Computer vision algorithms for detecting secondary tasks in naturalistic driving studies

F. Dellinger [1*], M. Robert-Seidowsky [1], E. Bernard [1],
L. Guyonvarch [2], A.Guillaume [2]

[1] WASSA, 5 rue de l'église, Boulogne, France
[2] LAB Renault PSA, 132 rue des Suisses, Nanterre, France
flora.dellinger@wassa.fr

International Conference on Driver Distraction and Inattention
03/22/2017

Motivations

- **Driver distraction** (in particular use of **mobile phones**): main concern for road safety studies.
- Importance of **NDS (Naturalistic Driving Studies)**: high amount of data but need to be annotated.
  - Manual annotation: expensive and **time-consuming**.
  - Alternative: **automatic detection based on computer vision** methods.
    - Helpful for manual annotators.

Data and challenges

Cockpit camera

Driver camera

Top view camera

Pedal camera

+

CAN files
(speed, wheel rotation, seat belts…)

Data and challenges

Passenger?

Cockpit camera

Driver camera

Top view camera

Pedal camera

+

CAN files
(speed, wheel rotation, seat belts…)

Data and challenges

Cockpit camera

Driver camera

Passenger?

Hands on wheel?

Top view camera

Pedal camera

+

CAN files
(speed, wheel rotation, seat belts…)

Data and challenges

Passenger?

Hands on wheel?

Texting?

+

CAN files
(speed, wheel rotation, seat belts…)

Data and challenges

Cockpit camera

Driver camera

Passenger?

Phone-to-the-ear use?

Hands on wheel?

Texting?

Top view camera

Pedal camera

+

CAN files
(speed, wheel rotation, seat belts…)

Data and challenges

Cockpit camera

Driver camera

Passenger?

Phone-to-the-ear use?

Hands on wheel?

Texting?

+

CAN files
(speed, wheel rotation, seat belts…)

Foot on gas/ clutch/brake?

Top view camera

Pedal camera

Data and challenges

Cockpit camera

Driver camera

Passenger?

Phone-to-the-ear use?

Hands on wheel?

Texting?

\+

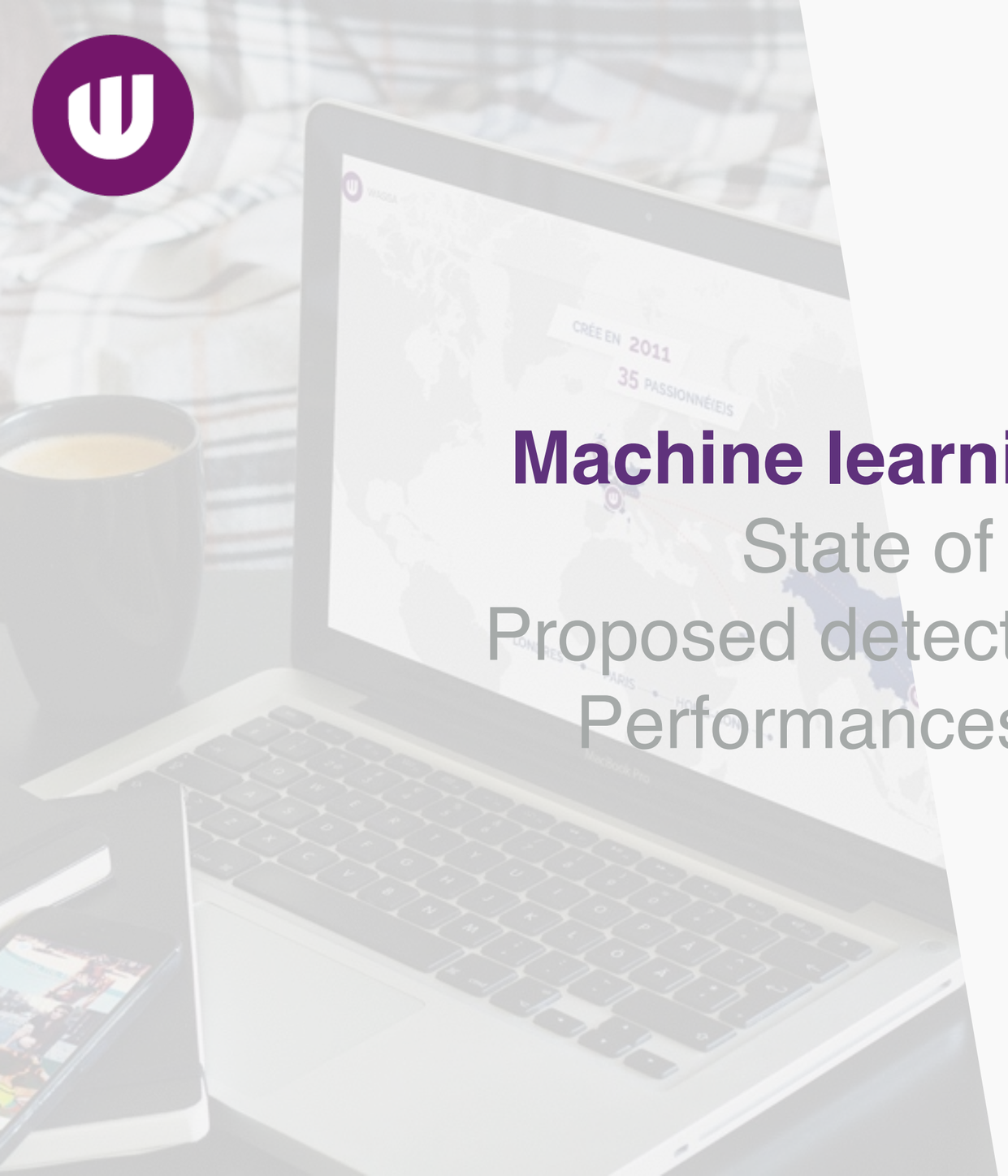
CAN files (speed, wheel rotation, seat belts…)

Foot on gas/ clutch/brake?

Top view camera

Pedal camera

Difficulties: - Low resolution data, gray level, strong illuminations…
- High amount of data (computing time).



