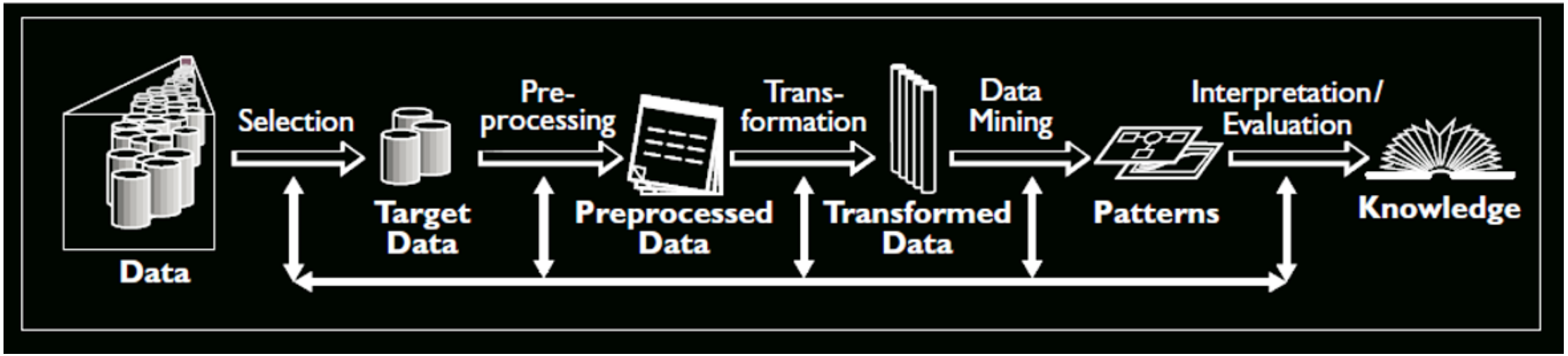
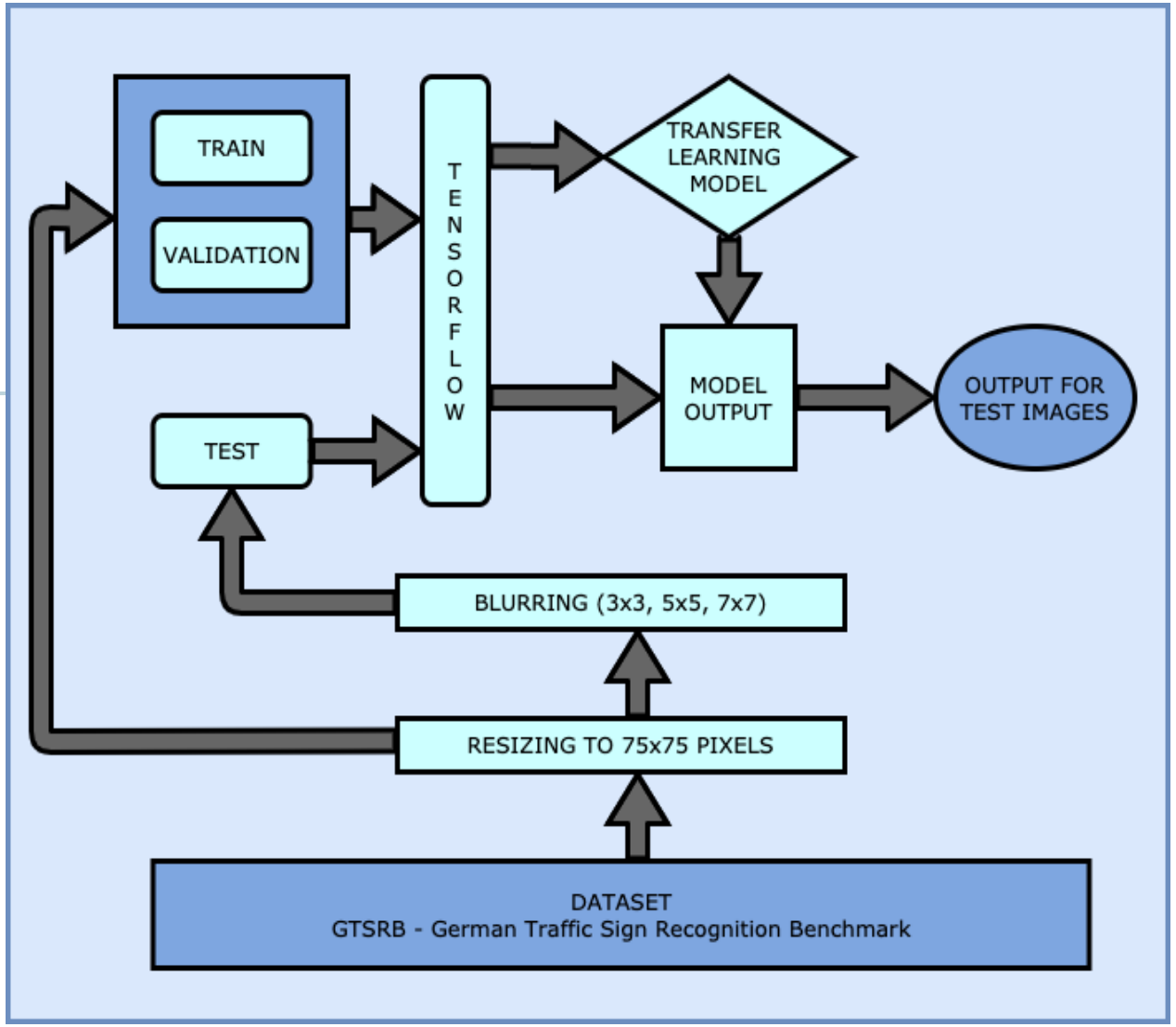
Methodology

