DE LA RECHERCHE À L'INDUSTRIE

Co-clustering de séries temporelles multivariées pour la validation du véhicule autonome

21 septembre 2021

Loïc GIRALDI

avec Etienne GOFFINET (LIPN, Renault), Anthony COUTANT (LIPN, HephIA), Mustapha LEBBAH (LIPN, HephIA), Hanane AZZAG (LIPN)

Table of contents

Industrial context

AD / ADAS: Huge increase of the number of systems

2015 2022

ADAS SYSTEMS

2015 2022

VEHICLE APPLICATIONS

2015 2022

PLANTS & MARKETS

AD / ADAS: Huge increase of car complexity

ADAS ECU

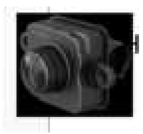
Front camera

Radar

HD Map

Lidar

Around view camera



Ultrasound barrier

Flight recorder

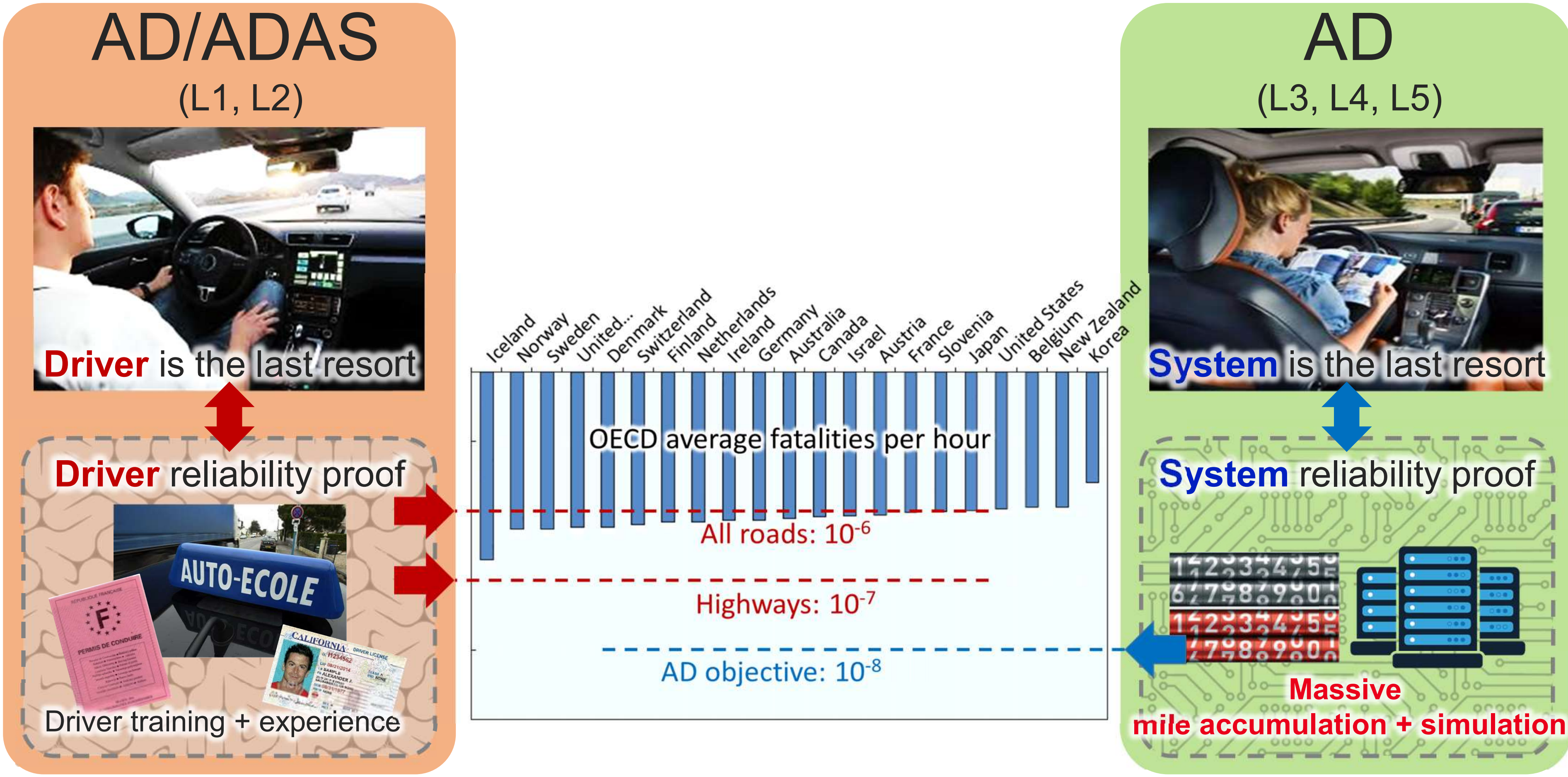
Redundant steering

Redundant braking

AD / ADAS: Huge increase of car complexity

