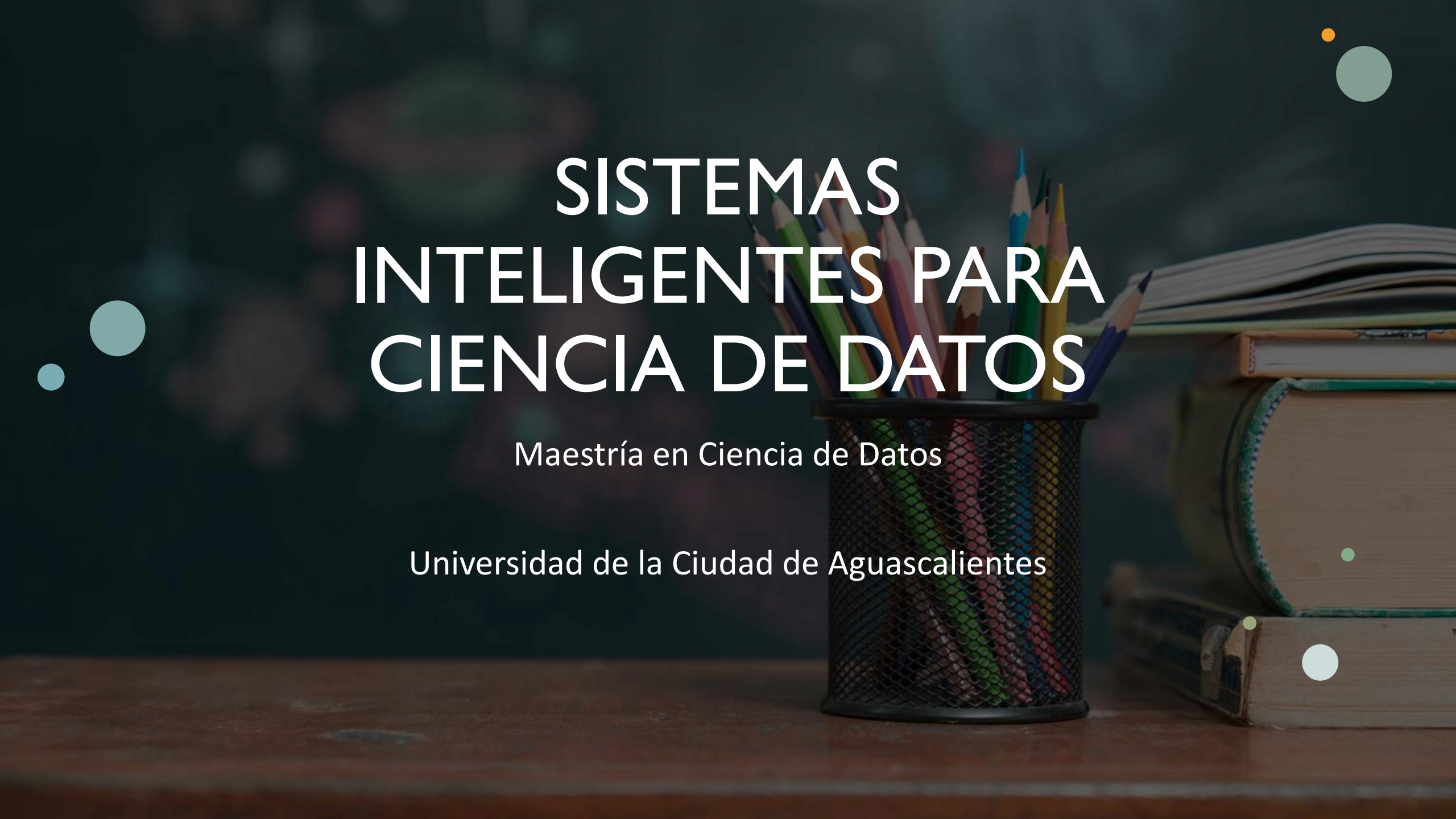


SISTEMAS INTELIGENTES PARA CIENCIA DE DATOS

Maestría en Ciencia de Datos

Universidad de la Ciudad de Aguascalientes

Técnicas de procesamiento de datos

¿Qué es el preprocesamiento?

- Conjunto de
  - técnicas y
  - herramientas
- Previo de cualquier análisis de datos o modelado

¿Para que hacer preprocesamiento?

- El 99% de los casos los datos no están listos para ser analizados
- Los datos pueden
  - No estar en la forma deseada
  - Tener errores
  - No tener algunos datos de interés
- Crucial en el análisis de datos

Tipos de preprocesamiento de datos

- Selección
- Tratamiento de valores faltantes
- Tratamiento de valores atípicos
- Transformación de formatos
- Normalización
- Tratamiento de duplicados
- Integración de múltiples fuentes
- Reducción de dimensionalidad
- Creación de características

Selección de datos

- Selección de columnas
- Selección de registros
  - Muestreo
  - Por propiedades

Tratamiento de valores faltantes

- ¿Qué hacer cuando tenemos?
  - Na, NaN, “”, NULL
- Eliminar los registros
  - Sencillo
  - Se pierde información
- Imputación
  - Valor aleatorio
  - Media/Mediana general
  - Media por grupos
  - Vecino mas cercano
  - Regresión lineal
  - Algoritmos especializados de imputación

¿Valores atípicos? (outliers)