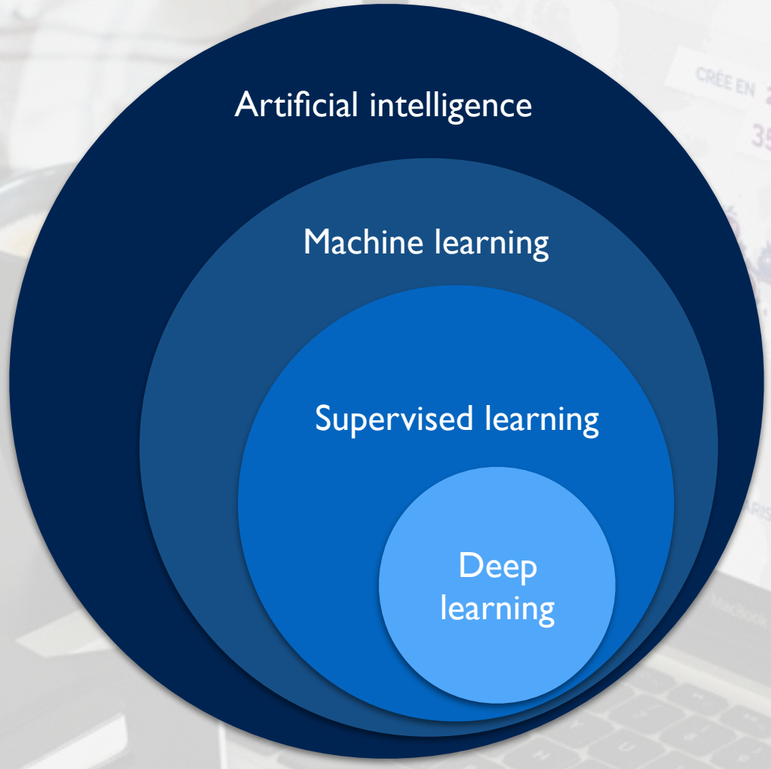
Machine learning concepts

State of the art

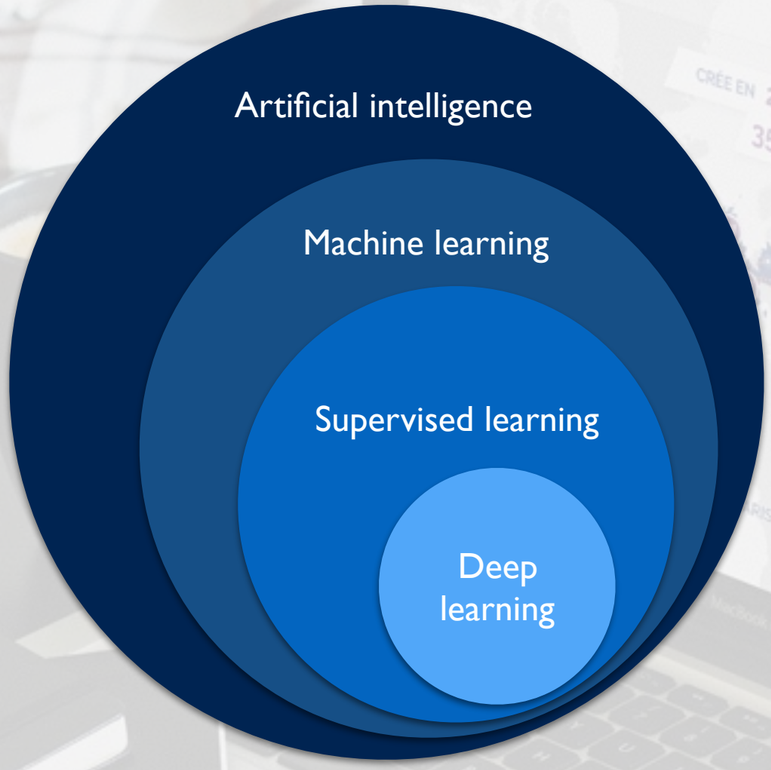
Proposed detection algorithms

Performances evaluation

Machine learning concepts



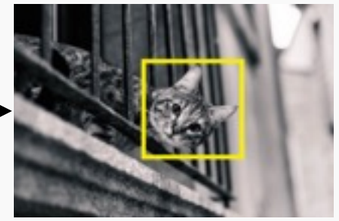
Machine learning concepts



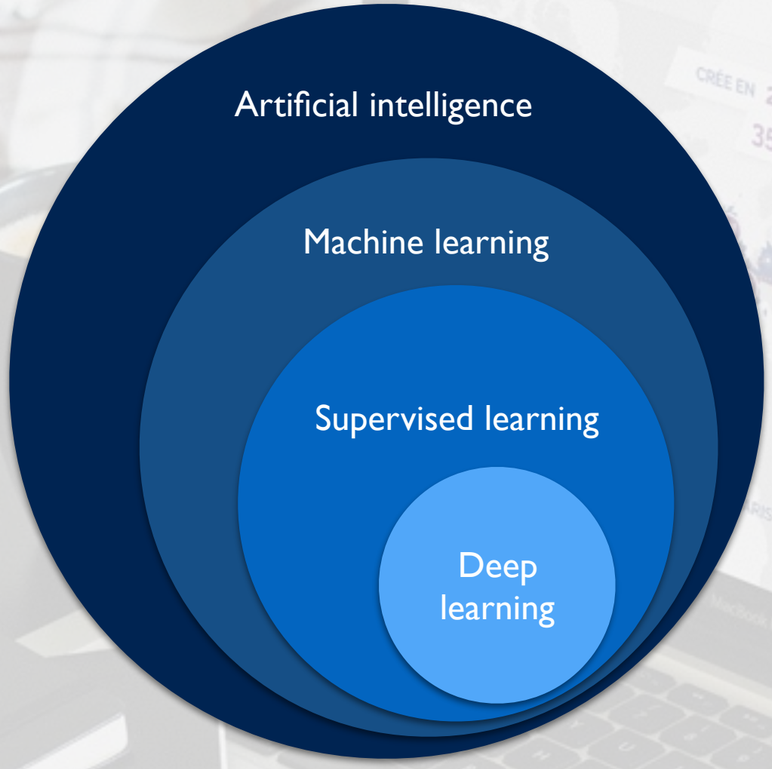
Traditional Machine learning

Feature extraction

Classifier

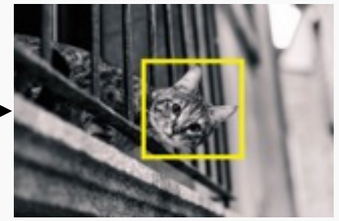


Machine learning concepts



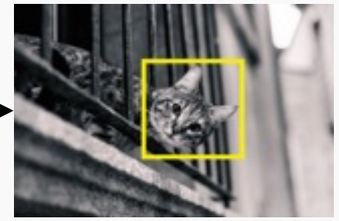
Traditional Machine learning

Feature extraction → Classifier →

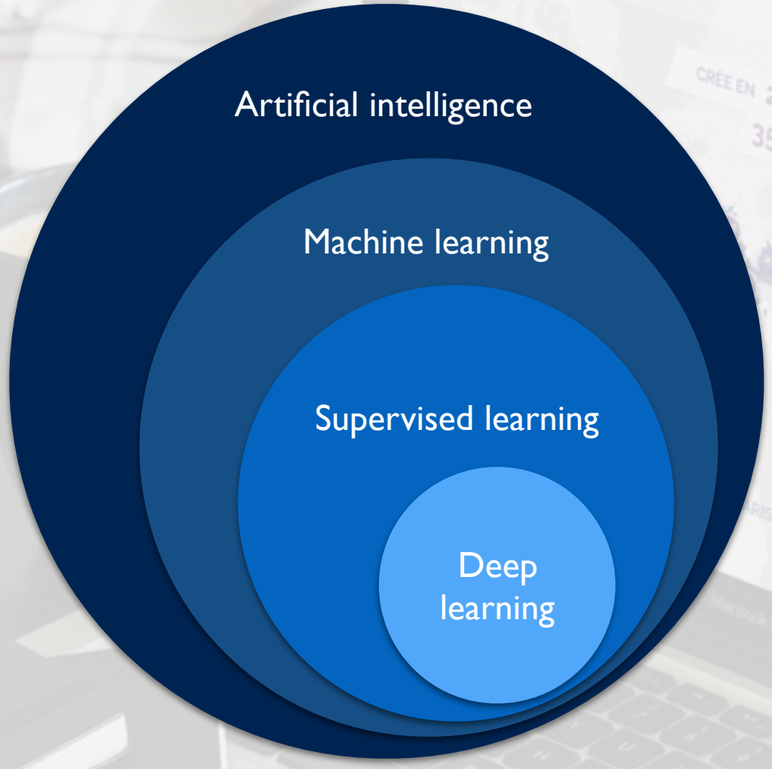


Deep learning

Trains detector model →



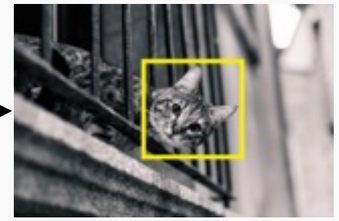
Machine learning concepts



Traditional Machine learning

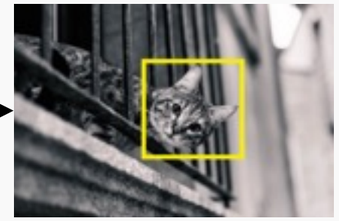
- Fast to train and test
