4th February 2025

Applied AI and Data Science on Emergent Technologies

Invited Talk VIT Chennai

Professor Dr. rer. nat. habil. Sven Groppe
https://www.ifis.uni-luebeck.de/~groppe

Stations of my academic life

Research Areas

- Artificial Intelligence, Machine Learning and Data Science
  - LLMs, Agentic Workflows, Mathematical Optimizations, Graph Neural Networks, Chatbots, Reasoning
- Data Management Tasks
  - Query Processing & Opt., Indexing, Mapping, Compression, Replication, Caching, Transaction Handling
- Data Models
  - Knowledge Graphs, Semantic Web, Property Graphs, Relational Data, XML
- Types of Data
  - Big Data, Data Streams
- Emergent Hardware Technologies
  - Many-Core CPU, GPU, FPGA, Quantum Computer
- Platforms
  - Internet, Internet of Things, Cloud, Post-Cloud (Fog/Edge/Dew Computing), P2P, Mobile, Parallel and Main Memory Servers
- Advanced Applications
  - Citizen Science, Customer Communications, Pandemics like Covid-19, Software Vulnerability Prediction
- Sustainability
  - Sustainable Computing/AI, Applications for Sustainability

Lectures by Sven Groppe

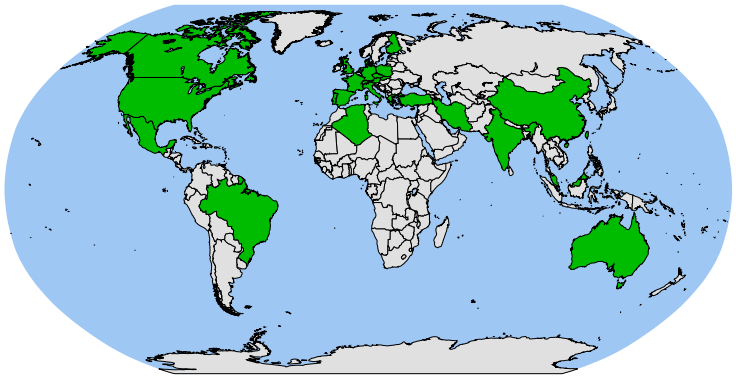
50 lectures in the areas of

- Information Systems N T (inclusive Knowledge Graphs, LLMs and Graph Neural Networks) (since 2024)
- Quantum Computing N T (inclusive Quantum Machine Learning) (since 2022)
- Semantic Web N T (since 2011)
- Databases (2014, since 2024)
- Mobile and Distributed Databases N (since 2008)
- Cloud and Web Technologies N T (since 2015)
- Algorithms and Datastructures (2014)
- XML Databases N (2007)
- Compiler Design N T (2006/2007)
- Next Web Generation (2006, together with M. Zaremba)

N Newly designed T Online Tutorials

Supervision and Publication Record

- 3 supervised dissertations, 6 current PhD students
- ≈ 100 bachelor/master/diploma thesis/student projects
- > 191 publications
  - 16 publications at A/A1 ranked conferences[1]
  - 195 co-authors affiliated with organizations in 28 countries on 6 continents



  - 48 publications (25%) are co-authored by bachelor/master students (being students at time of writing)

[1] according to Conference Ranks

Project Grants (≈ 2M €)

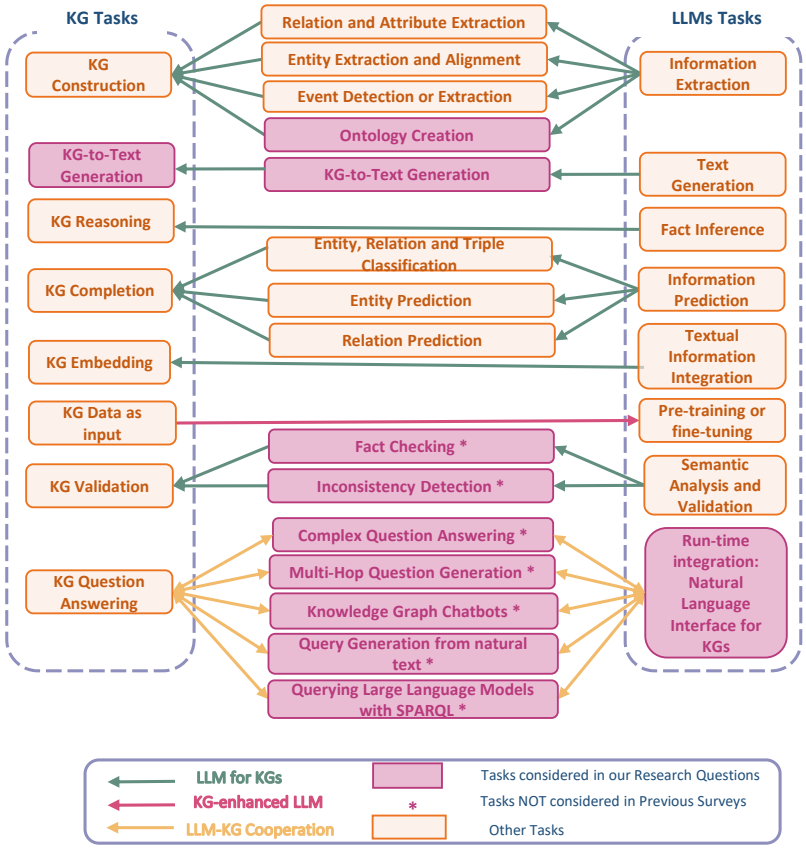
1. D I High Quality Knowledge Graphs from recent English, French and German Emergent Trends with the example of COVID-19 (2022-2025)
2. B QC4DB: Accelerating Relational Database Management Systems via Quantum Computing (2022-2025)
3. D Hybrid^2-Indexstrukturen für Hauptspeicherdatenbanken (2019-2024)
4. D BigSIoT: Big Data Management for the Semantic Internet of Things (2020-2023)
5. D Hardwarebeschleunigung von Semantic Web Datenbanken durch dynamisch rekonfigurierbare FPGAs (2013-2015)
6. W Beschleunigung relationaler Datenbanken mittels laufzeitadaptiver FPGA-Cluster (2013-2015)
7. D Logisch und Physikalisch Optimierte Semantic Web Datenbank-Engine (2007-2009)

