

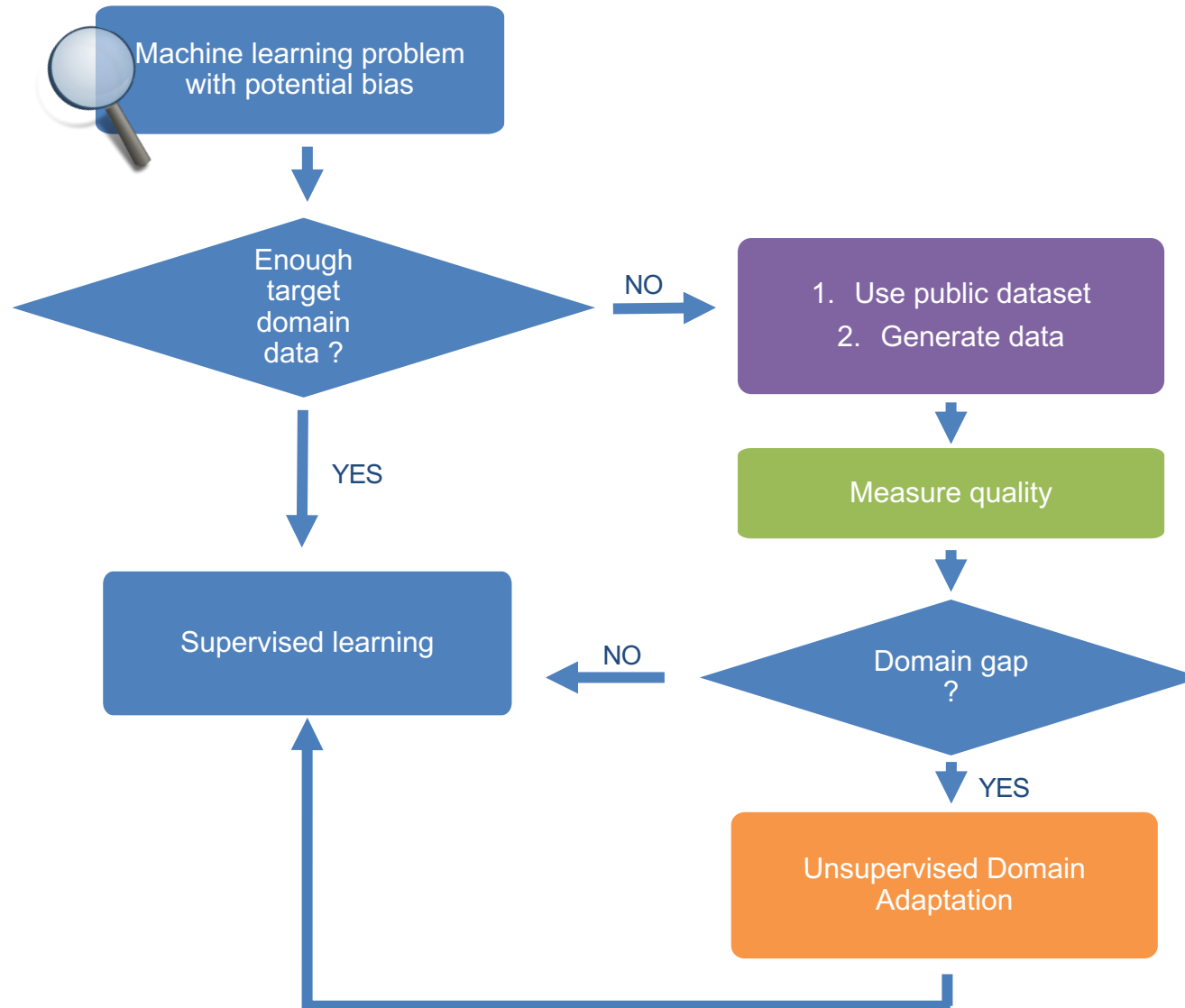
# TOWARDS A METHODOLOGY FOR SYNTHETIC DATA GENERATION, DOMAIN GAP CHARACTERIZATION AND MITIGATION ?