- Light architecture

Feature extraction → Classifier →

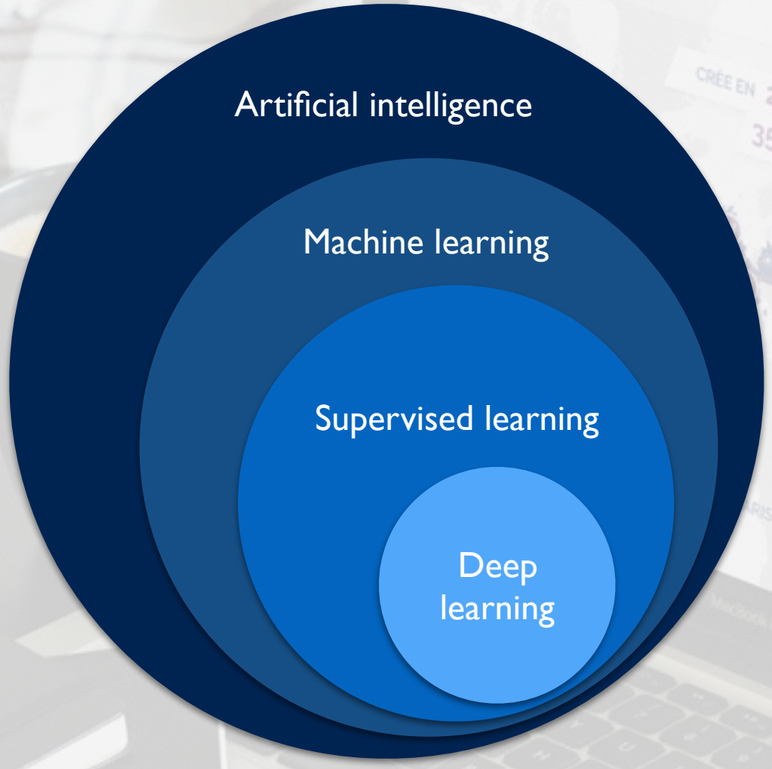


Deep learning

Trains detector model →



Machine learning concepts

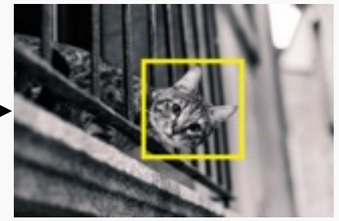


Traditional Machine learning

- Fast to train and test
- Light architecture

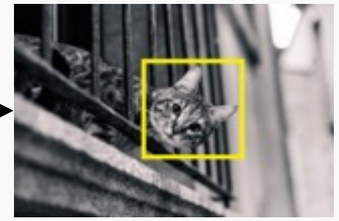
- Performances limited for complex objects
- Difficulty to find distinctive features
- Tuning of parameters

Feature extraction → Classifier →

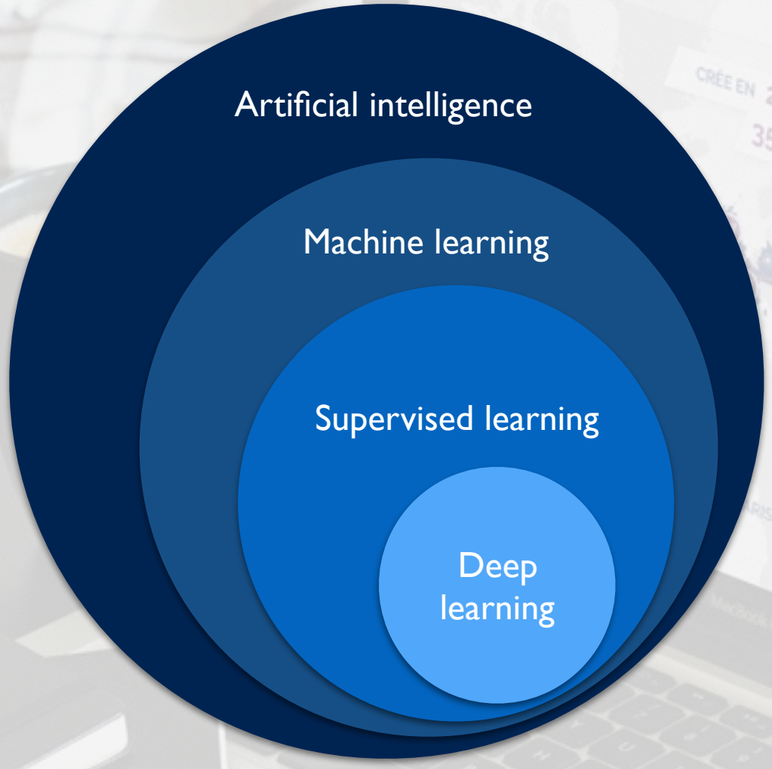


Deep learning

Trains detector model →



Machine learning concepts

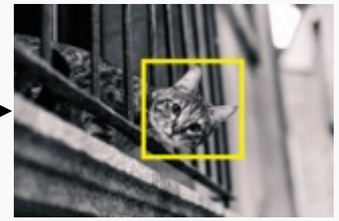


Traditional Machine learning

- Fast to train and test
- Light architecture

- Performances limited for complex objects
- Difficulty to find distinctive features
- Tuning of parameters

