Aalto University
School of Chemical
Engineering

HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI

singpro

Combining Data Analytics and Scheduling – First Results and Open Challenges

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

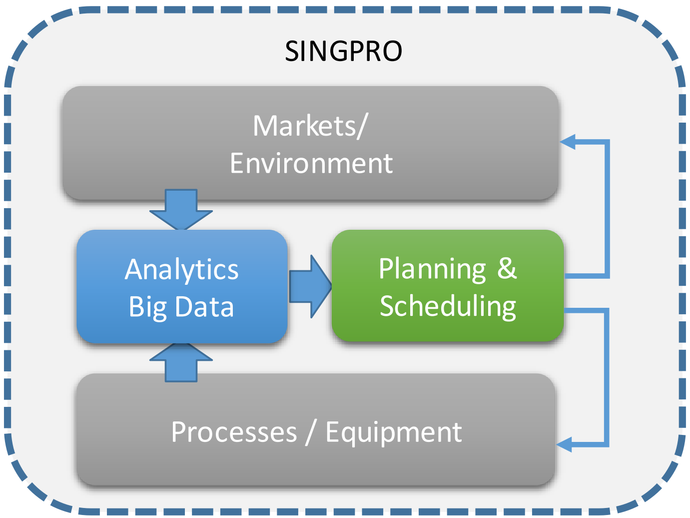
[1)]Aalto University, [2)]ABB Power Grids Germany

AIChE Annual Meeting, Orlando, FL, 14.11.2019

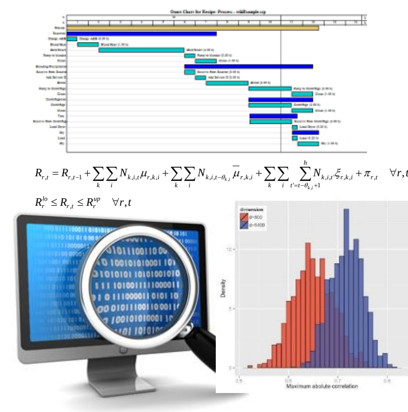
2019 AIChE® ANNUAL MEETING
19AIChE Annual Meeting, Orlando, FL
Transforming the Future through Chemical Engineering
November 10-15, 2019
Hyatt Regency, Orlando

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



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)

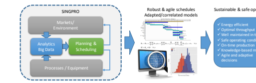
Prof. Heljanko (University of Helsinki)

Dr. Tewodros Deneke (University of Helsinki)

Dr. Teemu Ikonen (Aalto CHEM)

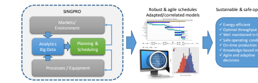
Dr. Hossein Mostafaei (Aalto CHEM)

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)



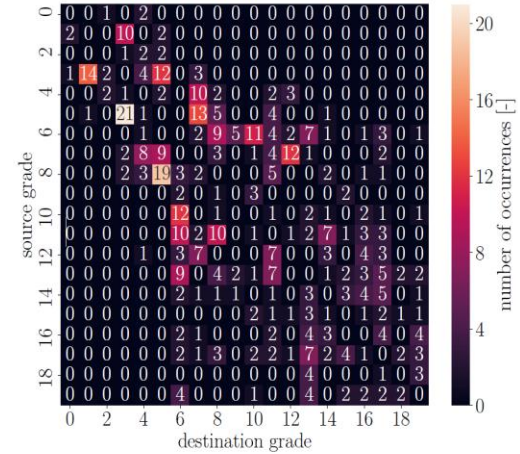
- **Often we focus on the most obvious data assuming simple causality**
  4. What information actually is relevant for root-cause analysis? Are there hidden relationships?
- **Many decisions in optimization add to the complexity**
  5. Are there decisions that can be excluded from the optimization scope, 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?

