Roundtable Artificial Intelligence, Machine Learning and Regulation

Session 2: Data issues - acquisition, quality, interpretability and biases

Latifa Oukhellou, Université Gustave Eiffel

26-27 January 2023, Paris

International Transport Forum (ITF), OECD

Session 2 - Data Issues

At the beginning of the century

Data issues

Big Data Era

- Increasing number of available data
- Data-driven models instead of conventional approaches (physical or simulation)

Data-driven paradigm

- Learn **generalizable** models from large amounts of datasets of experience (every experiment)
- Scale to **real-world problems** such autonomous transports
- What are the requirements in the data to success in open world settings?

Data acquisition : Quality issues

- Large amount of data collected by sensors in autonomous transports
- Perception of the environment
- In railway domain, **Unprotected mainline network** versus protected environment
- Sensors, one of the **costly components** in autonomous vehicles

⇒ Equipped with cutting edge sensors

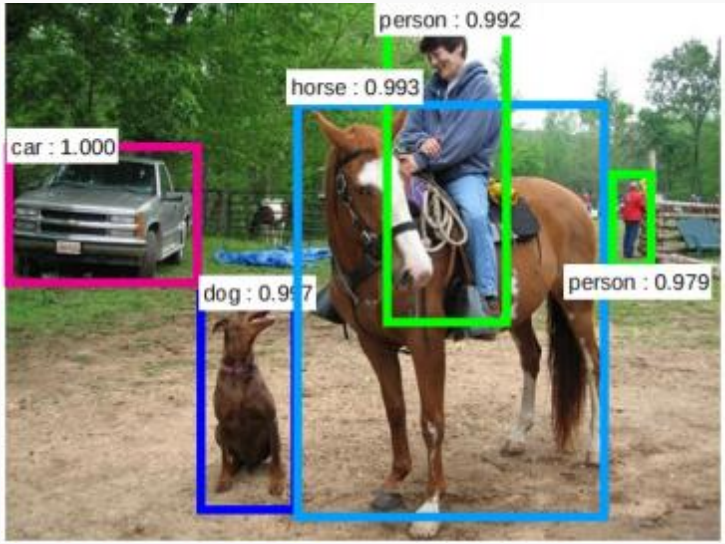
⇒ Advanced pre-processing tools, real-time processing

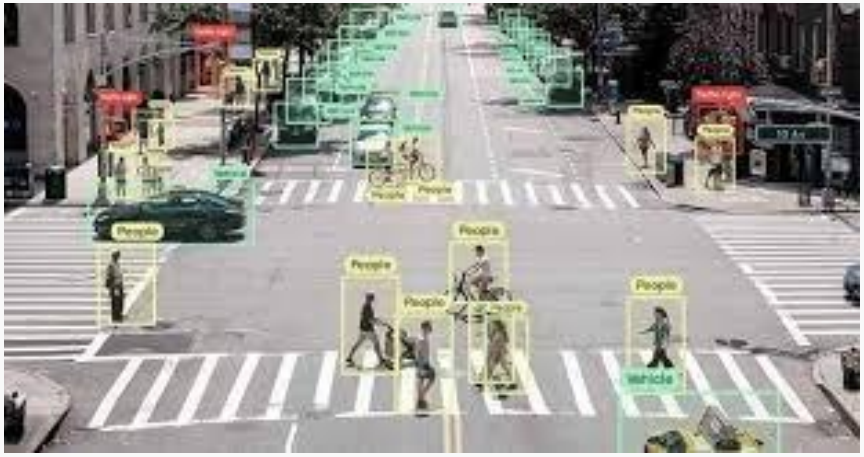
https ://blogs.nvidia.com/blog/2018/08/15/autonomous-trains-deep-learning-dgx-drive/

https ://blogs.ischool.berkeley.edu/w231/files/2021/02/autonomous*V ehicleS ensors* – 768*x*390.*jpg*

Representativeness of Data

- Success of Machine Learning approaches in **natural language processing (NLP)** and **image recognition** in **open world** settings
- Need for large representative datasets to avoid **Overfitting**
- Example : Object detection