Knowledge Discovery in Databases (KDD)



Process Design and Tasks:

Reading	Reading Images and Labels
Resizing	Resizing Images
Blurring	Blurring Images
Saving	Saving the Images in Arrays
Splitting	Splitting Data in Train, Test, and Validation Sets.

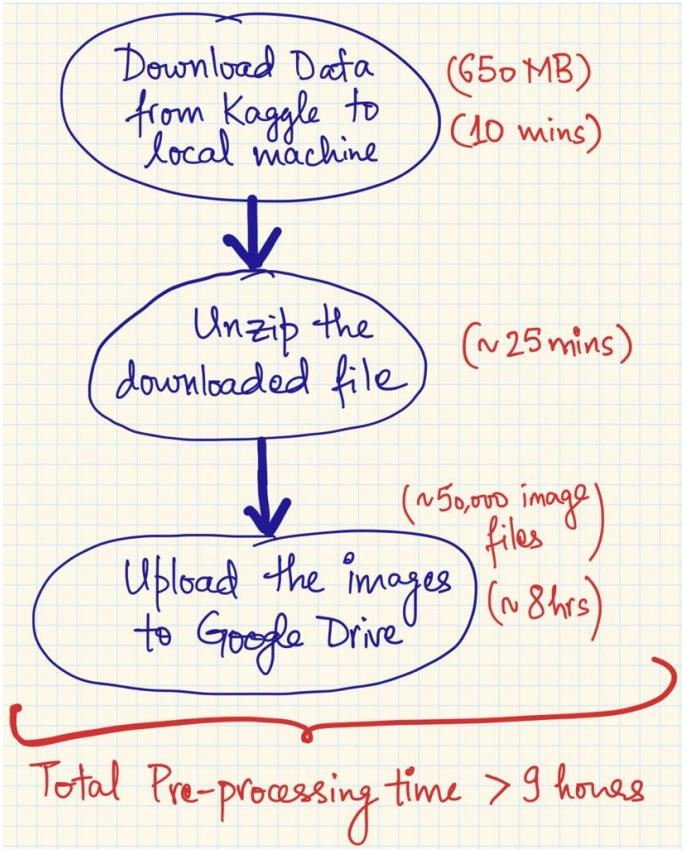


Reading the Images

- Library: CV2
- Three Approaches:
  - Approach 1
  - Approach 2
  - Approach 3
- Required time for pre-processing is improved from more than 9 hours to less than 1 minute.