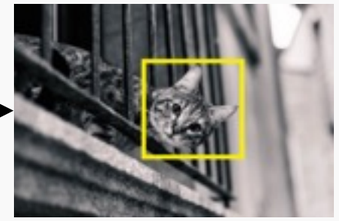
Feature extraction → Classifier →



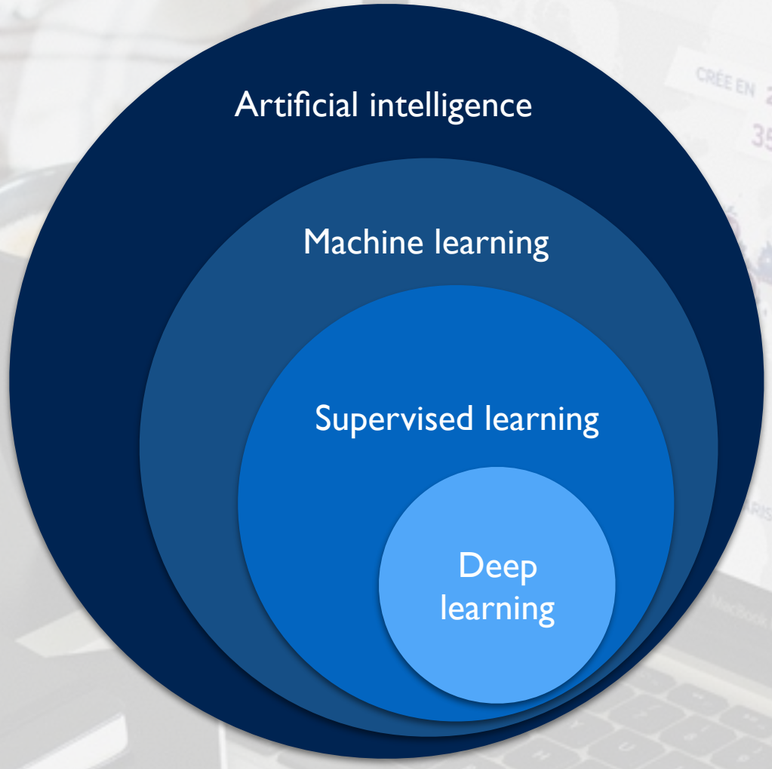
Deep learning

- Can model complex objects
- Computer selects itself distinctive features

Trains detector model →



Machine learning concepts

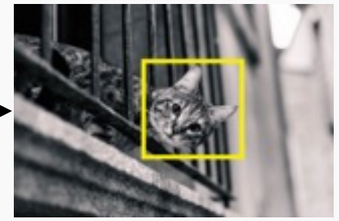


Traditional Machine learning

- Fast to train and test
- Light architecture

- Performances limited for complex objects
- Difficulty to find distinctive features
- Tuning of parameters

Feature extraction → Classifier →

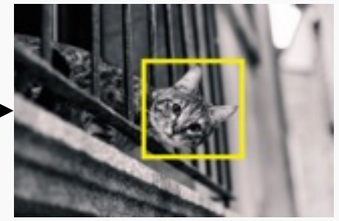


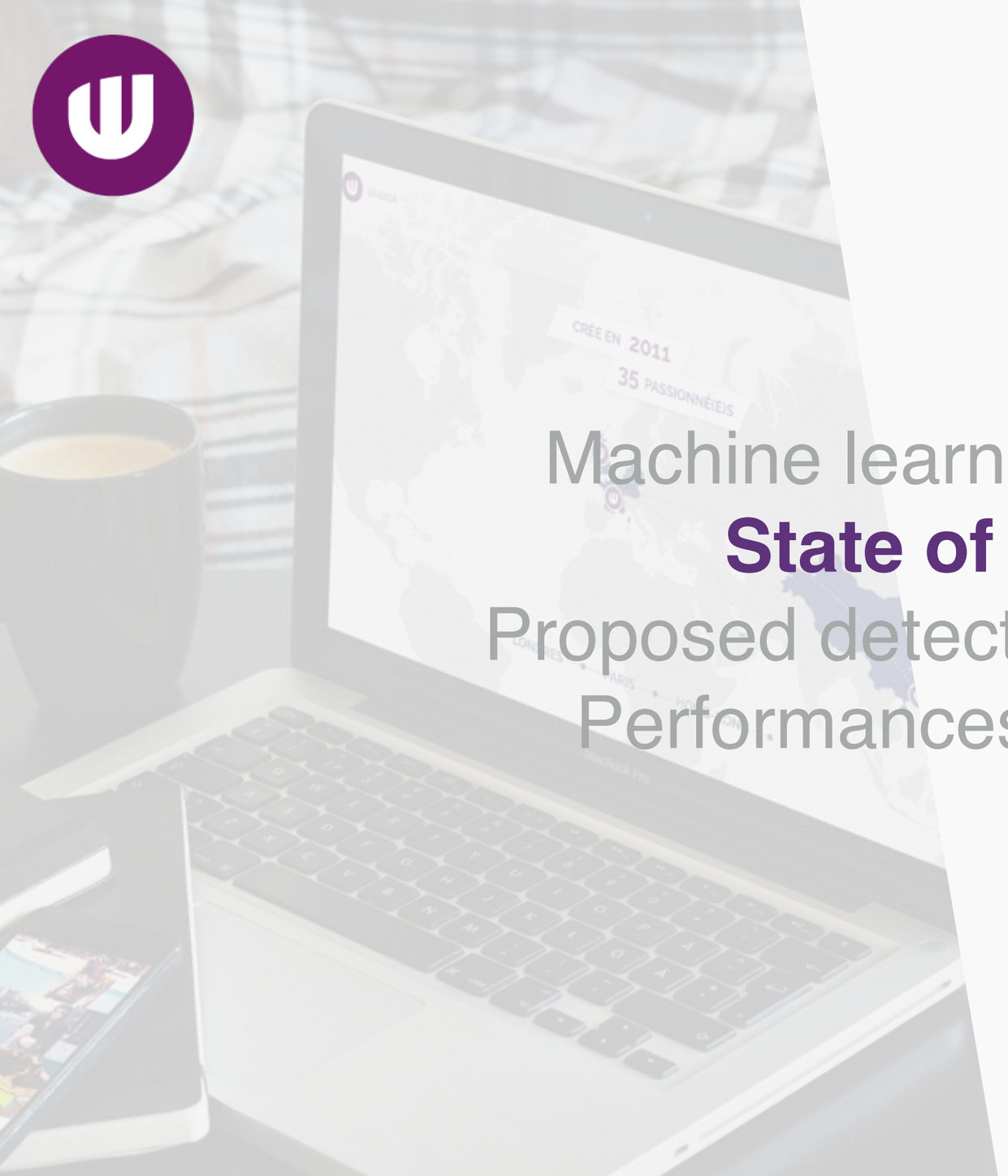
Deep learning

- Can model complex objects
- Computer selects itself distinctive features

- Strong requirements: high amount of training data, GPU
- Long training and computing

