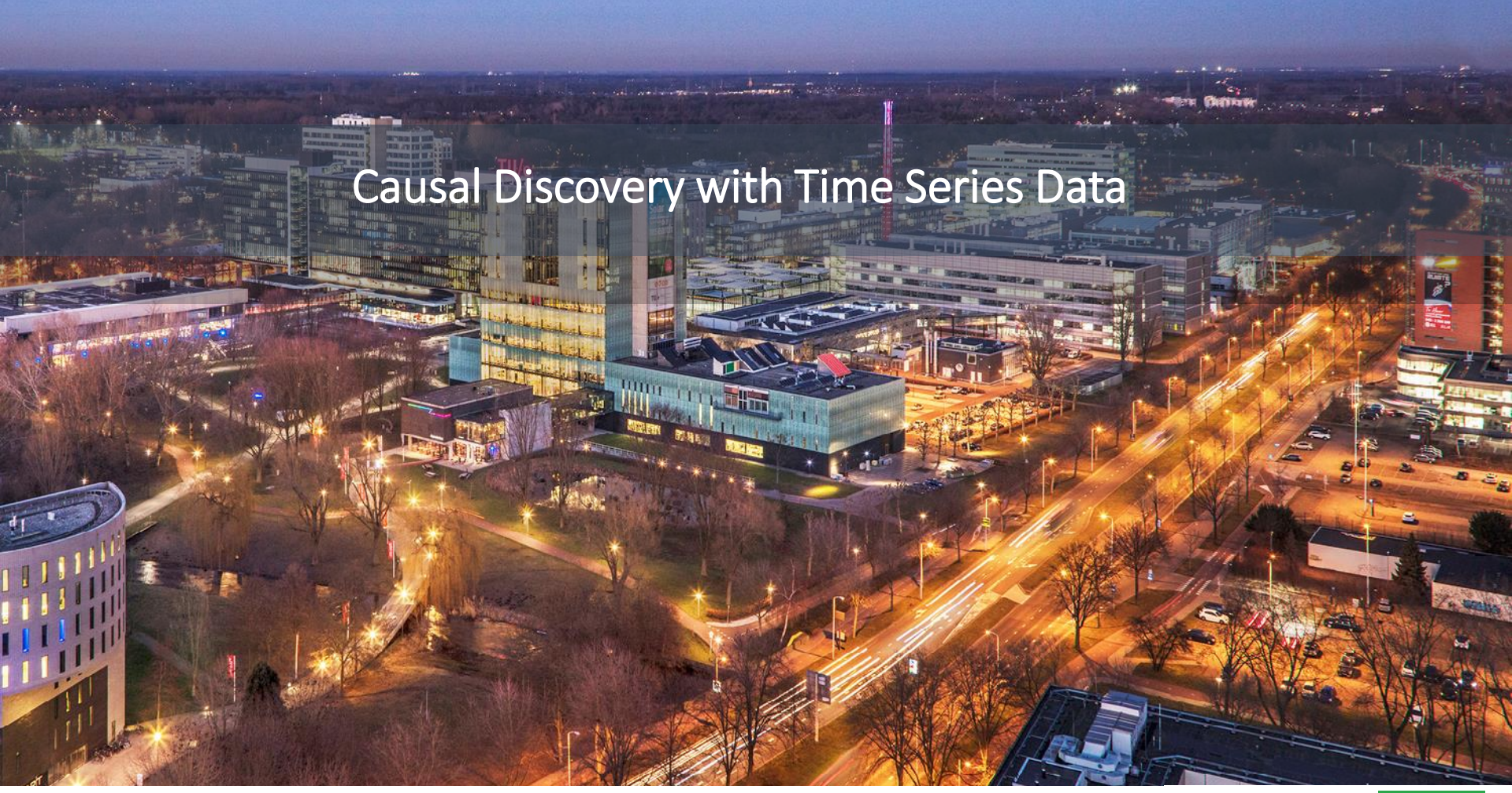


Causal Discovery with Time Series Data

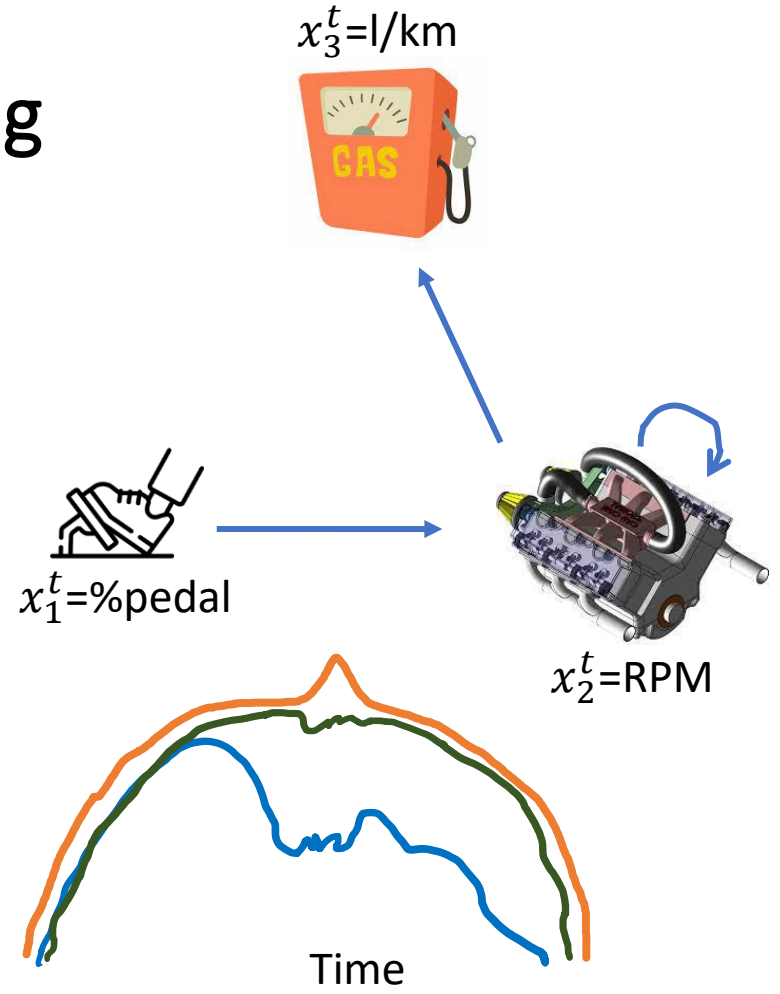
- Alexander Mey, 01/02/2022

About me

- Alexander Mey, PostDoc Tu Eindhoven/ASML
- Background: Reinforcement Learning (PostDoc),
  Machine Learning (PhD), Mathematics (Bachelor + Master)
- Supervisors: Rui Castro (TU/e), Hans Onvlee (ASML)

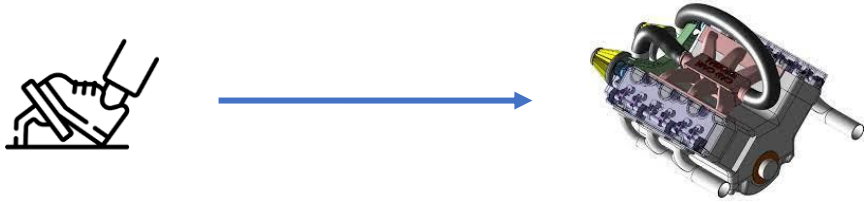
Problem Setting

- Observe time-series data $x^t = (x_1^t, \dots, x_d^t)$, t is time, $x_i^t \in \mathbb{R}$
- Goal: Find a causal graph, i.e. a directed graph with causal interpretation, between variables
- ASML context: Complex machines with many signals of different frequency
- Uses of graph:
  1. General understanding of process
  2. Root cause analysis
  3. Online fault detection/prevention



Our Starting Points

1. Use predictive power as a surrogate for causal direction (Granger-Causality)



The past of the gas pedal is better to predict the future of the engine than the past of the engine is to predict the future of the pedal.

Our Starting Points

1. Use predictive power as a surrogate for causal direction (Granger-Causality)

The past of the engine is better to predict the future of the gas pedal than the past of the pedal to predict the future of the engine.