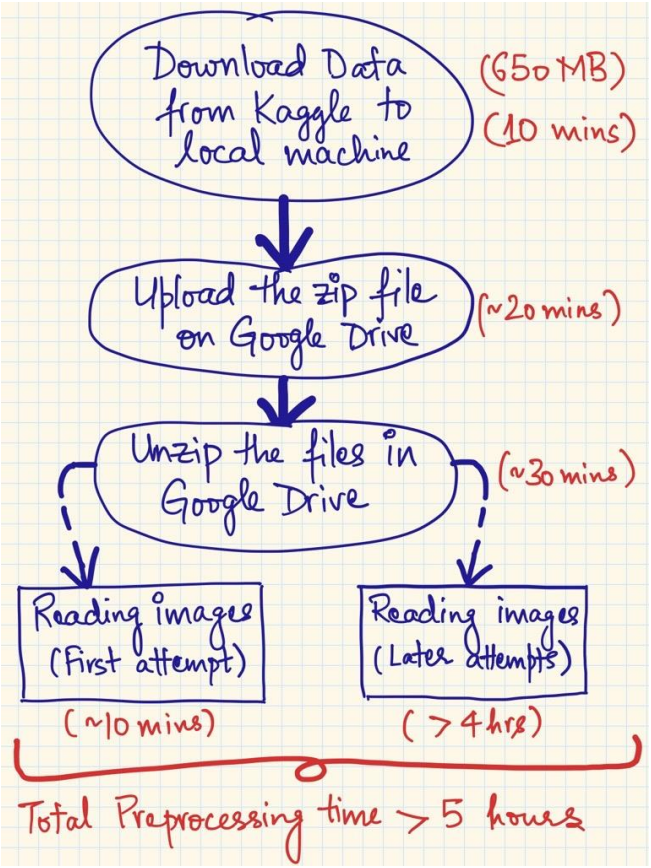
Approach 1 of Reading Images

- Download data from Kaggle to local machine.
  - Data Size: Around 650 MB
  - Time required: Around 10 minutes
- Unzip the downloaded file.
  - Time required: Around 25 minutes.
- Upload the images to Google Drive.
  - Number of files: 51,888
  - Time Estimated: >8 hours
- Total Pre-processing Time: >9 hours



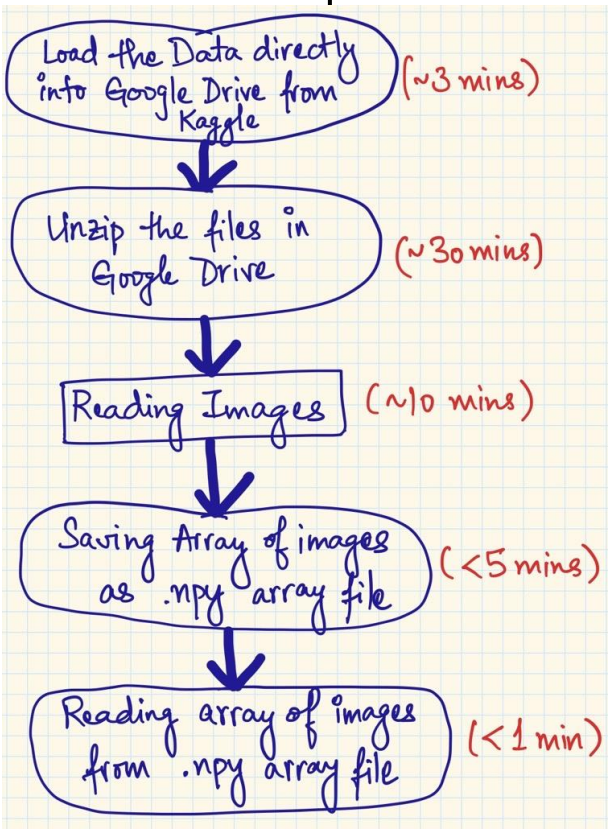
Approach 2 of Reading Images

- Challenge: Reduce the time required for uploading images to Google Drive
- Download data from Kaggle to local machine
  - Data Size: Around 650 MB
  - Time required: Around 10 minutes
- Upload the .zip data file to Google Drive
  - Time required: Around 20 minutes
- Unzip the files in Google Drive
  - Time required: Around 30 minutes
- Reading images (First Attempt):
  - Time required: Around 10 minutes
- Reading images (Second Attempt):
  - Estimated Time: >4 hours
- Total Pre-processing Time (Second Attempt): >5 hours



Approach 3 of Reading Images

- Challenge: Reduce the images reading time on second onward attempts
- Load data to Google Drive from Kaggle
  - Time required: Around 3 minutes
- Unzip the files in Google Drive
  - Time required: Around 30 minutes
- Reading images
  - Time required: Around 10 minutes
- Saving Array of Images as .npy file
  - Time required: <5 mins
- Reading array of Images from .npy file
  - Time required: < 1 min



