Aalto University
School of Chemical
Engineering

HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI

singpro

Combining Data Analytics and Scheduling – SINGPRO Results and Open Challenges

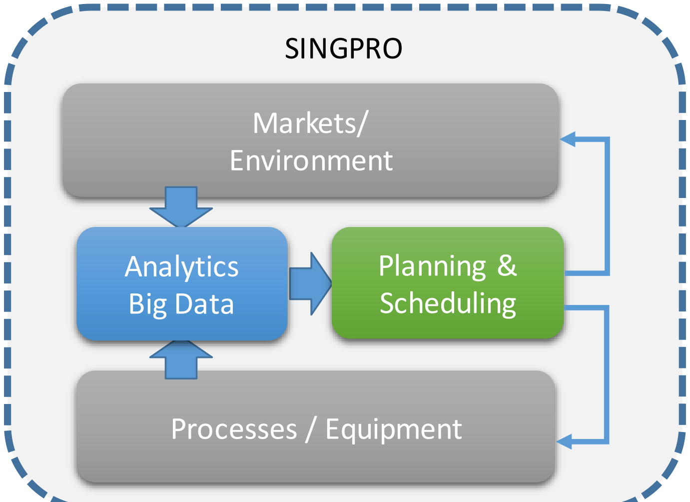
Iiro Harjunkoski[1,2], Teemu Ikonen[1], Hossein Mostafaei[1] , Keijo Heljanko[3] and Tewodros Deneke[3].

[1)]Aalto University, [2)]ABB Power Grids Germany, [3)] University of Helsinki

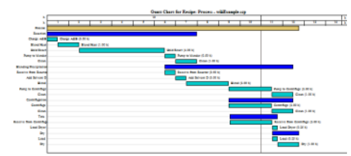
SINGPRO Seminar, 19.11.2019

SINGPRO Project (2018-2019)
Synergistic and intelligent process optimization

Academy of Finland project: Adj. Prof. Harjunkoski (Aalto CHEM) & Prof. Heljanko (University of Helsinki)

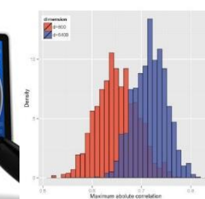


Robust & agile schedules
Adapted/correlated models



$$R_{r,t} = R_{r,t-1} + \sum_k \sum_i N_{k,i,t}\mu_{r,k,i} + \sum_k \sum_i N_{k,i,t-\theta_{k,i}}\bar{\mu}_{r,k,i} + \sum_k \sum_i \sum_{t'=t-\theta_{k,i}+1}^{h} N_{k,i,t'}\xi_{r,k,i} + \pi_{r,t} \quad \forall r,t$$

$$R_r^{lo} \le R_{r,t} \le R_r^{up} \quad \forall r,t$$



Sustainable & safe operations

- ✓ Energy efficient
- ✓ Optimal throughput
- ✓ Well maintained in time
- ✓ Safe operating conditions
- ✓ On-time production
- ✓ Knowledge-based models
- ✓ Agile and adaptive decisions

SINGPRO Targets
Create and prove novel concepts in real life

- Show that big data technologies can be deployed together with optimization strategies, to close the decision loop in automation
  - The results can help defining future research needs within systems-level integration of process control systems and data-driven decision making
- Collaborate with Finnish industry on piloting the methodology
  - Get access to real data, process information and the opportunity to discuss, test and demonstrate the solution approaches in practice
  - Create collaborative concepts that are re-usable across various industries

SINGPRO Project Team

Adj. Prof. Harjunkoski (Aalto CHEM)

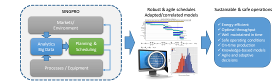
Prof. Heljanko (University of Helsinki)

Dr. Tewodros Deneke (University of Helsinki)

Dr. Teemu Ikonen (Aalto CHEM)

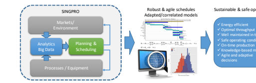
Dr. Hossein Mostafaei (Aalto CHEM)

Questions to be Answered (1/2)



- **Often a production plan is already ”old” soon after being rolled out to the plant floor**
  1. Could I do better planning by knowing more about the process, i.e. utilizing the real-time data?
- **Schedules are usually based on average durations (tables)**
  2. Is it better to dynamically generate accurate statistics on process behaviour every time I want to schedule?
- **Disturbances and breakdowns often come as a surprise**
  3. How many incidents can actually be predicted and avoided?

Questions to be Answered (2/2)



- **Often we focus on the most obvious data assuming simple causality**

4. What information actually is relevant for root-cause analysis? Are there hidden relationships?

- **Many decisions in optimization add to the complexity**

5. Are there decisions that can be excluded from the optimization scope, based on what we know from the data?

- **Data is mostly collected and stored only for troubleshooting**

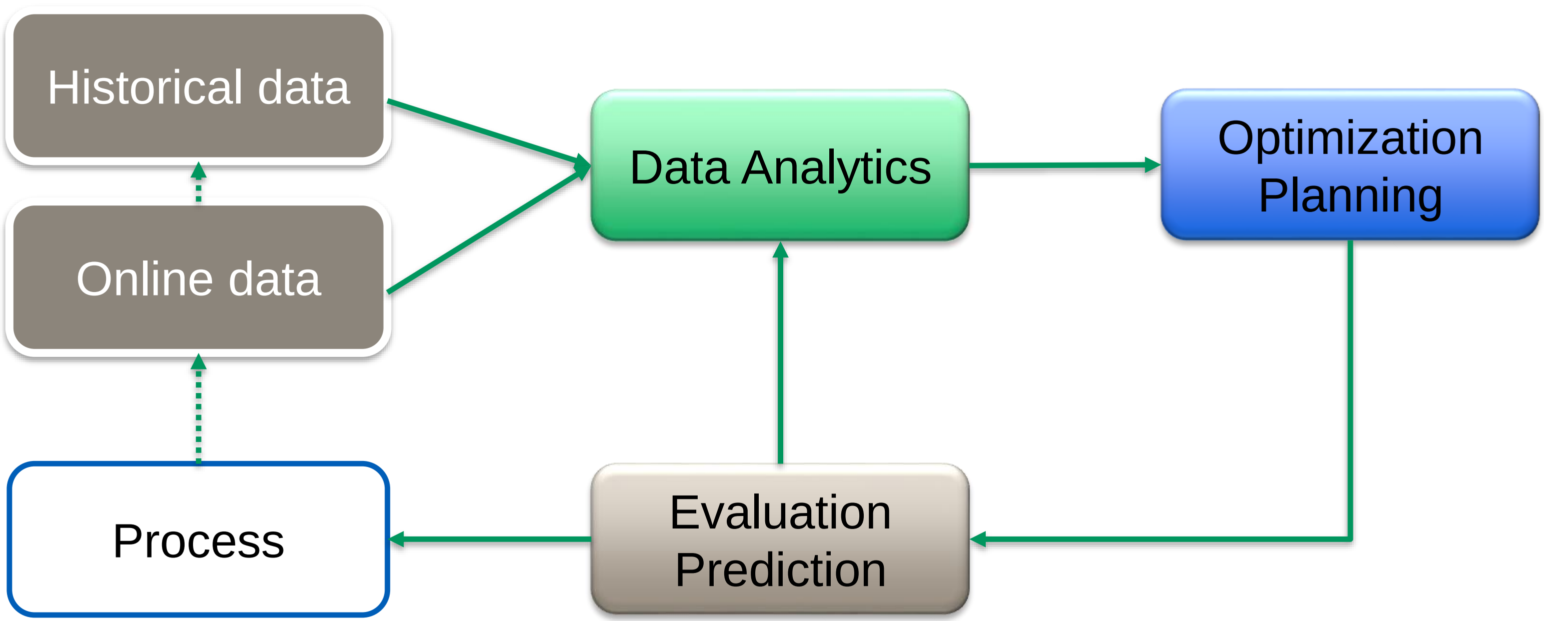
6. What is the actual value of this data?