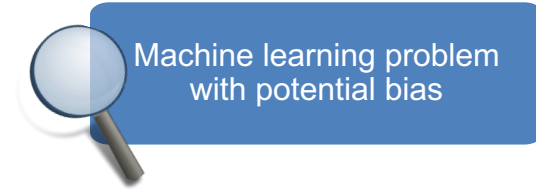
**Amélie BOSCA**

**Data scientist, Sopra Steria**

# General methodological workflow

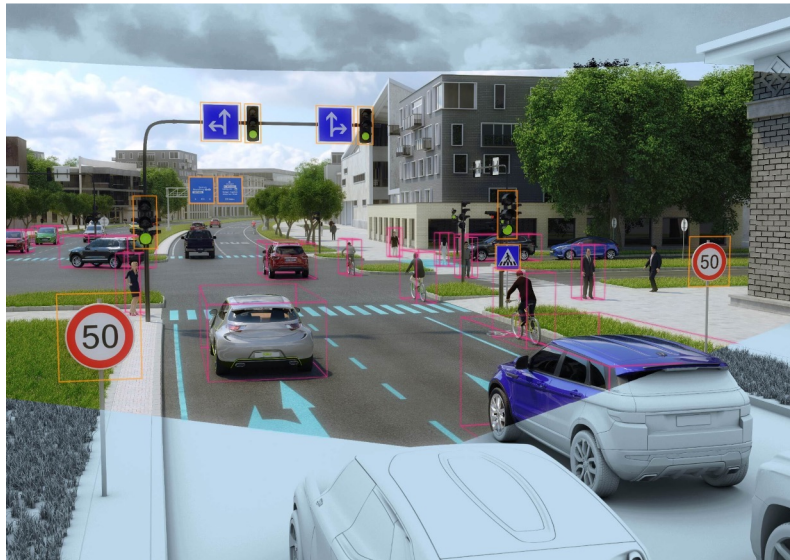


# Application Example



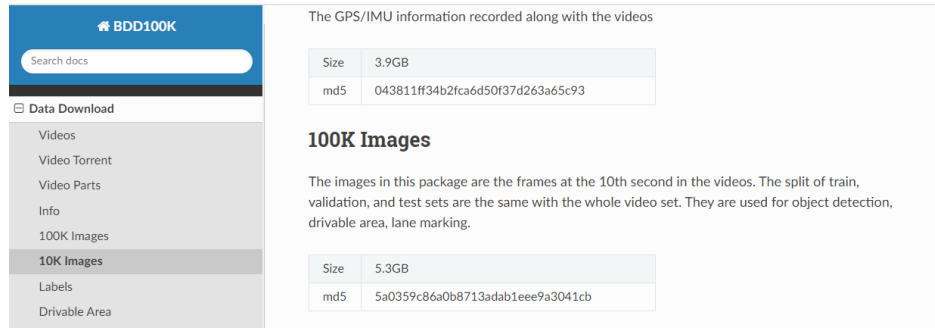
## Context: Autonomous driving

- Goal: recognize pedestrian & vehicles
- Use case : Valéo Scene Understanding
- Test images (ex: VDP): target



# Public Dataset

## BDD100k



The screenshot shows the BDD100k website interface. On the left, there is a sidebar with a search bar and a 'Data Download' section containing links for Videos, Video Torrent, Video Parts, Info, 100K Images, 10K Images, Labels, and Drivable Area. The main content area displays the title 'BDD100K' and a search bar. Below this, there are two data tables. The first table is for '100K Images' and the second is for '10K Images'. Both tables show 'Size' and 'md5' values. The '100K Images' table shows a size of 3.9GB and an md5 of 043811f34b2fca6d50f37d263a65c93. The '10K Images' table shows a size of 5.3GB and an md5 of 5a0359c86a0b8713adab1eee9a3041cb. A paragraph of text explains that the images are frames from videos at the 10th second, used for object detection, drivable area, and lane marking.