D DFG projects I International Project W BMWi/ZIM B BMBF

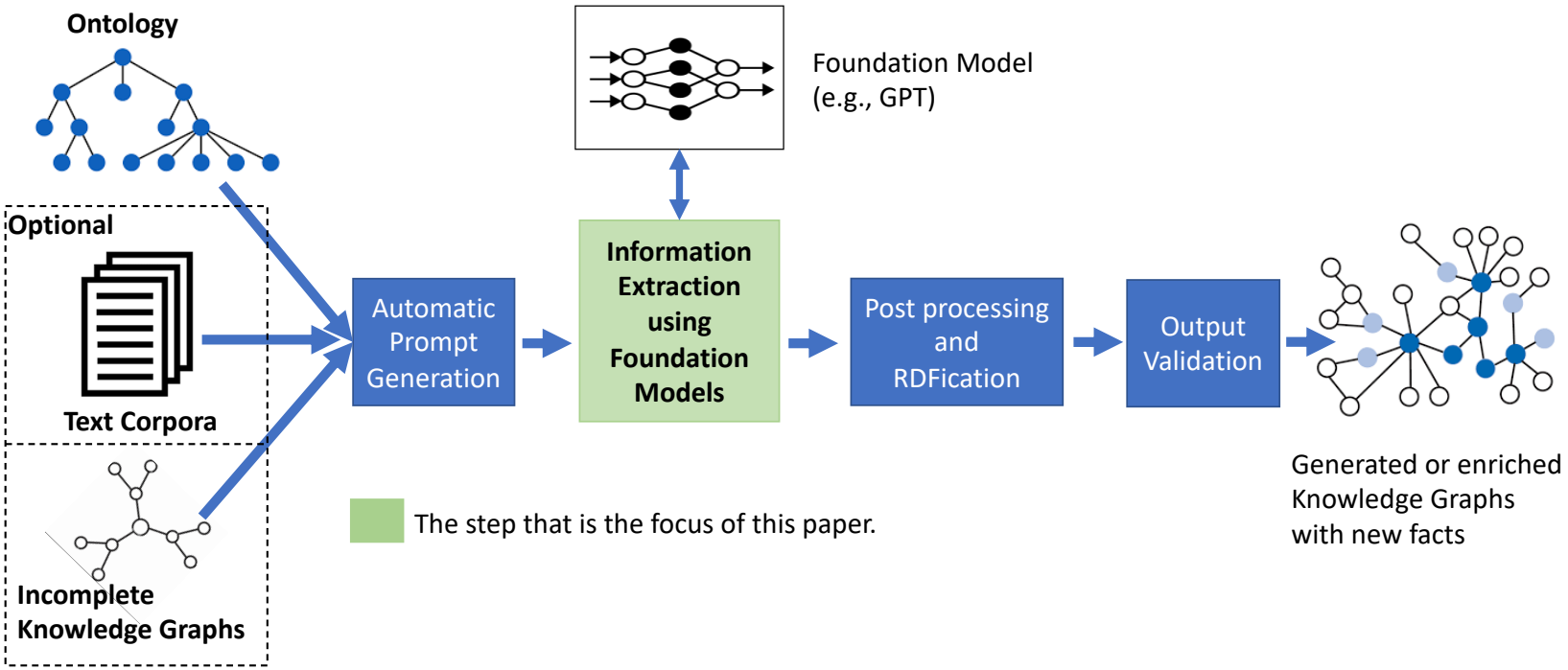
Scientific Services

- General Chair
  - The International Conference on Applied Machine Learning and Data Analytics (AMLDA) '23
  - International Semantic Intelligence Conference (ISIC) ('21-'22)
  - International Health Informatics Conference (IHIC) ('22-'24)
  - International EdTech Conference (IEdTC) '23
- Workshop Chairs
  - Quantum Data Science and Management (QDSM)@VLDB ('23-'24)
  - Semantic Big Data (SBD)@SIGMOD ('16-'20)
  - Big Data in Emergent Distributed Environments (BiDEDE)@SIGMOD ('21-'23)
  - Very Large Internet of Things (VLIoT)@VLDB ('17-'22)
- many other scientific services
  - ≈ 128 PC memberships
  - reviewer of ≈ 42 journals
  - editor of 4 journals
  - ...

