Overcoming poor data quality: Optimizing validation of precedence relation data

Alena Otto
TU Munich, Campus Heilbronn, Heilbronn Data Science Center

Scheduling Seminar, April 2026

Technical University of Munich, Campus Heilbronn

Introductory slides: Overview

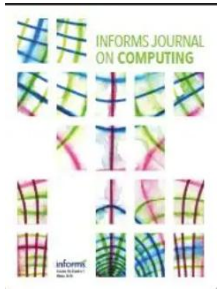
Academic affiliations

Chair of Advanced Analytics in Manufacturing Management, Technical University of Munich, Campus Heilbronn

Collaborating partner of CIRRELT – Interuniversity Research Centre on Enterprise Networks, Logistics and Transportation, Canada

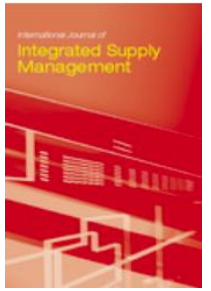
Till 2025: Chair of Management Science / Operations and Supply Chain Management, Univerisity of Passau

Editorial duties

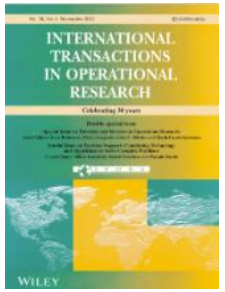


INFORMS J on Computing

OMEGA



Int J of Integrated Supply Mgmt



Special Issue on Transportation & Logistics

Special Issue on Assembly Systems

Selected awards, third-party funding, projects

- OptWare award of the „Initiative Wissenschaft und Automobile Industrie“ (IWA e.V.)
- Dissertation award of the Operations Research Society (GOR e.V.), 2012

DFG

KIMoNo
New Ways of Mobility

Bayerische Forschungsallianz

Programm committee and organizational work at conferences and workshops

IFORS

European conference on Operational Research

OR 2026 Passau

TRISTAN

New Challenges in Scheduling Theory
Aussois

ODYSSEUS

Internationale Tagung Wirtschaftsinformatik

Research topics

Scheduling in manufacturing

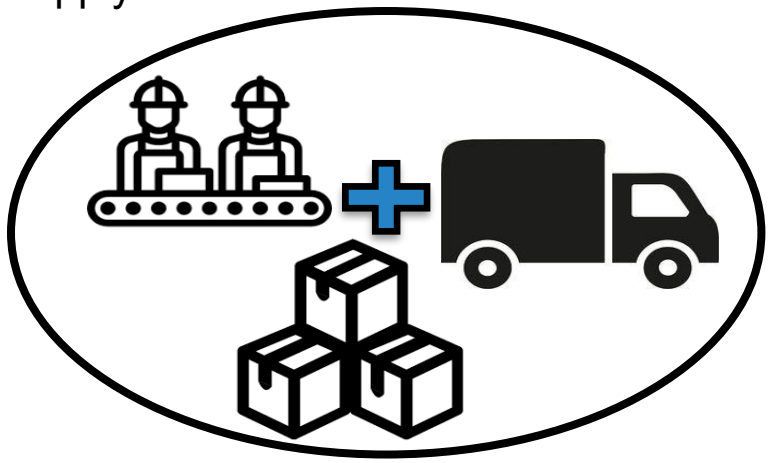
Assembly line balancing

Warehouse operations

Healthcare logistics

Drone operations

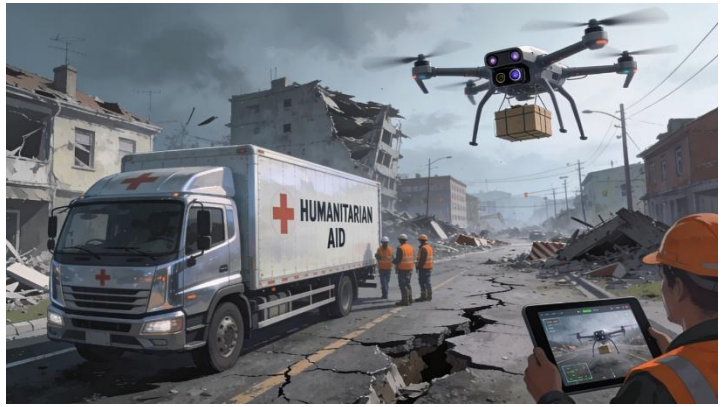
Integrated planning for resilient supply chains



Railway logistics

Human aspects, ergonomics

Disaster relief under incomplete information



Overcoming poor data quality: Optimizing validation of precedence relation data

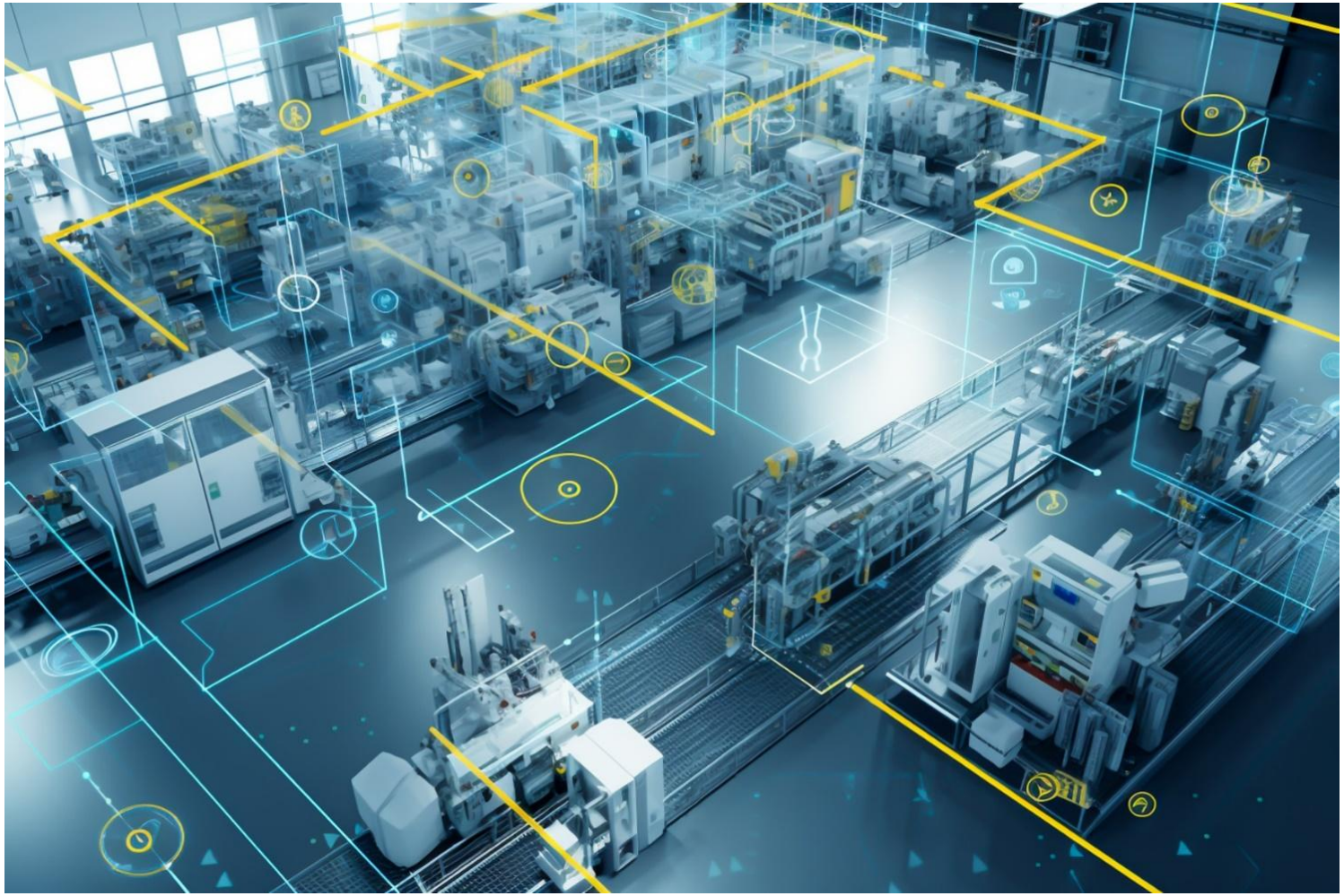
Dr. Benedikt Finnah
University of Duisburg-Essen