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

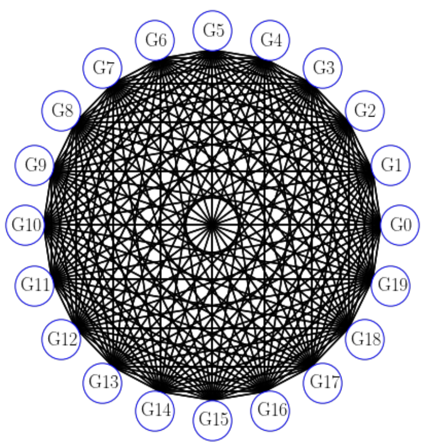


Occurrence table obtained from historical data

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

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

I_i = Set of grades that can follow grade i

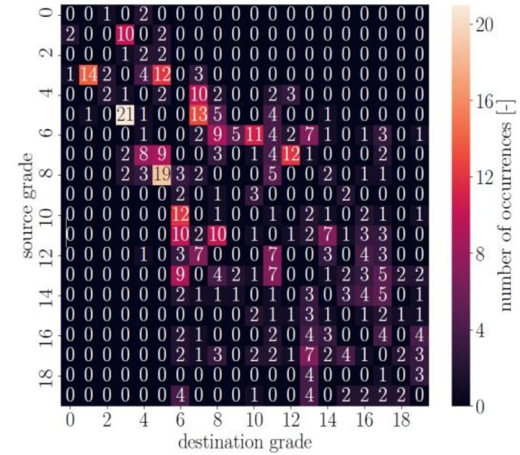


(a)

Data Driven Model for Grade Change in P&P Process

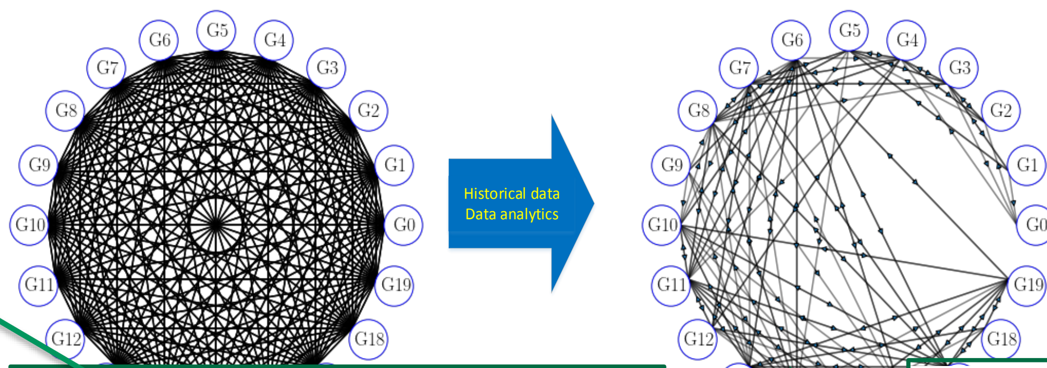
Combining data analytics and machine learning with a rigorous scheduling model in an integrated fashion

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)

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

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

(b)

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

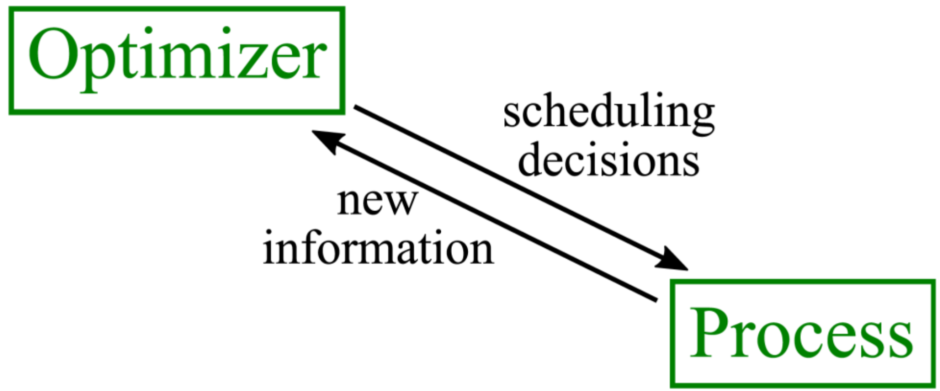
	Full-space		Data-driven	
#Production runs	16	17	16	17
CPU(s)*	18000	18000	4143.4	6974.6
#Constraints	8773	9340	5263	5596
Objective ($)	58536.1	50613.8	53988.8	50338.8 (0.5% ↓)
Relative gap	76.08	72.14	0	0