SINGPRO Highlights
Combine big data analytics with optimization

Need to develop and focus on:

- Online, reactive and anticipative tools for sustainable and efficient operations
- Collaboration interfaces between scheduling optimization and big data analytics / machine learning resulting in more agile, self-aware and flexible decisions
- Combine first-principle models with machine learning in an efficient way to reduce the modeling complexity and efforts
- Create - in a fully data driven fashion - models of normal process behaviour and predictive models of process disruptions

Example Collaboration Loop
Process → Analytics → Optimization

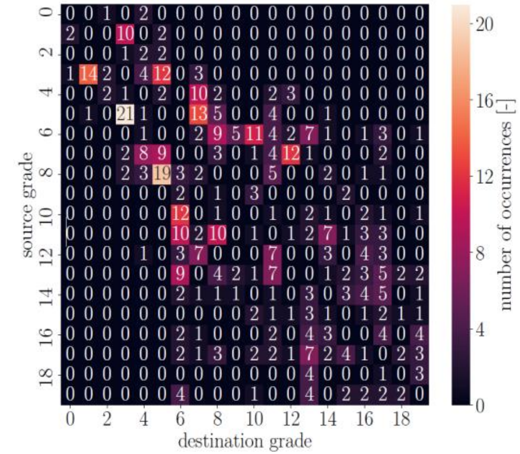


1. Could I do better planning by knowing more about the process, i.e. utilizing the real-time data?

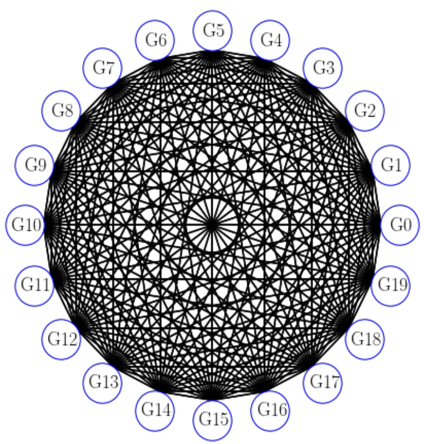
Data Driven Model for Grade Change in P&P Process

Combining data analytics and machine learning with a rigorous scheduling model in an integrated fashion

Heuristic constraints derived from the data analytics methods allow faster performance



Occurrence table obtained from historical data



(a)

$X_{i,r} = 1$ if production run r processes grade i

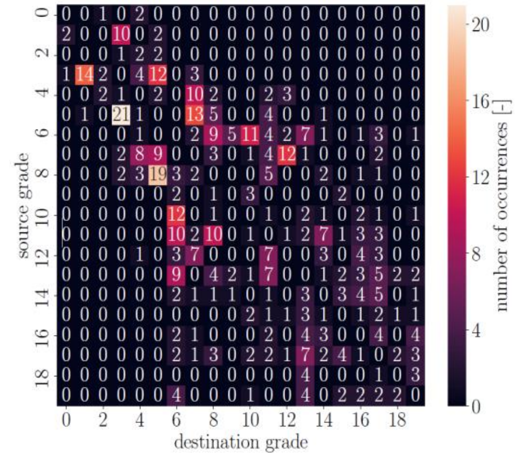
$$X_{i,r-1} \leq \sum_{i' \in I_i} X_{i',r} \quad \forall i \in I, r \in R|_{r\geq 2}$$

I_i = Set of grades that can follow grade i

Data Driven Model for Grade Change in P&P Process

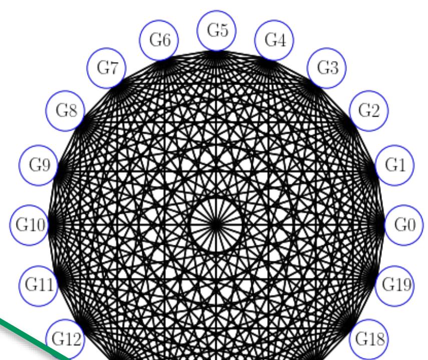
Combining data analytics and machine learning with a rigorous scheduling model in an integrated fashion

Heuristic constraints derived from the data analytics methods allow faster performance



Occurrence table obtained from historical data

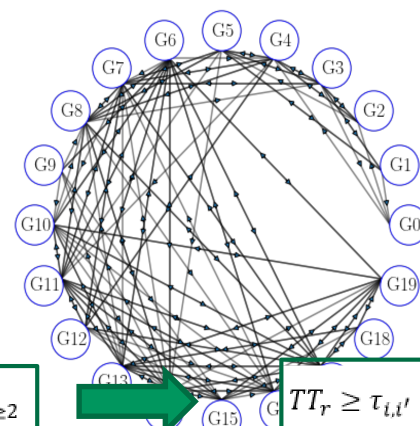
Parameter: Transition time between grades *i* and *i'*



$$TT_r \geq \tau_{i,i'}(X_{i',r-1} + X_{i,r} - 1) \quad \forall i, i' \in I, r \in R|_{r\geq 2}$$

(a)

Historical data
Data analytics



(b)

$$X_{i,r-1} \leq \sum_{i' \in I_i} X_{i',r} \quad \forall i \in I, r \in R|_{r\geq 2}$$

$$TT_r \geq \tau_{i,i'} \times (X_{i',r-1} + X_{i,r} - 1) \quad \forall i \in I, i' \in I_i, r \in R|_{r\geq 2}$$

Considerable reduction in the number of constraints

Case Study Results

Two weeks: due date 1= 168 h, due date 2= 336 h for 20 grades

Objective: minimize grade change transition time and production runs

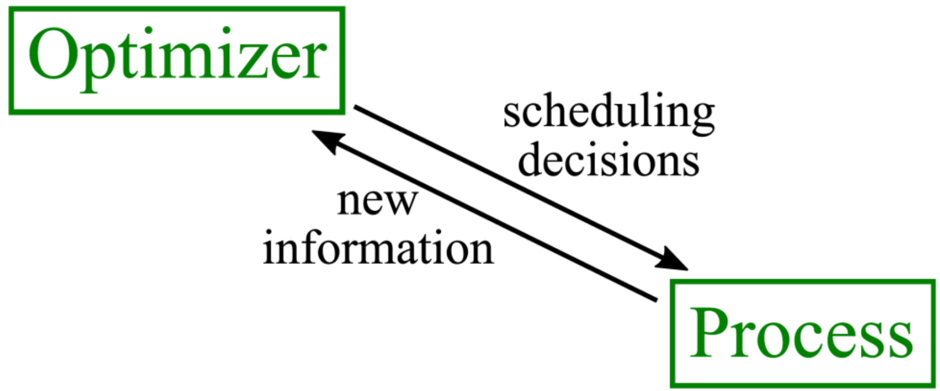
	Full-space		Data-driven	
#Production runs	16	17	16	17
CPU(s)*	18000	18000	4143.4	6974.6
#Constraints	8773	9340	5263	5596
Objective ($)	58536.1	50613.8	53988.8	50338.8 (0.5% ↓)
Relative gap	76.08	72.14	0	0

*GAMS/CPLEX 12.7.1 (Intel i5-7300U, 2.60 GHz, 8 GB of RAM, Windows 10, 64-bit)

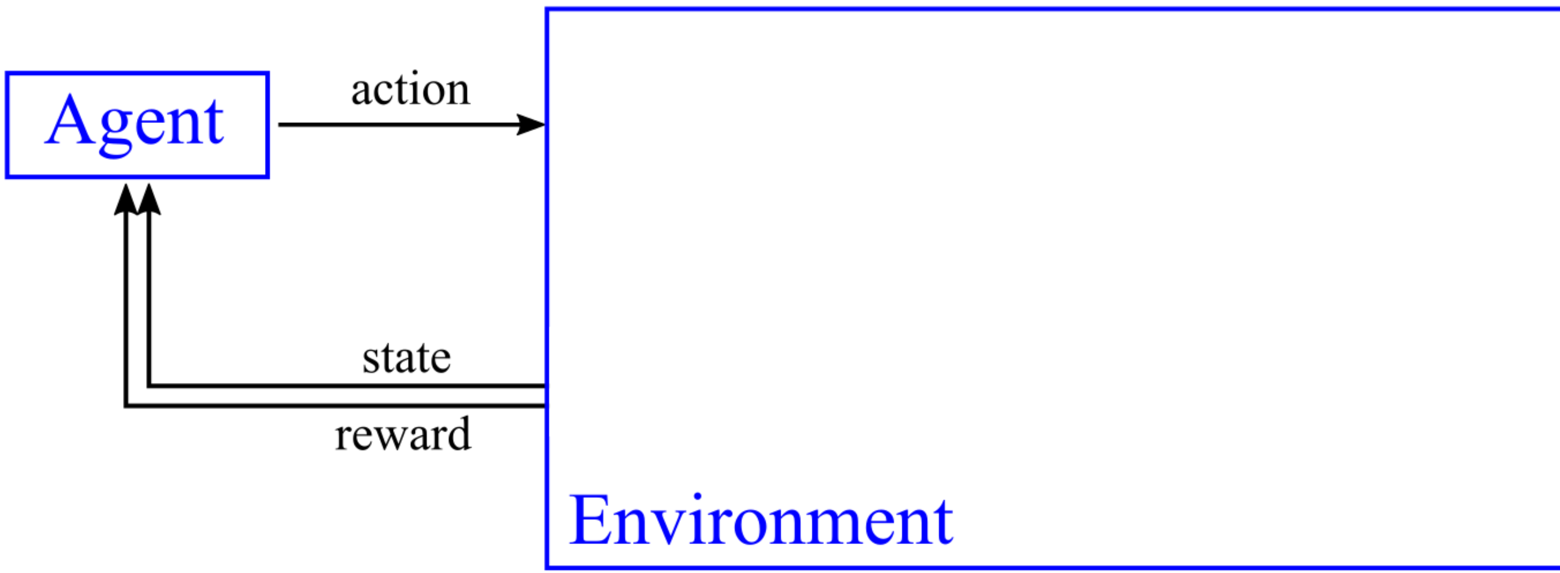
Reinforcement Learning (RL) of Online Rescheduling Decisions

