

# Data Preparation II

Adrián Matute, Pablo Martínez, Osvaldo Del Valle, Andres Callealta, Jorge  
Martínez

*Tecnológico de Monterrey Campus Querétaro, México*

## 1. Descripción de los datos

Se partió del dataset generado en la primera fase de [DATA PREPARATION](#) el cual consta de 8,087 imágenes JPG con sus respectivas anotaciones de bounding boxes, cajas que engloban a los objetos de la imagen (vacas).

## 2. Selección de Datos

La fase previa de modelado nos mostró un efecto adverso en la precisión del modelo al haber generado un dataset combinado con fotos de día y noche. Esto nos llevó a decidir crear modelos separados para la detección de vacas en el día y en la noche.

Para la realización de los nuevos datasets tomamos las fotos y etiquetas seleccionadas previamente ya que estas ya fueron procesadas previamente para remover las fotos sobreexpuestas y subexpuestas.

## 3. Construcción de los datasets

Se decidió hacer una separación inicial tomando en cuenta la hora de captura, ya que este atributo se conserva en el nombre de la imagen. Esta se realizó asignando las fotos tomadas entre horarios de 5:50 y las 19:35 hrs al dataset diurno, y las imágenes capturadas fuera de este horario al dataset nocturno por medio del siguiente script: [LauCowsDayNightDatasets](#).

Posteriormente se revisaron manualmente los datasets generados para mover todas las fotos asignadas incorrectamente por las variaciones de los ciclos diurnos a través del tiempo. Ya con las fotos separadas se juntaron con sus etiquetas correspondientes, por medio del nombre de archivo, para terminar con 4,681 imágenes para el día y 3,406 de noche.

Finalmente se realizó la separación en conjuntos de entrenamiento y validación para ambos datasets. Resultando en: 3,744 imágenes de entrenamiento y 937 de validación para el dataset de día y 2724 imágenes de entrenamiento y 682 de

validación para el dataset nocturno.

## 4. Formateo de Datos

Los datos no recibieron ninguna modificación en cuanto a los archivos. Las modificaciones fueron únicamente la separación en el dataset diurno y nocturno.