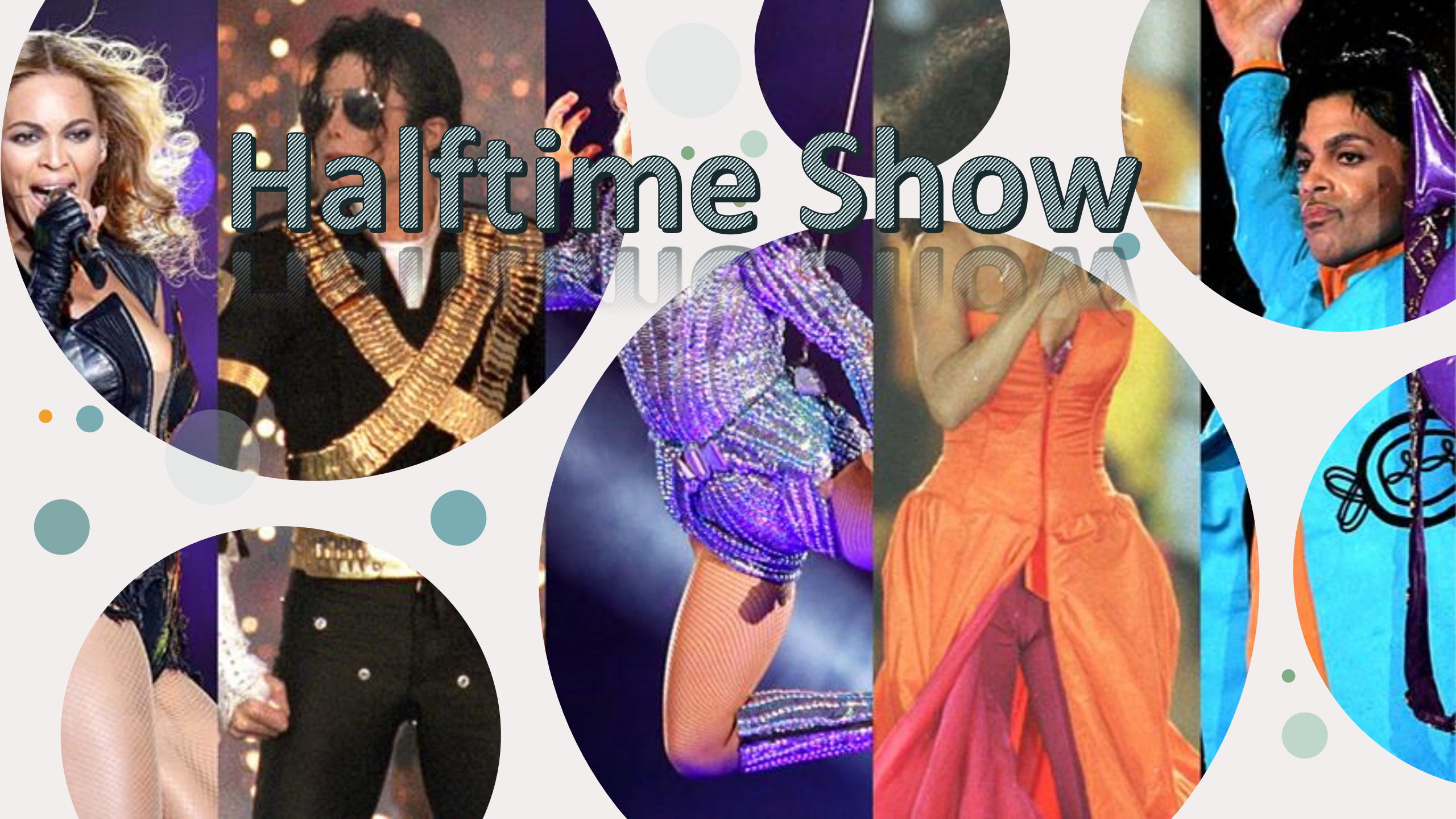
- Valores son inusuales o extremos en comparación con el resto
- Causados por errores en
  - Medición
  - Captura
- También pueden ser legítimos

- Métodos para detectar atípicos
  - Análisis visual
  - Boxplot
  - Análisis de densidad
  - Regresión
  - Medidas de distancia
- Eliminar los registros
- Imputación

Transformación de formatos

- Todo lo que ya vimos de
  - Texto
  - Audio
  - Imagen
- Codificación de variables categóricas
- Agregación de datos
- Agrupamiento
- Separación de datos

Halftime Show



Normalización

- Datos tienen distintos origines e interpretaciones
  - Escalas son distintas
- Transformar los datos a una escala común
- Métodos
  - Min-Max
  - Z-score
  - Logarítmico
- Estas transformaciones afectan la interpretación

Tratamiento de duplicados

- ¿Qué consideramos un duplicado?
  - Registro completo
  - Identificador único
  - Cierta cantidad/selección de valores
- Tratamientos
  - Conservar solo uno
    - Primero
    - Ultimo
    - Mas alto
  - Agregar los datos
  - Conservar todos

Integración de múltiples fuentes

- ¿Qué hacemos cuando tenemos los datos en distintas tablas?
- Unión de tablas
  - Left
  - Right
  - Inner
  - Outer
- Cardinalidad
  - 1-1
  - 1-muchos
  - Muchos-1
- Revisar que las uniones son correctas
  - Duplicados
  - Faltantes

Reducción de dimensionalidad

- Gran cantidad de variables/columnas
  - Información repetida
  - Información no importante
- Utilizar todas las variables
  - Mas no es mejor
  - Modelos mas difíciles de entrenar
  - Ruido
  - Relaciones espurias
  - Menor desempeño
- Selección de variables
  - Conocimiento del dominio
  - Análisis estadístico
- Componentes principales

Creación de características

- Crear nuevas características a partir de las existentes
  - Porcentajes
  - Tasas
  - Máximos/Mínimos


colab

Notas finales

- La aplicación de técnicas de preprocesamiento
- Puede afectar la calidad de los resultados
- Puede desbloquear nuevos análisis
- Es una tarea iterativa
- Puede ser fácilmente la tarea que demande mas tiempo/esfuerzo