Questions:

- When to trigger a new rescheduling procedure?
- Mathematical programming or a heuristic algorithm?
- How far ahead to schedule (i.e. horizon length)?
- How much computing resource to allocate?

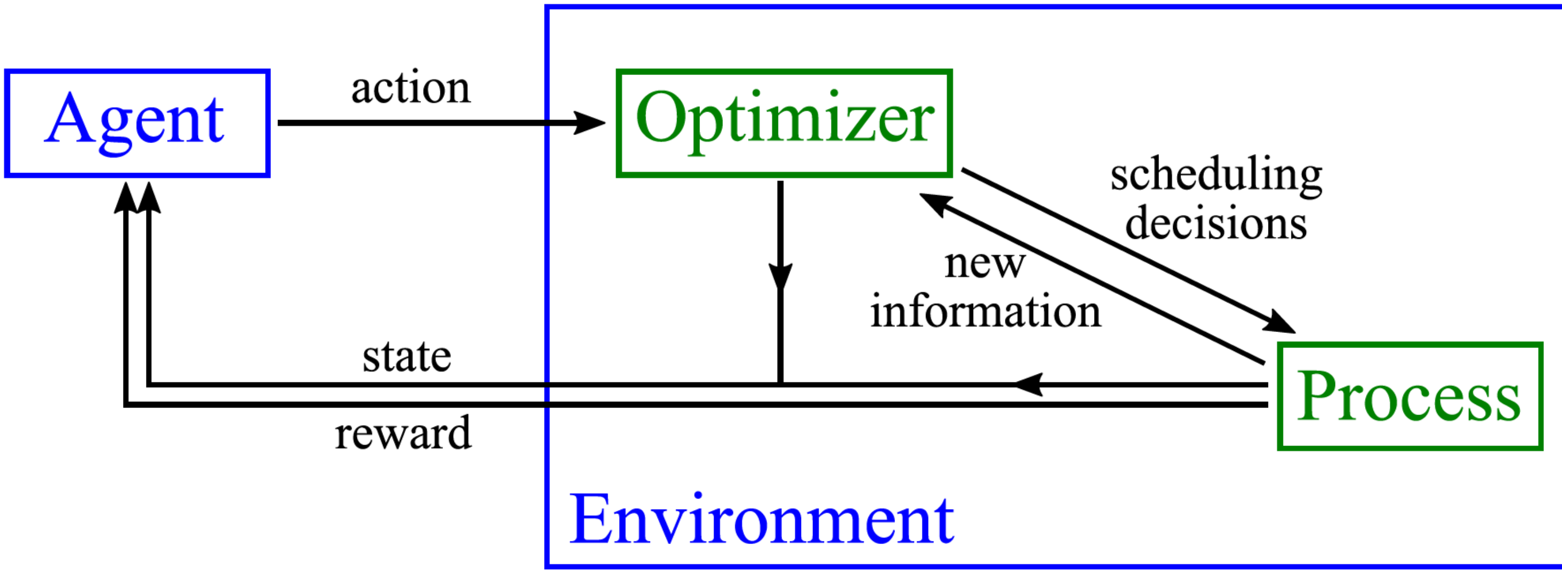


Reinforcement Learning (RL) of Online Rescheduling Decisions



Aalto University
School of Chemical
Engineering

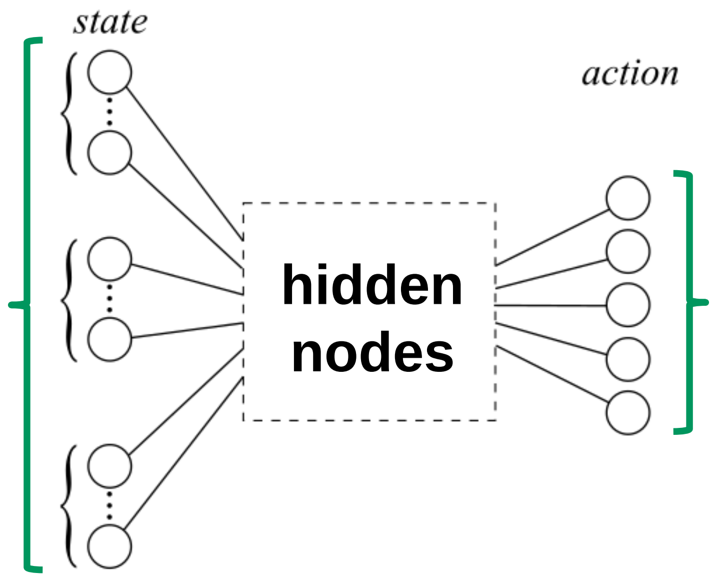
Reinforcement Learning (RL) of Online Rescheduling Decisions