*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

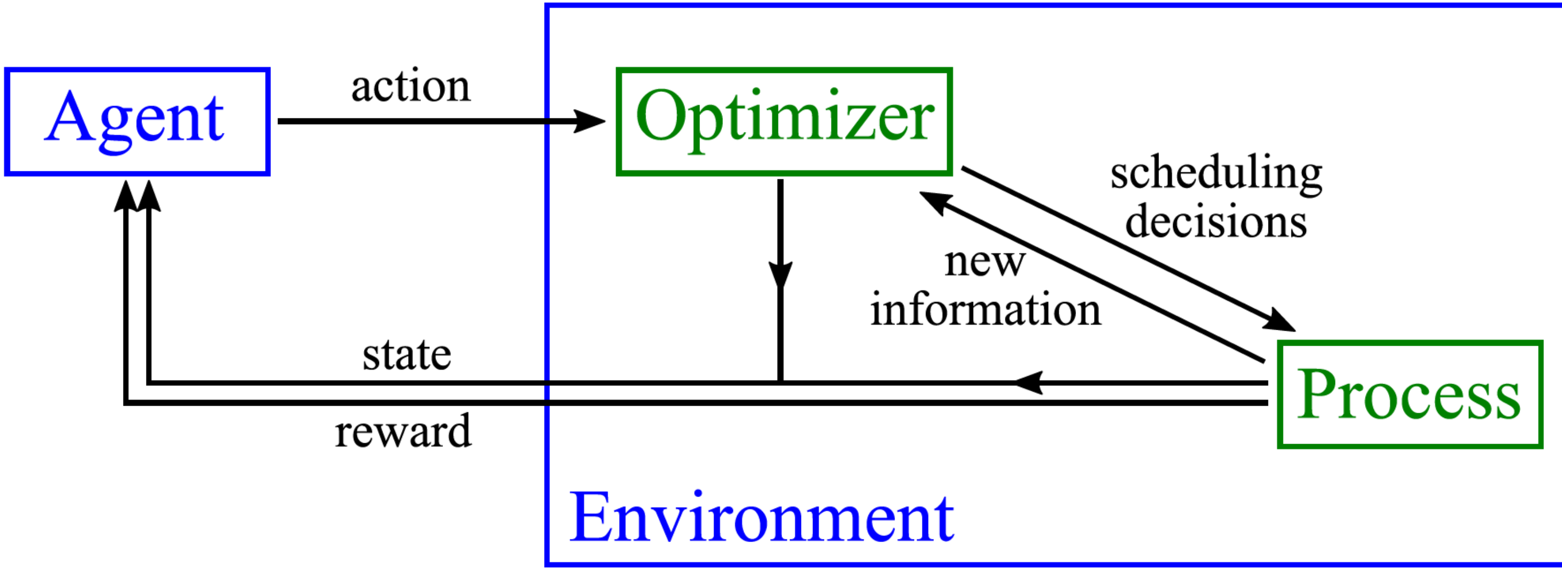
Questions:

- 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?



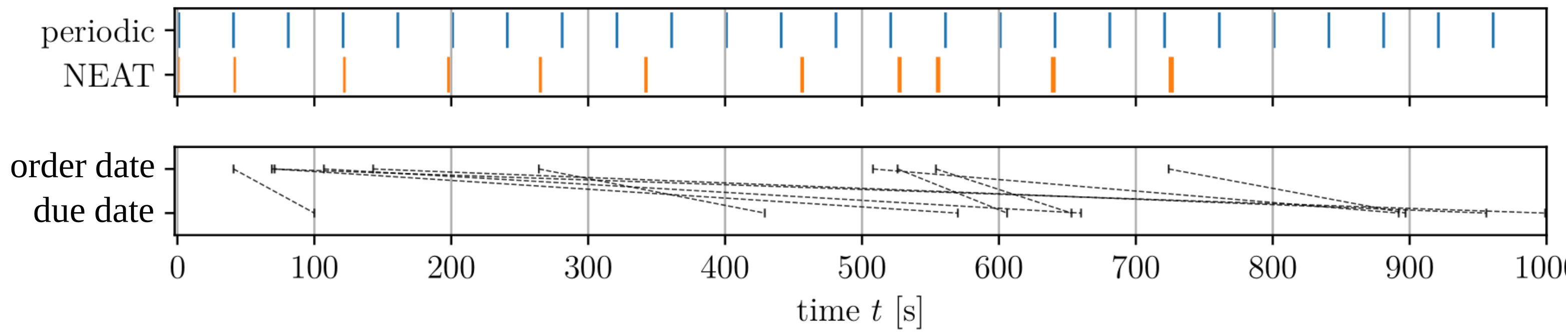
Aalto University
School of Chemical
Engineering

Reinforcement Learning (RL) of Online Rescheduling Decisions



RL of rescheduling: 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?