Prof. Dr. Jochen Gönsch
University of Duisburg-Essen

Prof. Dr. Alena Otto
Technical University of Munich

Finnah, Gönsch, Otto,
European Journal of Operational Research 2025

Our vision for the future: Digital transformation in manufacturing

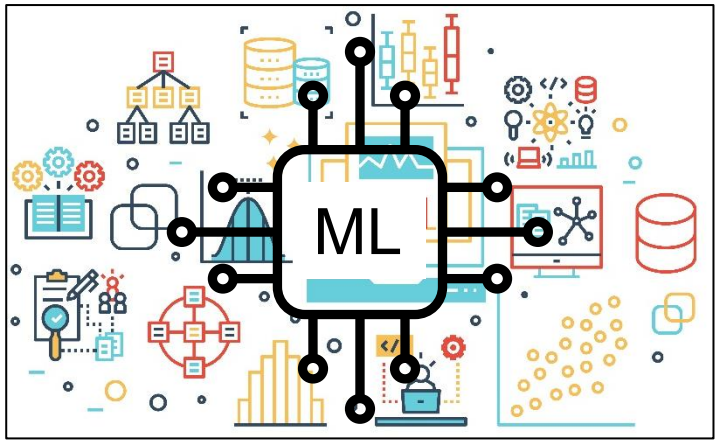


Our topic: Overcoming poor data quality

- Experts validate data…
- ..but their **time is limited**

Case study: car assembly line

- Certain types of **data** is **missing** for automated planning (optimization)
- Some of this data has **high requirements on accuracy**



- ML can retrieve large amounts of data..
- …but this **data** is **noisy**

Case study: Final assembly of cars at assembly lines

- After the start of production, assembly line operations have to be re-optimized repeatedly[1)]
- Required data
  - Assembly operations (Tasks)
    - Duration
    - Precedence relations
    - …
  - Cycle time
  - …

Data on precedence relations is often missing

1) See, e.g., Boysen et al. 2008, Otto & Otto 2014

The data on precedence relations has high requirements on accuracy

Illustrative example

The data on precedence relations has high requirements on accuracy

Illustrative example

The data on precedence relations has high requirements on accuracy

Requirements on data

Illustrative example

The data on precedence relations has high requirements on accuracy

Requirements on data

- **Required**
  - Data on **all** existing precedence relations

Otherwise:
infeasible plans may result

Illustrative example

The data on precedence relations has high requirements on accuracy

Requirements on data

- **Required**
  - Data on **all** existing precedence relations

Otherwise:
infeasible plans may result

- Desired
  - **No superfluous** precedence relations in the data

Otherwise:
unnecessary constraints → unnecessary idle time in the plan

Illustrative example

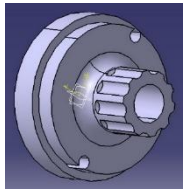
Data on precedence relations is not available

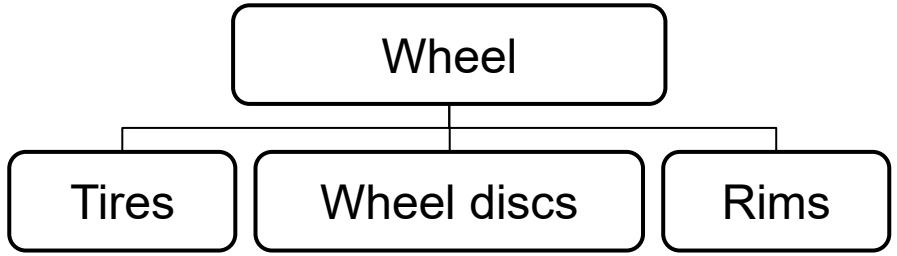
Information on precedence relations is in the heads of the planners

Data is missing

Sources of data on precedence relations in the car assembly

- Machine learning algorithms can assists to extract the data from various data sources, e.g.,
  - Design data: CAD (Rabemanantsoa & Pierre 1996, Worner et al. 2021)
  - Bill of materials (Niu et al. 2003)
  - Historical production plans, plans for similar production (Kashkoush & ElMaraghy 2014)





Data is noisy
We cannot identify unnecessary precedence relations with certainty
➔ Infeasible plans are possible

Challenge: Data validation by experts is extremely time intensive

- Representation of an interview with the expert
  - Specific questions to each pair of tasks („whether a precedence relation between task i and task j is unnecessary?“)

Required expert time to collect data on precedence relations			
Literature:	30 tasks	435 task pairs	3-4 hours

The expert time is limited ➔
use experts to validate **selected** data entries

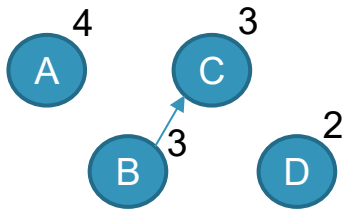
Schematic illustration of the proposed data collection and validation concept

