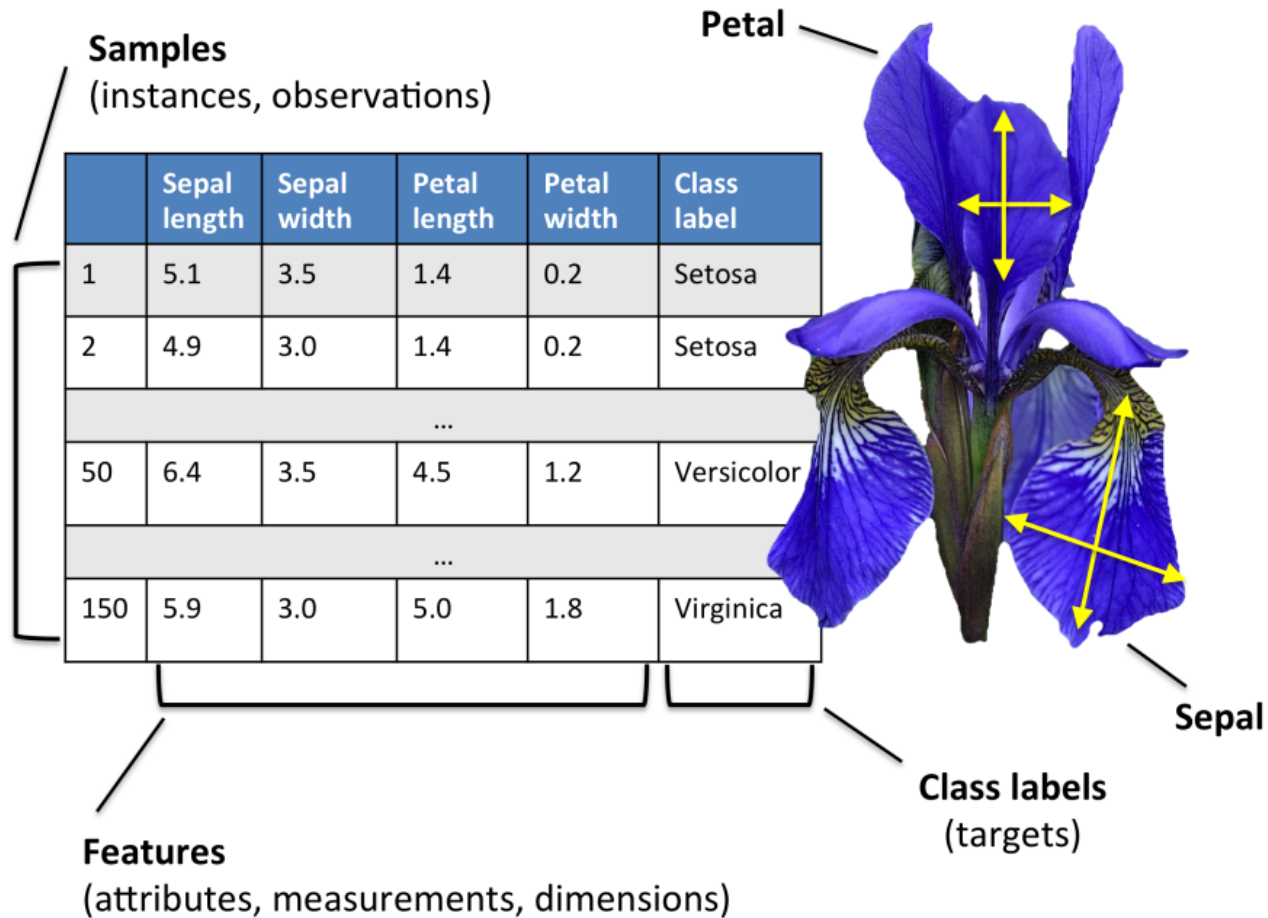
- Gráfico de caja o boxplot
- Gráfico de puntos o scatter plot
- Gráfico de percentiles ajustado