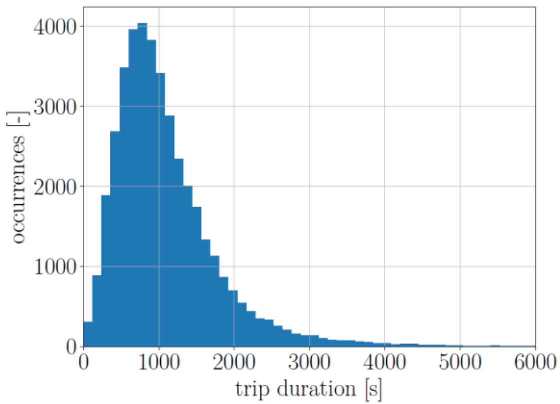
Scheduling problem based on the NYC taxi data

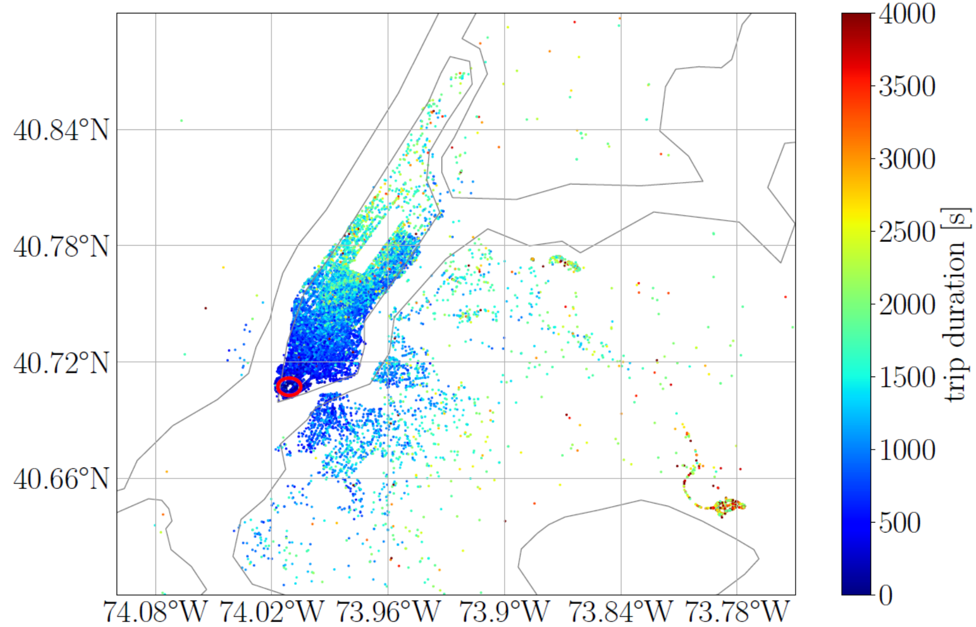
Openly available dataset at https://www.kaggle.com/c/nyc-taxi-trip-duration

Data of over 1.4 million taxi tips in NYC

The ground truth is the trip duration

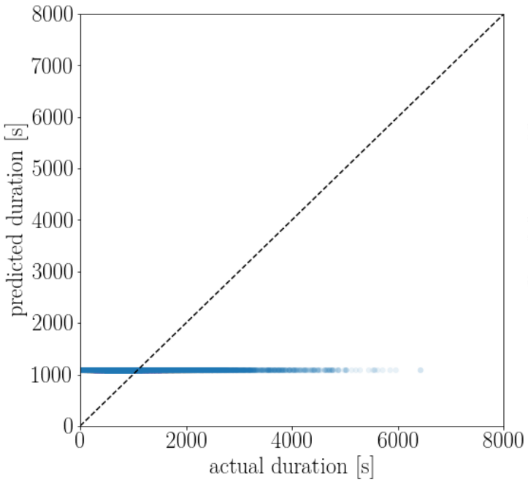
Examples of features: Passenger count, pick-up and drop-off dates, time and coordinates