Concept

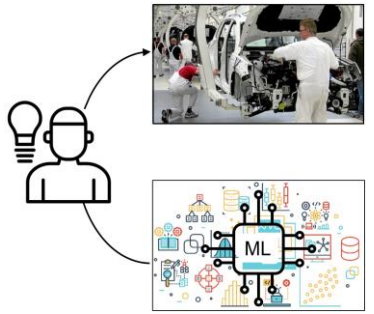
Illustration

Precedence relations:

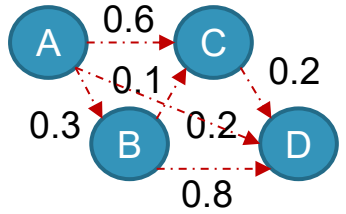
Ground truth
(complete precedence relations set E^*)



UNKNOWN

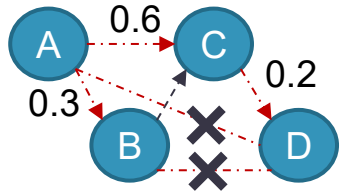


Initial database

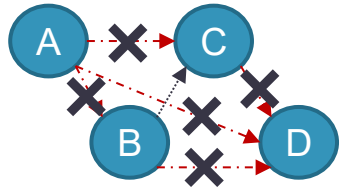


NOISY

Selective interview:
3 queries are enough to recover an optimal solution



Baseline:
Exhaustive interview (6 queries)



TOO COSTLY

Schematic illustration of the proposed data collection and validation concept

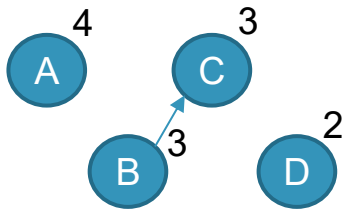
Concept

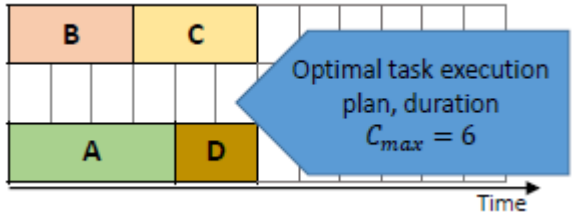
Illustration

Schedule for two parallel machines ($p = 2|prec|C_{max}$):

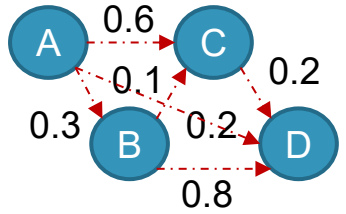
Ground truth (complete precedence relations set E^*) is unknown

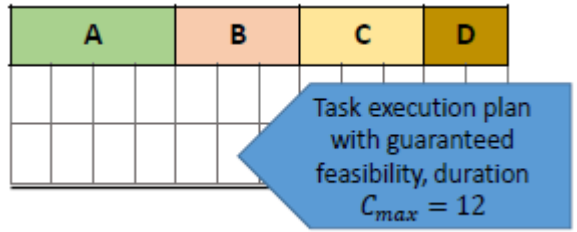
Precedence relations:



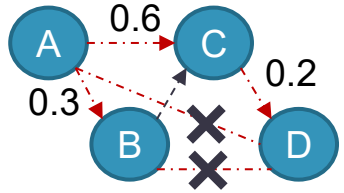


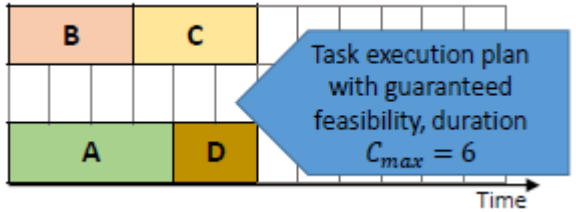
Initial database is noisy



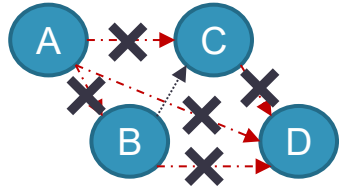


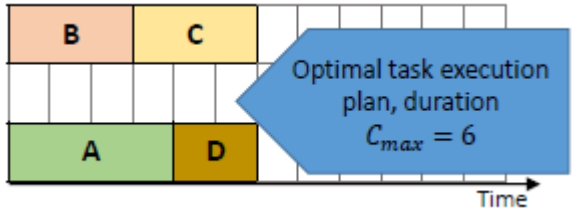
Selective interview:
3 queries are enough to recover an optimal solution





Baseline:
Exhaustive interview (6 queries)





Schematic illustration of the proposed data collection and validation concept

Concept

Ground truth (complete precedence relations set E^*) is unknown

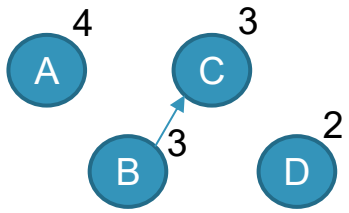
Initial database is noisy

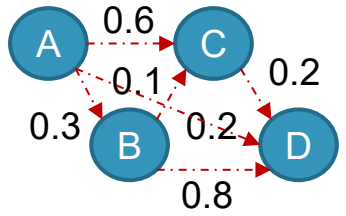
Selective interview:
3 queries are enough to recover an optimal solution

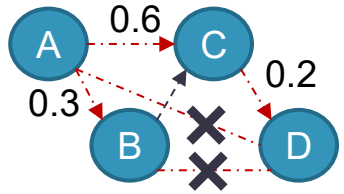
Baseline:
Exhaustive interview (6 queries)

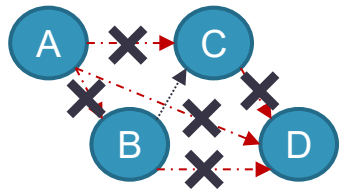
Illustration

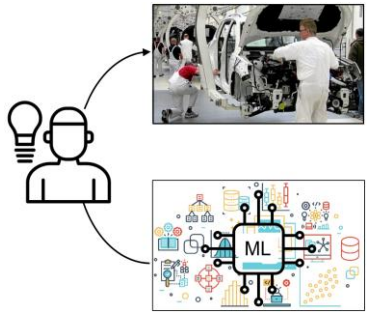
Precedence relations:



