Generation of training datasets for ML methods for autonomous vehicles from simulations

Dr. Bill Roungas, ICCS

Bilbao, 24 September 2023

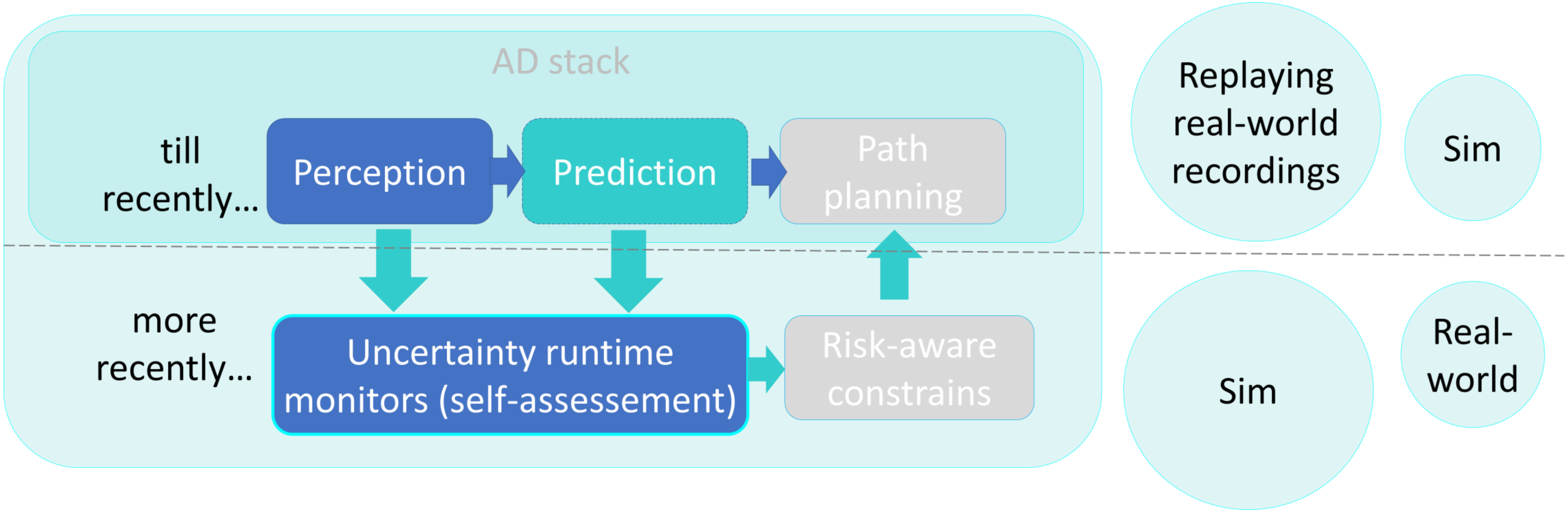
EVENTS

Outline

- Intro
- Synthetic data generation: Theory- vs Data- Driven Problem Solving
- Machine Learning (ML) for Autonomous Vehicles (AVs)
- Augmenting an Existing Image Dataset
- Image Generation Issues
- Simulated Data
  - Data Generation
  - Data Utilization
- Conclusion

Funded by the European Union

Intro: Training the AD stack layers



Funded by the European Union

Synthetic data generation: Theory- vs Data- Driven

EVENTS

- Theory-driven approaches
  - ✓ Utilize existing theory on the subject of interest (image mathematical transformations/filters).
    - ➢ Very often theory is inadequate or completely lacking.
  - ✓ Strive to develop theory, if it doesn't exist.
    - ➢ Developing problem-solving theory takes time.