# 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

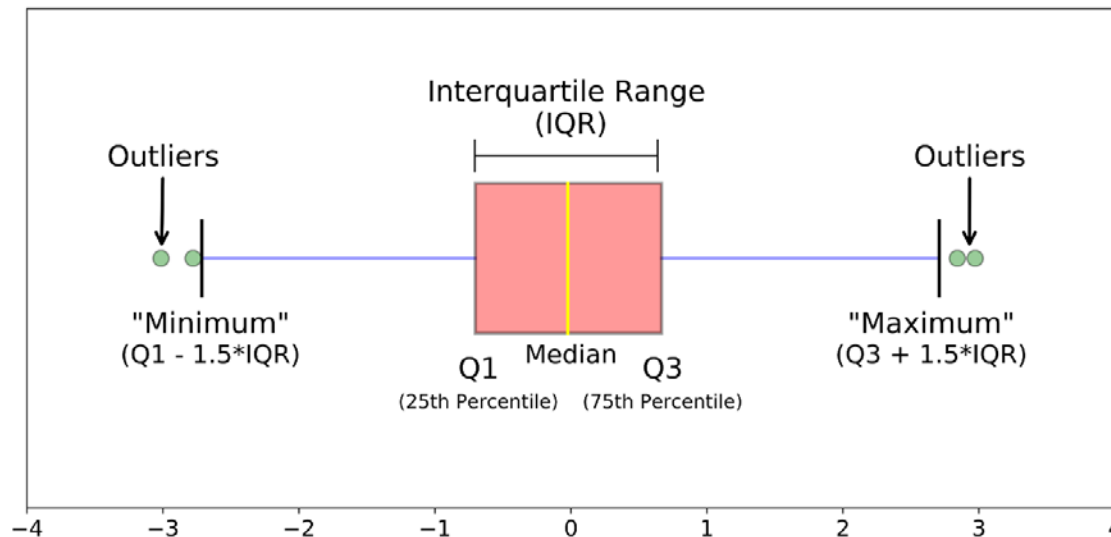


# Dataset Iris



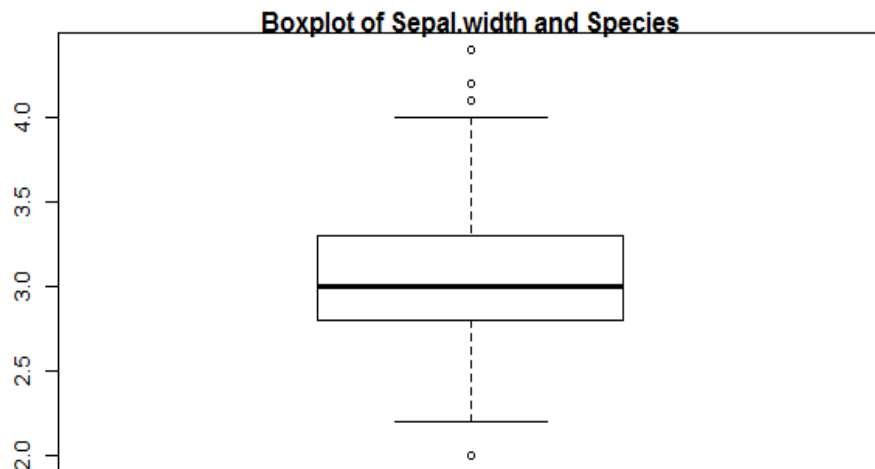
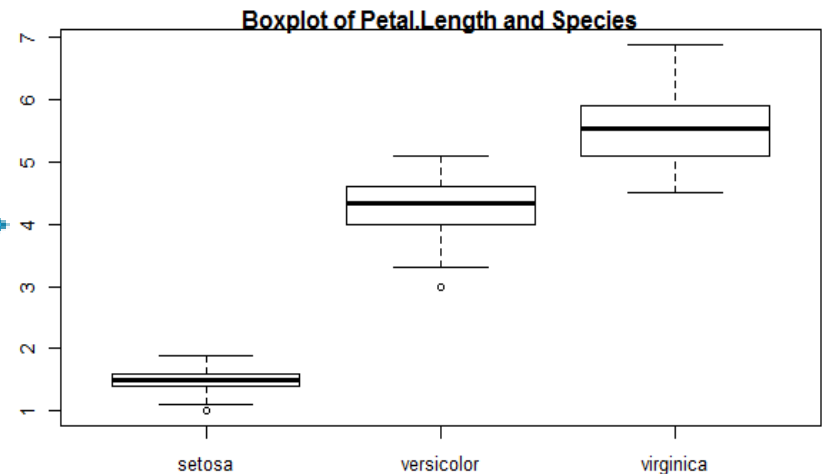
# Gráfico de caja