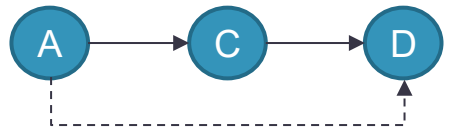








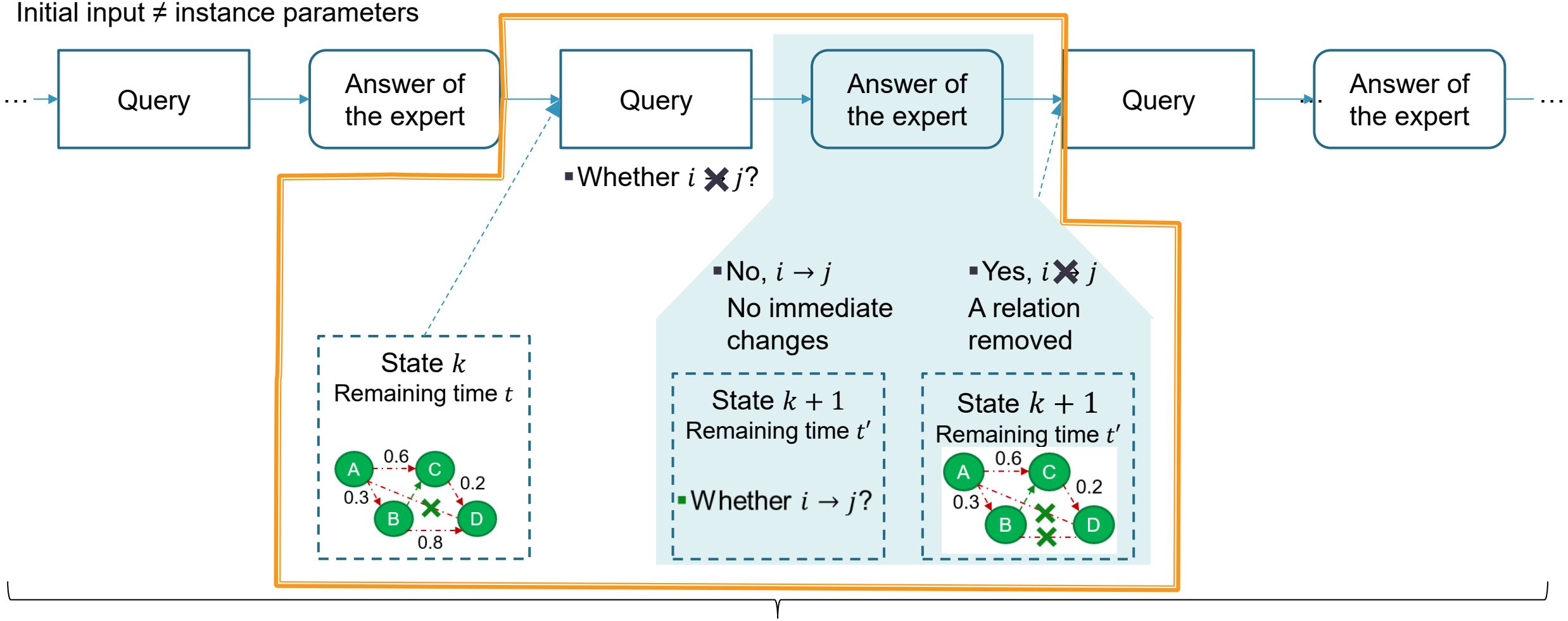
The question is not trivial because of *transitivity*



Data Validation Problem (DVP)

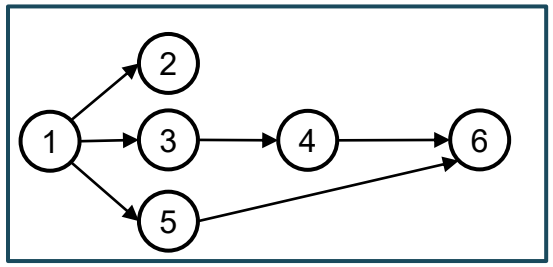
- Dynamic optimization under uncertainty with a restricted total time budget

Incomplete information
Initial input ≠ instance parameters

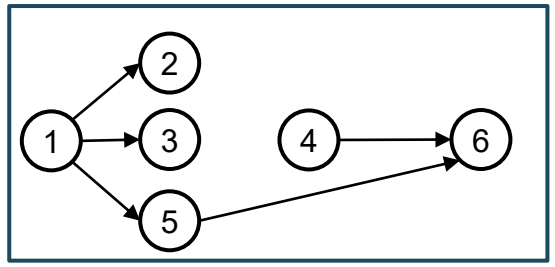


Objective: Maximize the expected total weighted number of removed precedence relations within time budget T

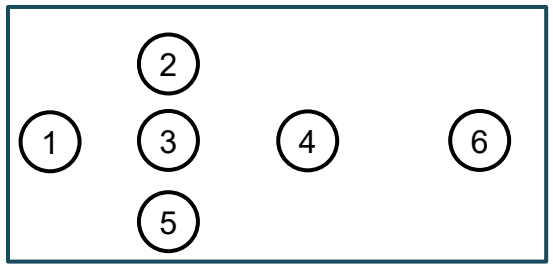
Initial definitions: concepts of min-, max-, and target precsets



- **Maximum precset**
  - $\bar{E}$, with $E^* \subseteq \bar{E}$. $\bar{E}$ contains all real precedence relations



- **Target precset**
  - E^* contains exactly the real precedence relations
  - In practice, the target precset is unknown

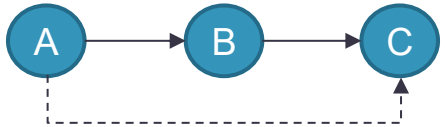


- **Minimum precset**
  - $\underline{E}$, with $\underline{E} \subseteq E^*$. $\underline{E}$ does not contain unnecessary precedence relations

See Otto, C., & Otto, A. (2014). Multiple-source learning precedence graph concept for the automobile industry. European Journal of Operational Research, 234, 253-265.

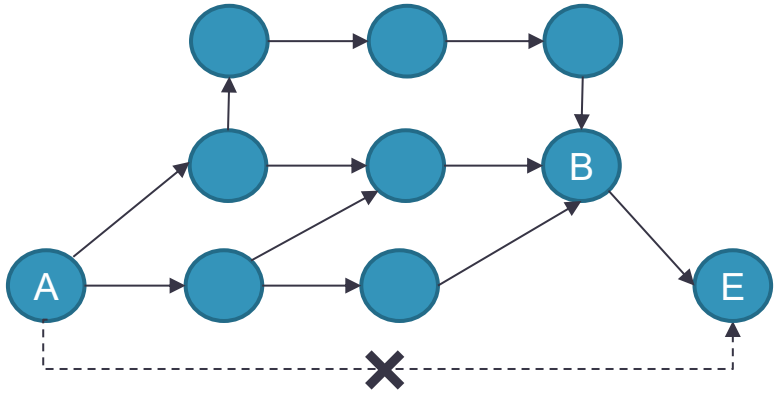
Initial definitions: simple, error-resistant questions

- Transitivity: if A→B and B→C, then A→C
- Direct precedence relations: A→B, B→C
- Indirect precedence relations: A→C