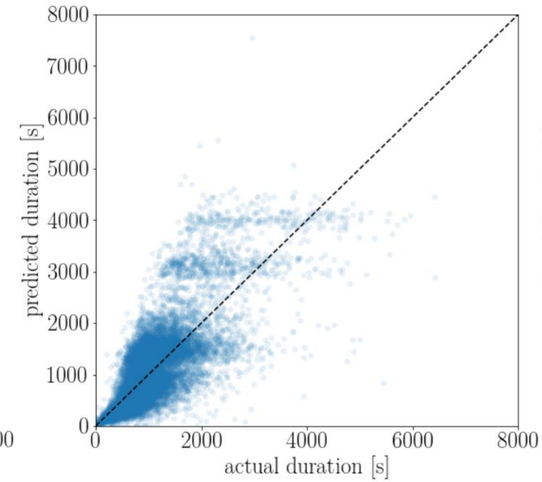


Aalto University
School of Chemical
Engineering

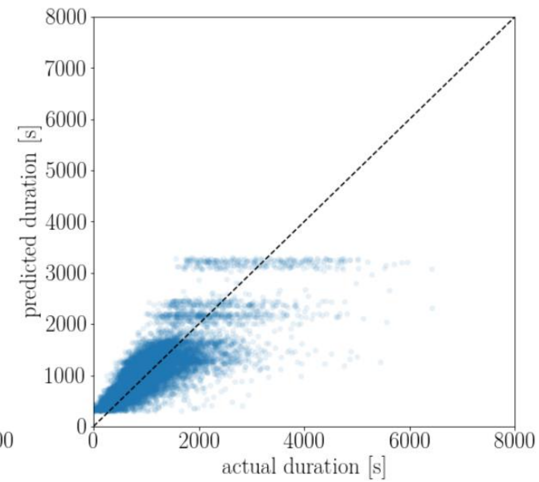
Scheduling problem based on the NYC taxi data



1) Average model



2) Rate model



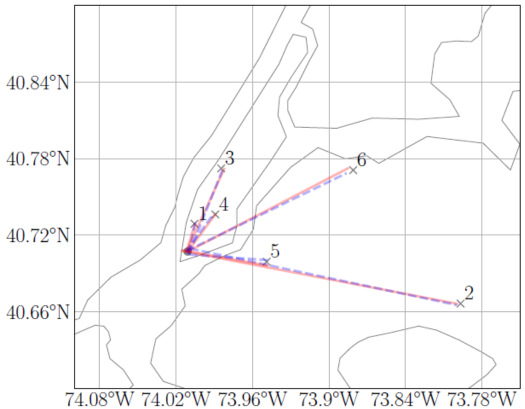
3) Gaussian process (GP) model

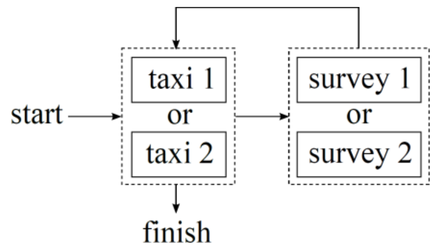
prediction model	RMSE	r^2
average	701.8	-2.37e-4
rate	556.5	0.371
GP	438.4	0.610

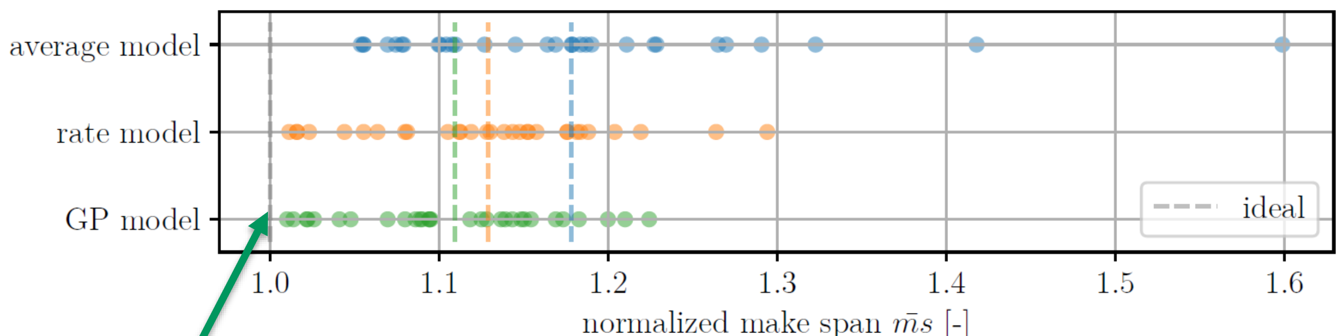
RMSE = the root mean square error
r^2 = the coefficient of determination