⇒ Labeling large datasets, data annotation

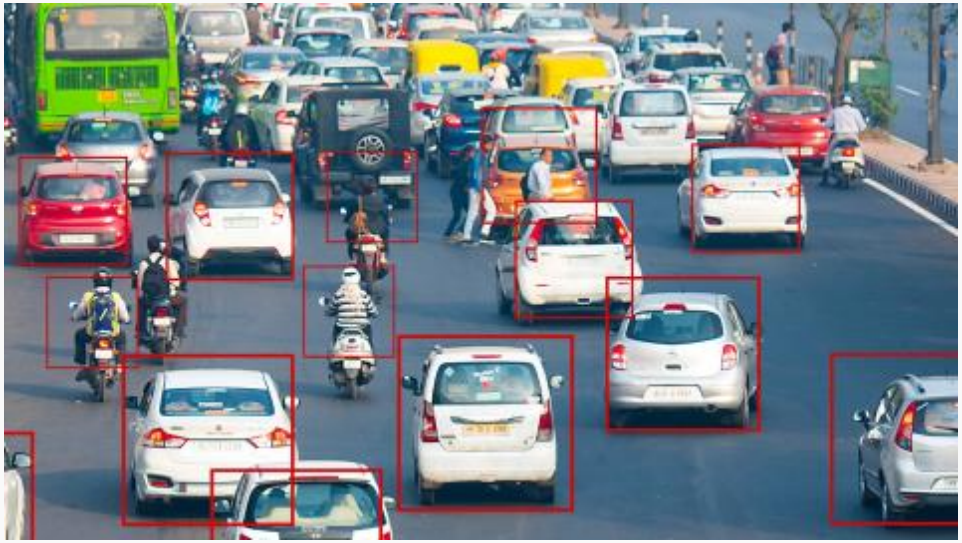
⇒ Urban scenes under different contexts (weather, calendar, spatial ....)

⇒ Representative data, rare events

https ://pyimagesearch.com/2021/08/02/pytorch-object-detection-with-pre-trained-networks/
https ://imerit.net/data-annotation/

Dataset construction

- **Operational design domain (ODD)** : specification of the domain automated driving system is designed to operate
- Iterative Dataset construction
- Adapt the dataset to the application domain



https ://www.numerama.com/tech/321536-ce-que-conduire-aux-etats-unis-nous-dit-de-la-voiture-autonome.html
https ://www.lexpress.fr/monde/pourquoi-les-routes-indiennes-sont-les-plus-mortelles898917.*html*

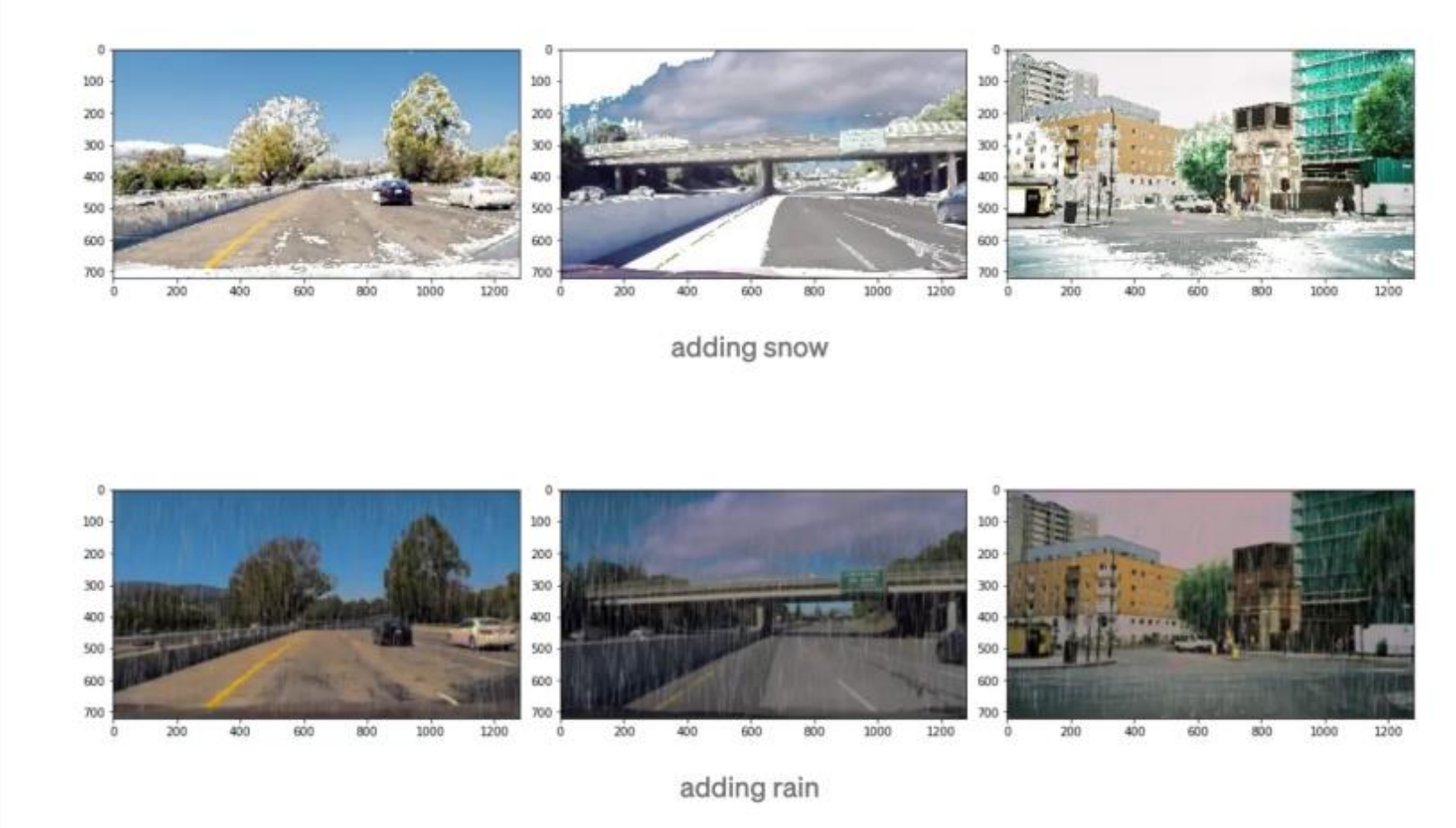
Data augmentation

- Original images

https ://towardsdatascience.com/automold-specialized-augmentation-library-for-autonomous-vehicles-1d085ed1f578