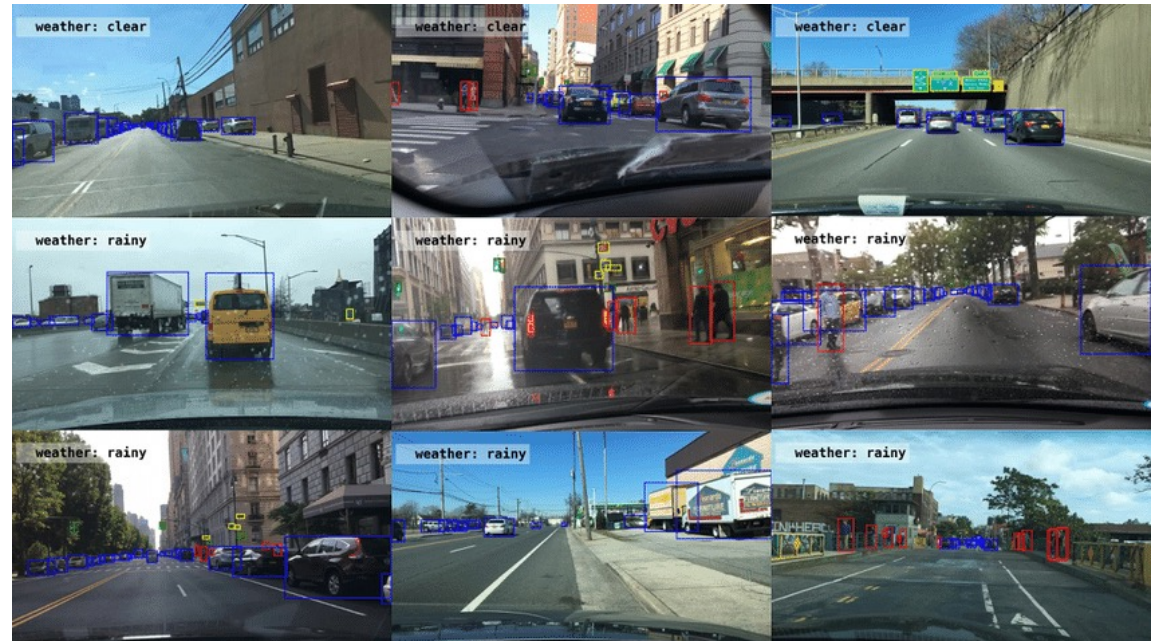
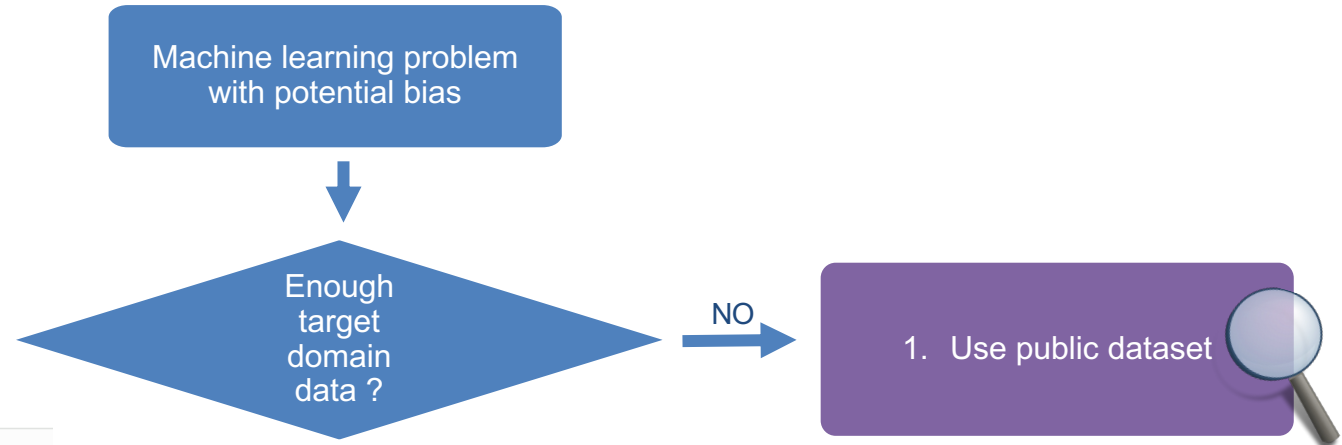
Size	md5
3.9GB	043811f34b2fca6d50f37d263a65c93

**100K Images**

The images in this package are the frames at the 10th second in the videos. The split of train, validation, and test sets are the same with the whole video set. They are used for object detection, drivable area, lane marking.