Aalto University
School of Chemical
Engineering

Scheduling problem based on the NYC taxi data







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

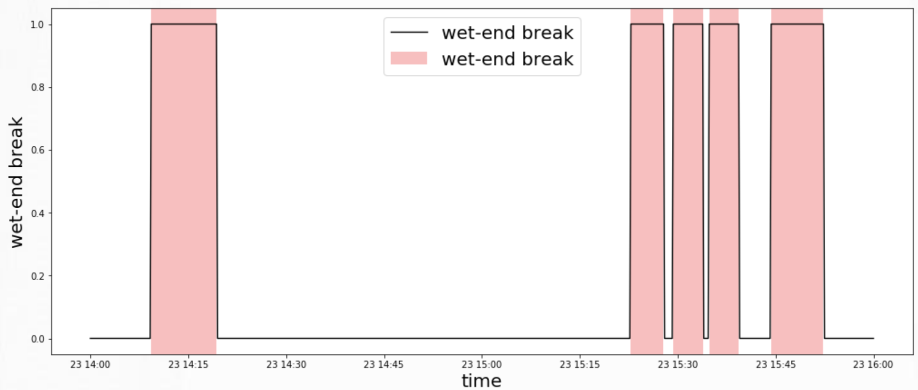
Aalto University
School of Chemical
Engineering

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

OVERVIEW: BREAK PREDICTION MODELING

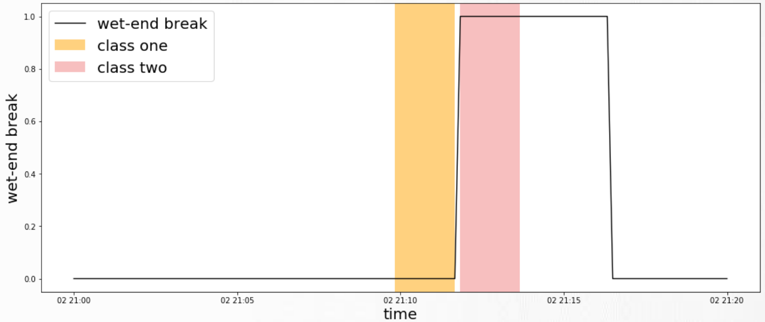
PREPROCESSING

- Data preprocessing
  - Cleaning: removing ambiguous break signals
  - Resampling: to have a common rate among signals
  - Slicing: selecting near-break and normal operation regions
  - Scaling: balancing differences in amplitude

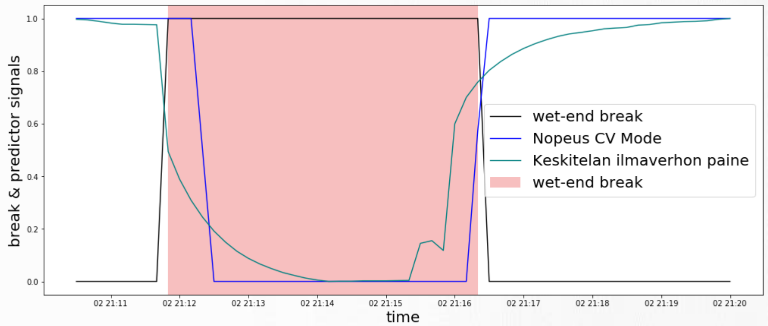


PREDICTION

- 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

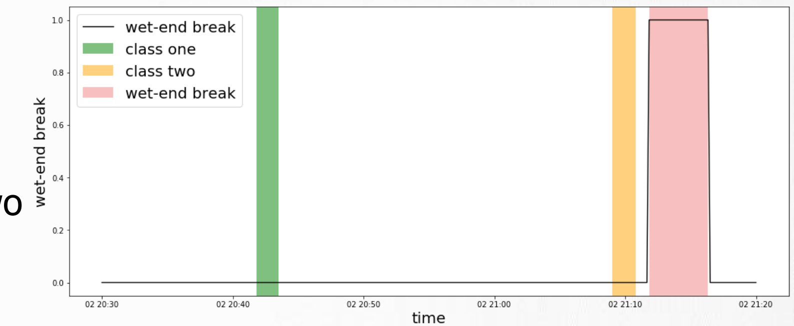


- Result:
  - 96.7% accuracy
  - Highly influenced by correlated signals
  - Fails to identify root-causes
  - Not valuable in practice