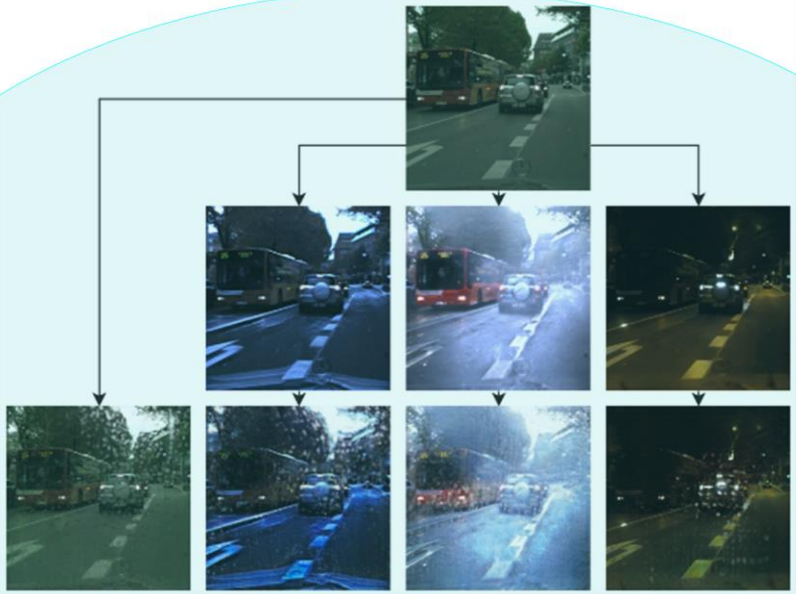
- Data-driven approaches
  - ✓ Minimizes the reliance on existing theory
  - ✓ Focus on building solutions directly from available data.
    - ➢ Large amounts of data can be compiled relatively easy by suitable sensor setups.
- However:
  - ➢ State-of-the art data-driven methods (i.e. Machine Learning) are data hungry.
  - ➢ More often than not, there is a large variety of corner cases which require special care during data collection.
  - ➢ Despite indicating a faster path towards a solution than developing theory, (annotated) data collection remains an intensely time-consuming and tedious process.

Funded by the European Union

Two examples of data-driven dataset generation

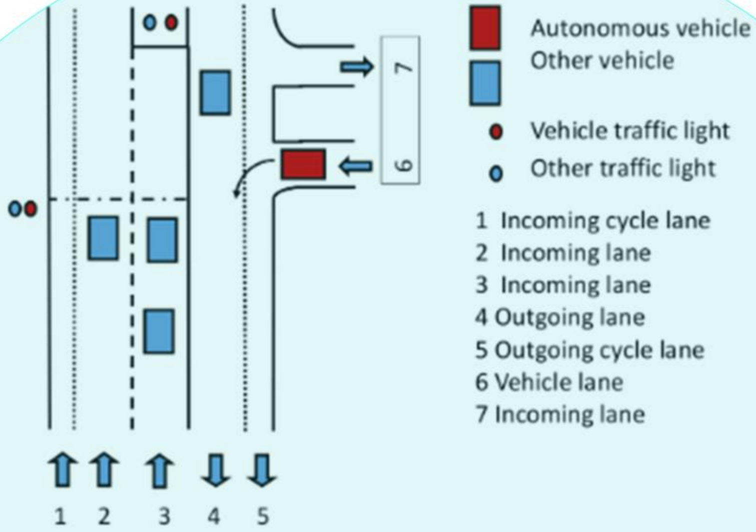
EVENTS

Creating artificial bad weather images from original images using ML



*Source: Multi-weather city: Adverse weather stacking for autonomous driving, ICCV, Oxf Brookes Uni

Annotating events in videos using ML



*Source: ROAD: The ROad event Awareness Dataset forAutonomous Driving, PAMI, 2022

Funded by the European Union

ML for AVs – What kind of data?

EVENTS

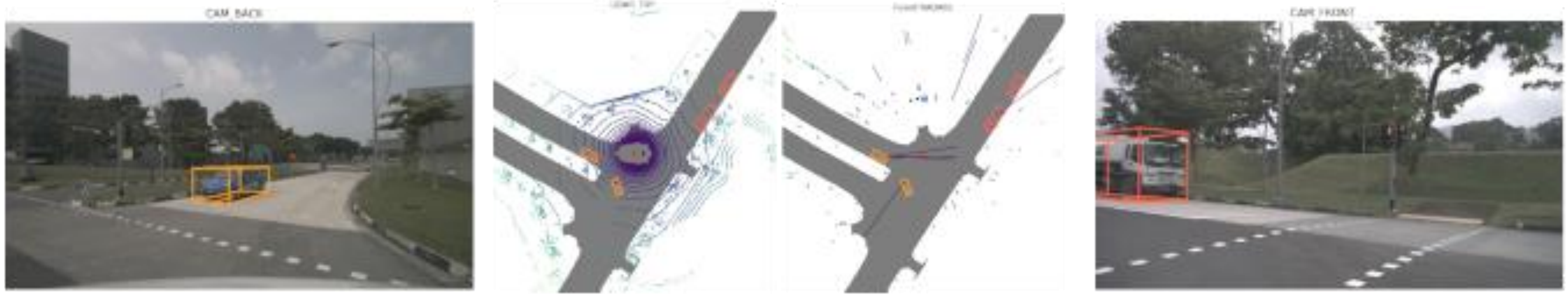
Current research on AVs develops perception and decision mechanisms on a variety of sensor suites. → Different datasets required for each layer of the AD stack!