Trains detector model →





Machine learning concepts

State of the art

Proposed detection algorithms

Performances evaluation

State of the art

- **Traditional machine learning techniques**
  - **Hand detection**
    - Aggregated Channel Features (based on color and shape informations) [Das et al., 2015] [Rangesh et al., 2016]
    - Histogram of Oriented Gradient (HoG) + Support Vector Machine (SVM) [Ohn-Bar, 2014]
  - **Phone-to-the-ear detection**
    - Detection of ear area based on face detection and landmarks, followed by HoG + SVM [Seshadriv et al., 2016]
- **Deep learning techniques**
  - **Hand and phone-to-the-ear detections**
    - Detect face, hand, cell-phone and steering wheel based on approach called Multiple Scale Faster-RCNN [Hoang Ngan Le et al., 2016]

Machine learning concepts

State of the art

Proposed detection algorithms

Performances evaluation

Proposed detection algorithms: Framework

Proposed detection algorithms: Framework

Proposed detection algorithms: Framework

Small objects, with moderate complexity and fixed localization

Proposed detection algorithms: Framework

Small objects, with moderate complexity and fixed localization

Traditional machine learning

Proposed detection algorithms: Framework

Small objects, with moderate complexity and fixed localization

Traditional machine learning

Objects with strong complexity and variability, fast and unpredictable movements

Proposed detection algorithms: Framework

Small objects, with moderate complexity and fixed localization

Traditional machine learning

Objects with strong complexity and variability, fast and unpredictable movements

Deep learning

Proposed detection algorithms: Traditional Machine learning

Crop an image region

Proposed detection algorithms: Traditional Machine learning

Crop an image region → Compute feature image descriptor HoG (Histogram of Oriented Gradient)

Proposed detection algorithms: Traditional Machine learning

Crop an image region

Compute feature image descriptor HoG
(Histogram of Oriented Gradient)

SVM model pre-trained with labeled videos

Apply a classifier called SVM
(Support Vector Machine)

3 SVM models pre-trained with CAN files

Proposed detection algorithms: Traditional Machine learning

Crop an image region

Compute feature image descriptor HoG
(Histogram of Oriented Gradient)

SVM model pre-trained with labeled videos

3 SVM models pre-trained with CAN files

Apply a classifier called SVM
(Support Vector Machine)