- Consider a question to A→E (indirect precedence relation in the current data base)
  - Let the Expert confirm that A and E are independent, can you use this information? No, not immediately
  - Would the Expert be able to answer the question to A→E at all? Prohibitively expensive

Current database:



Assumption: Expert can only answer questions on direct precedence relations

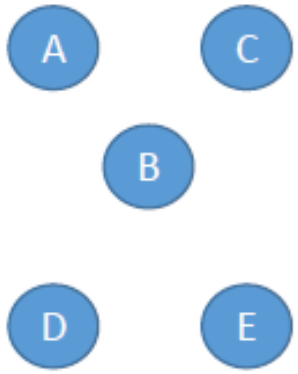
Data Validation Problem (DVP)

- Given:
  - Set of tasks $V = \{1, \dots, n\}$
  - *[Initially unknown]* target precset $E^* = \{(i,j) \in V \times V \mid i \text{ is predecessor of } j\}$, which is transitive and acyclic
  - Minimum precset $\underline{E}$ and maximum precset $\bar{E}$, such that $\underline{E} \subseteq E^* \subseteq \bar{E}$
  - Oracle accepts questions $(i,j) \in t^-(\bar{E})$: "Is i independent from j?"
    - Answer: Bernoulli random variable Ω_{ij} with success probability p_{ij}
  - Cost to state a question $\tau_{ij} \geq 0$
  - Weight (importance) of a question w_{ij}
  - Total time budget T_0
- Find:
  - Dynamic interview policy π that maximizes the expected weighted number of positive answers of the oracle

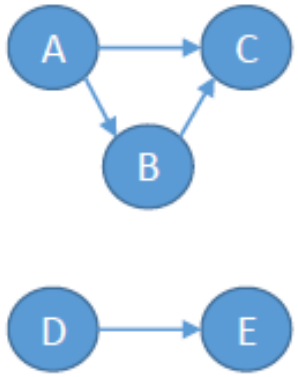
$$\max_{\pi} \mathbb{E}^{\pi} \left[\sum_{q=0}^{Q_\pi - 1} w_{\pi(q)} \Omega_{\pi(q)} \right]$$

Lookahead: On policies for DVP

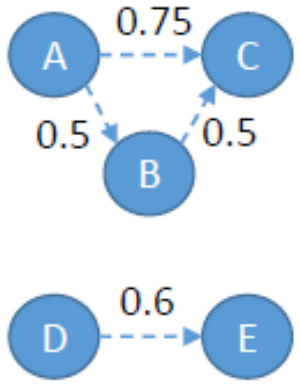
Minimum precedence relations set $\underline{E}_0$



Maximum precedence relations set $\overline{E}_0$



Forecast (probabilities p_{ij})



Expected number of uncertain precedence relations removed with budget $T_0 = 2$ queries

Naive (random):	1.089
Myopic (largest p_{ij} first):	1.100
Optimal:	1.175

Optimal policy is expected to remove about 10% more precedences than intuitive alternative policies

Dynamic program

- State

$$S = (\bar{E}, \underline{E}, T)$$

- $T \leq T_0$ Remaining time budget for the interview
- $\bar{E}$ Current maximum graph
- $\underline{E}$ Current minimum graph
- $p_{i,j}$ Probability of removing (i,j) from the current maximum graph
- $\tau_{i,j}$ Answering time of question to precedence relation (i,j)
- $X(S)$ Set of feasible questions $X(S) = t^-(\bar{E}) \backslash t^+(\underline{E}) : \tau_{i,j} \leq T$
- Question's contribution

$$C(S,(i,j),\Omega_{ij}) = \begin{cases} w_{ij}, & \text{if } \Omega_{ij} = 1, \\ 0, & \text{if } \Omega_{ij} = 0. \end{cases}$$

- Bellman equation:

$$B(S) = \begin{cases} \max\limits_{(i,j)\in X(S)} p_{ij}(w_{ij} + B(\bar{S}_{ij})) + (1 - p_{ij})B(\underline{S}_{ij}), & \text{if } X(S) \neq \emptyset, \\ 0, & \text{if } X(S) = \emptyset, \end{cases}$$

with

$$\bar{S}_{ij} = (\underline{E}, \bar{E} \setminus \{(i,j)\}, T - \tau_{ij})$$

$$\underline{S}_{ij} = (\underline{E} \cup \{(i,j)\}, \bar{E}, T - \tau_{ij})$$

Selected analytical results: Merits of expert data validation – Sufficient time budget

Proposition 4. *Consider a DVP(-gen) instance. To find the target precset E^* with certainty, it is necessary and sufficient to state queries to the following set of task pairs $\mathcal{I}^* = (t^-(E^*) \setminus \underline{E}_0) \cup (\bar{E}_0 \setminus E^*)$. Any feasible policy in the DP stated in Section 3.2 (provided the initial budget T_0 is set to a sufficiently large number) poses queries exactly from set $\mathcal{I}^*$.*

- The number of required interview questions is much lower than generally perceived
  - Consider example with $n = 50$ tasks. E.g., naïve estimate: $\binom{50}{2} = 1225$
  - Alternative naïve estimate (in a typical case, see paper): $|\bar{E}_0 \setminus \underline{E}_0| = 0.3 \cdot \binom{50}{2} \approx 368$
  - Observation: the number of direct precedences in “real-world” precsets is usually proportional to n
  - Number of required interview questions in a typical case (see paper): 197

Selected analytical results: Merits of expert data validation – Limited time budget

- Share of independencies (nominator) in the total number of queries $|\mathcal{I}^*|$ required to derive the unknown target graph

$$\tilde{p} \approx \frac{\left(OS(\bar{E}_0) - OS(E^*)\right)\binom{n}{2}}{|\mathcal{I}^*|} = \frac{\left(OS(\bar{E}_0) - OS(E^*)\right)\binom{n}{2}}{|t^-(E^*) \setminus \underline{E}_0| + \left(OS(\bar{E}_0) - OS(E^*)\right)\binom{n}{2}}.$$

	$OS(E^*) = 0.2$				$OS(E^*) = 0.6$				$OS(E^*) = 0.9$			
# tasks n =	20	50	100	500	20	50	100	500	20	50	100	500
$\tilde{p}$ [%]	83.5	92.9	96.4	99.3	71.7	86.7	93.0	98.5	38.8	62.0	76.7	94.3
# required queries $\lvert\mathcal{I}^*\rvert$ [thousands]	0.2	1.1	4.1	100.6	0.1	0.6	2.1	50.7	0.05	0.2	0.6	13.2
Share of $\lvert\mathcal{I}^*\rvert$ in all task pairs [%]	95.8	86.1	83.0	80.6	55.8	46.1	43.0	40.6	25.8	16.1	13.0	10.6

Note. $|t^-(E^*)| = 1.5n$, $OS(\bar{E}_0) = 1$, $OS(\underline{E}_0) = 0$.