- Most commonly included sensors for perception:
  - Lidar.
  - Set of radar sensor(s).
  - Set of RGB, stereo and/or RGB-D (depth) camera sensors.
- Most commonly included data for path planning:
  - Set of trajectory points
  - Topology/Map data
  - Traffic rules contextual data

→ Cross-annotating AV perception/motion data even for a simple scenario can be extremely time consuming!

Funded by the European Union

ML for AVs – Required data sample



Source: NuScenes Dataset

Funded by the European Union

Augmenting an existing image dataset

- Simply augmenting an existing dataset is quite standard and can be done via classic Computer Vision tools including
  - ✓ Geometric (perspective/affine/mirror/rotating) transformations.
  - ✓ Blurring plus combinations of morphological filters.
  - ✓ Color transfer between images via suitable of color spaces.
- The above have been shown to be effective in improving performance of object detection algorithms, <u>but only up to a point</u>.
- Questions like *how many data are required*, or *what are the limits of a resulting perception/decision algorithm trained on that data* <u>remain largely intractable.</u>

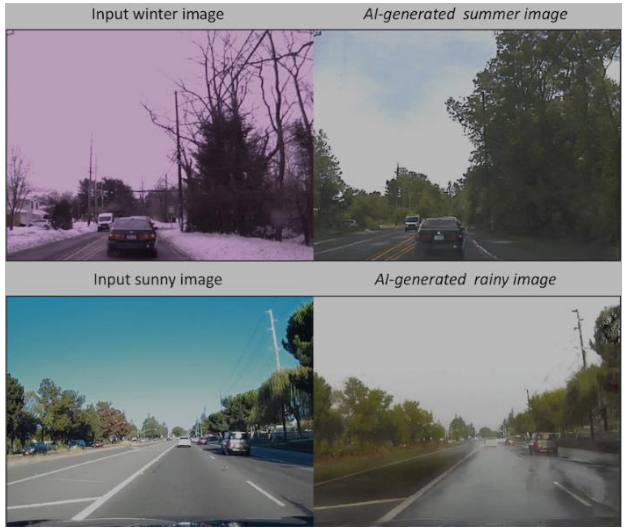
Funded by the European Union

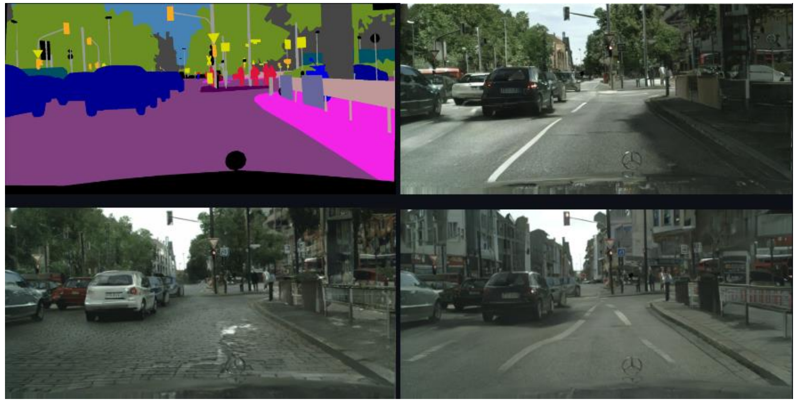
Augmenting an existing image dataset

- Currently under exploration:

- Image translation techniques, particularly for the adverse weather conditions case (e.g. via MUNIT-UNIT *).