Interplay of LLMs and KGs



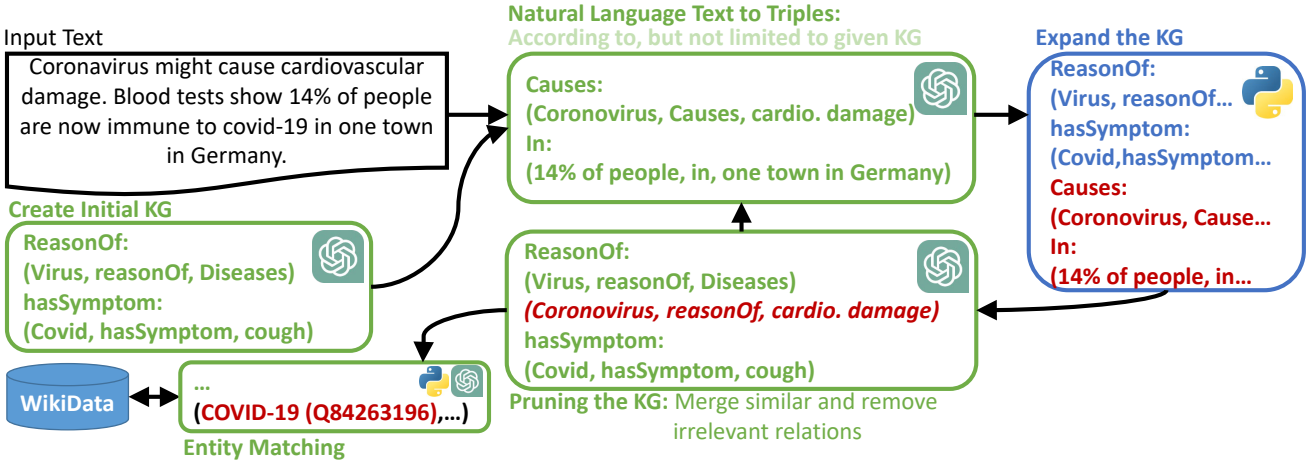
KG Construction with LLMs
(Project with Uni Paris/Toulouse)



- Knowledge graph (KG) construction, completion, fact validation, extraction facts from unseen text* (limitations due to hallucination)

* H.Khorashadizadeh et al., Exploring In-Context Learning Capabilities of Foundation Models for Generating KGs from Text, Text2KG@ESWC 2023

KG Construction with LLMs
(Project with Uni Paris/Toulouse)



- Result (comparison to manual labeling)

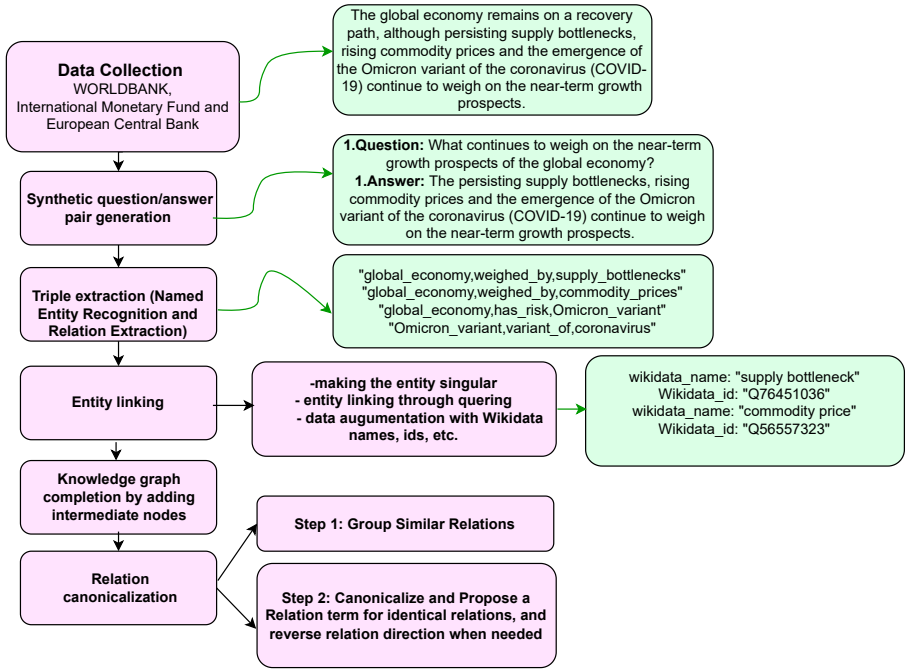
	Cov	Gen	Db
Macro F1-score	66.60	**76.93**	72.85