Decision

Passenger presence

Yes

No

Foot on pedal

Yes

No

Proposed detection algorithms: Deep learning

Training step

Proposed detection algorithms: Deep learning

Training step

Standard Architecture

Proposed detection algorithms: Deep learning

Training step

Standard Architecture

Labeled samples

Proposed detection algorithms: Deep learning

Training step

Standard Architecture

Labeled samples

Training step

Proposed detection algorithms: Deep learning

Training step

Standard Architecture

Labeled samples

Training step

Object detector

Proposed detection algorithms: Deep learning

Training step

Standard Architecture

Labeled samples

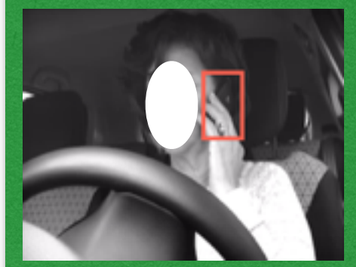
Training step

Object detector

Phone-to-the-ear

Texting

Hands on wheel



Proposed detection algorithms: Deep learning

Testing step

Proposed detection algorithms: Deep learning

Testing step

Object detector

Proposed detection algorithms: Deep learning

Testing step

Object detector

Input image

Proposed detection algorithms: Deep learning

Testing step

Object detector

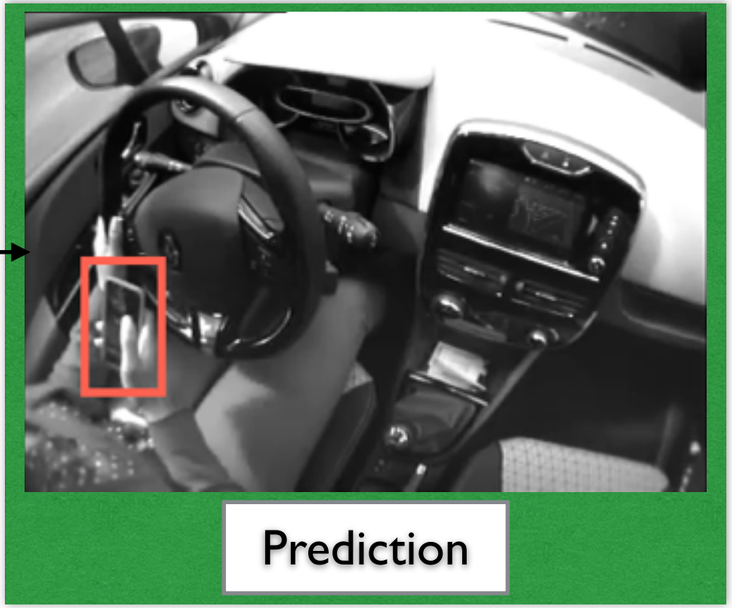
Testing step

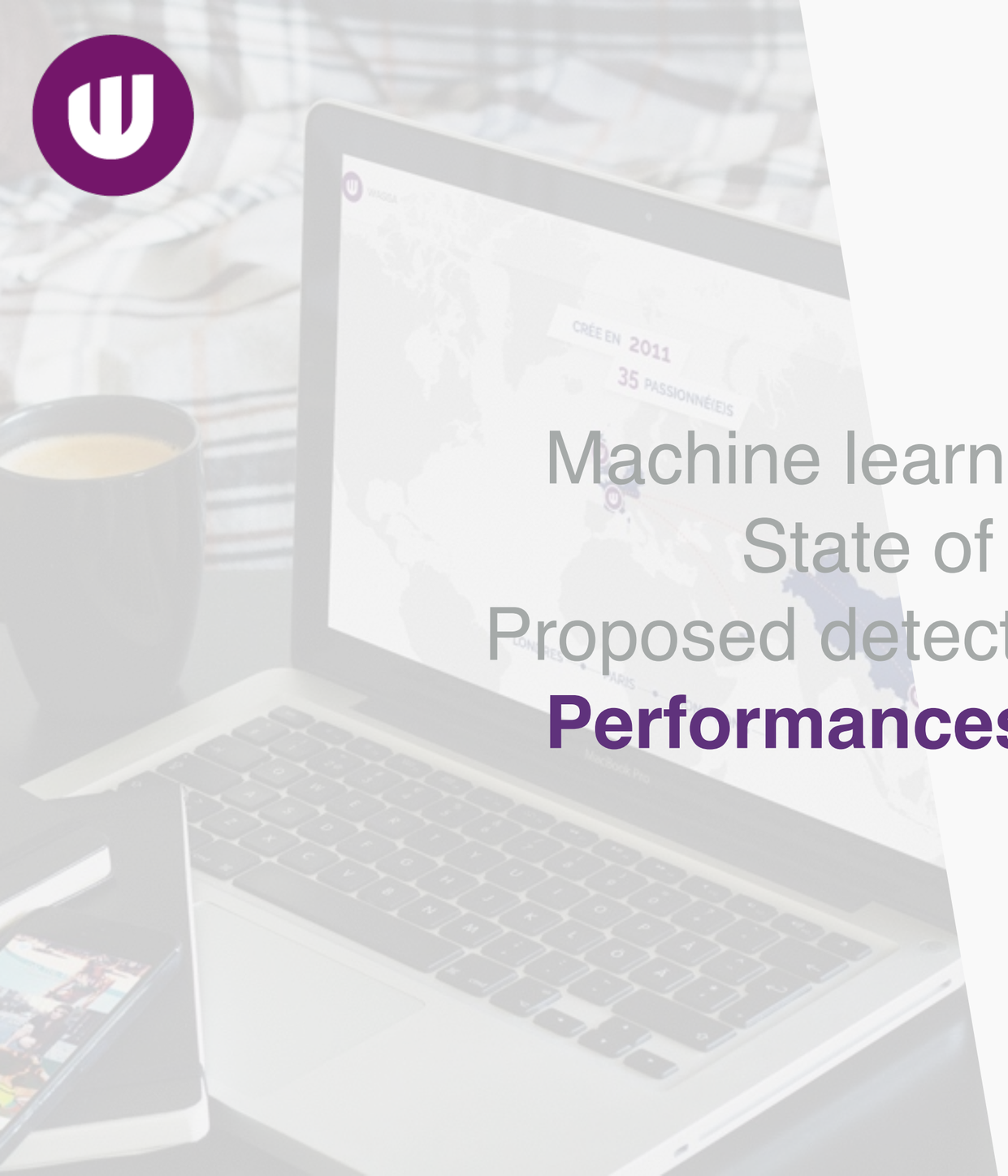
Proposed detection algorithms: Deep learning

Testing step

Object detector

Testing step





Machine learning concepts
State of the art
Proposed detection algorithms
Performances evaluation

Performances evaluation: Frame-by-frame

Secondary tasks	Testing data	Precision rate	Recall rate
Passenger	21 videos of 51 minutes in total (balanced set)	95,6 %	99,8 %
Feet on pedals	4 videos of 55 minutes in total (balanced set)	Close to 100 %	94 to 98%
Hands on wheel	6 vidéos of 81 minutes in total (80% with hands)	99,5 %	85,4 %
Texting	11 videos of 155 minutes in total (7.5% of phone)	17 %	67 %
Phone-to-the-ear	12 videos of 122 minutes in total (3.2% of phone)	13 %	67 %

Performances evaluation: Frame-by-frame

Secondary tasks	Testing data	Precision rate	Recall rate
Passenger	21 videos of 51 minutes in total (balanced set)	95,6 %	99,8 %
Feet on pedals	4 videos of 55 minutes in total (balanced set)	Close to 100 %	94 to 98%
Hands on wheel	6 vidéos of 81 minutes in total (80% with hands)	99,5 %	85,4 %
Texting	11 videos of 155 minutes in total (7.5% of phone)	17 %	67 %
Phone-to-the-ear	12 videos of 122 minutes in total (3.2% of phone)	13 %	67 %

+ Algorithm rarely misses passenger presence
- Some false detections.