Size	md5
5.3GB	5a0359c86a0b8713adab1eee9a3041cb

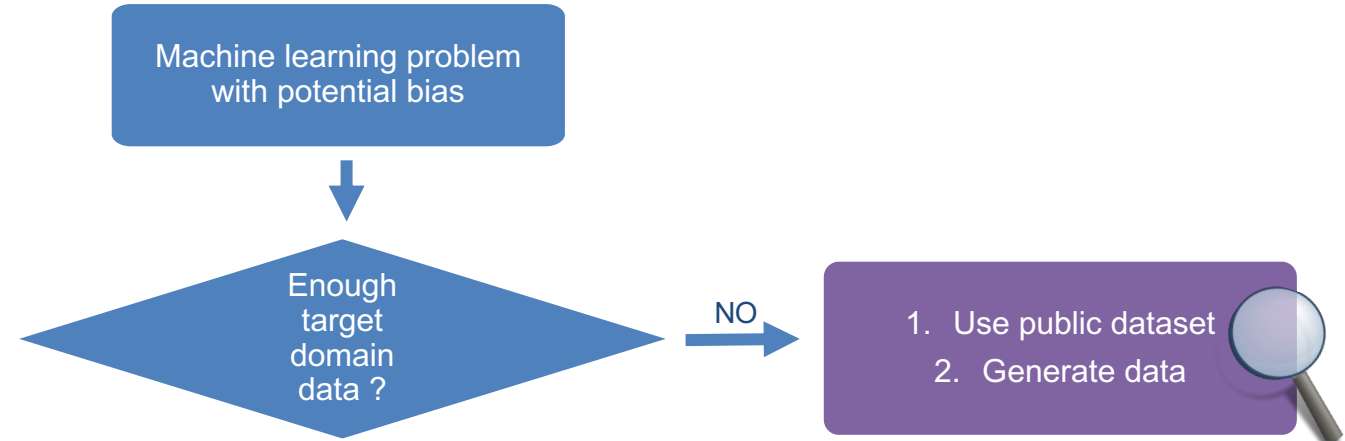
- Well-known dataset from literature
- Available annotations
- Different weather conditions



# Corner cases

## Generating Corner cases:

- Rare conditions
- Rare Scenarios



Snow(CORNER CASE)

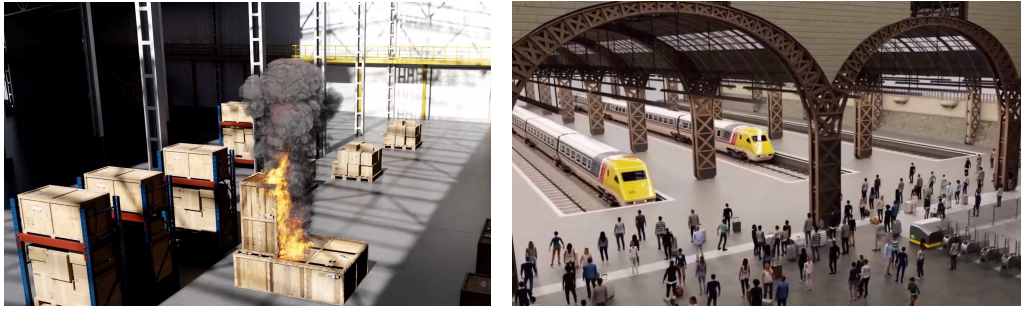


High Illumination (CORNER CASE)



# Generating synthetic and AI-generated images

## Full synthesis (3D-renderers)



 SYNSET

- + Complete generation control
- Recommended for limited contexts
- Often significant domain gap