Aalto University
School of Chemical
Engineering

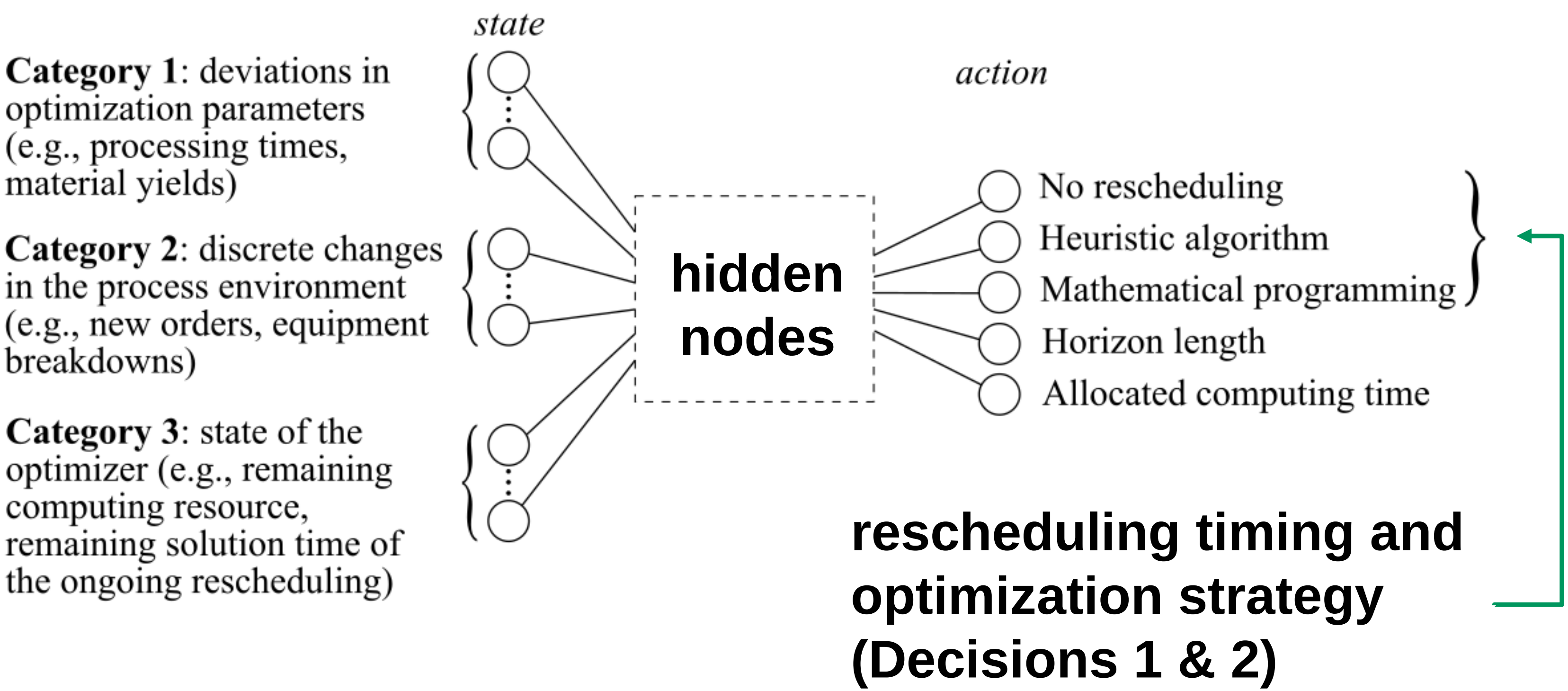
State and Action Spaces

changes in the environment (i.e. the process and the optimizer



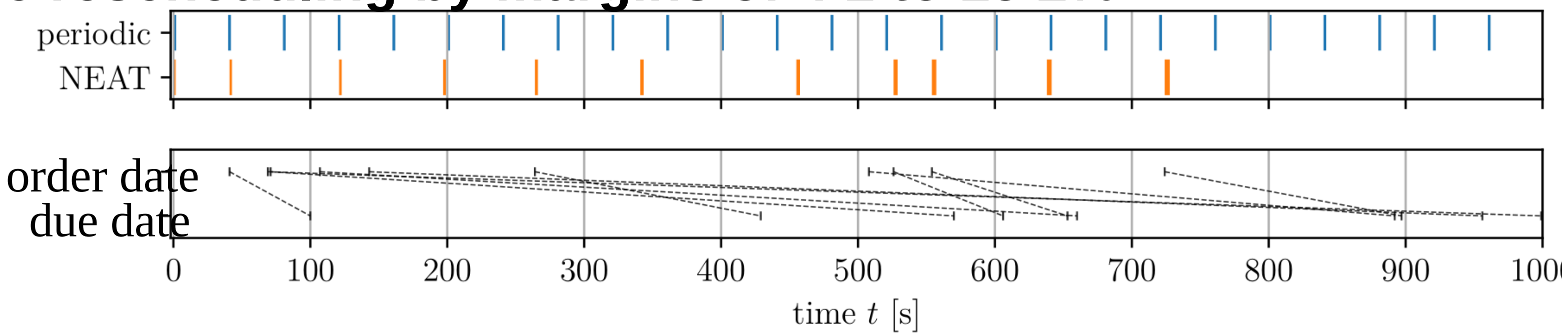
rescheduling decisions

State and Action Spaces



First Results

- **RL algorithm: Neuroevolution of Augmenting Topologies (NEAT) (Stanley and Miikkulainen, 2002)**
- **Simplified decision space:**
    - Rescheduling timing
    - Computing resource allocation
- **On three test cases, better closed loop schedules than by periodic rescheduling by margins of 4.1 to 15.2%**



2. Is it better to dynamically generate accurate statistics on process behavior every time I want to schedule?

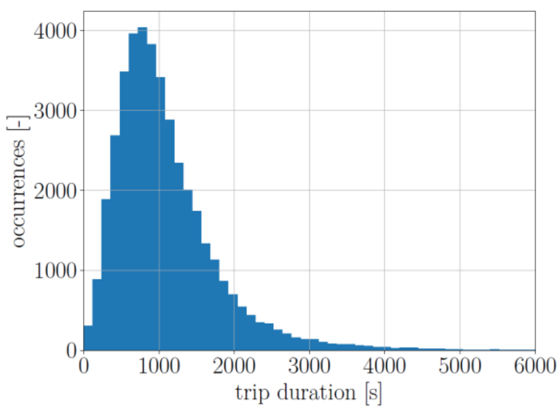
Scheduling problem based on the NYC taxi data

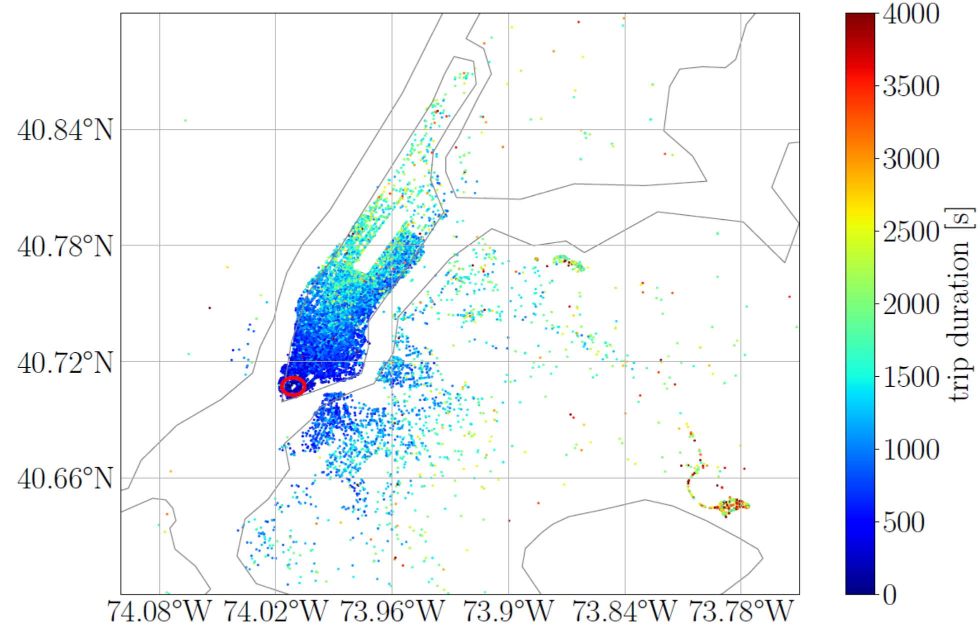
Openly available dataset at https://www.kaggle.com/c/nyc-taxi-trip-duration

Data of over 1.4 million taxi tips in NYC

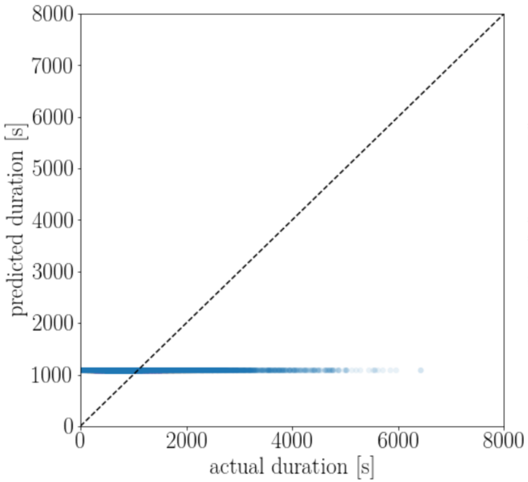
The ground truth is the trip duration

Examples of features: Passenger count, pick-up and drop-off dates, time and coordinates

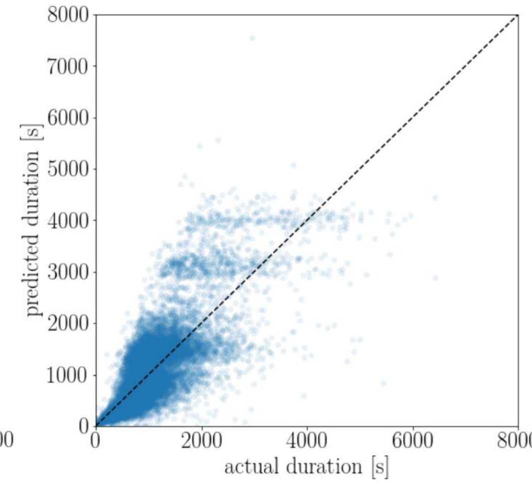




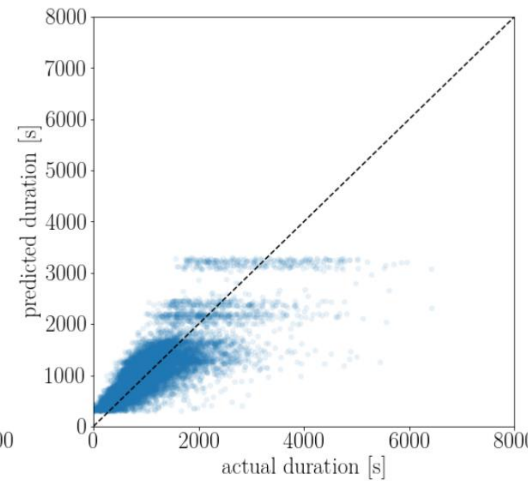
Scheduling problem based on the NYC taxi data