Our Starting Points

2. Use data driven predictive models (time-series forecasting)

- Simple (linear) models can scale very well, fast to compute

- Many models can be online updated/validated, opportunity for online monitoring

- Possibility of independent of model validation

Our Starting Points

Use data driven predictive models

1. Data $(x^1, \dots, x^T)$, $x^t \in \mathbb{R}^d$
2. Model h: $\mathbb{R}^d \to \mathbb{R}^d$ for $h \in H, h(x^t) \simeq x^{t+1}$, e.g. h linear, h $\in \mathbb{R}^{d \times d}$
3. Predictive loss l: $\mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$, e.g. $l(h(x^t), x^{t+1}) = ||h(x^t) - x^{t+1}||$
4. A map g: $H \to G$, where G are (weighted) directed graphs on d nodes, e.g. if $h \in \mathbb{R}^{dxd}$ then $g(h) = h$

1. Find h, such that $E\left[l\left(x^{t+1}, h(x^t)\right)\right]$ is small (Machine Learning task)
2. Use $g(h)$ as surrogate for causal graph
3. Add constraints on g (no feedback loop), e.g. via penalty on g [1]

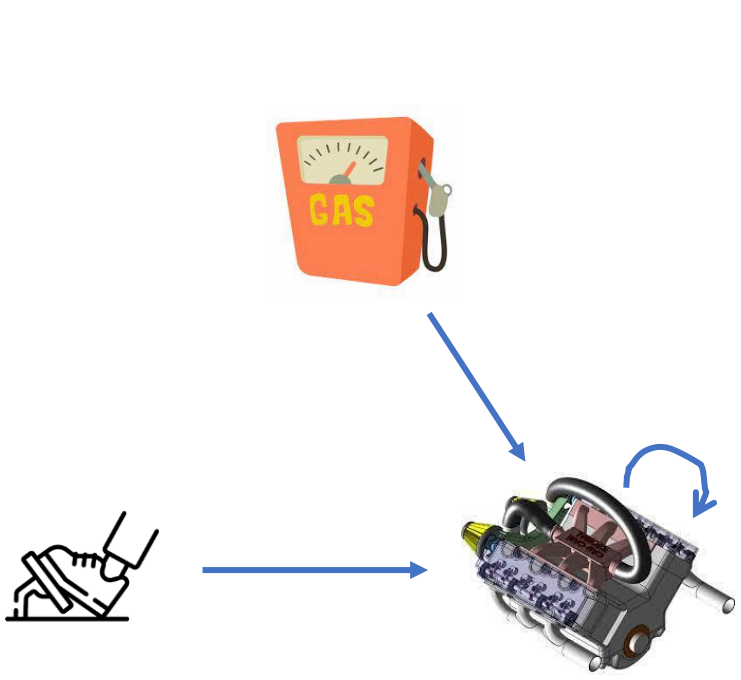
$$E\left[l\left(x^{t+1}, h(x^t)\right)\right] + penalty(g(h))$$