LSTD: Approximate Dynamic Programming Heuristic

Algorithm 1 LSTD (Input: discount factor λ, the number of iterations M)

Initialize $K \times 1$ vector of regression parameters $\theta := 0_K$;
Initialize $K \times K$ matrix $A := 0_{KK}$; initialize $K \times 1$ vector $b := 0_K$;
for $m = 1$ to M **do**
  Initialize $S := S_0$; initialize counter $l := 0$;
  $\Phi^0 := \Phi(S)$;
  while $X(S) \neq \emptyset$ **do**
    $l := l + 1$;
    $(i^*, j^*) := \underset{(i,j) \in X(S)}{\arg\max}\, p_{ij}(w_{ij} + \tilde{B}(\bar{S}_{ij})) + (1 - p_{ij})\tilde{B}(\underline{S}_{ij})$;
    $b := b + \Phi^{l-1} p_{i^*,j^*} w_{i^*j^*}$;
    Stochastic state transition;
    $\Phi^l := \Phi(S)$;
    $A := A + \Phi^{l-1}(\Phi^{l-1} - \lambda\Phi^l)^T$;
  end while
  $\theta := A^{-1}b$;
end for

$$\mathcal{A}(S) = \sum_{(i,j) \in X(S)} \frac{p_{ij} w_{ij}}{\tau_{ij}}.$$

- We selected the following $\mathrm{K} = 5$ features for the value approximation function (VFA): $\phi_1(S) = \mathcal{A}(S), \phi_2(S) = \sqrt{T}\mathcal{A}(S), \phi_3(S) = T\mathcal{A}(S), \phi_4(S) = \sqrt{T}, \phi_4(S) = T$

Numerical study

- Compare Dynamic data collection with Myopic and Naive data collection
  - Dynamic data collection (LSTD): DP solved with an approximate dynamic programming approach
  - Myopic data collection: State a question with the largest probability to discover an independence
  - Naïve data collection: State a question randomly
- Data-Collection instances based on realistic SALBP instances (Otto et al., 2013)
  - Number of task $N = 50$
  - 525 instances with order strength (OS) of the target graph in $\{0.2, 0.6, 0.9\}$
  - Initial maximum graph based on initial sequence $1{:}N$
  - Initial minimum graph is the empty set
  - $\tau_{i,j} = 1, T = 150 \Rightarrow 150$ questions
  - $p_{i,j}$ based on a similar SALBP instance
    - If a precedence relation exists in the similar instance, it is likely to exists (90%)
    - If a precedence relation does not exist in the similar instance, it is unlikely to exists (10%)

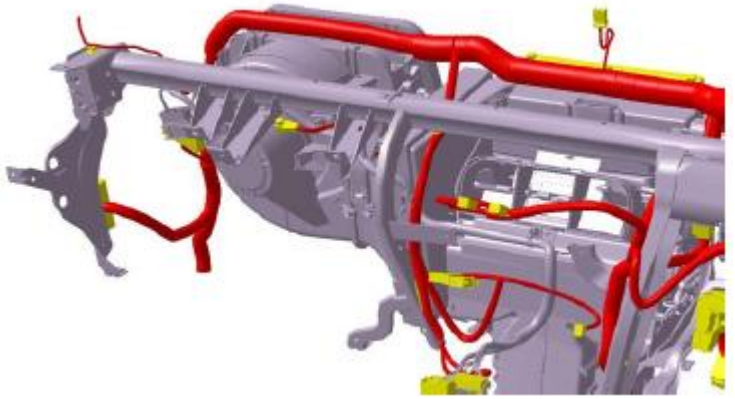
Numerical study

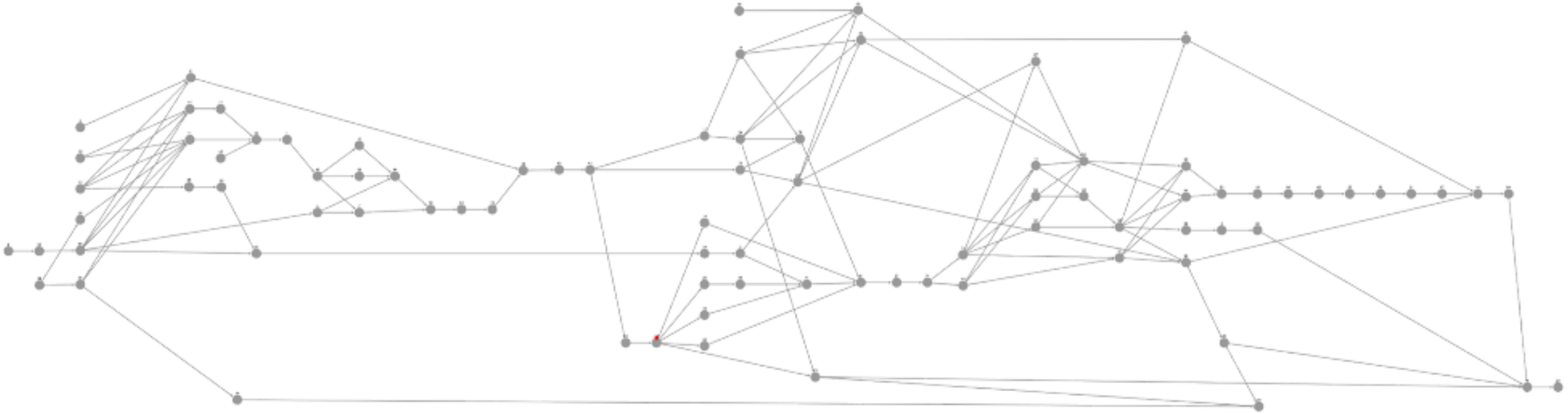
Policy	$OS(E_I^*) = 0.2$	$OS(E_I^*) = 0.6$	$OS(E_I^*) = 0.9$
ALPLib instances with $n = 20, T_0 = 24$			
Naive	22.1 (15.0 %)	20.1 (23.1 %)	14.3 (40.8 %)
Myopic	23.6 (16.1 %)	22.9 (26.3 %)	20.2 (57.7 %)
LSTD	23.6 (16.1 %)	22.9 (26.3 %)	20.5 (58.5 %)
ALPLib instances with $n = 50, T_0 = 150$			
Naive	140.6 (14.3 %)	126.2 (25.6 %)	81.2 (64.7 %)
Myopic	147.8 (15.0 %)	144.4 (29.3 %)	96.6 (76.9 %)
LSTD	147.7 (15.0 %)	144.4 (29.3 %)	116.6 (92.8 %)
ALPLib instances with $n = 100, T_0 = 600$			
Naive	564.7 (14.2 %)	523.1 (26.3 %)	383.1 (74.8 %)
Myopic	591.9 (14.9 %)	580.6 (29.1 %)	418.6 (81.8 %)
LSTD	591.9 (14.9 %)	583.7 (29.3 %)	466.6 (91.1 %)