1) Average model



2) Rate model

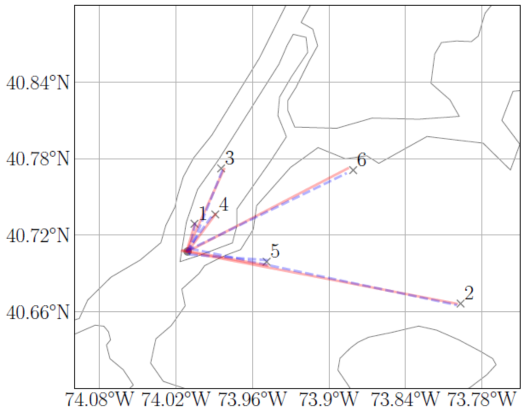


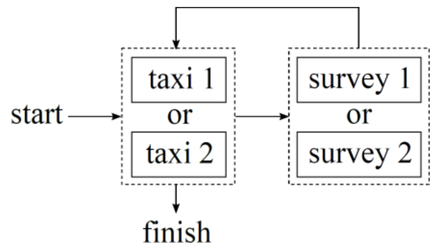
3) Gaussian process (GP) model

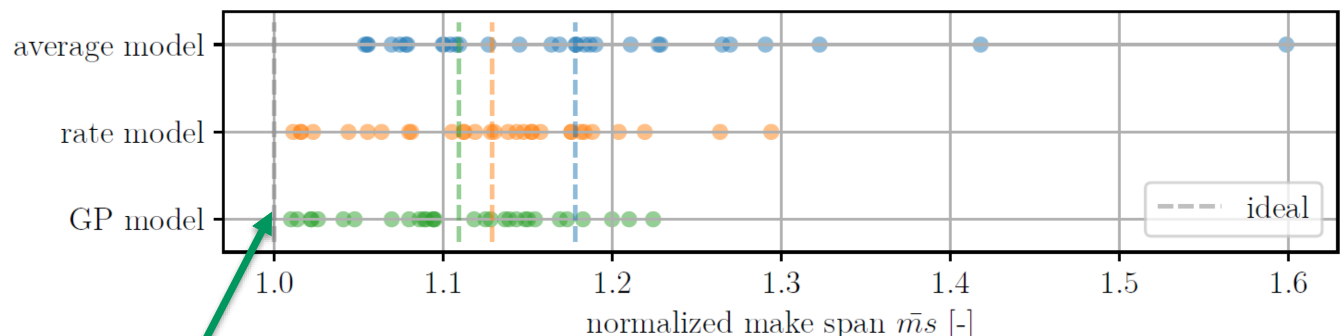
prediction model	RMSE	r^2
average	701.8	-2.37e-4
rate	556.5	0.371
GP	438.4	0.610

RMSE = the root mean square error
r^2 = the coefficient of determination

Scheduling problem based on the NYC taxi data



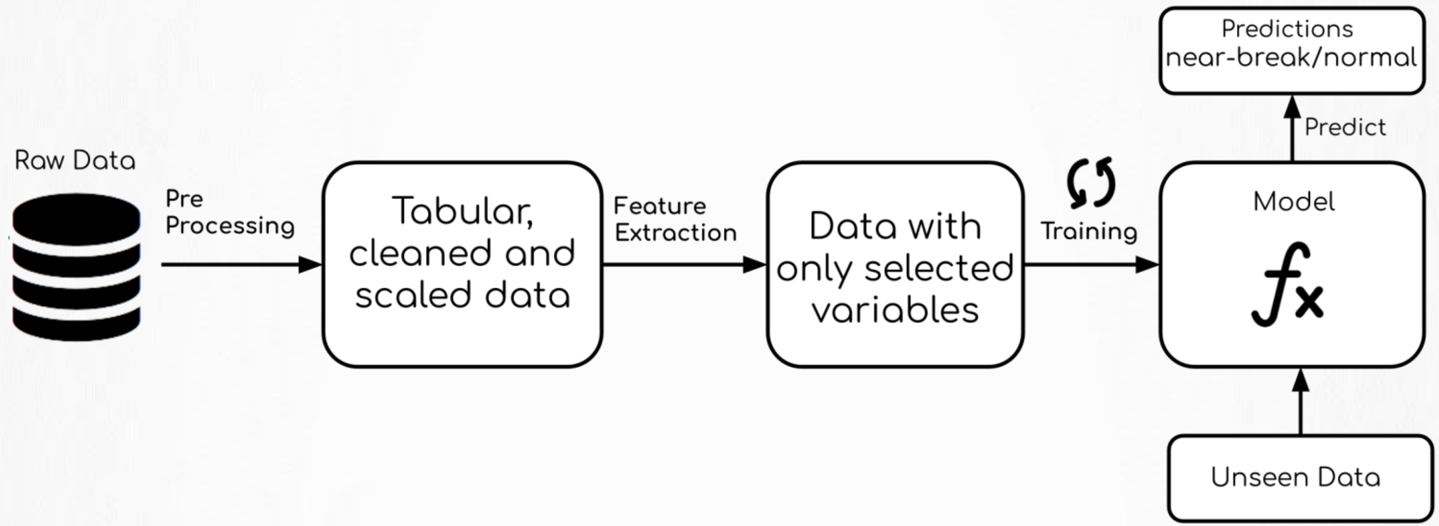




The reference model is the *ideal predictor*, for which $r^2 = 1$ and RMSE $= 0$.

3. How many incidents can actually be predicted and avoided?

OVERVIEW: BREAK PREDICTION MODELING



PREPROCESSING: DATA REPRESENTATION

- Ideally we want the data be arranged as:

	sensor 1	sensor 2	...	sensor n	status
ts 1					
ts 2					
ts 3					
...					
ts n					

DATASET DECOMPRESSION

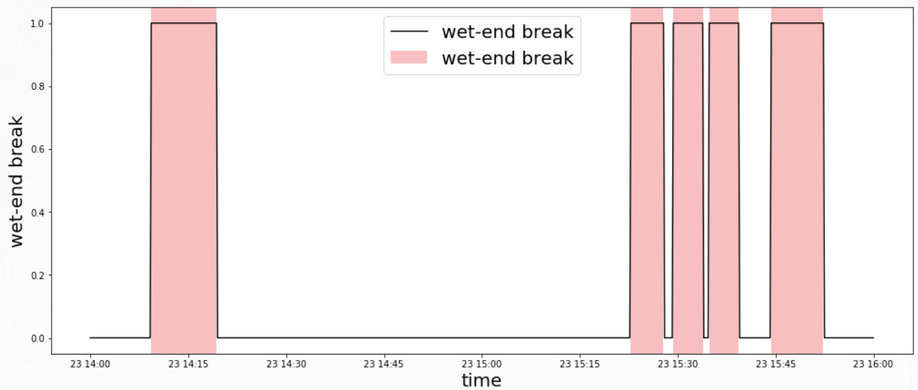
- Plant data dump is delta-compressed as:

tag 1	val 1	ts 1	tag 2	val 2	ts 2	...	tag n	val n	ts n

- Variable no. columns at different timestamps
- Around 3.5M tags over a day, 322M over two months and 1.4B over a year
- More than 1000 sensors
- **Translate data representation**
- **Decode ground truth** (status of the machine)
- **Null values, Different sampling rates, etc.**