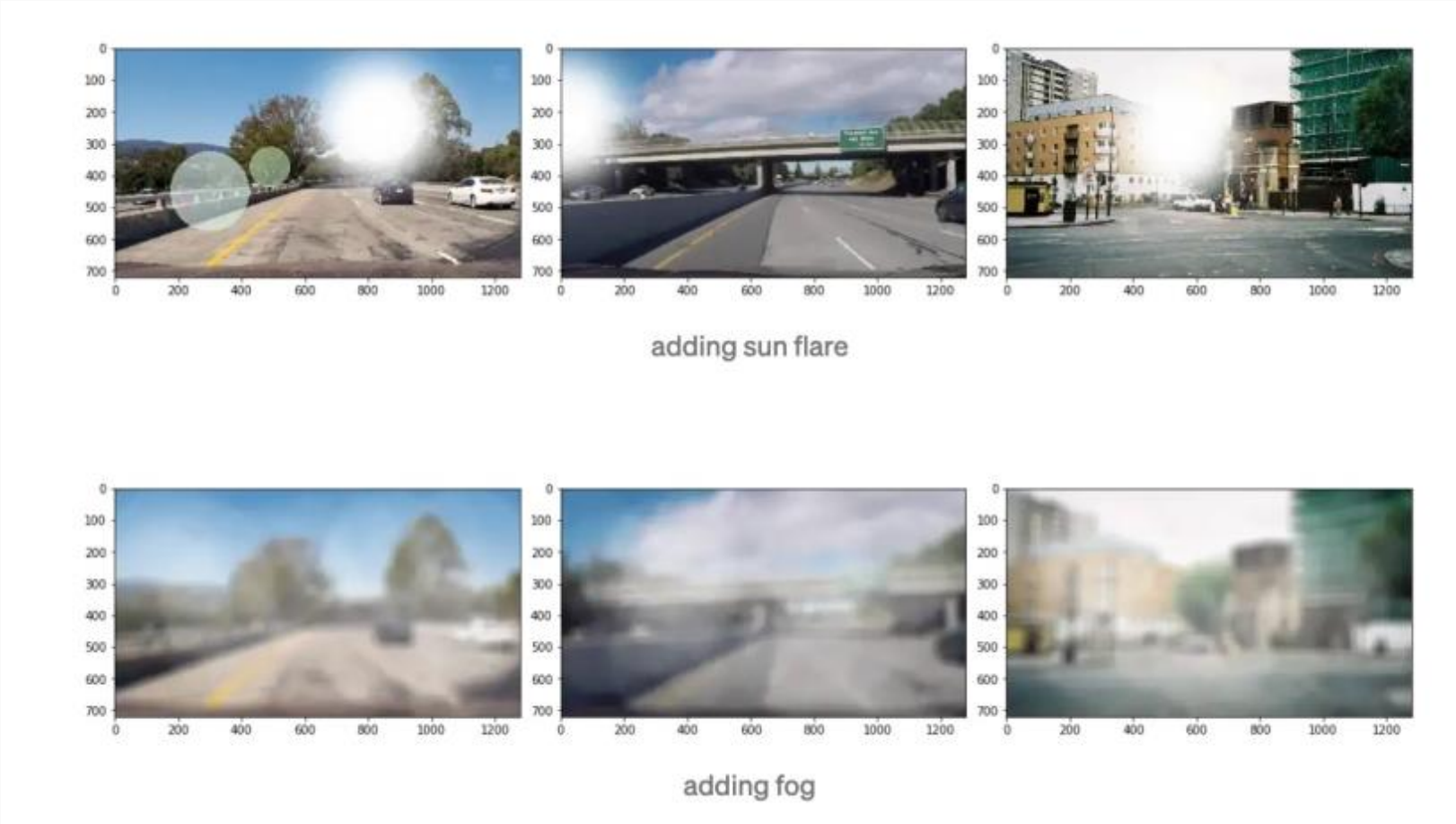
Data augmentation

- Adding Snow or rain to images



Data augmentation

- Adding sun flare or frog



Rountable session 2

Issues

- How to increase the quality and availability of development, verification and validation datasets?
- Which incentives and mechanisms help to share data freely?
- Can synthetic data be used as a substitute?
- How to avoid or address risks from using data?