Note. Percentages in the brackets report the share of removed unnecessary precedences $(|E_{I,0}| - |E_{I,\pi}|)/(|E_{I,0}| - |E_I^*|)$

Case study: Final assembly line balancing of a car manufacturer (1/4)

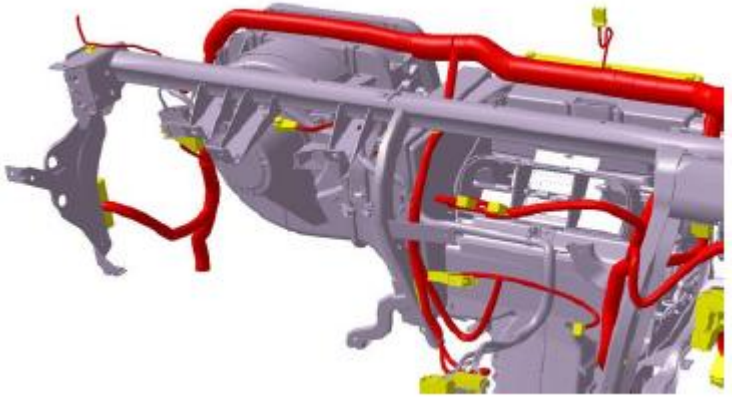
- Assembly line balancing for the assembly of the cockpit (86 tasks)
  - Time budget $T_0 = 300$ questions (5 hours, 8.2% of total task pairs)
  - Two objective functions examined
    - Minimize the number of stations
    - Minimize the cycle time



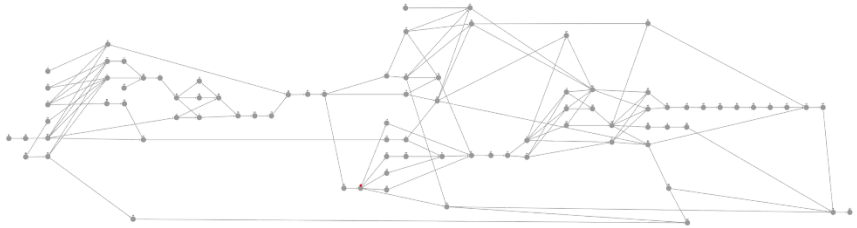


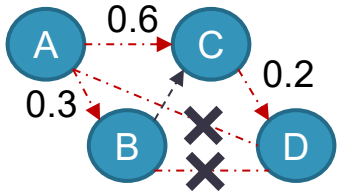
Case study: Final assembly line balancing of a car manufacturer (2/4)

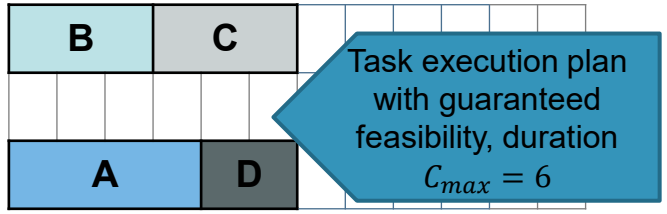
- Assembly line balancing for the assembly of the cockpit (86 tasks)
  - Time budget $T_0 = 300$ questions (5 hours, 8.2% of total task pairs)
  - Two objective functions examined
    - Minimize the number of stations
    - Minimize the cycle time



- Importance (weights) of revealed independencies: $(t_i + t_j)$







Case study: Final assembly line balancing of a car manufacturer (3/4)

- SALBP-1: Objective to minimize the idle time

Policy	Cycle time $c =$ $1.00 \cdot c_{min}$	$1.02 \cdot c_{min}$	$1.04 \cdot c_{min}$	$1.06 \cdot c_{min}$	$1.08 \cdot c_{min}$	$1.10 \cdot c_{min}$
Naive	2.00	2.00	2.00	1.20	2.15	1.80
Myopic	2.00	2.00	2.25	1.90	3.00	2.00
LSTD	2.25	2.20	2.60	2.05	3.00	2.00
Max. possible reduction in # of stations $F(\cdot\|\bar{E}_{I,0}) - F(\cdot\|E_I^*)$	4.00	4.00	4.00	3.00	3.00	2.00

The results show the average number of the removed stations from the initial solution $F(\cdot|\bar{E}_{I,0})$.

LSTD removes 0.25 stations (or 0.25/2=12.5%) more on average than other policies
Incomplete data validation reduces the number of stations by 11.25% on average

Case study: Final assembly line balancing of a car manufacturer (4/4)

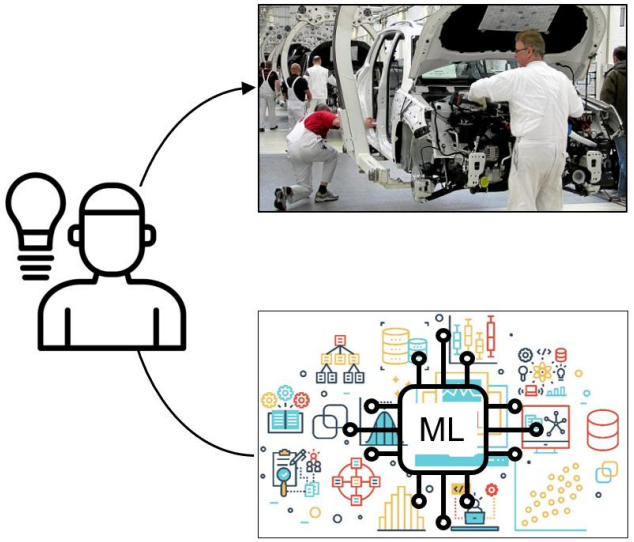
- SALBP-2: Objective to minimize cycle time