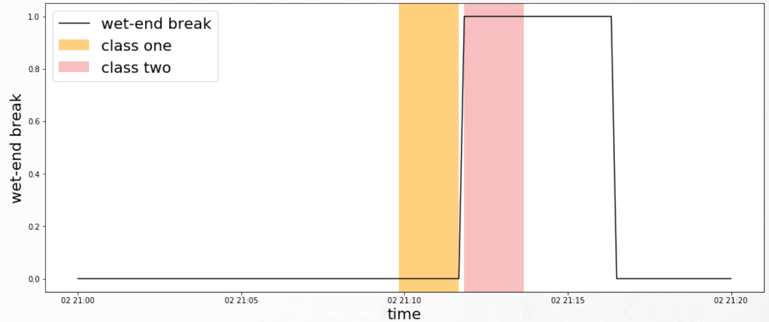
PREPROCESSING

- Next data preprocessing steps:
  - Cleaning: removing secondary breaks caused by a primary one
  - Resampling: to have a common sampling rate among signals
  - Slicing: selecting near-break and normal operation regions
  - Scaling: balancing differences in amplitude

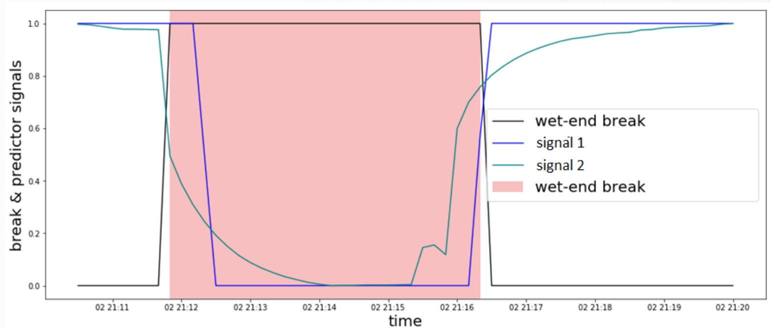


PREDICTION

- Model 1: Can we distinguish break from non-break?
  - Samples that are before a break are used as class one
  - Samples that are during the break are used as class two
  - The task is to predict (distinguish) these two classes

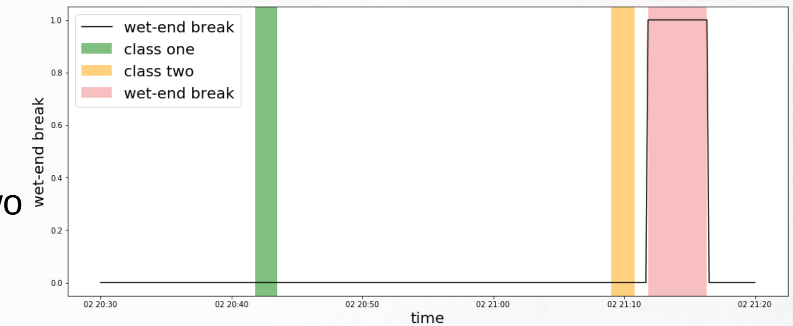


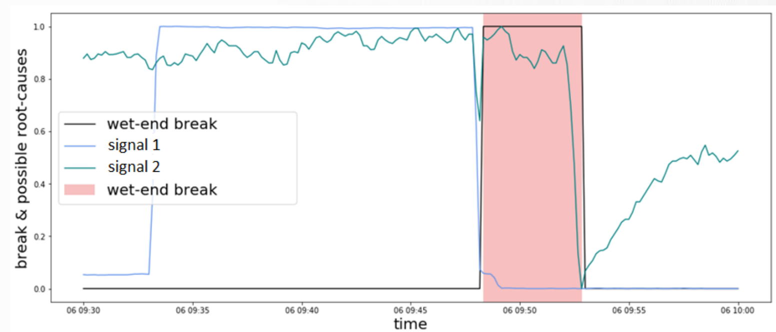
- Result:
  - Trivial: 96.7% accuracy
  - Highly influenced by signals highly correlated with break signal
  - Fails to identify root-causes
  - Not valuable in practice



PREDICTION

- Model 2: Can we distinguish near-break from normal?
  - Samples that are further away from a break are used as class one
  - Samples that are just before a break are used as class two
  - The task is to predict (distinguish) these two classes
- Result:
  - Fairly low prediction accuracy
  - Identifying main features used in prediction: Root-causes
  - Potential root-causes need domain expert analysis
  - Can still pick up highly correlated signals (often these are data cleaning problems)
  - Idea: Try to identify signals that have time varying correlations, and use only them in prediction

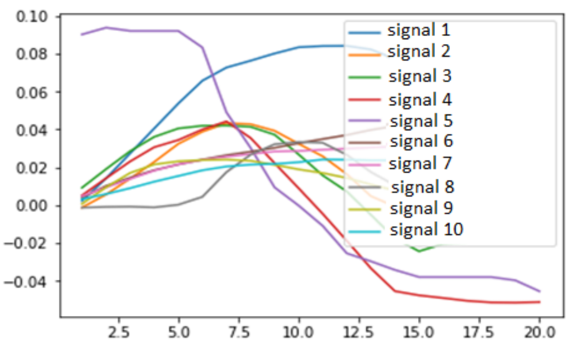




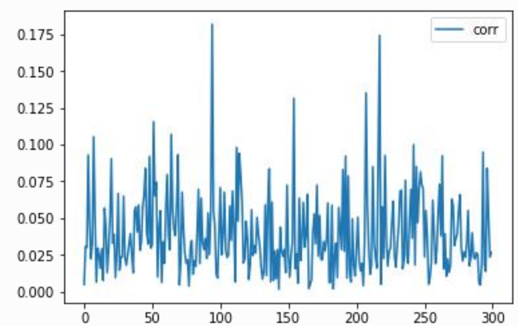
CORRELATION BASED FEATURE EXTRACTION

- Calculate correlation of all signals to wet-end break at various lags (0 to 20 min)
- Remove highly correlated signals at lag zero (probably not causes of the break)
- Keep signals that show the most increasing correlation trend (sorted at lag 8 min)
- Just use the top 20 signals in the prediction model: Improves accuracy

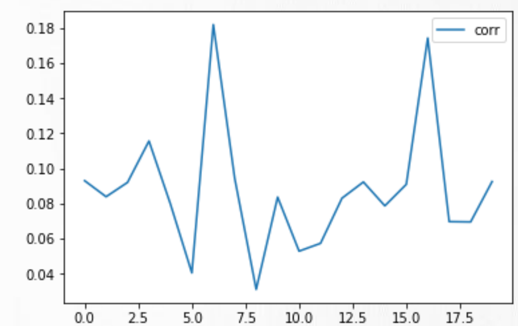
Correlation trend at various lags



Maximum correlation values out of 20 lags for top 300 interesting signals

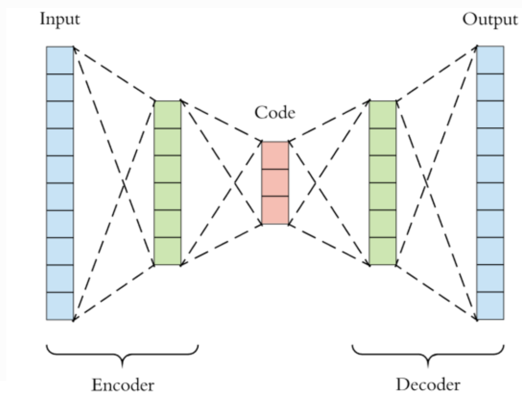


Maximum correlation values out of 20 lags for top 20 interesting signals



LATENT FEATURE WITH AUTOENCODERS

- Ongoing work
- Why the need?
  - Data understanding
  - Latent variables and dynamics discovery
  - Automatic data clean-up before prediction modeling
- How
  - Train an Autoencoder (approx. lossy compression) and visualize the encoder output



FINDINGS

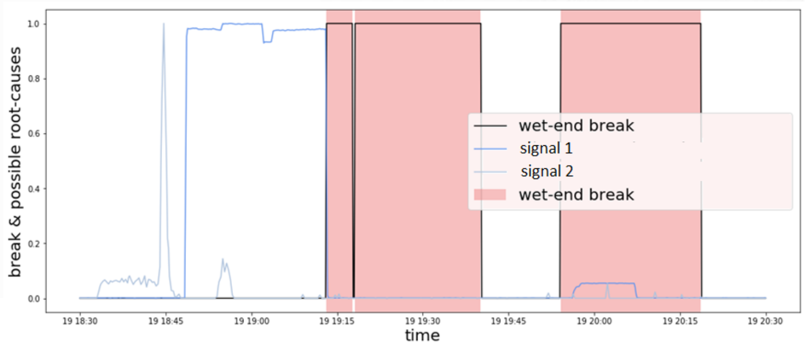
- Data cleaning is difficult without domain expert help
  - Breaks cause breaks: A break can be more easily predicted if it is preceded by a break
- Prediction accurary is still low: Very complex process with 1000+ sensors
- Machine learning algorithms can provide root cause candidates
  - Evaluating root cause candidates requires a lot of domain expertise
  - Many data cleanup problems were detected by looking at candidate root causes
    - Example: Break signal is sampled at low frequency, which was detected only by looking at highly correlated root cause