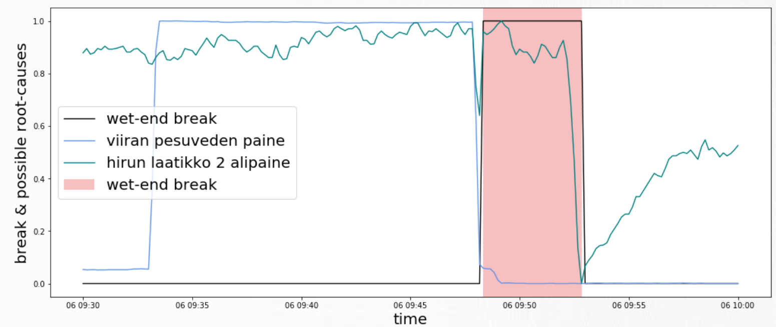


PREDICTION

- 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:
  - 50-58% accuracy
  - Identifies some potential root-causes
  - More likely valuable in practice
  - Still requires further study and improvement

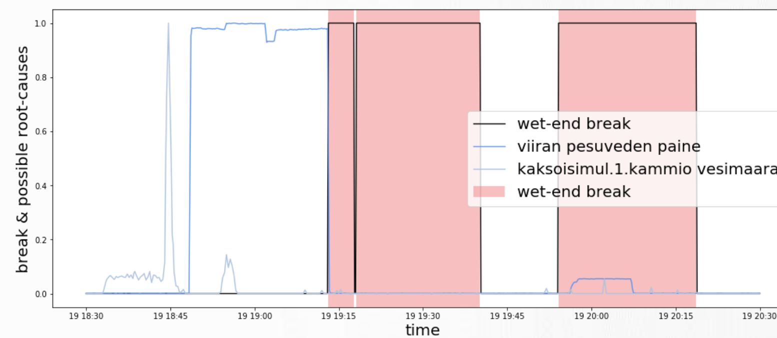


4. What information actually is relevant for root-cause analysis? Are there hidden relationships?

ROOT-CAUSE ANALYSIS

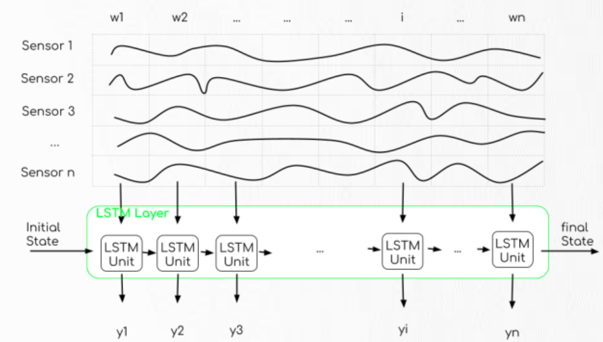
- Challenges
  - Identifying possible root-cause signals
  - Identifying delays between root-cause signals and break
  - Identifying and removing highly correlated signals (with no delay) to the break