Summary of Reading Images

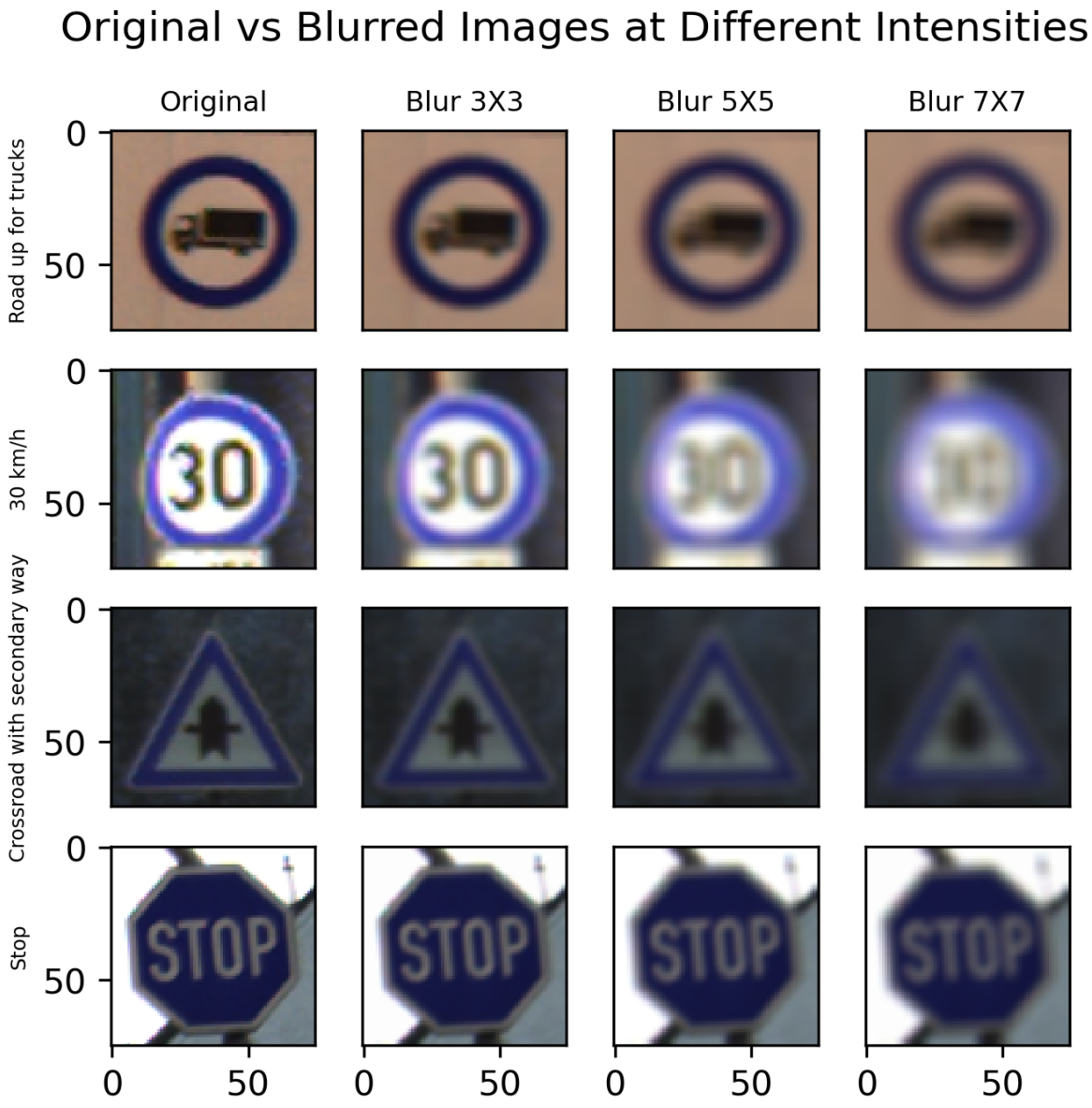
- First Attempt:
  - Load data to Google Drive from Kaggle: Around 3 minutes
  - Unzip the files in Google Drive: Around 30 minutes
  - Reading images: Around 10 minutes
  - Saving Array of Images as .npy file: Around 5 minutes
  - Total time: Around 50 minutes
- Subsequent Attempts:
  - Reading array of Images from .npy file: < 1 min

Resizing the Images

- Library: CV2
- Dataset Images has various sizes such as 28x29, 97x96, 125x136, 168x180, etc.
- Transfer Learning Models require minimum size of 32x32.
- Resized images to 75x75 pixels.

Blurring the Images

- Library: CV2
- Technique: Gaussian Blur
- Levels of Blurring:
  - 3x3
  - 5x5
  - 7x7



Splitting of Data

- Data available in Train and Test Subsets
- Train subset has 75% data
- Test subset has 25% data
- Train subset is further divided in two subsets
  - Train: 80% data
  - Validation: 20% data