4. What information actually is relevant for root-cause analysis? Are there hidden relationships?

ROOT-CAUSE ANALYSIS

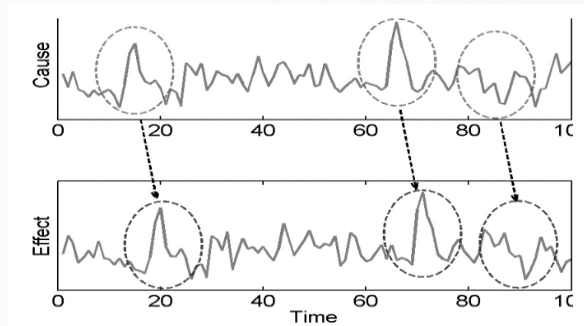
- Challenges
  - Identifying possible root-cause signals
  - Identifying delays between root-cause signals and break
  - Identifying and removing highly correlated signals (with no delay) to the break
    - A: Using the features reported to be used by the machine learning model, or
    - B: Using Granger Causality.

- Result:
  - A few interesting possible root-causes discovered
    - Pressure x
    - Fluid level y
    - Quality z



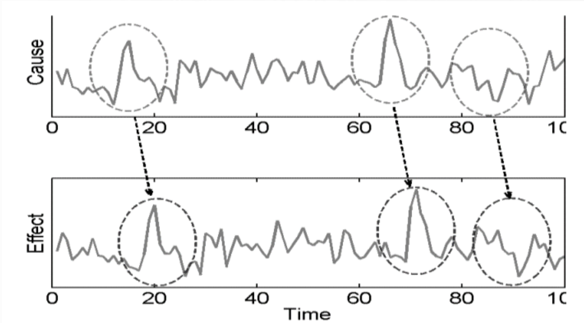
CORRELATION VS CAUSALITY

- (Granger) causality quantifies the extent to which one time series is predictive of another.
- Most classical methods of estimating Granger causality assume linear timeseries.
- More recent approaches of estimating Granger causality for nonlinear timeseries are not generic.
- Applicable for increasing predictive model interpretability, feature selection and root cause analysis (e.g. what caused a paper machine break?)



BACKGROUND & PROBLEM FORMULATION

- Granger causality
  - Cause is prior to effect.
  - The cause makes unique changes in the effect (i.e. contains unique info. about it).
  - If a signal ***X*** "Granger-causes" (or "G-causes") a signal ***Y***, then past values of ***X*** should contain information that helps predict ***Y*** above and beyond the information contained in past values of ***Y*** alone.



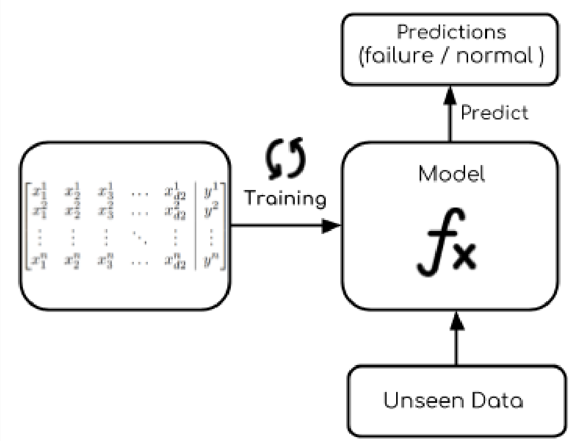
$$
\begin{aligned}
Y(t) &= \sum_{l=1}^{L} a_l Y(t-l) + \epsilon_1 \\
Y(t) &= \sum_{l=1}^{L} a'_l Y(t-l) + \sum_{l=1}^{L} b'_l X(t-l) + \epsilon_2,
\end{aligned}
$$

BACKGROUND & PROBLEM FORMULATION

- Quantifying Granger causality estimation
    - Accuracy difference between a model with or without a lagged version of a predictor variable.
    - But that means retraining our model ***no. lag * no. predictors*** times. Any better solution?
    - What if we permute or apply noise on each predictor during prediction (evaluation) instead of fully dropping it and retraining?
    - The features for which such permutation causes accuracy of prediction to drop seem to be used by the machine learning model to predict accurately
    - This gives us an alternative way to find root causes

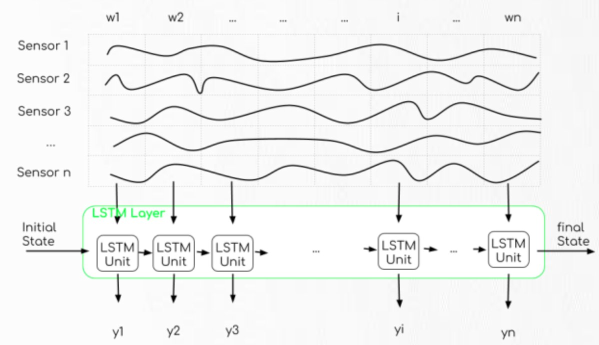
$$Y(t) = \sum_{l=1}^{L} a_l Y(t-l) + \epsilon_1$$

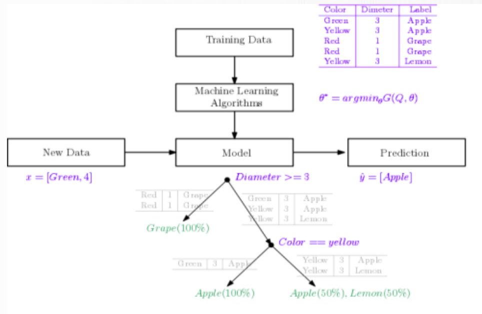
$$Y(t) = \sum_{l=1}^{L} a'_l Y(t-l) + \sum_{l=1}^{L} b'_l X(t-l) + \epsilon_2,$$



DEPLOYED LEARNING ALGORITHMS

- LSTM
  - Neural nets which have memory and feedback
  - Can capture trends more easily
  - Can be used to learn sequence to sequence problems
- Random forest
  - Constructs multiple decision trees on training data
  - Generates a model in parallel
  - Combines the results of all decision trees
  - Can provide information on feature importance
- Cross correlation and Granger causality

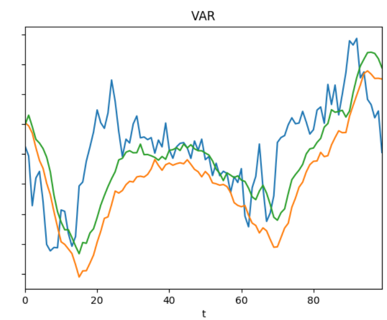




SIMULATION EXPERIMENTS

- Linear Vector Autoregressive Model
  - Simulated VAR model with known sparse ***A*** matrix.
  - Make prediction of each variable using all others as predictors.
  - Permute each predictor variable and make prediction again.
  - Calculate variable dependency based on reduction in accuracy (R2).

$$x_t = \sum_{k=1}^{K} A^{(k)} x_{t-k} + e_t,$$



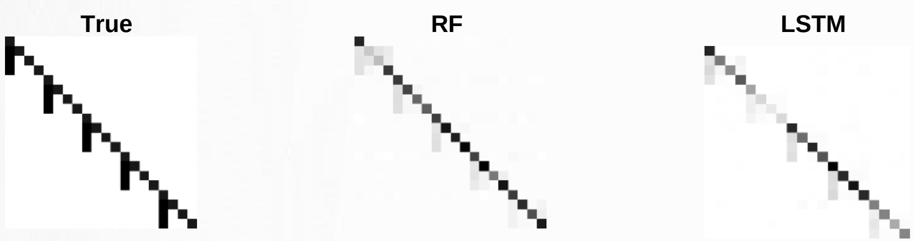


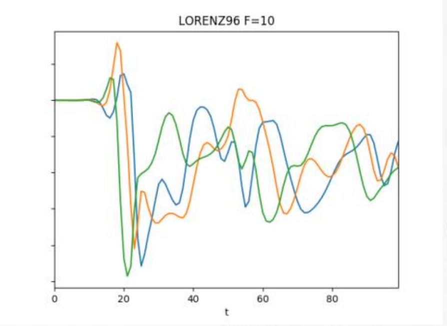
Table 2: VAR results (AUROC)

T	500	1000	5000
RF	97.62	99.80	100.0
LSTM	85.72	92.03	100.0

SIMULATION EXPERIMENTS

$$\frac{dx_i}{dt} = (x_{i+1} - x_{i-2})x_{i-1} - x_i + F$$

- Nonlinear Lorenz96 Model
  - Simulated data using Lorenz96 model with a given forcing value. Such model is often used to model climate dynamics.
  - Make prediction of each variable using all others as predictors.
  - Permute each predictor variable and make prediction again.
  - Calculate variable dependency based on reduction in accuracy (R2).