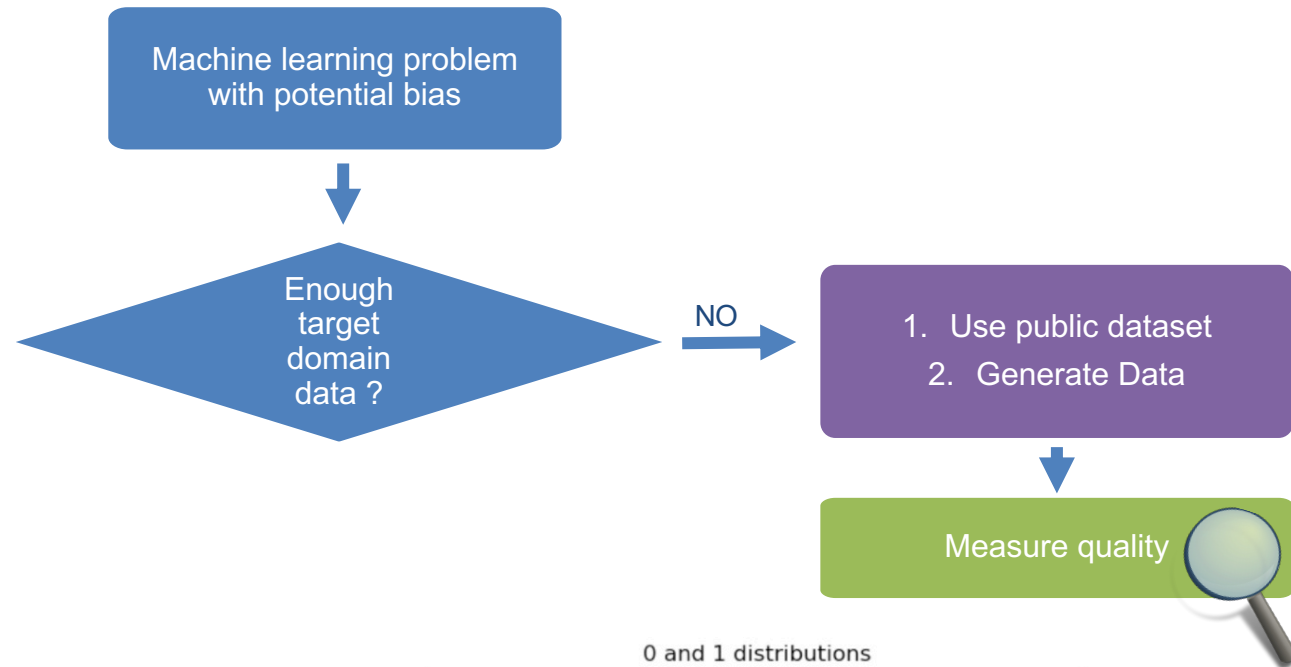
## Data Augmentation (AI)



 jojiGEN

- + Recommended for corner cases
- + Often reduced domain gap
- Significant volume of data for training

# Data quality



## Measurement of domain gap between:

- VDP datasets (target data)
- BDD100k dataset/ generated images (source data)

## Example of metric:

- Wasserstein Distance (WD)

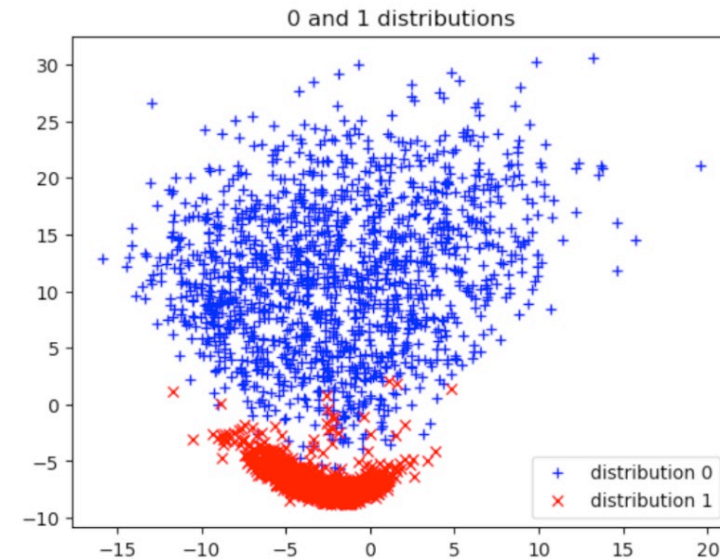
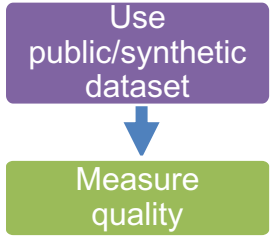