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

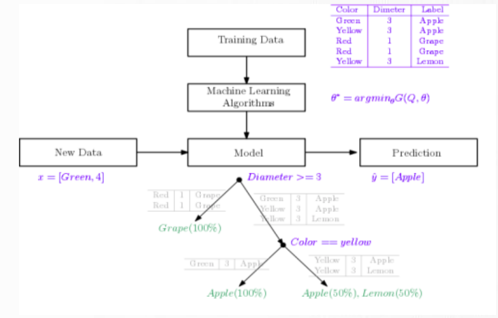


DEPLOYED LEARNING ALGORITHMS

- 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

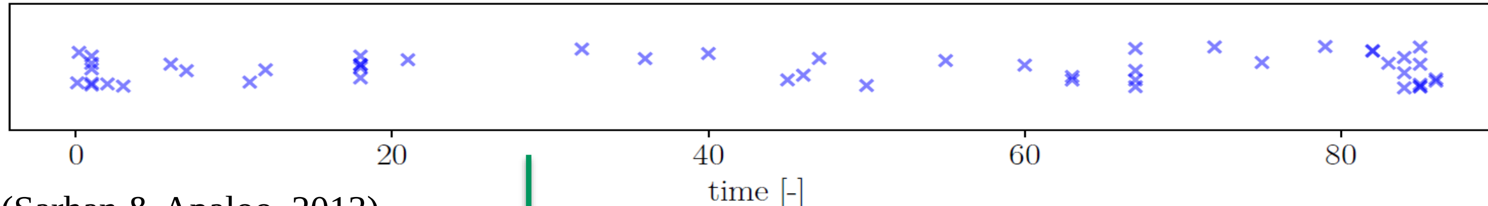


- Cross correlation and Granger causality

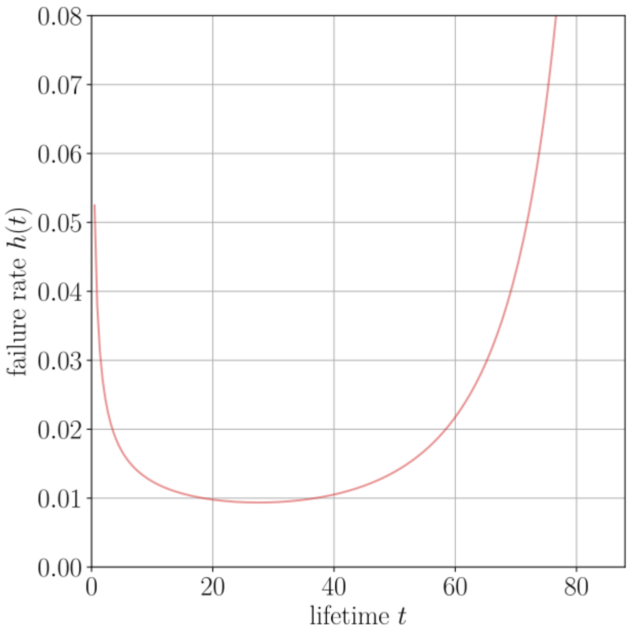
5. Are there decisions that can be excluded from the optimization scope, based on what we know from the data?

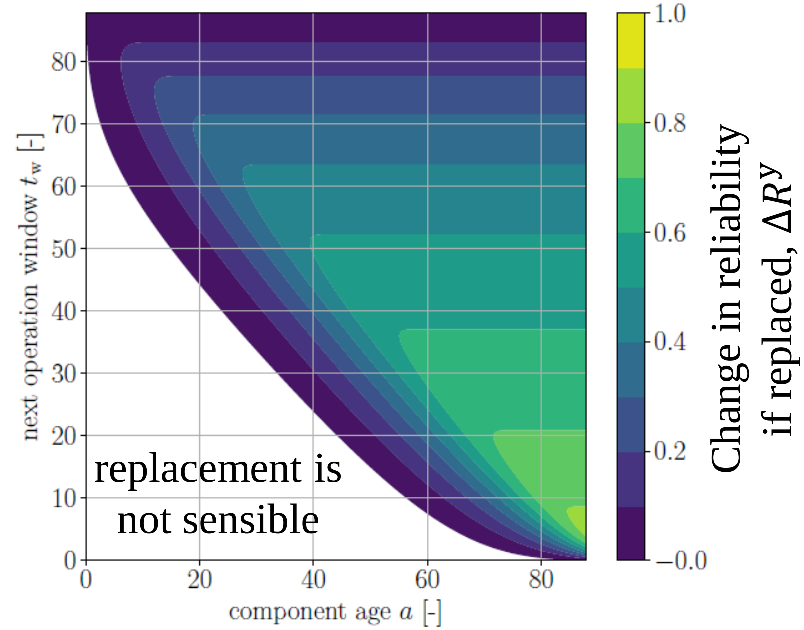
Selective maintenance optimization

Experimental data of failure times (Aarset, 1987)



Failure model (Sarhan & Apaloo, 2013)



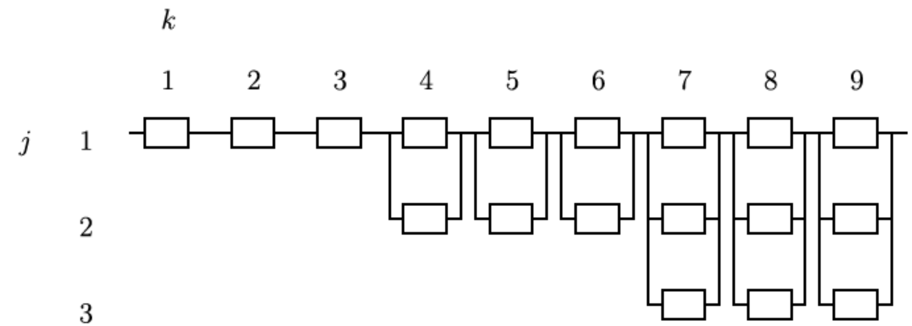


Selective maintenance optimization

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

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

An order of magnitude reduction in solution time



$y_{k,j}$ 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

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**

Acknowledgement

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

ACADEMY OF FINLAND

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