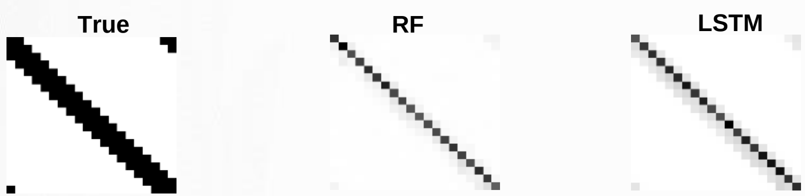


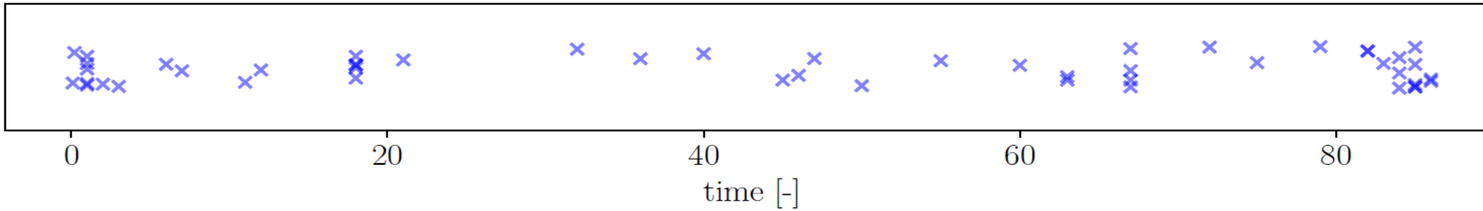
Table 1: Lorenz96 results (AUROC)

F	10	10	40	40
T	500	1000	500	1000
RF	98.46	99.95	88.44	98.89
LSTM	80.82	95.85	71.41	73.65

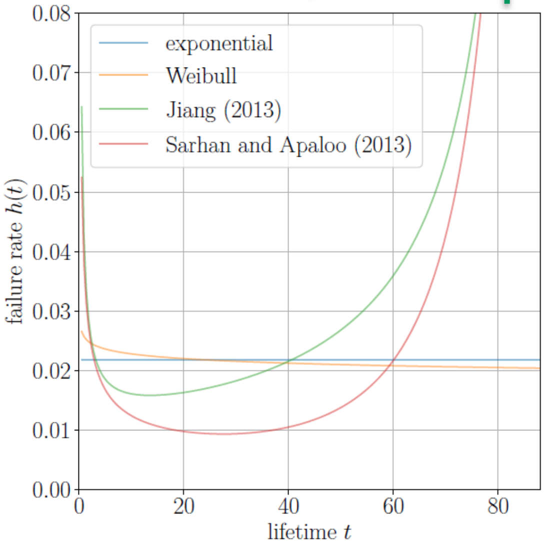
5. Are there decisions that can be excluded from the optimization scope, based on what we know from the data?

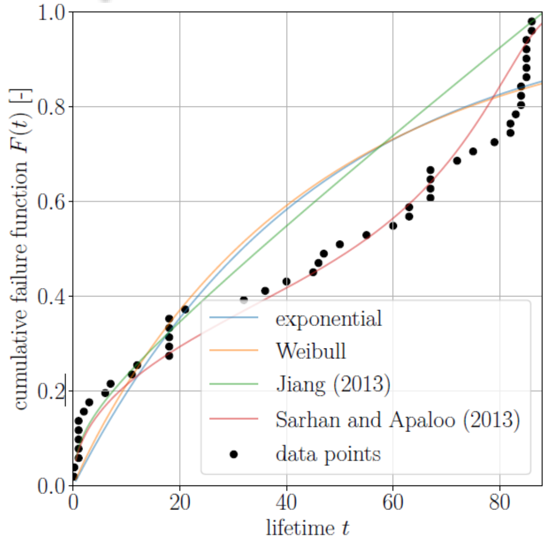
Selective maintenance optimization

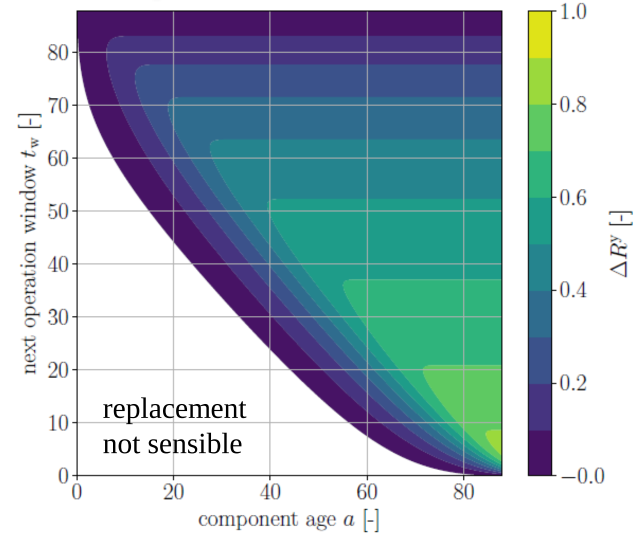
Experimental data of failure times (Aarset, 1987)



Sarhan and Apaloo (2013)



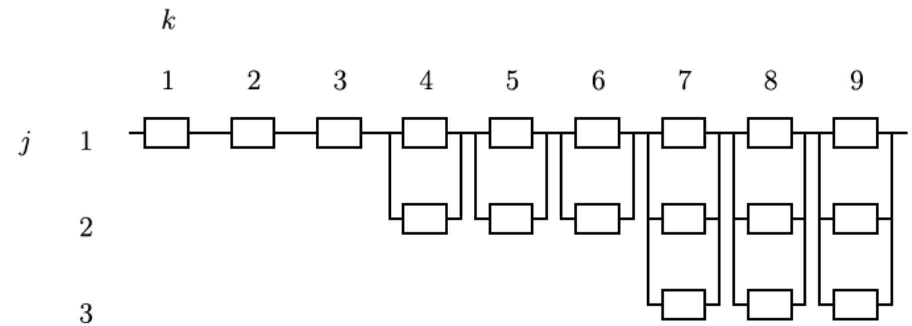




Selective maintenance optimization

- **The objective is to maximize the reliability of the system**
- **Subject to maintenance time and budget constraints**
- **We include the following bound:**

$$y_{k,j} = 0, \quad \forall k, j \in \{(k,j) | \Delta R^{y}_{k,j} \leq 0\}$$



$y_{k,j}$	**Binary variable defining whether unit (k, *j*) is replaced**
$\Delta R^{y}_{k,j}$	**Improvement in realiabity of the unit (k, *j*), if component is replaced**

→ An order of magnitude reduction in solution time

6. What is the actual value of this data?

Conclusions

- **We can see clear benefits from using more advanced methods to process historical / on-line data**
- **Applying AI/ML methods can**
  - *Improve the accuracy of scheduling*
  - *Improve the predictability of processes*
  - *Help reducing the search domain of large-scale problems*
- **The data-related work is still very problem-specific**
  - *A generic ”cookbook” still missing to reduce the efforts*
- **The value of data is difficult to estimate (industry looking into this) partially due to lack of access to business figures**
- **Collaboration across discplines is a prerequisite for success**

Break Prediction Conclusions

- **The paper machine breaks are difficult to predict**
- **Domain expert help is needed in many cases of the problem**
  - Specifying the right question to predict
  - Specifying data cleaning principles
  - Evaluating potential root causes
    - *Debugging machine data cleaning*
    - *Indentifying potential true root causes*
- **Machine learning cannot be done successfully without access to domain experts**

Acknowledgement

The financial support from Academy of Finland, through project SINGPRO, is gratefully acknowledged.

ACADEMY OF FINLAND

https://singpro.github.io/pages/about.html