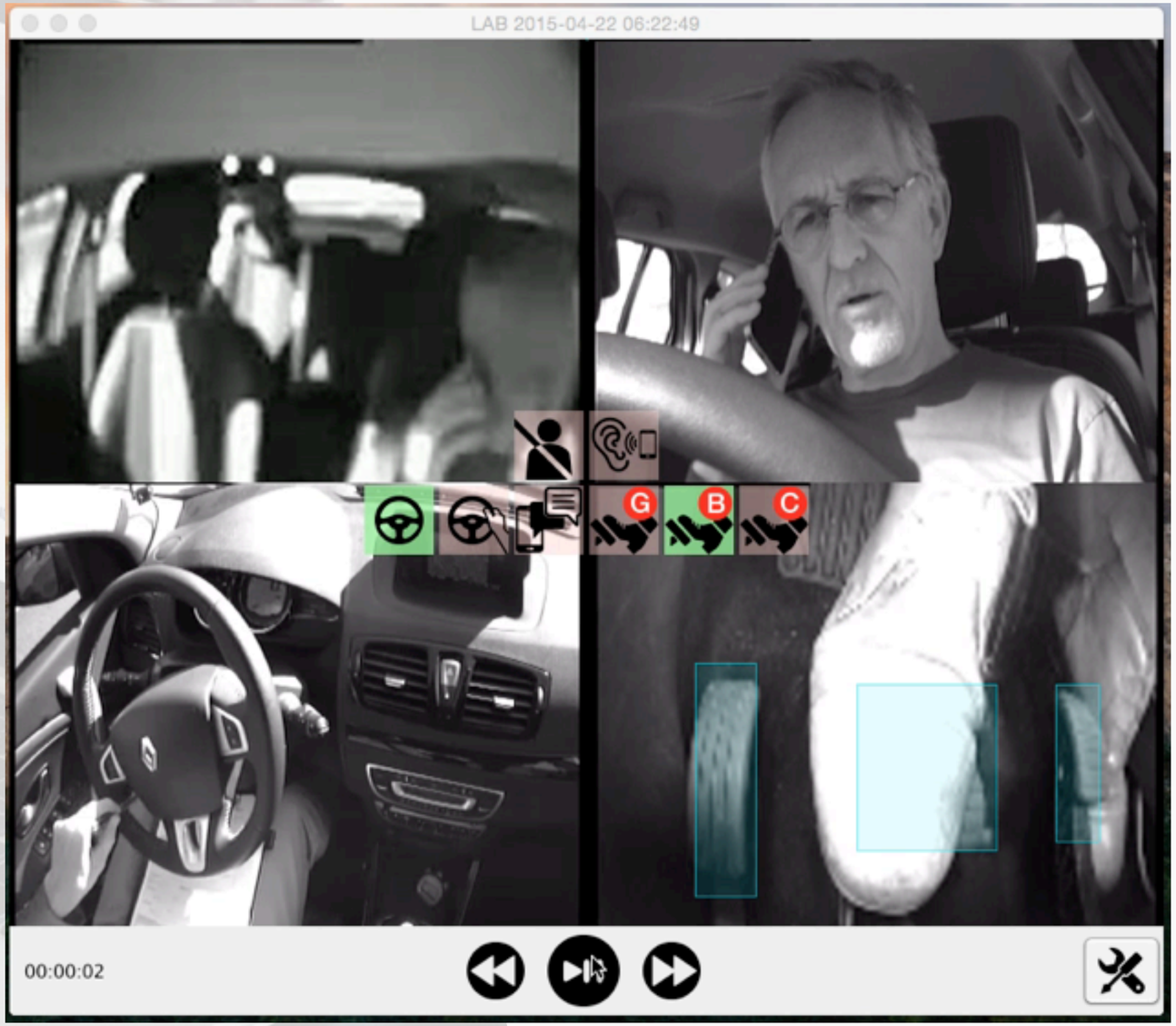
Performances evaluation: Frame-by-frame

Secondary tasks	Testing data	Precision rate	Recall rate	
Passenger	21 videos of 51 minutes in total (balanced set)	95,6 %	99,8 %	+ Algorithm rarely misses passenger presence - Some false detections.
Feet on pedals	4 videos of 55 minutes in total (balanced set)	Close to 100 %	94 to 98%	+ Algorithm is rarely wrong - Misses some detections.
Hands on wheel	6 vidéos of 81 minutes in total (80% with hands)	99,5 %	85,4 %	
Texting	11 videos of 155 minutes in total (7.5% of phone)	17 %	67 %	
Phone-to-the-ear	12 videos of 122 minutes in total (3.2% of phone)	13 %	67 %	

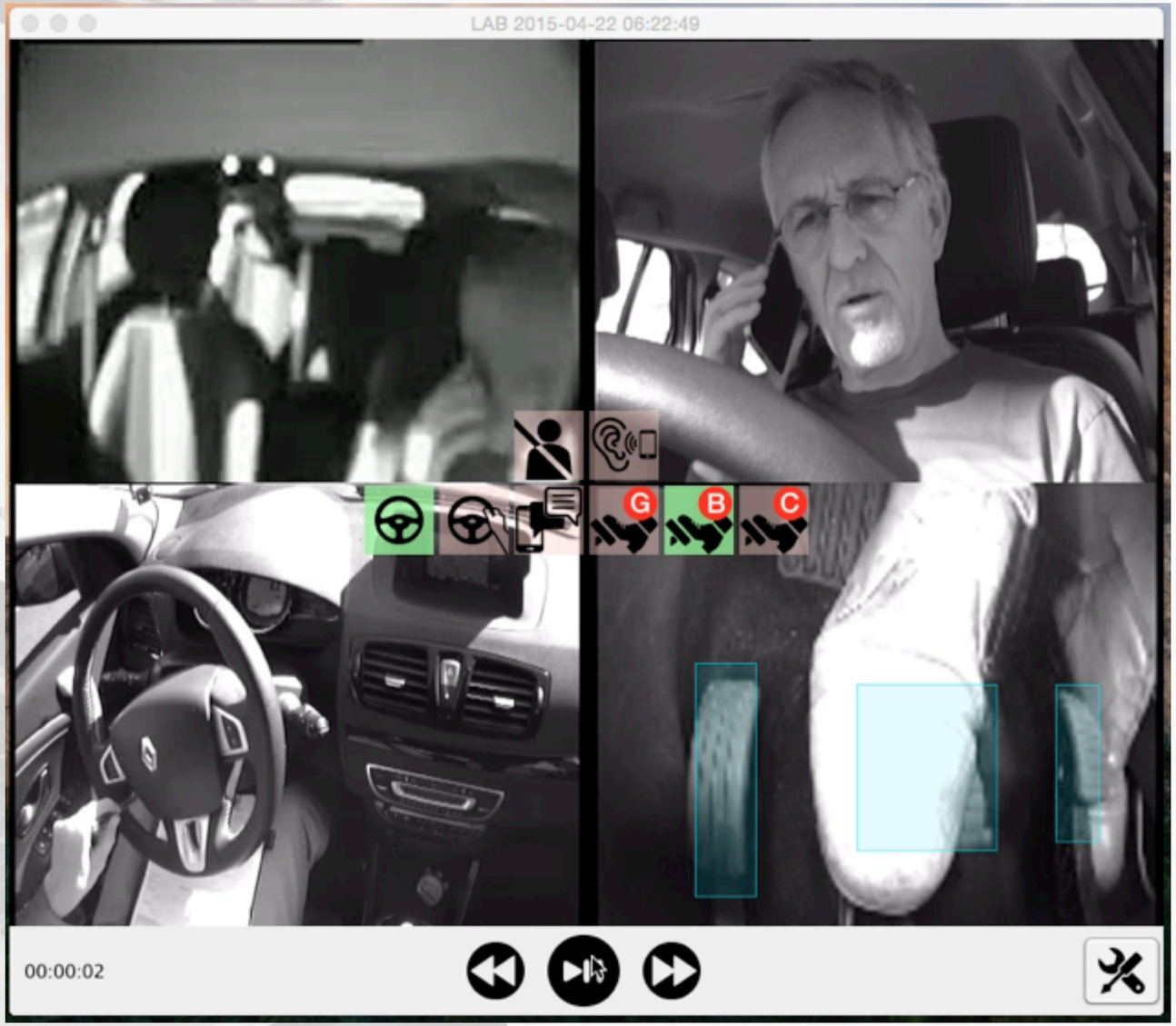
Performances evaluation: Frame-by-frame

Secondary tasks	Testing data	Precision rate	Recall rate	
Passenger	21 videos of 51 minutes in total (balanced set)	95,6 %	99,8 %	+ Algorithm rarely misses passenger presence - Some false detections.
Feet on pedals	4 videos of 55 minutes in total (balanced set)	Close to 100 %	94 to 98%	+ Algorithm is rarely wrong - Misses some detections.
Hands on wheel	6 vidéos of 81 minutes in total (80% with hands)	99,5 %	85,4 %	
Texting	11 videos of 155 minutes in total (7.5% of phone)	17 %	67 %	+ Algorithm correctly detects 2 out 3 frames with phone presence - A lot of false detections.
Phone-to-the-ear	12 videos of 122 minutes in total (3.2% of phone)	13 %	67 %	

Performances evaluation: Example



Performances evaluation: Example



Conclusions

- **Use of machine learning approaches for NDS**
  - Promising results for secondary tasks detection.
  - Allows to strongly reduce manual annotations computing time.
  - False detections still need to be lowered.
- **Improvements**
  - Post-processing filtering, add object tracking.
  - Add more training samples.
  - Try other frameworks or network architectures.