2015: https://informationisbeautiful.net/visualizations/million-lines-of-code/

AD / ADAS: High reliability requirements



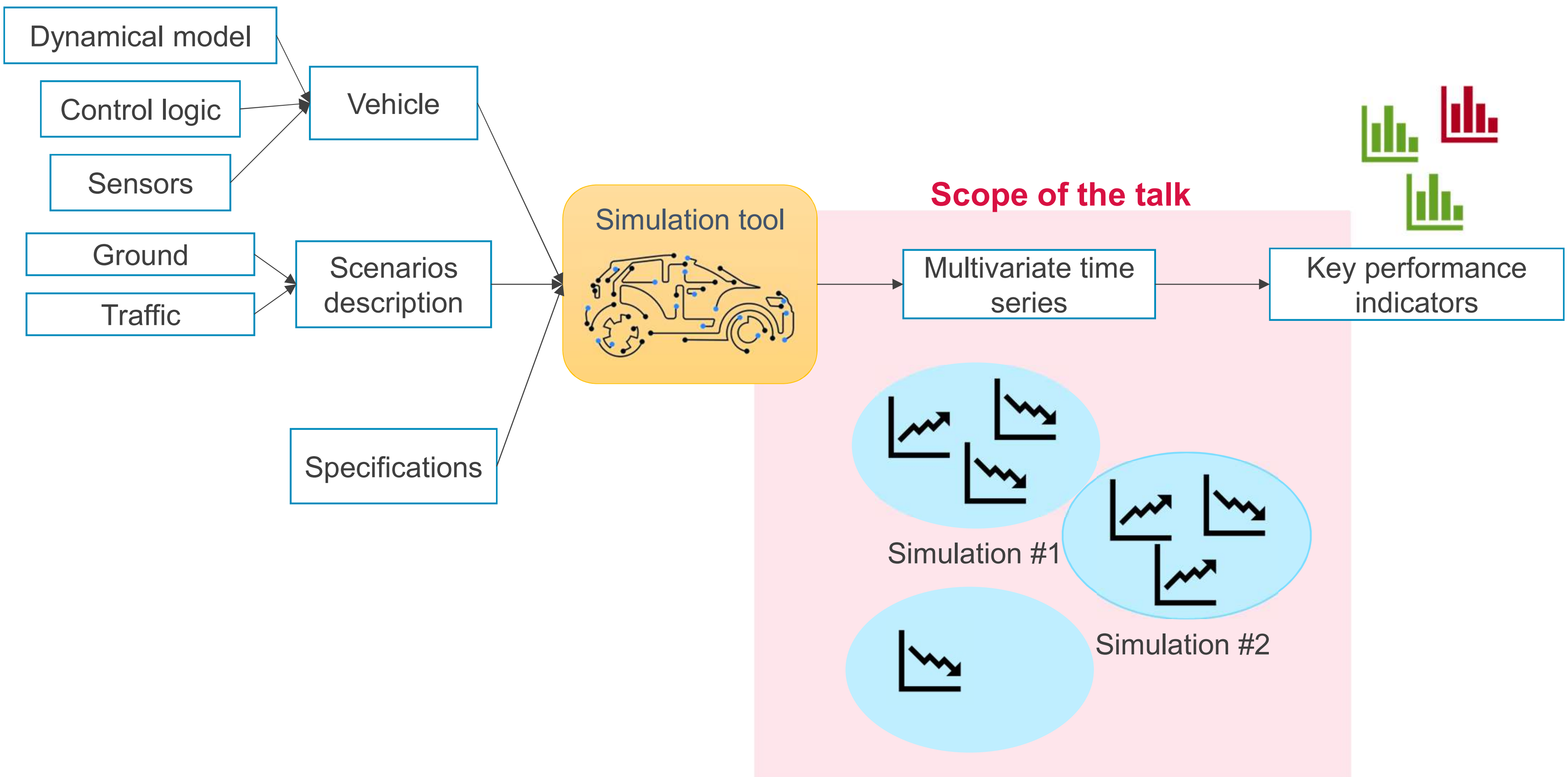
Naive estimation of the required number of cars for validation

- For a reliability of 10^{-8} and a confidence of 95%, using a Poisson distribution we find
  - Required driving time $= 3 \times 10^8$ hours
  - Number of kilometers at 50kph $= 1.5 \times 10^{10}$ km

$\approx 10^5$ cars should be dedicated to AD/ADAS validation

- A (partial) solution: numerical validation

Typical design of experiment for AD/ADAS validation



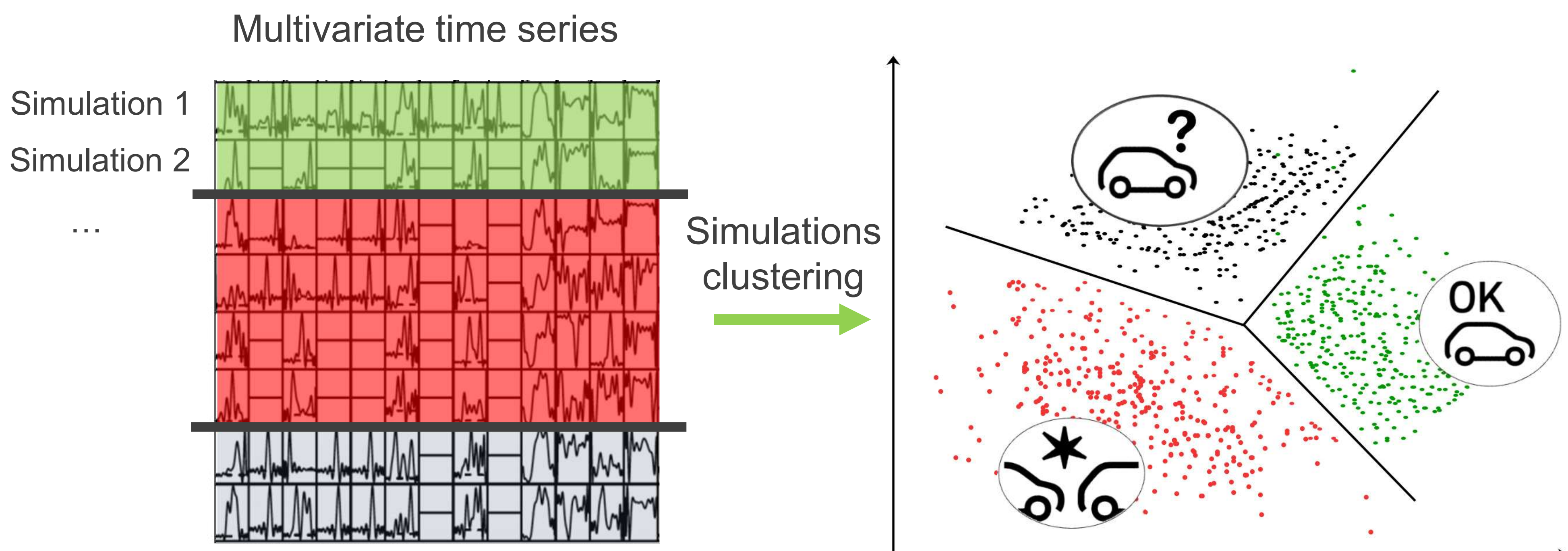
Opportunities and challenges

- **Business opportunities from time series mining:**
  - Identify operating modes of the vehicle – multivariate time series (co-)clustering problem
  - Identify anomalies