- Utilization of segmented images for relevant image generation. Possible pipeline: Image -> Segmented Image -> Image synthesis via px2pixHD **





- * (Multimodal) UNsupervised Image-to-image Translation
- ** This could also exploit the segmentation camera provided in many simulation environments

Funded by the European Union

Image generation issues

- Based on Generative Adversarial Networks.
- Training process can be expected to be unstable, unpredictable and time consuming.
- Required computational resources are highly intensive in terms of time and hardware.
- Resulting images can be of questionable usability in terms of resolution and image quality, especially in VRUs.



Funded by the European Union

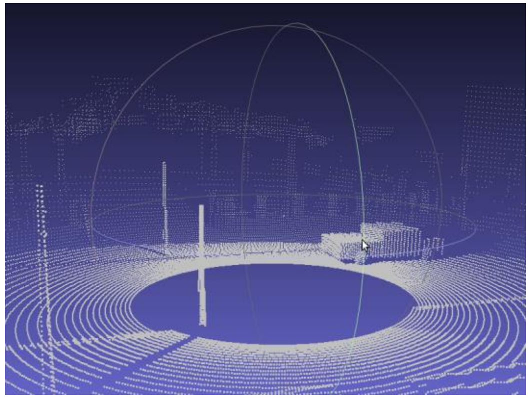
So, what about…

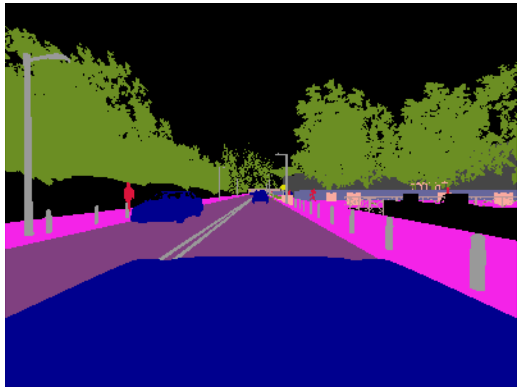
- Lidar data?
- Radar data?
- Is there a way to *rationally* augment – enrich respective datasets?
  1. Maintaining data realism
  2. Preserving the soundness of the annotations
- (Open) questions like:
  - ***Q1***: How much and what kind of data can be considered satisfactory to train an algorithm on a specific scenario or corner case?
  - ***Q2***: How much and what kinds of noise/uncertainty/variability can be filtered out and/or tolerated by an algorithm trained on a specific dataset before it fails?

Funded by the European Union

Simulated data – Generation

- Simulation software offers a fully controllable environment where a large variety of the parameters involved in an experiment can be pre-defined or arbitrarily tuned.
- Besides the ground truth, simulation software offers adjustable models for the entire sensor suite of AVs, including lidar, radar and cameras.



- Collected data are readily annotated by the simulation’s contextual ground truth and simulated scenario.
- Flexibility in scenario building and parameter tuning implies greater ease in considering data collection pertaining to corner cases.

Funded by the European Union

Simulation data – Utilization

- Simulations facilitate the benchmarking of designed algorithms solutions by:

1) Being able to exactly replicate the original experiment and/or scenario
2) Being able to include various sources and levels of uncertainty/variability to the original experiment and/or scenario, ranging from uncertainty in sensor measurements to large deviations from the original scenario.

- Recall (open) questions ***Q1*** and ***Q2.***

Funded by the European Union

Conclusion

- We cannot claim that, in absolute terms, data generated from simulations can replace real-world data

BUT

- They can greatly enhance incomplete real-world data
- Produce data for extreme, high-risk or rare events
- Provide 100% accurate goundtruth data (skipping the need for the cumbersome task of data semantic annotation)

Funded by the European Union

www.events-project.eu

EVENTSproject22

@EVENTSproject22

EVENTS project

EVENTS

Thank you for your attention!

Dr. Bill Roungas, ICCS
vroungas@iccs.gr

Funded by the European Union

This project has received funding under grant agreement No 101069614. It is funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or European Commission. Neither the European Union nor the granting authority can be held responsible for them.