Thank you for your attention.

flora.dellinger@wassa.fr

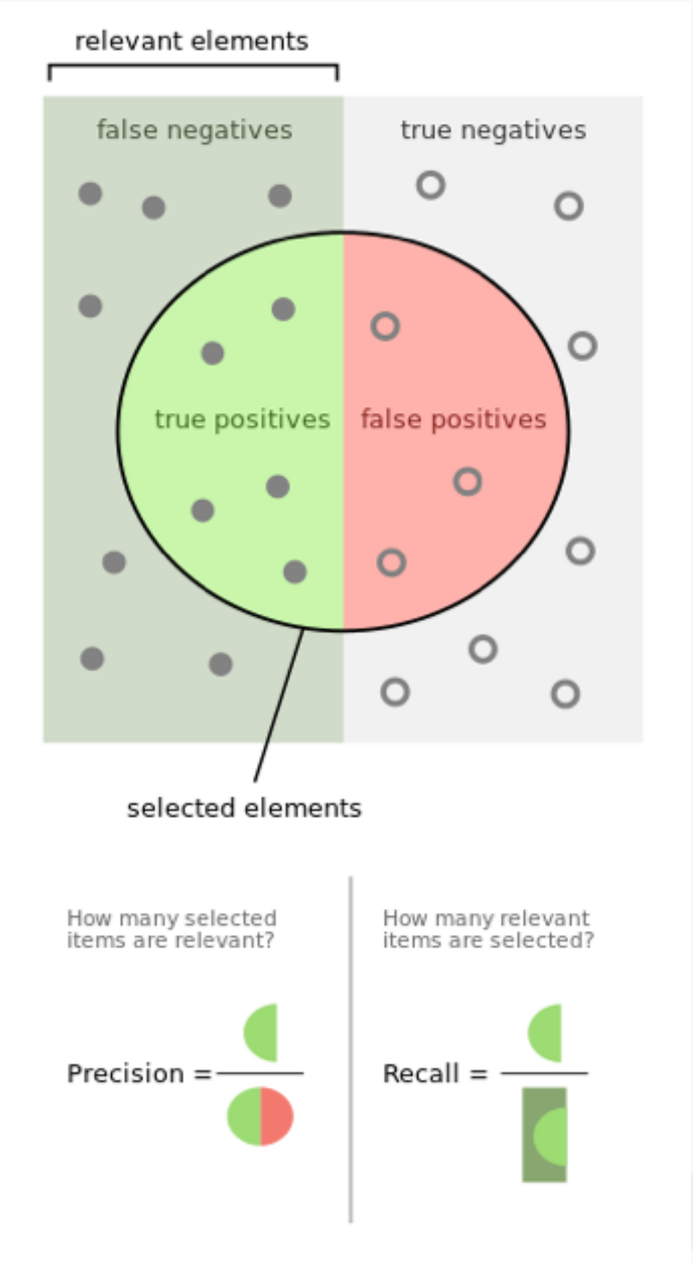
5 rue de l'église
92100 Boulogne-Billancourt
France

(+33)1 83 64 44 70

http://www.wassa.fr

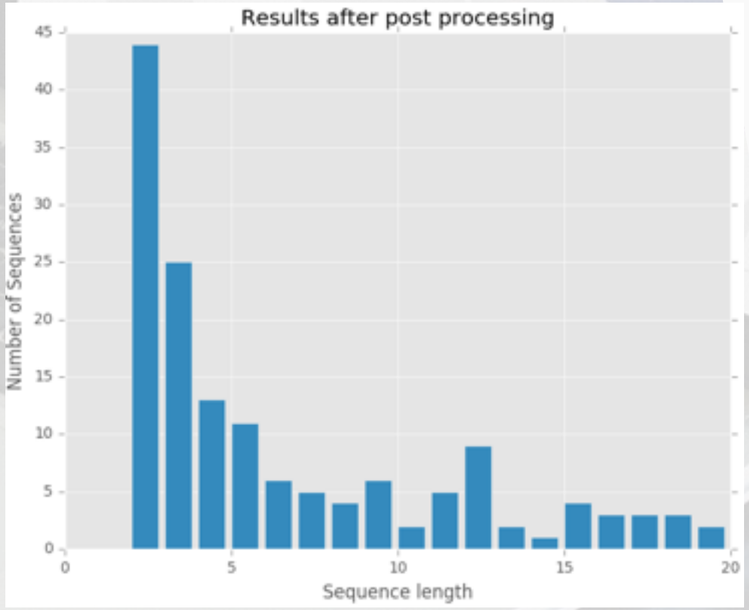
Performances evaluation: Protocol and metrics

- **Videos labeled manually** through sequences for each feature:
  - **Positive and negative** sequences.
  - **Several videos with different conditions:** day/night situations, different drivers, wearing gloves or not, different types of cellphones etc.
- **Studied metrics:**
  - Evaluation frame-by-frame: Recall and Precision
  - Evaluation by sequence (Texting and phone-to-the-ear).

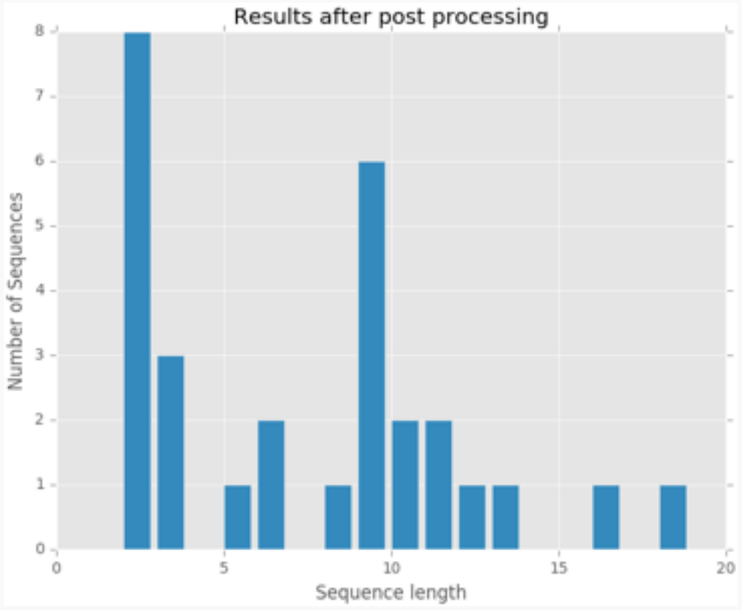


Performances evaluation: Sequences

Metrics	Positive sequences correctly detected	False positive sequences detected
Texting	25 out of 32	352 out of 377
Phone-to-the-ear	12 out of 14	103 out of 115



Histogram of sequence length for Texting



Histogram of sequence length for Phone-to-the-ear