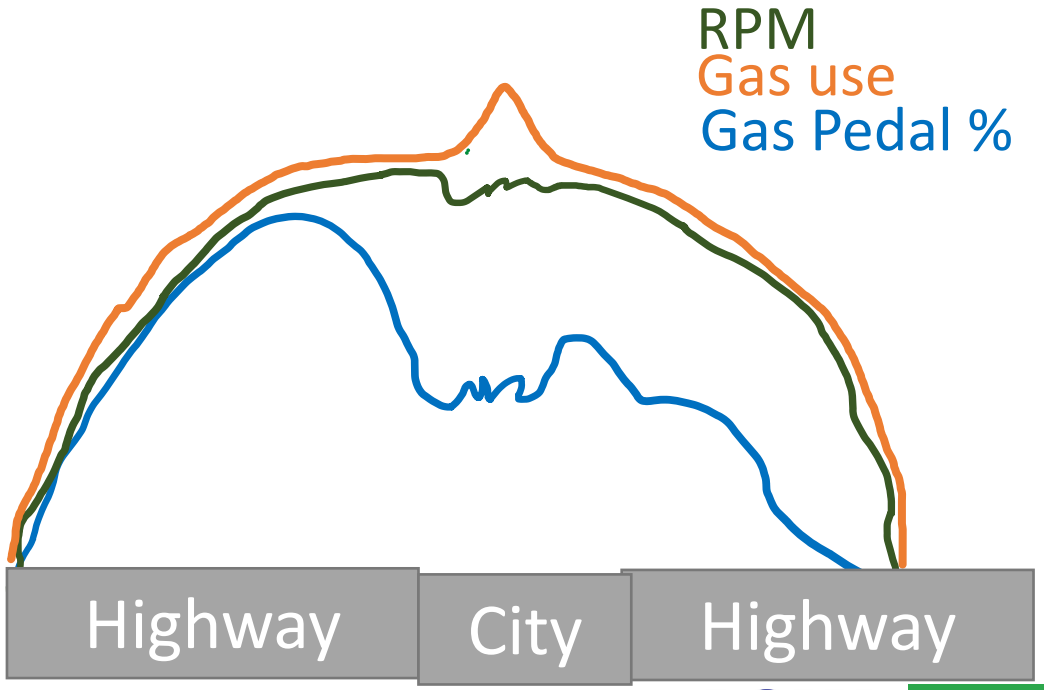
ASML TU/e

Invariances and Causality

Biggest problem: Machine Learning models pick up (spurious) correlations

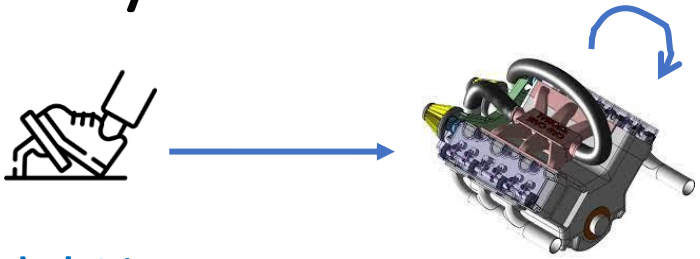
Promising recent work on using invariances to avoid that [2],[3],[4]

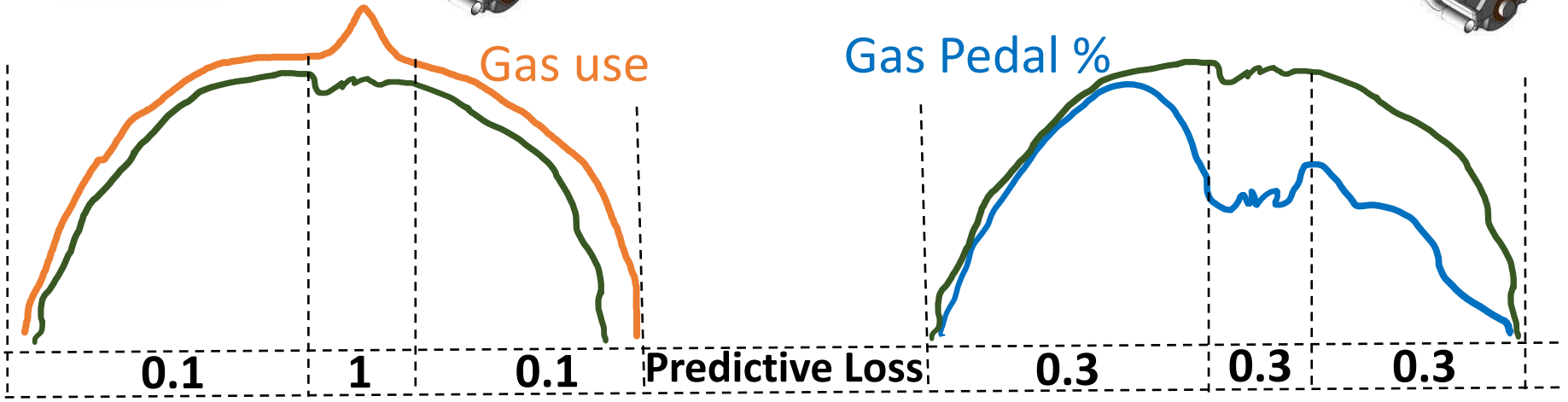




ASML TU/e

Invariances and Causality





Performance not invariant in different environments

Performance is invariant

ASML TU/e

Invariances and Causality

- Assumption 1: **Causes** are invariant over time, and across different environments, e.g. different drivers