Illustration of two distributions with WD

# Characterizing and Measuring Domain Gap



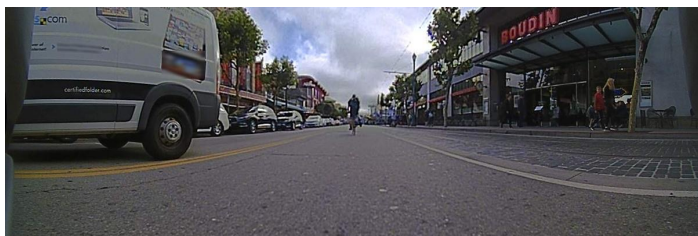
Source data (BDD 100k)



Is there a GAP ?



Target Data (VDP Use Case)



Definition (images): difference in semantic, texture and shapes between two distribution of images.

Two approaches:

- Proxy distance
- Feature extraction on layers

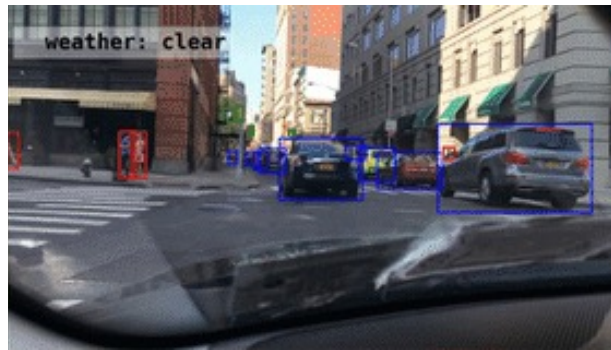
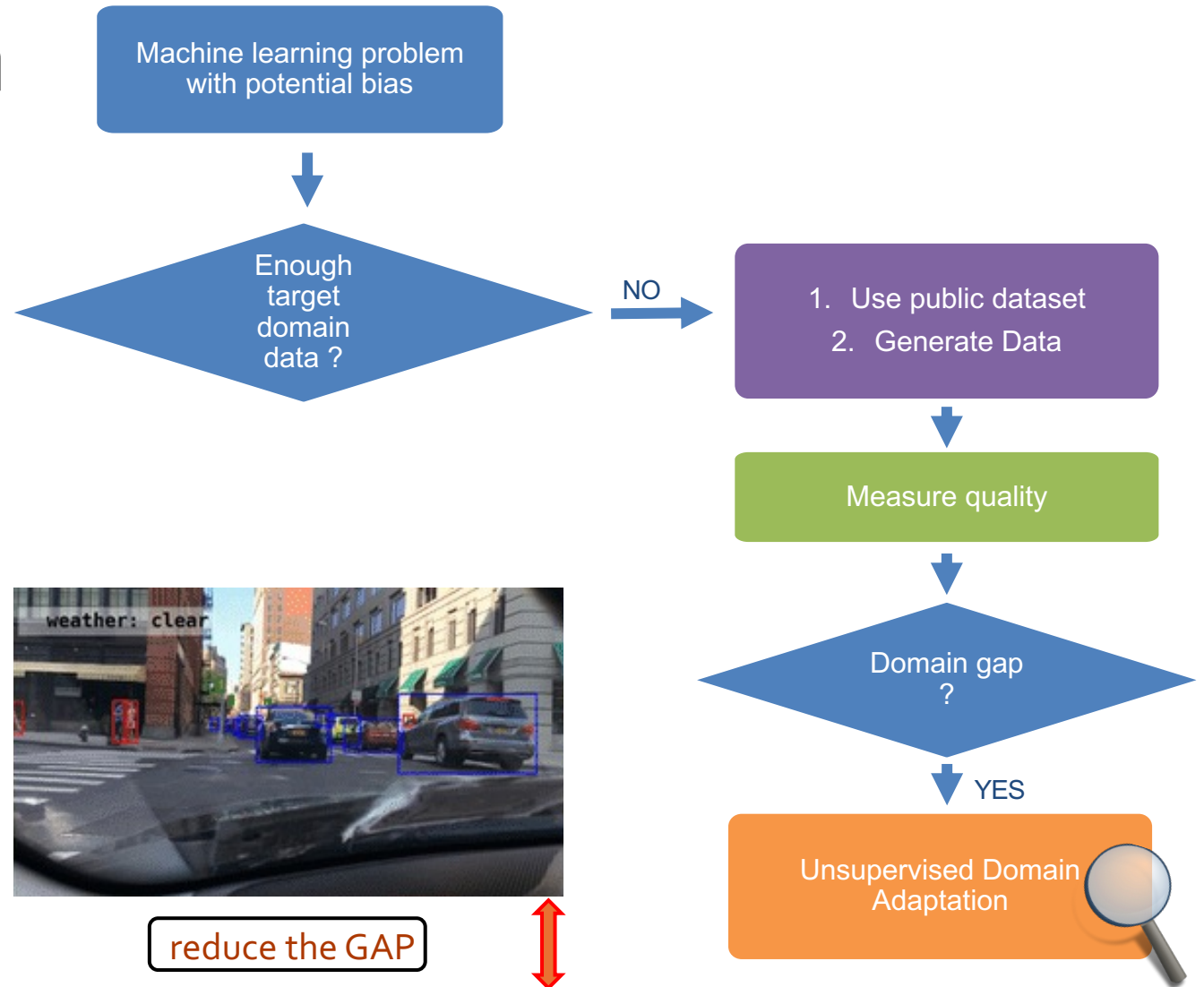
**DQM**