  - COV: Represents the original triples extracted using GPT-4
  - GEN: COV with augmented data generated by GPT-4
  - DB: Merges the original triples from COV with triples from an external data source (with daily statistical updates and policy actions)

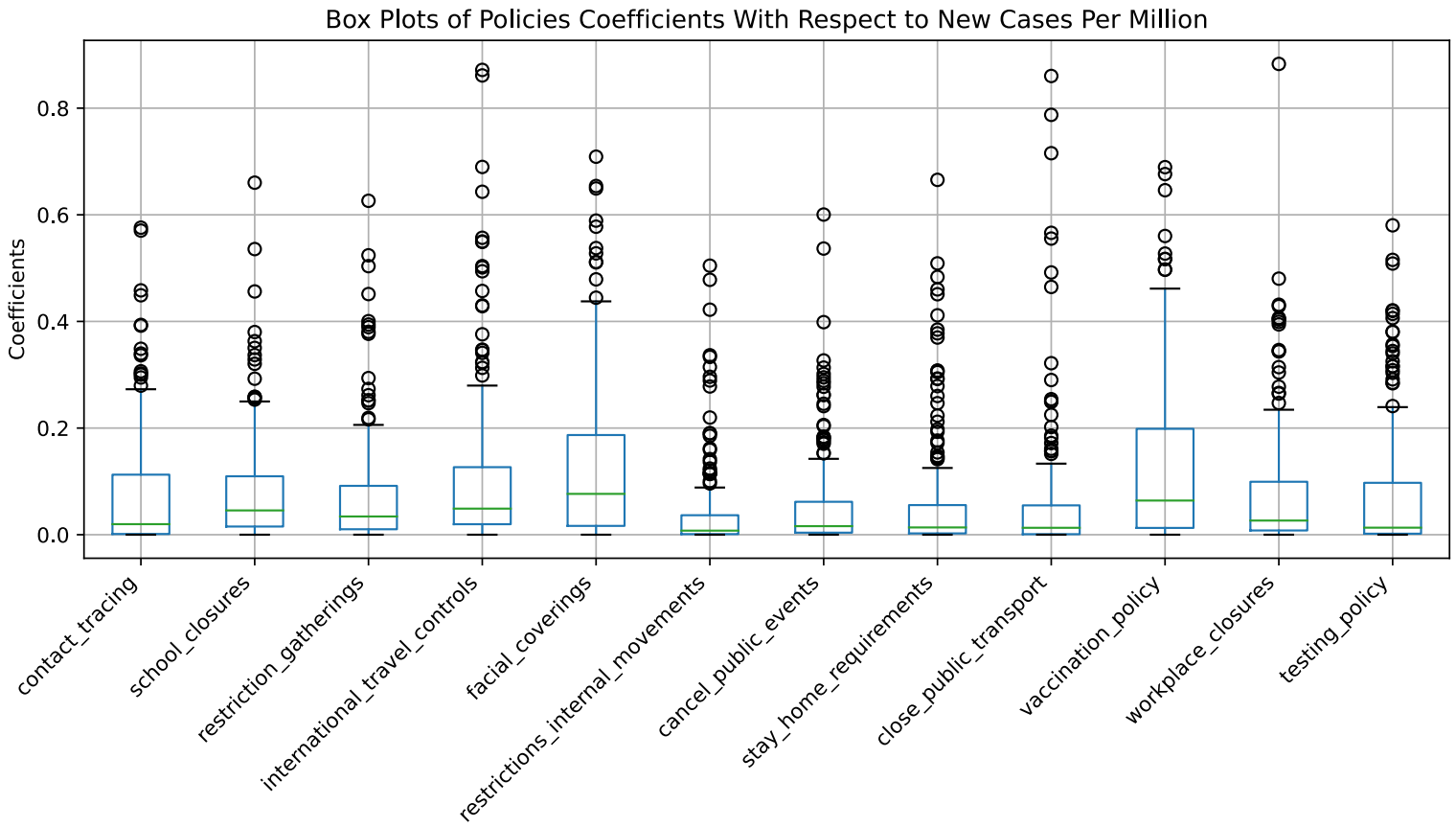
KG Construction with LLMs
(Project with Uni Paris/Toulouse)

- Approach to **Open Information Extraction**
  - extracting structured information directly from unstructured text **without a predefined schema**