- 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



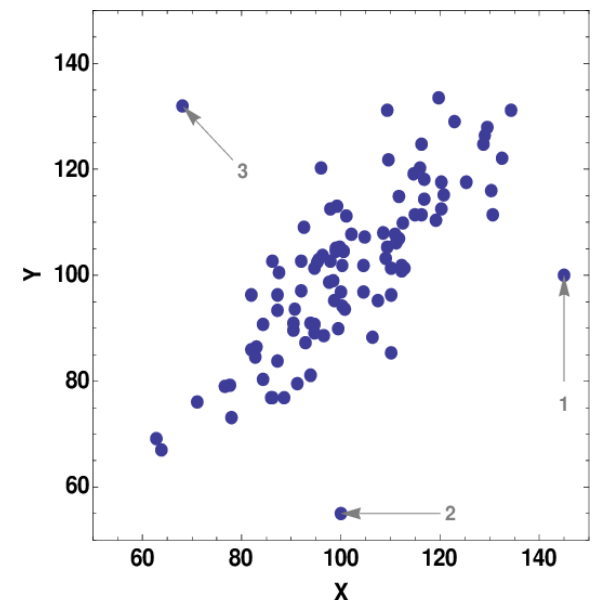
# Ejemplo en R

```
### 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 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.

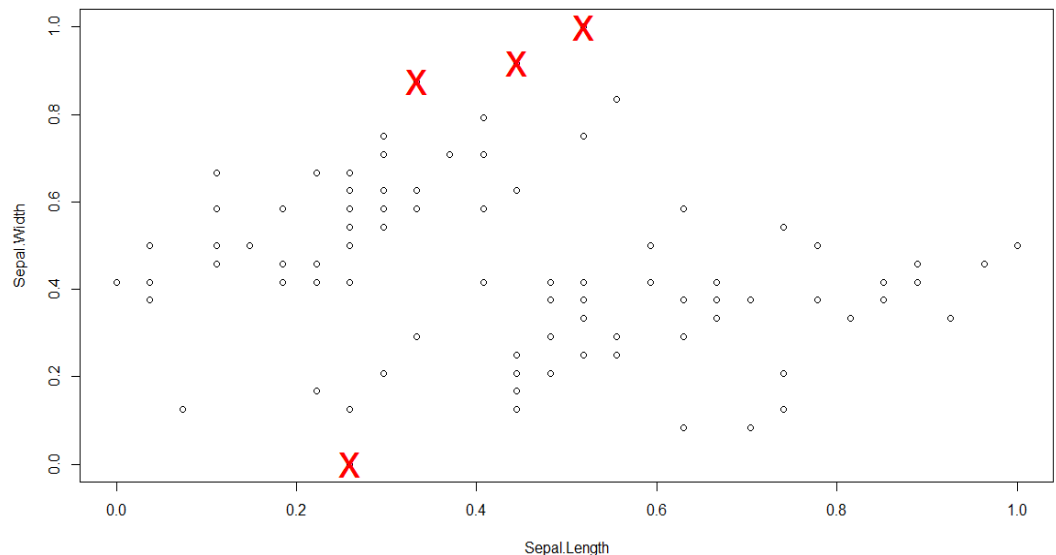




# Ejemplo en R

```
### 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

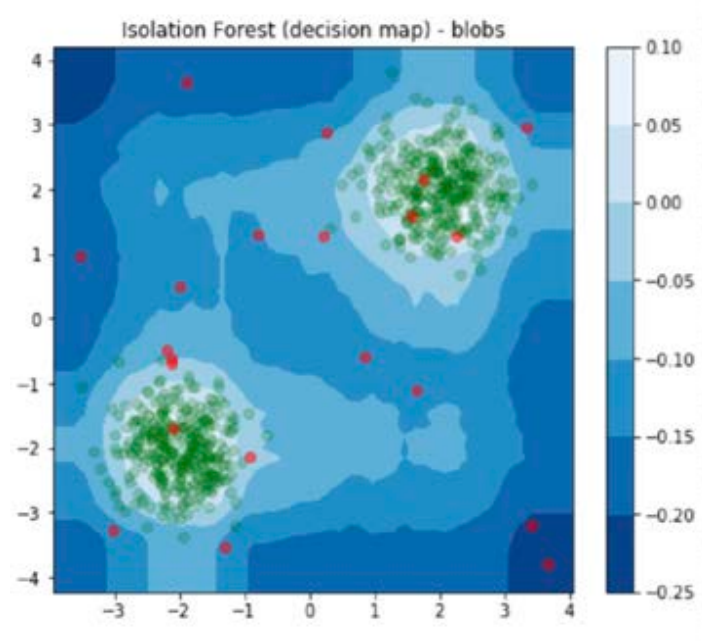
- Test de hipótesis
  - Test de Grubb
- Uso de scores
  - Distribución normal
  - T-student
  - IQR ( $Q3 - Q1$ )