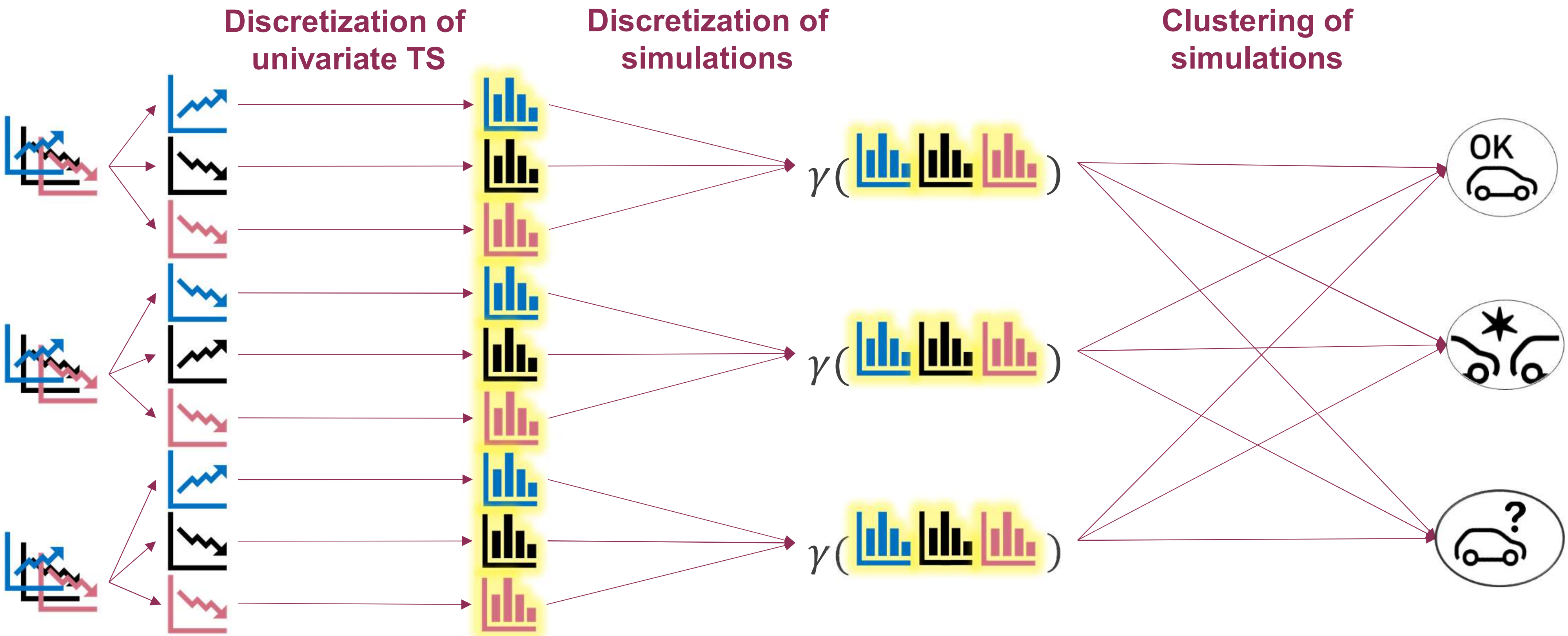
- **Scientific challenges**
  - **Many simulations** (e.g. 10k)
  - **Many signals** (e.g. 300)
  - **Many timesteps** (signal sampling @ 20Hz)
  - Dataset size **up to 1Tb**
  - **Different time series lengths** per simulation

Clustering of AD / ADAS simulations

Clustering of simulations via multivariate time series analysis

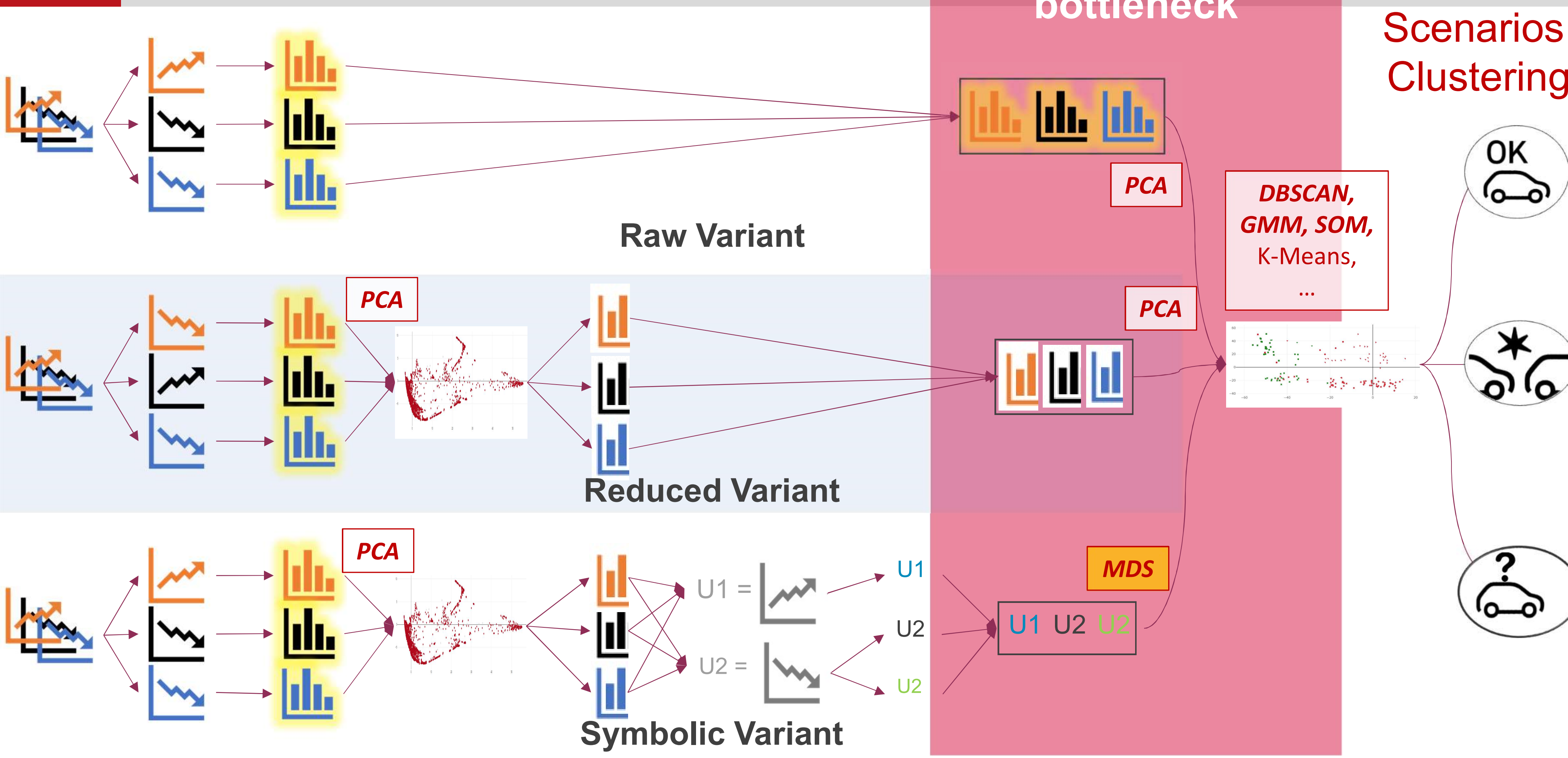


From raw time series to simulation clustering

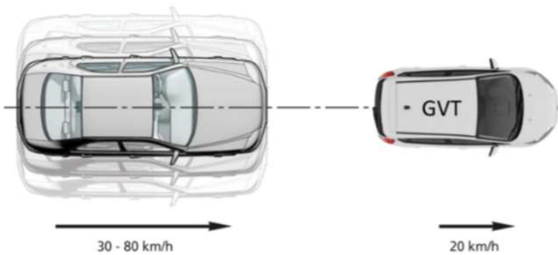


FFT + log-periodogram interpolation
Caiado, J., Crato, N., & Peña, D. (2009). Comparison of times series with unequal length in the frequency domain. *Communications in Statistics—Simulation and Computation*