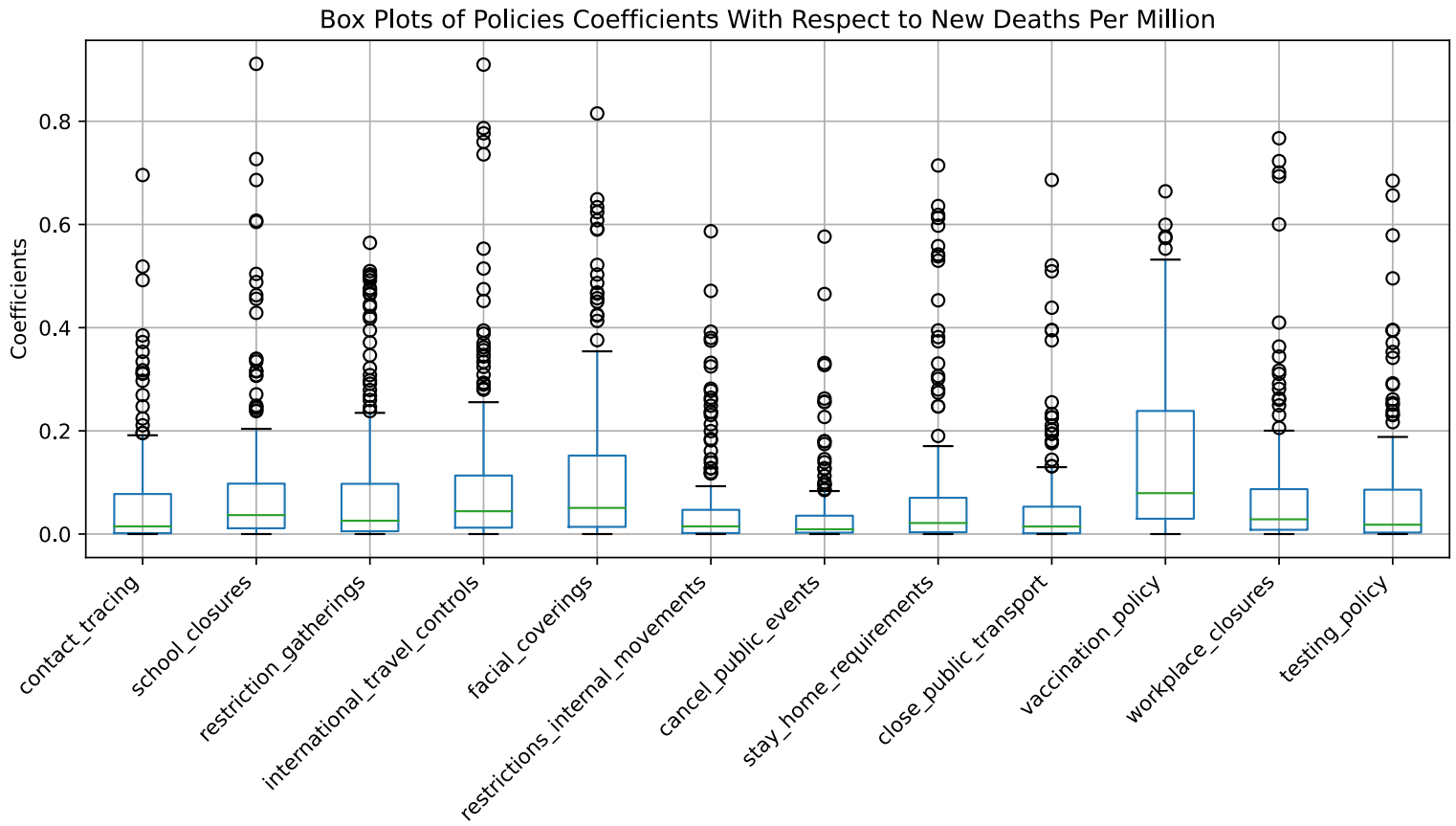
Pipeline:



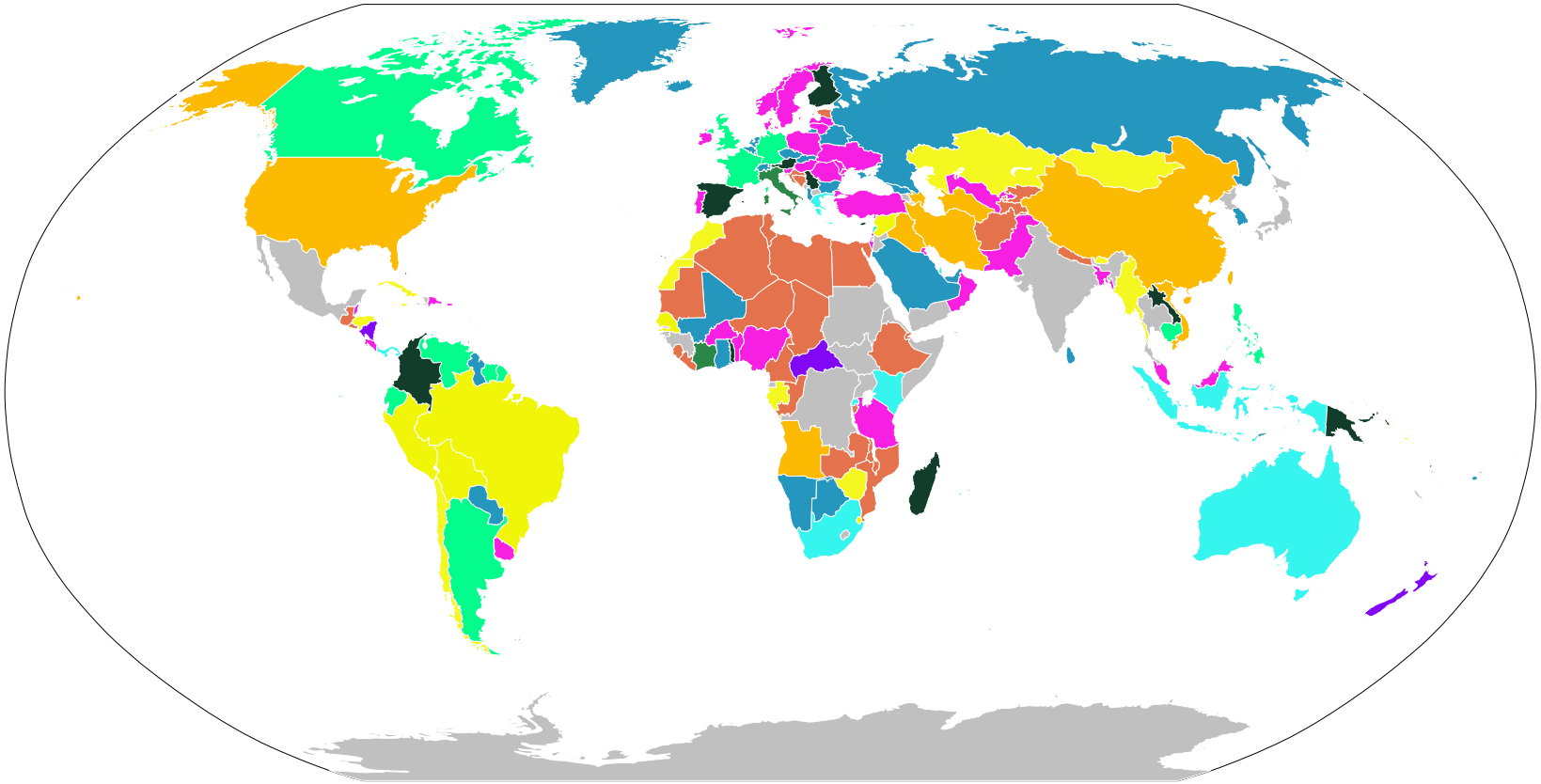
Knowlege Graph for Analysis of COVID-19 Policies (Policy Importance in 183 Nations)



Knowlege Graph for Analysis of COVID-19 Policies (Policy Importance in 183 Nations)



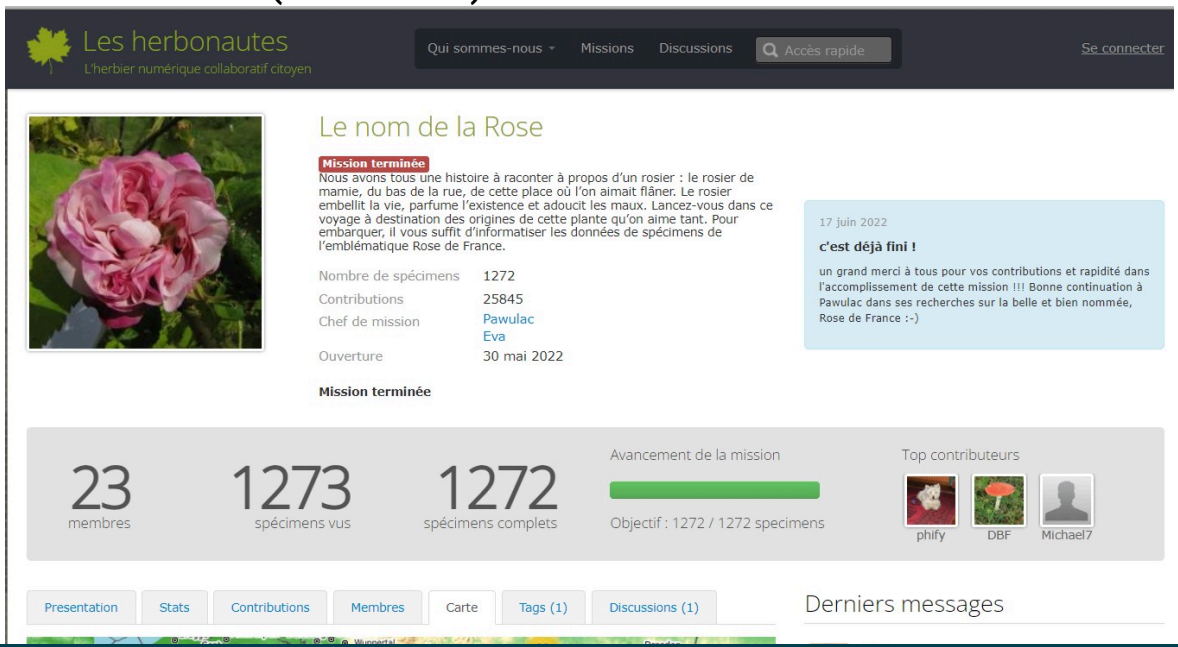
Nations clustered according to COVID-19 Policies Importance (Policy Importance in 183 Nations)