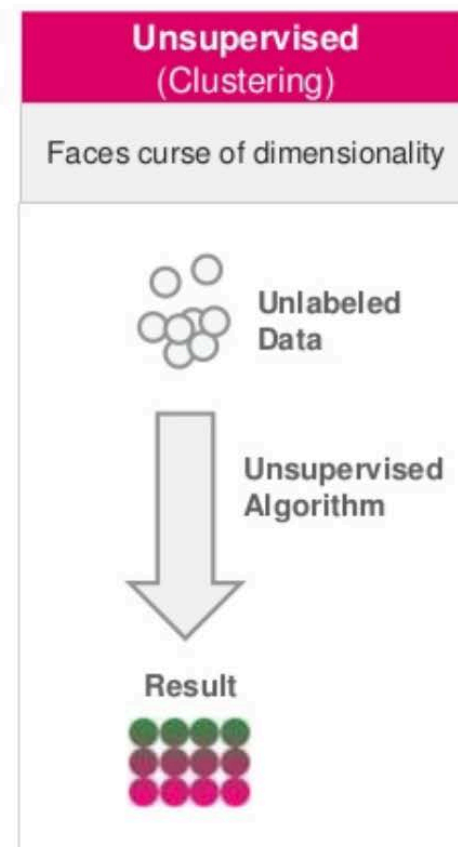
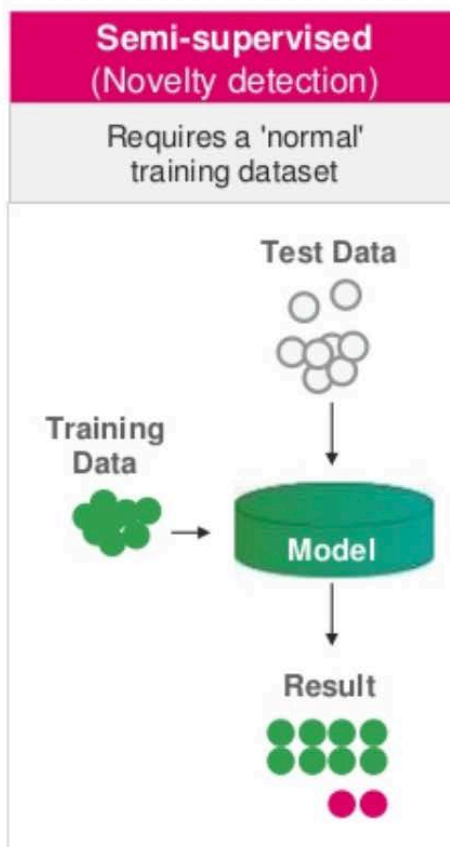
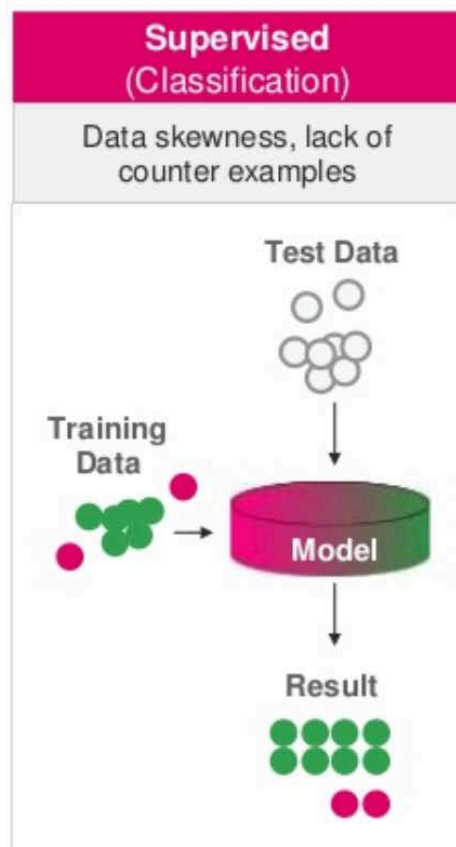
# Análisis basado en ML