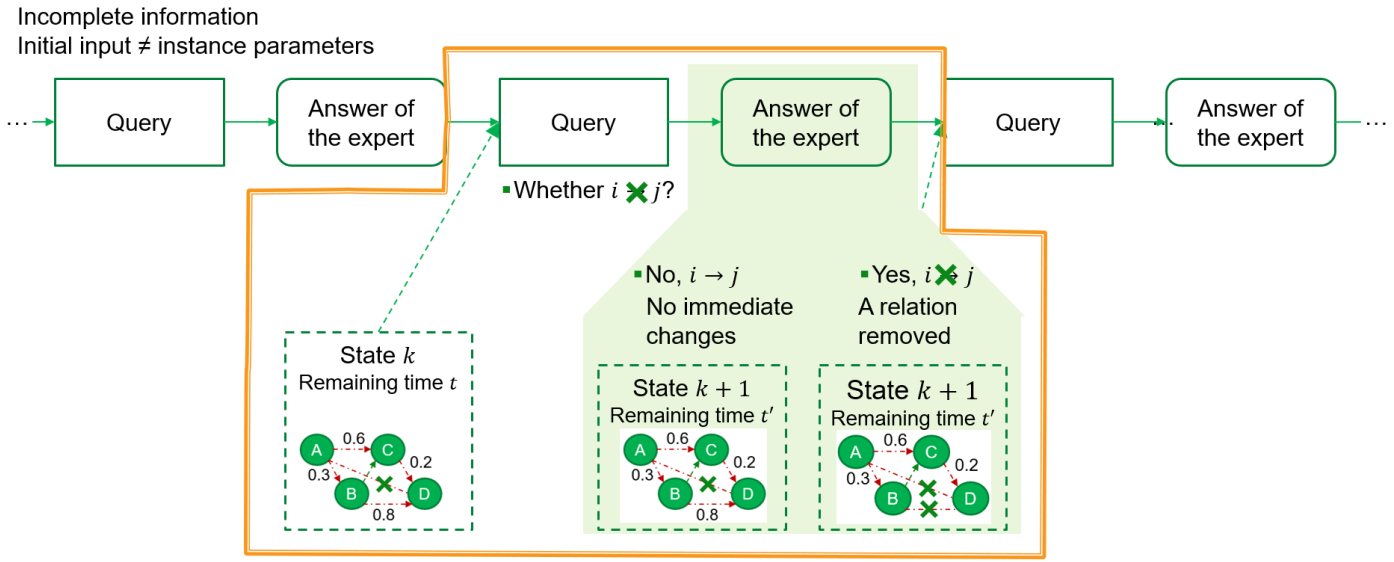
Policy	The number of stations $R =$ 13	14	15	16
Naive	4.79	8.32	10.27	1.98
Myopic	5.39	8.75	11.33	3.19
LSTD	5.63	9.13	11.45	3.77
Max. possible reduction in the cycle time $F(\cdot\|\bar{E}_{I,0}) - F(\cdot\|E_I^*)$	9.10	10.35	14.22	5.65

The results show the average number of the cycle time reduction compared to the initial solution $F(\cdot|\bar{E}_{I,0})$.

LSTD outperforms Myopic by up to 18.2% and Naïve by up to 90.4%
Incomplete data validation leads to a significant reduction in cycle time

Conclusions

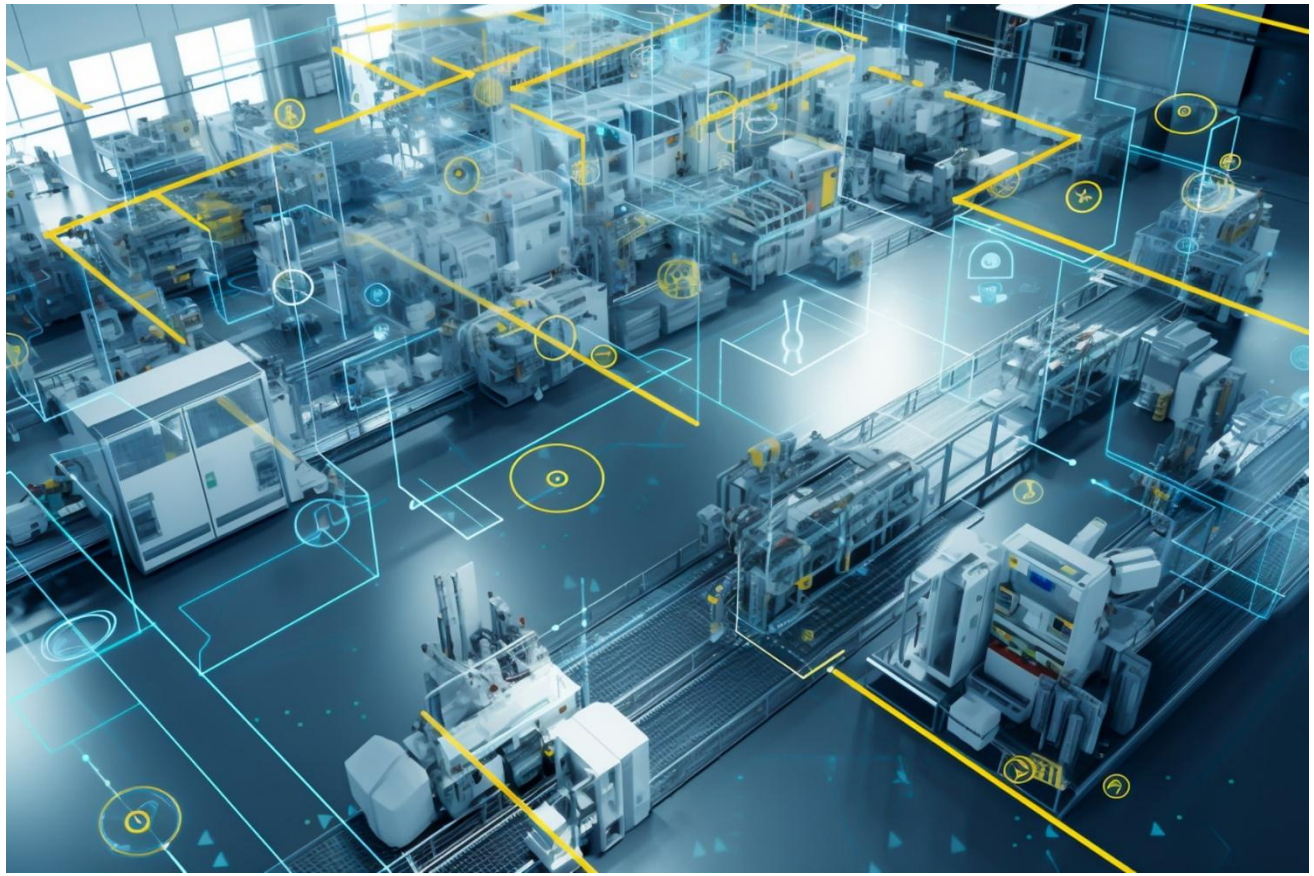




- Further adapt and extend methodology to specific applications (e.g., parallel machines)

- Further types of queries
  - E.g., several queries at ones

Our envisioned contribution



Extend the optimization of data collection and validation to other types of data

Thank you for your attention!