From raw time series to simulation clustering

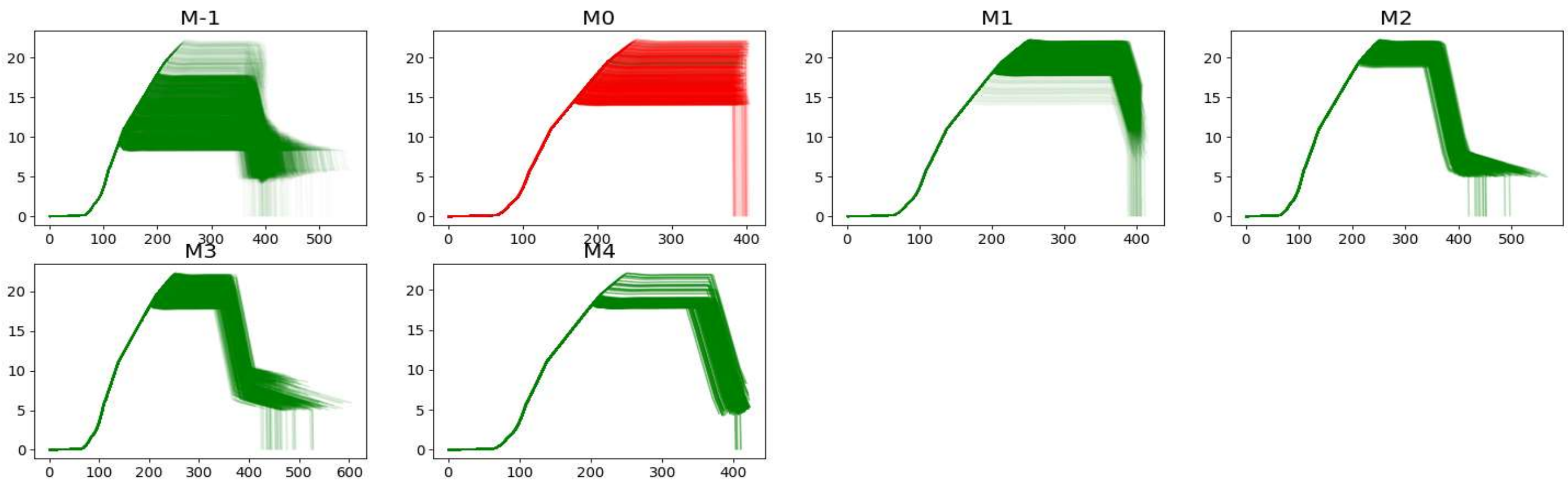


- **Advanced Emergency Braking – Car to Car Rear Moving (AEB-CCRm)**
  - 20000 simulations
  - Varying speeds
  - Varying overlaps
  - Analysis based on 30 signals

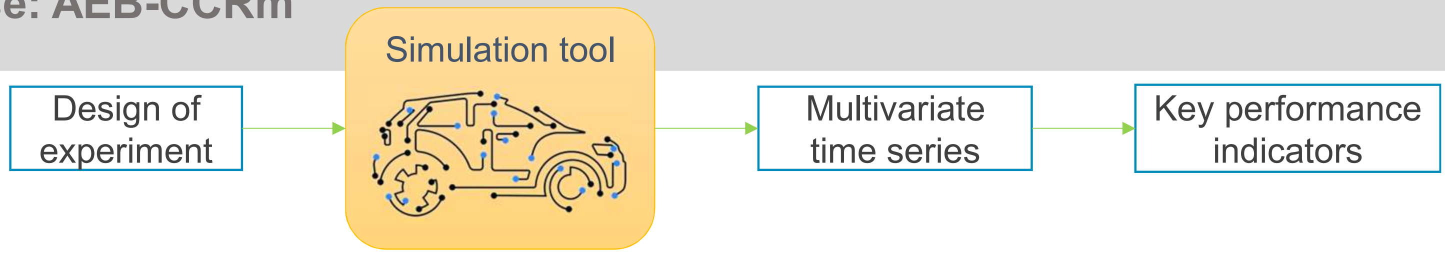


- **Analysis performed with the reduced variant of the pipeline**

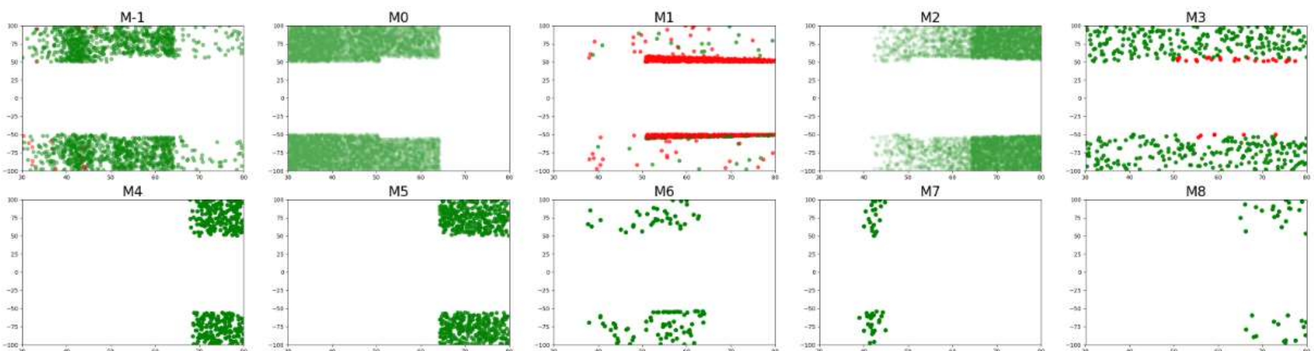
Car speed w.r.t. time



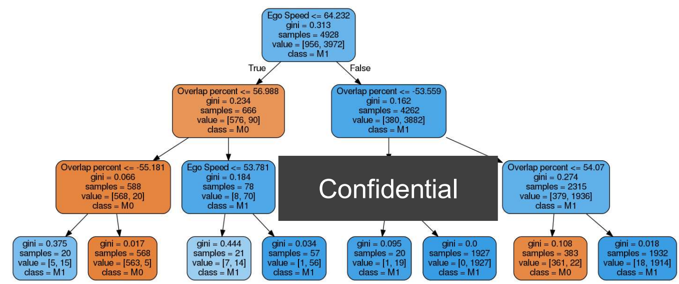
Use case: AEB-CCRm



- **Focus on multivariate time series: can we extract more informations?**
- **Yes: combine clustering with classification methods:**
  - feature: design of experiment / input simulation parameters
  - label: assigned cluster