Estimate Domain Gap :

- Maximum Mean Discrepancy (MMD)
- Central Moment Discrepancy (CMD)
- Wasserstein Distance
- H – Divergence



# Domain Adaptation



reduce the GAP



## Domain Adaptation between:

- VDP datasets (target data)
- BDD100k dataset/generated data (source data)

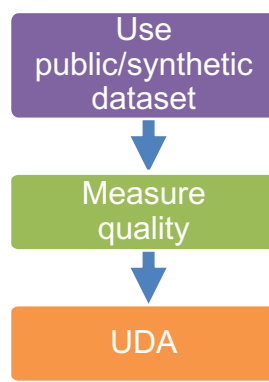
## Example of approach:

- Image Level Adaptation

# Reducing domain gap using UDA

## Unsupervised Domain Adaptation

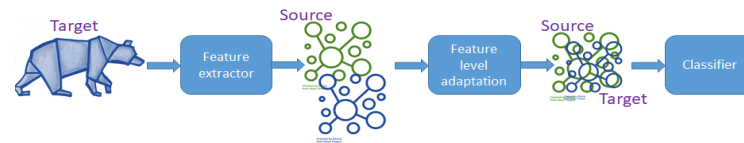
Using a model trained on source dataset on other target dataset using **only unlabeled target data**.



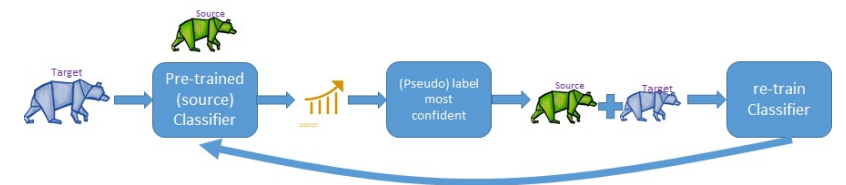
### Image Level Adaptation



### Feature Level Adaptation



### Pseudo-Labeling



- Cheap performance boost on trained models on public and/or synthetic data
- Bootstrap a project before acquiring (more) data

# Takeaway

Domain Gap must be taken into account using data sources like existing datasets or synthetic data.