Existen múltiples métodos de ML





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

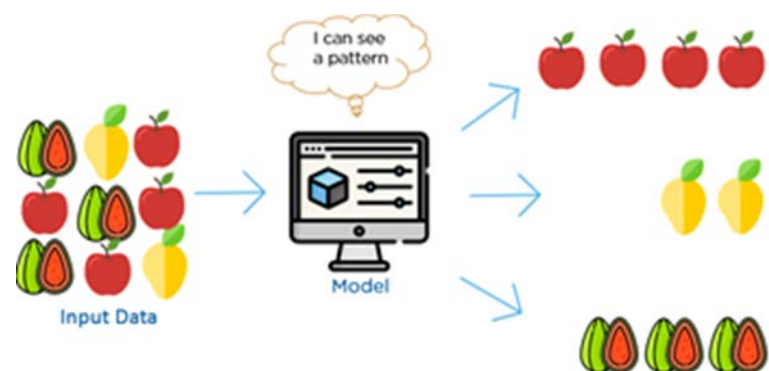


# Análisis basado en ML



# Análisis basado en ML

Input	Label	Prediction
	CAT	
	NOT CAT	
	CAT	
		 ?



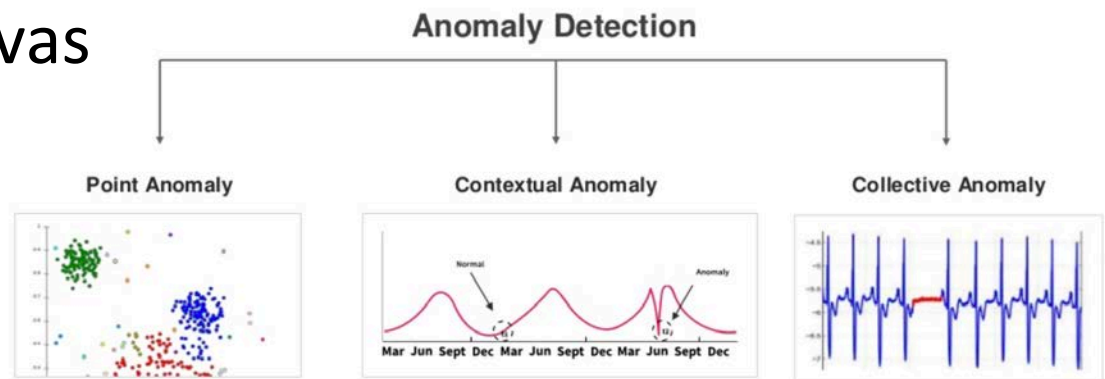
Clase 03: Anomalías