Car speed vs Overlap vs Cluster



Decision tree:
DOE classification
predicting cluster labels

Co-clustering of AD / ADAS simulations

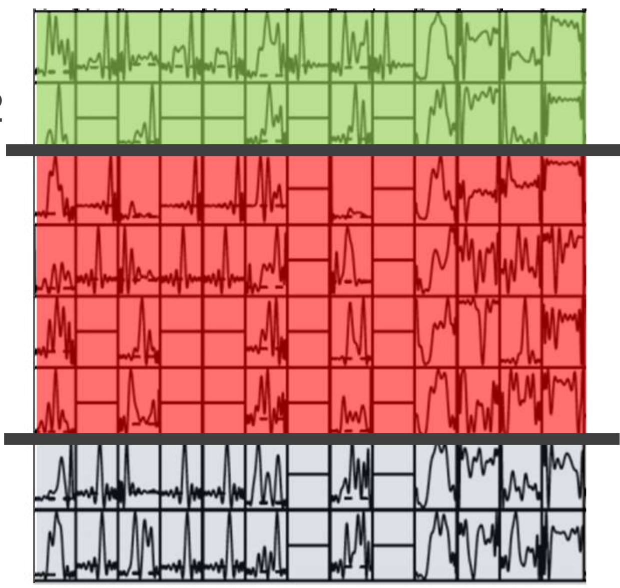
Simulation clustering

Multivariate Time series

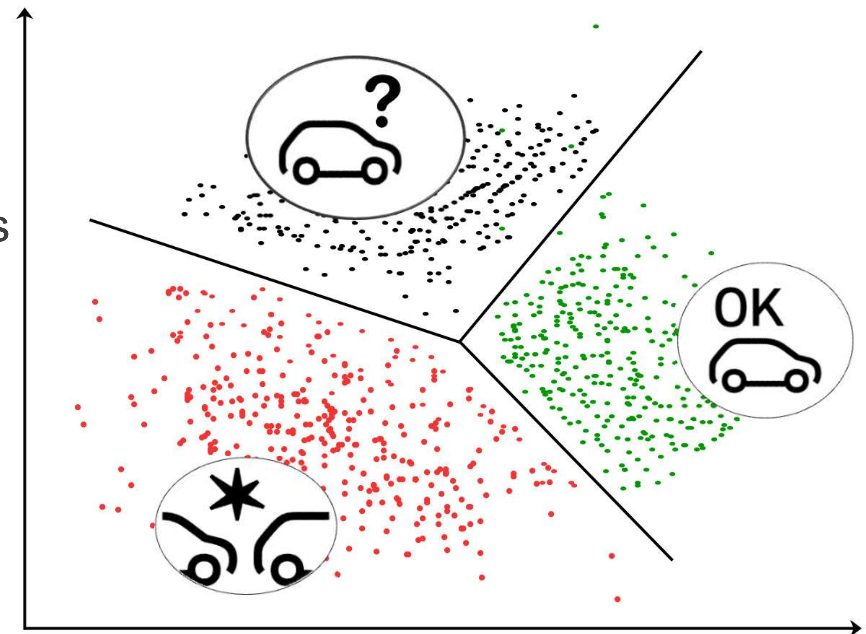
Simulation 1

Simulation 2

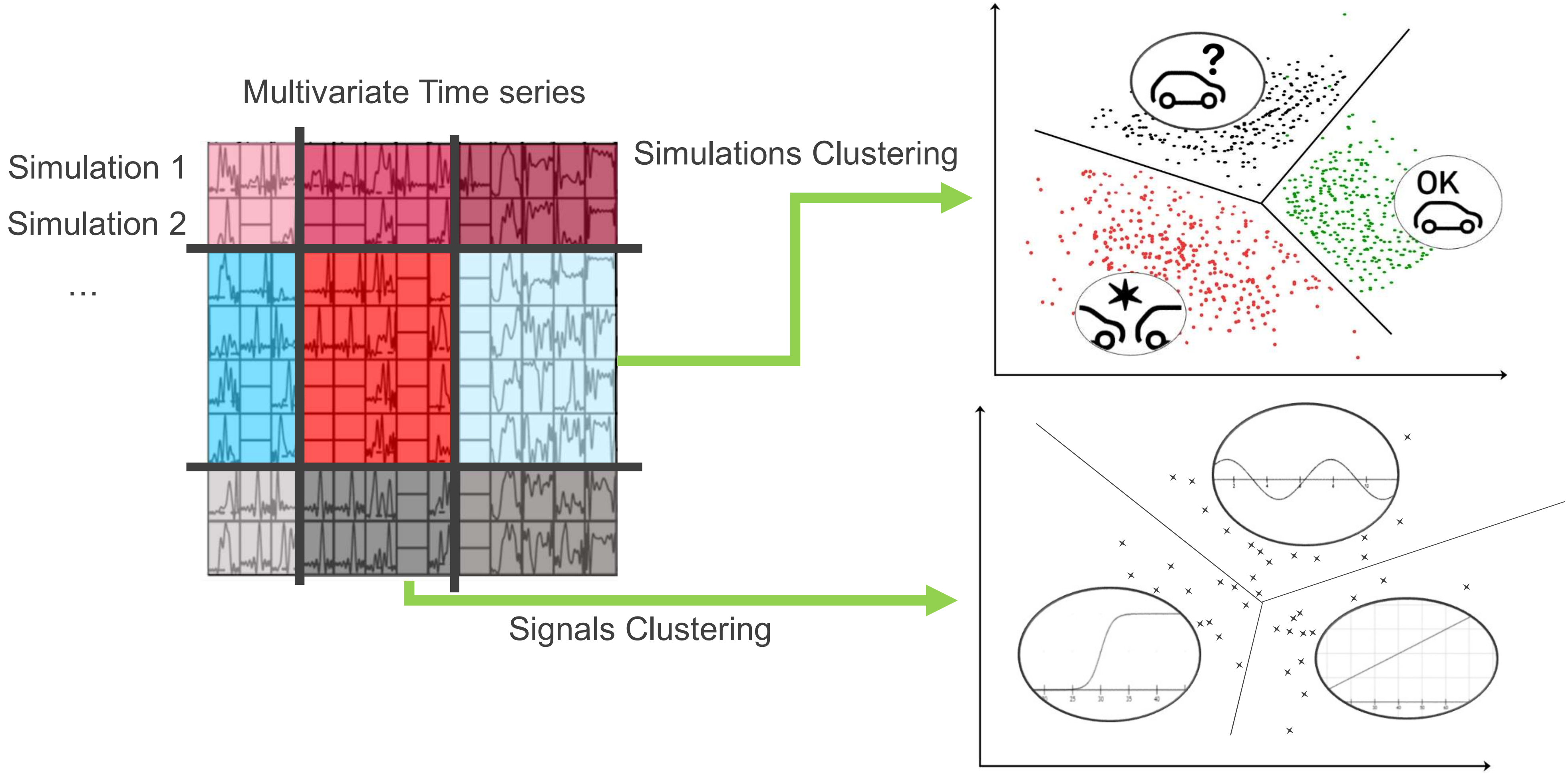
…



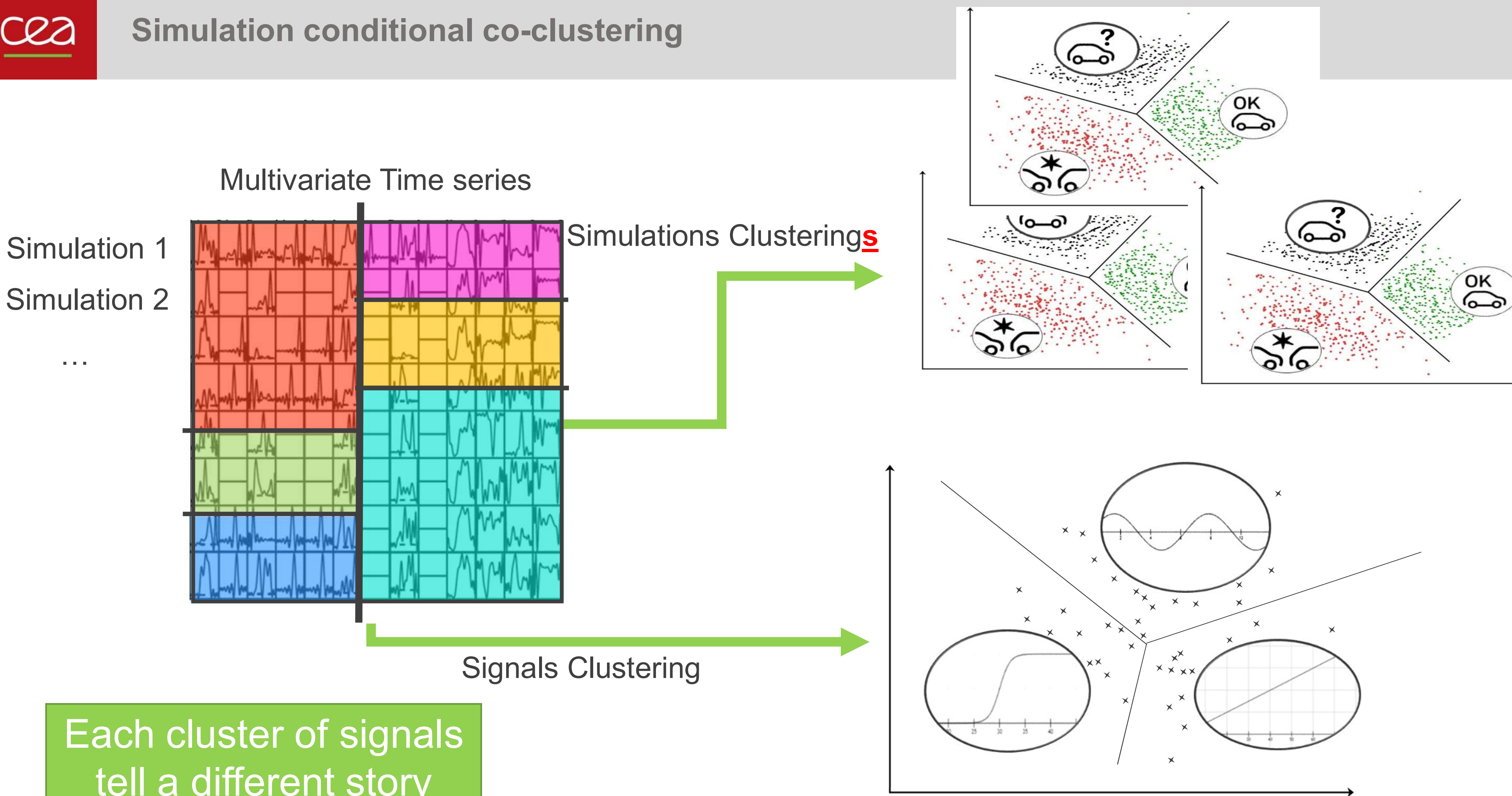
Simulations Clustering



Simulation co-clustering



Simulation conditional co-clustering



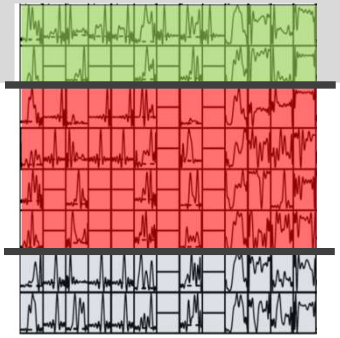
Each cluster of signals tell a different story

Model-based formulation

- **Notations:**
  - $(x_{ijs})_{ijs}$, dataset with $\boldsymbol{n}$ **observations** of $\boldsymbol{p}$ **features** in $\boldsymbol{d}$ **dimensions** (ie. After FFT + PCA + PCA)
  - Abuse of language and notations
    - the slice $(x_{\boldsymbol{i}js})_{js}$ is called a *"**row**"*
    - the slice $(x_{i\boldsymbol{j}s})_{is}$ is called a *"**column**"*
    - the index s will be omitted
  - $(z_{ik})_{ik}$ row cluster assignment variable
  - $(w_{jl})_{jl}$ column cluster assignment variable
  - θ hyperparameters

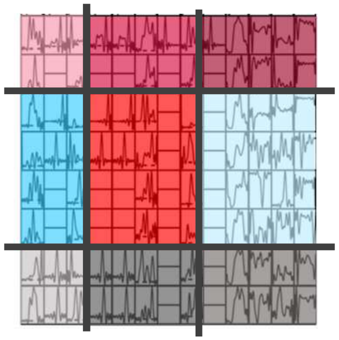
- **Model based methods for clustering**
  - Partitions will be represented with a mixture model
  - Cluster assignment uncertainty
  - Probabilistic outlier detection