- Assumption 2: Causal **mechanisms** are invariant in different environments: For example, the causal mechanism h between mass m and acceleration a and force F is given by $\mathrm{F} = h(m, a) = ma$

- Other viewpoint: Different environments, e.g. time intervals, may be seen as accidental interventions on the system

- Using causes for prediction is optimal under interventions

Summary

- Use predictive power as a surrogate for causal direction
- Use data driven models to find predictive power
- Steer data driven models to focus on causes and not correlations (invariance)

References

[1] Xun Zheng, Bryon Aragam, Pradeep Ravikumar, Eric P. Xing:
DAGs with NO TEARS: Continuous Optimization for Structure Learning. NeurIPS (2018)

[2] Peters, P. Buehlmann, and N. Meinshausen. Causal inference by using invariant prediction: Identification and confidence intervals. Journal of the Royal Statistical Society, (2016)

[3] Martín Arjovsky, Léon Bottou, Ishaan Gulrajani, David Lopez-Paz:
Invariant Risk Minimization. CoRR abs/1907.02893 (2019)

[4] Yoav Wald, Amir Feder, Daniel Greenfeld, Uri Shalit:
On Calibration and Out-of-domain Generalization. CoRR abs/2102.10395 (2021)

ASML TU/e

Other Challenges

1. Machine Learning models pick up (spurious) correlations
   - Promising recent work on using invariances to avoid that [1],[2],[3]
2. Deciding directionality of arrows
   - May force the model to chose based on performance
3. Time-series data, not i.i.d.
   - Actually useful? Often think of causal as time preceding
4. Not any model will scale
   - Simple (linear) models mostly will
5. The need of the map g means we have to restrict to interpretable models
   - Simple (linear) models are

Enforcing Invariance

- Let e $= \{x_e^1, \dots, x_e^{T_e}\}$ be a sample from environment e
- E the set of all environments
- For $h \in H$ define $R_e(h) = \sum_{x_e^t} l(x_e^{t+1}, h(x_e^t))$

Structural Invariance

Find $\{h_e\}_{e \in \mathrm{E}}$ that minimizes

$$\sum_{e \in E} R_e(h_e)$$

s.t.

$g(h_{e_1}) = g(h_{e_2})$

for all $e_1, e_2 \in E$

- Fitting one global model might not work well
- One local (e.g. linear) model per environment means less model mismatch
- Assumption 1: The resulting graph is invariant

ASML TU/e

Enforcing Invariance

- Let e $= \{x_e^1, \dots, x_e^{T_e}\}$ be a sample from environment e
- E the set of all environments
- For $h \in H$ define $R_e(h) = \sum_{x_e^t} l(x_e^{t+1}, h(x_e^t))$

Model Invariance

Find $h \in H$ that minimizes

$$\sum_{e \in E} R_e(h)$$

s.t.

$$h = \arg\min_{\bar{h}\ \text{s.t.}\ g(\bar{h}) = g(h)} R_e(\bar{h})$$

- $F = h(m, a) = ma$ optimal in most environments
- If h models a causal relationship, it does not change in any environment
- There should be no environment where for example $\bar{h}(m, a) = a + m$ is suddenly optimal

ASML TU/e

Enforcing Invariance

- Let e $= \{x_e^1, \dots, x_e^{T_e}\}$ be a sample from environment e
- E the set of all environments
- For $h \in H$ define $R_e(h) = \sum_{x_e^t} l(x_e^{t+1}, h(x_e^t))$

Loss Invariance

Find $h \in H$ that minimizes

$$\max_{\substack{\{\lambda_e\}_{e\in E} \\ \sum \lambda_e = 1 \\ \lambda_e \geq 0}} \sum_{e\in E} \lambda_e R_e(h)$$

- Similar to model invariance

ASML TU/e