# **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

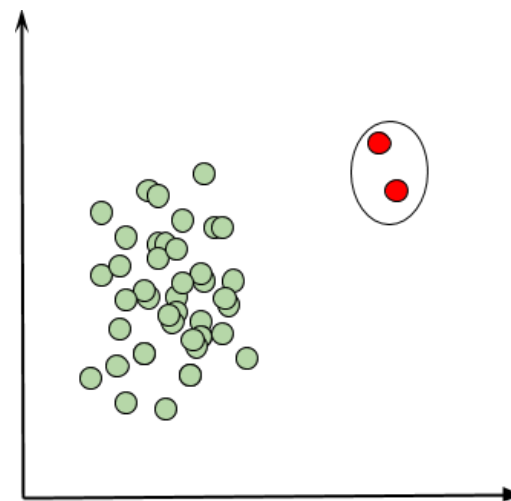
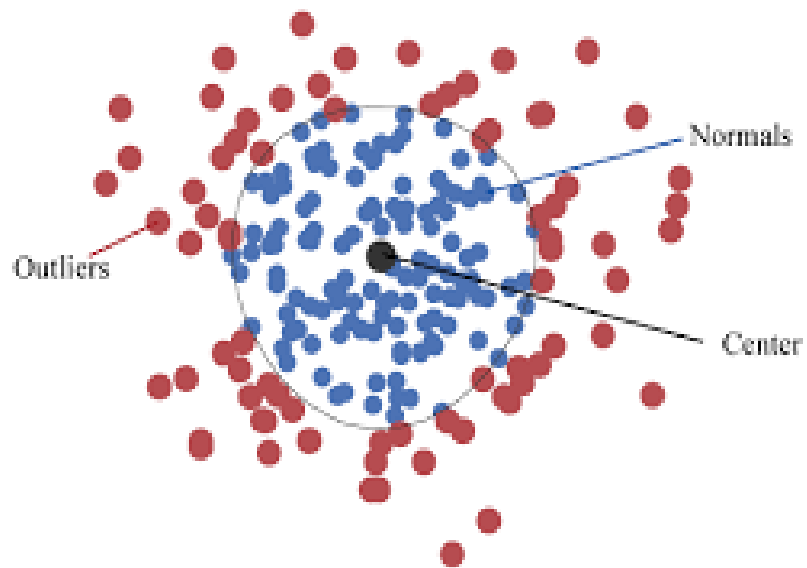


# Anomalías puntuales

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



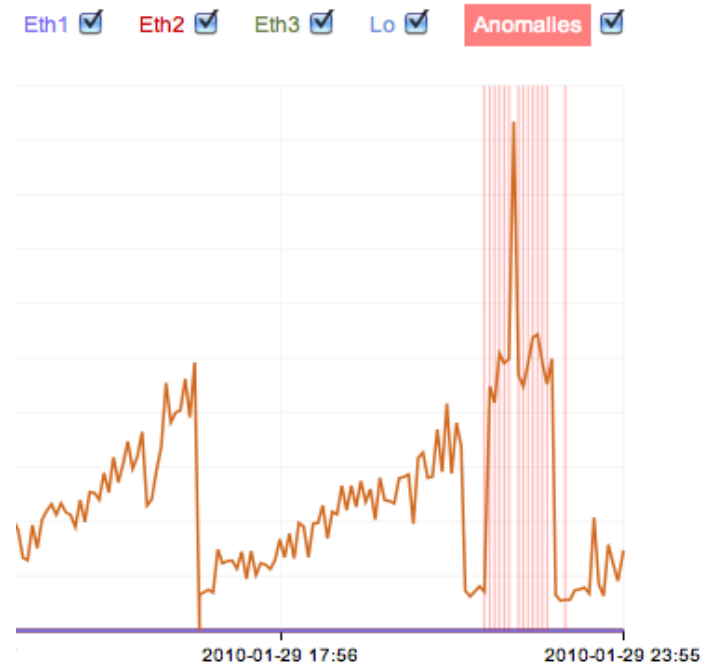
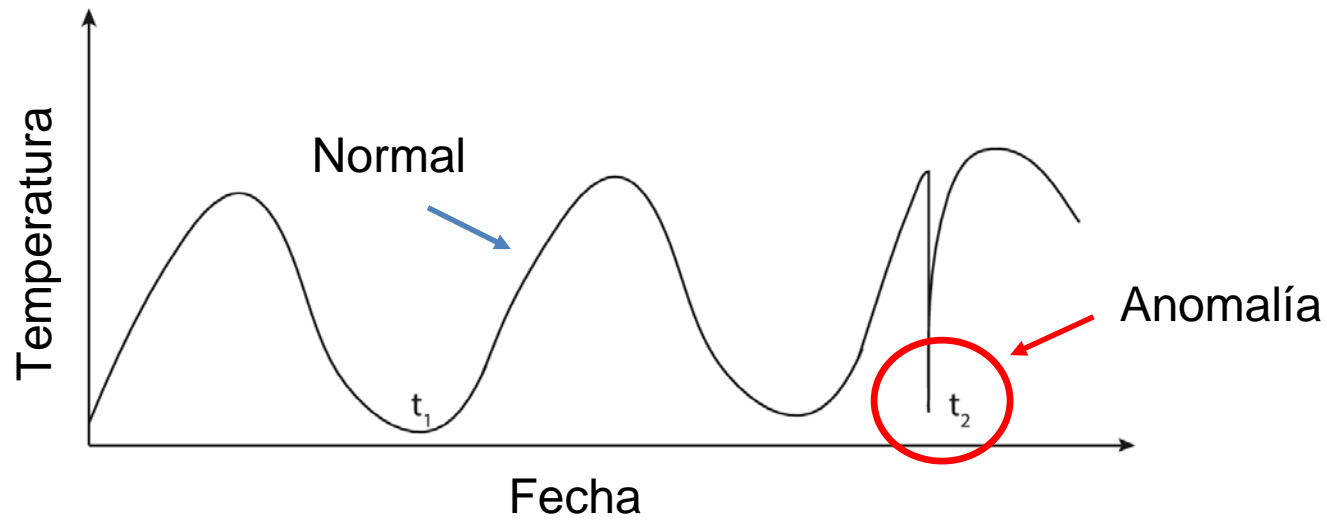