Model based formulation



- Clustering with mixture models

$$p(x \mid \theta) = \sum_{z} p(x \mid z;\theta)p(z;\theta)$$

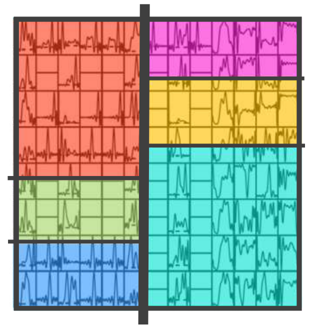


- Co-clustering with latent block models

$$p(x \mid \theta) = \sum_{z,w} p(x \mid z,w;\theta)p(z;\theta)p(w;\theta)$$

- Multi-clustering with latent block models

$$p(x \mid \theta) = \sum_{z,w} p(x \mid z,w;\theta)p(z \mid w;\theta)p(w;\theta)$$



- Gaussian assumption à la GMM

$$p\left(x_{ij} \mid z_i = k, w_j = l;\theta\right) \sim N(\mu_{kl}, \Sigma_{kl})$$

Inference process & model selection

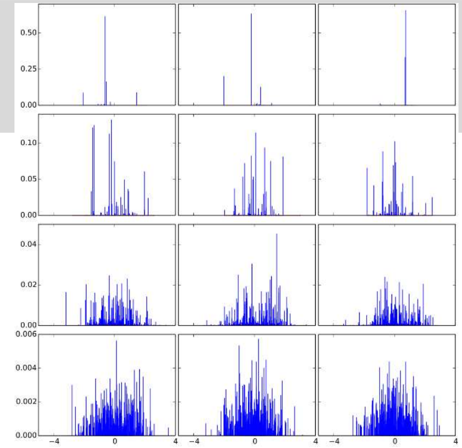
- **Inference process for latent block models: Stochastic Gibbs EM**
  - SE step: sample $p(z, w \mid x, \theta)$ with a Gibbs sampler
    - Sample $p(z \mid w, x, \theta)$
    - Sample $p(w \mid z, x, \theta)$
  - M step:
    - Update θ given (z, w)

- **Model selection (MS) using the integrated classification likelihood**

- **Issues**
  - Model selection is expensive
  - Without MS, the user must input additional parameters

- **Possible solution: introduce non-parametric Dirichlet Process**

Dirichlet process introduction



- **Intuitive formulation** of the Dirichlet process $DP(\alpha, G_0)$ (Chinese Restaurant Process)

$$p(Z_{n+1} \mid Z_n, \dots, Z_1) \propto \alpha\, G_0 + \sum_{k=1}^{K} n_k^* \delta_{Z_k^*}$$

Sampling from the distribution G_0

Sampling from the previous classes

- **Useful formulation** of the Dirichlet process $DP(\alpha, G_0)$ (Stick Breaking Process)

$$g_k \sim G_0, \qquad k = 1, \dots$$

$$\pi_k(r) = r_k \prod_{h=1}^{k-1} (1 - r_h), \qquad r_h \sim Beta(1, \alpha)$$

$$G = \sum_{k=1}^{\infty} \pi_k(r) \delta_{g_k} \sim DP(\alpha, G_0)$$

Symmetric Dirichlet process for co-clustering

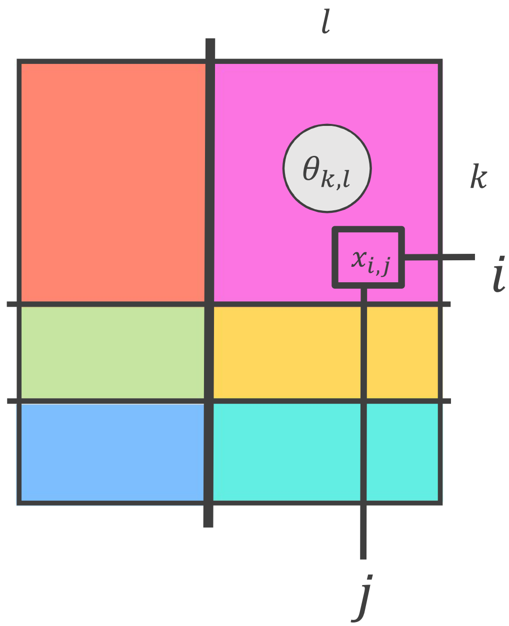
- **Model formulation**

$$x_{i,j} \mid \{z_i = k, w_j = l, \theta_{k,l}\} \sim F(\theta_{k,l}),\ \theta_{k,l} \sim G_0$$

$$\boldsymbol{z} \sim Mult(\boldsymbol{\pi}), \boldsymbol{w} \sim Mult(\boldsymbol{\rho})$$

$$\pi_k(\boldsymbol{r}) = r_k \prod_{h=1}^{k-1}(1 - r_h)\,, r_h \sim Beta(1, \alpha)$$

$$\rho_l(\boldsymbol{s}) = s_l \prod_{h=1}^{l-1}(1 - s_l)\,, s_l \sim Beta(1, \beta)$$



- **Bayesian Inference of $p(\boldsymbol{z}, \boldsymbol{w} \mid \boldsymbol{x}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{G_0})$ with a Gibbs sampler**
  - Sample row assignments $p(z \mid x, w, \alpha, \beta, G_0)$ row by row with the CRP
  - Sample column assignments $p(w \mid x, z, \alpha, \beta, G_0)$ column by column with the CRP