Monto de compra

# 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.

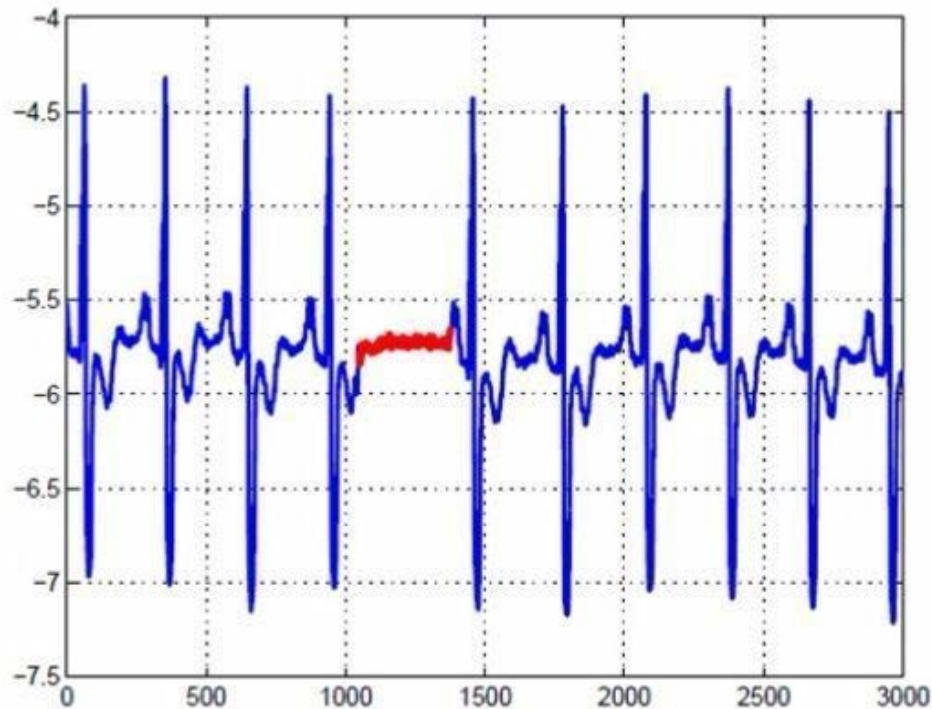
- Ejemplo: Si bien se considera que 30 °C están dentro del rango de temperaturas posibles, dado el contexto de Julio en la ciudad de Santiago, este punto de datos es ciertamente una anomalía.



# 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.

Clase 03: Anomalías

# **TALLER GRUPAL**

# Taller grupal

## Taller de Detección de anomalías

<https://colab.research.google.com/drive/1hHiN5LCG5dhIMamatXCVD7FuBh2FpC3j?usp=sharing>

## Formulario respuestas

<https://forms.gle/R5Uet1Ugte2XCpzS6>



ESCUELA DE INGENIERÍA  
FACULTAD DE INGENIERÍA

EDUCACIÓN  
PROFESIONAL

# Diplomado en Big Data y Ciencia de Datos

## Curso: *Ciencia de Datos y sus Aplicaciones*

Educación Profesional  
Escuela de Ingeniería UC

✉ regonzar@uc.cl

✉ rmunoz@uc.cl

✉ jcaiceo@uc.cl

Roberto González, Roberto Muñoz, Jaime Caiceo