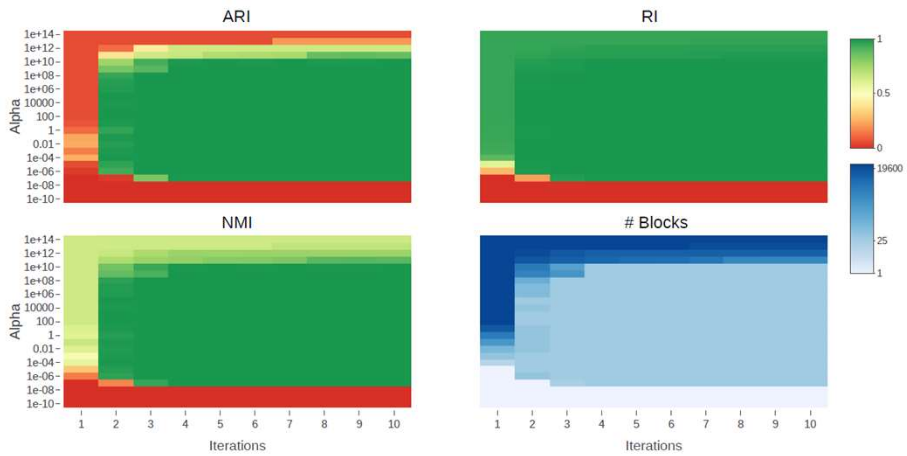
- **Complexity of the inference**

$$O(npd^2 + (n+p)\overline{K}\overline{L}d^3)$$

Benchmark on a synthetic dataset

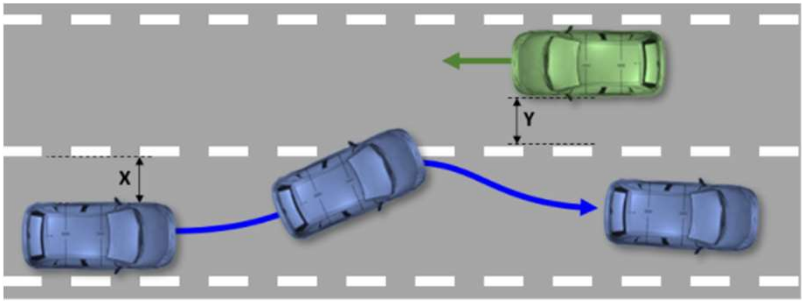
- **Synthetic dataset**
  - 5 row clusters with (20, 30, 40, 30, 20) observations
  - 5 column clusters with (40, 20, 30, 20, 30) observations
  - Each block has a Gaussian distribution
  - 19600 time series
- **Computer**
  - 12 processors Intel(R) Core(TM) i7-8850H CPU @ 2.60GHz
  - 32 Gb RAM

Method	ARI	RI	NMI	K	runtime (s)
$B\text{-}GMM$	0.825	0.982	0.946	25	18.6
$B\text{-}GMM_{MS}$	0.942	0.994	0.9803	30	85.9
$B\text{-}GMM_{14}$	0.979	0.997	0.994	25	89.3
LBM	0.823	0.913	0.887	25	17.5
LBM_{MS}	0.940	0.994	0.979	25	494.4
LBM_{49}	1	1	1	25	603.2
$B\text{-}DPMM$	0.670	0.958	0.906	16	25.3
$NPLBM$	**1**	**1**	**1**	25	**40**



Use case: ELK

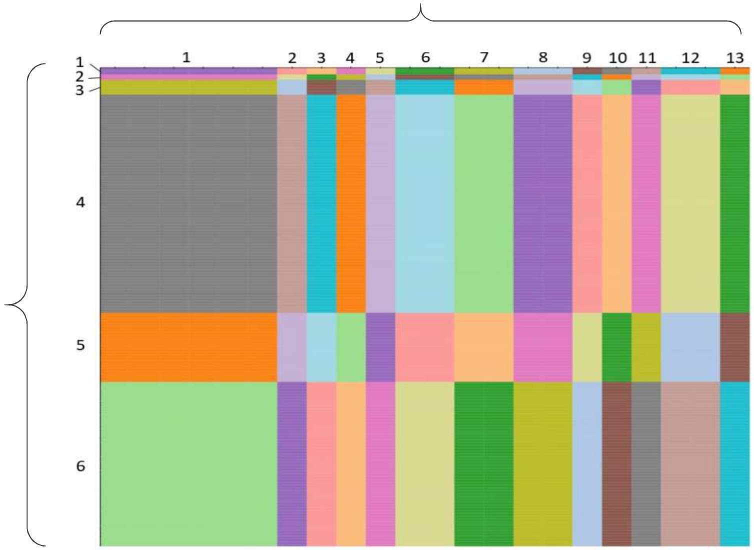
- **Emergency Lane Keeping (ELK)**
  - 400 simulations
  - 22 variables
  - Varying speeds
  - Varying decentering
  - Varying drifting angle
  - ...



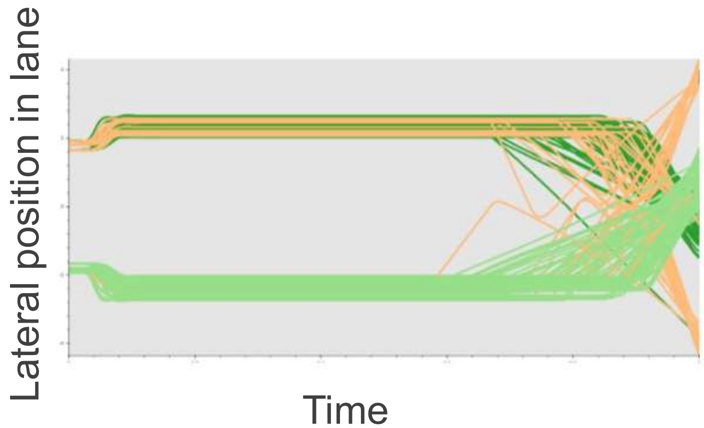
- **Result representation**

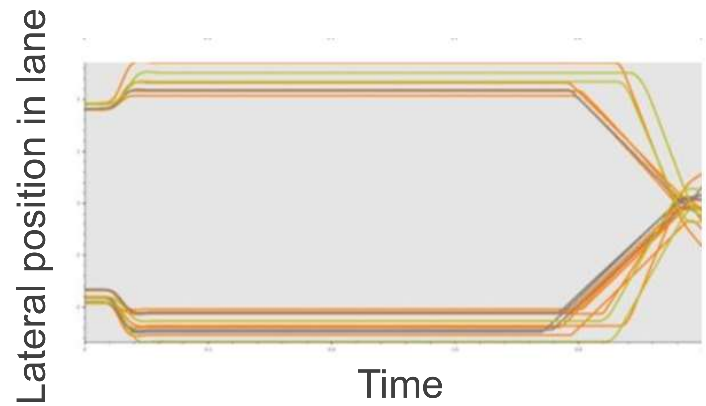
22 variables

400 simulations



Row clusters interpretations





Main row clusters:

- *Light Green:* drifting left and ELK fails
- *Dark Green:* drifting right and ELK fails
- *Orange:* ELK works

Other row clusters are outliers:

- ELK system activates too late

Conclusions & perspectives

- **Summary Part I**
  - Industrial simulation clustering workflow from large time series database
  - Main procedure in 4 steps
    - Vector representation of univariate time series
    - Dimension reduction of univariate vectors
    - Dimension reduction of multivariate components
    - Clustering in the reduced feature space
  - Classification based methodology in order to interpret clusters

- **Summary Part II**
  - Use pre-processing pipelines for co-clustering
  - Application of non-parametric co-clustering methodology for joined clustering of simulations and signals

- **Perspectives**
  - Better time series representation – replace FFT with wavelets, polynomials, ...
  - Scalability of the Dirichlet Process method
  - Multi-clustering visualization

- **For more details:**
  - E. Goffinet, M. Lebbah, H. Azzag, L. Giraldi and A. Coutant. A New Multivariate Time Series Co-clustering Non-Parametric Model Applied to Driving-Assistance Systems Validation, AALTD 2021
  - E. Goffinet, M. Lebbah, H. Azzag, L. Giraldi and A. Coutant. Multivariate Time Series Multi-Coclustering. Application to Advanced Driving Assistance System Validation. ESANN 2021

Merci de votre attention