LLMs in Citizen Science Platforms

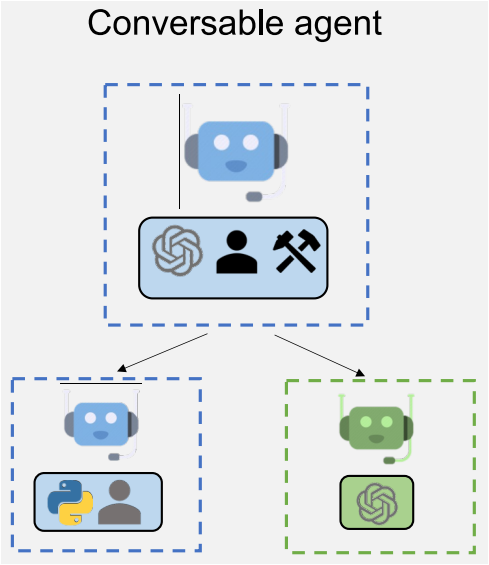
(Les Herbonautes, project with MNHN/Uni Paris)

- launched by Muséum national d'Histoire naturelle (MNHN)/Paris
- collaborative way of contributing to the creation of a scientific database based on the millions of photos of plants in the national network of naturalist collections (Récolnat)

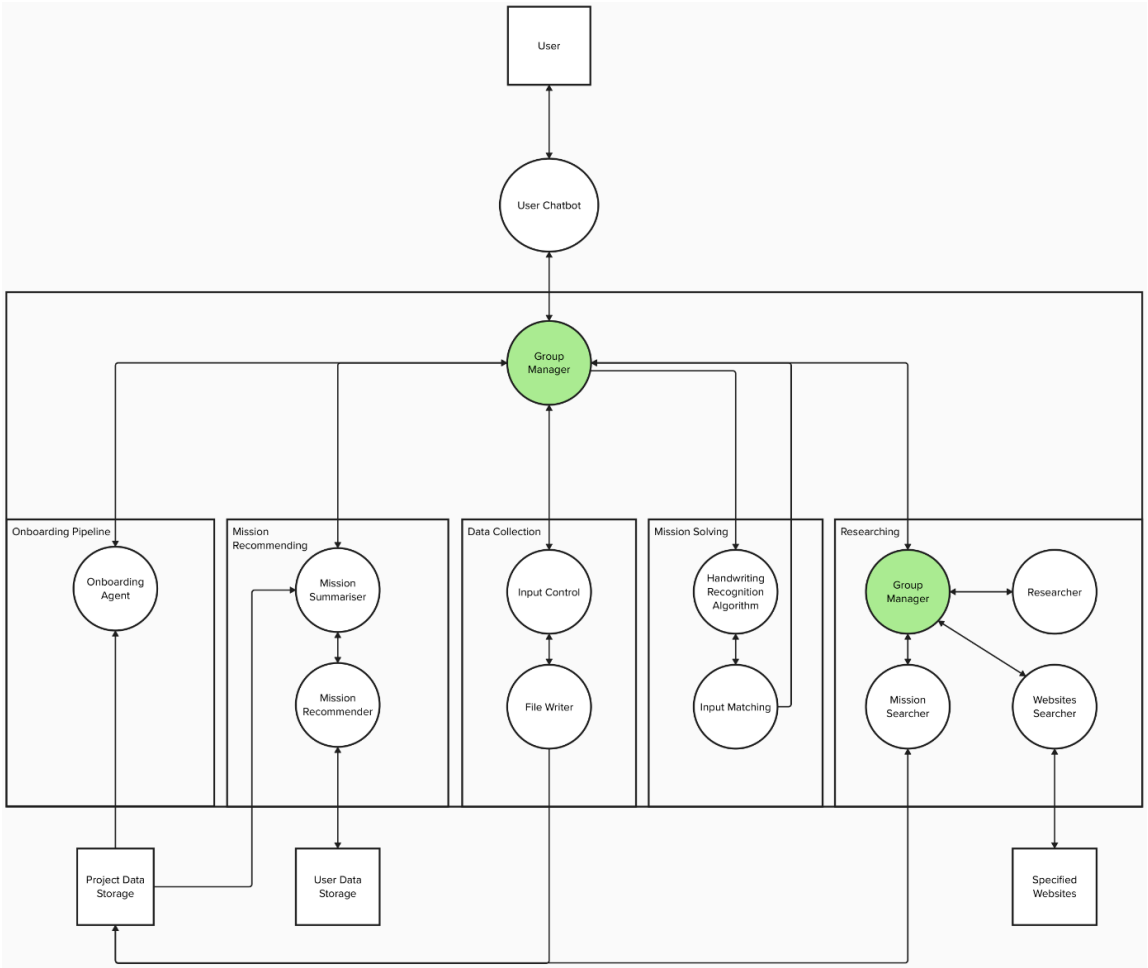


LLMs in Citizen Science Platforms

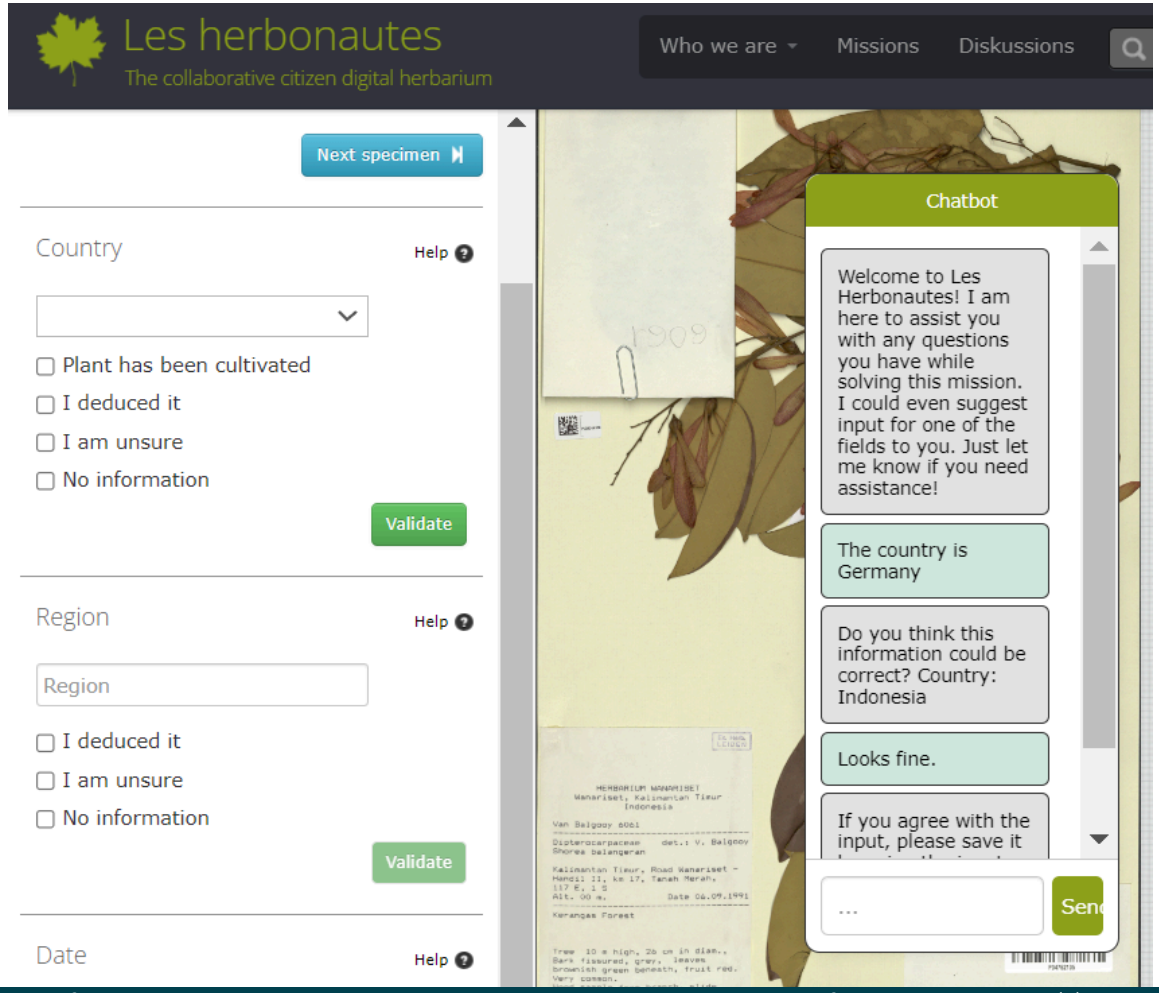
- LLM in form of a network of (LLM, human, code) agents as basis for an easy-to-use chatbot
  - for guiding users through the data validation process and with which users can retrieve important information by posing related questions for adding missing information (as some kind of easy-to-use tool for their research)
  - for high-level analysis with visual presentation of the analysis results in the form of maps, charts and/or in text
  - for detecting anomalies and marking any wrong data, or data and results of the chatbot they do not trust, as a trigger for data validation
  - for multi-lingual support



Multi-Agent Network in Citizen Science



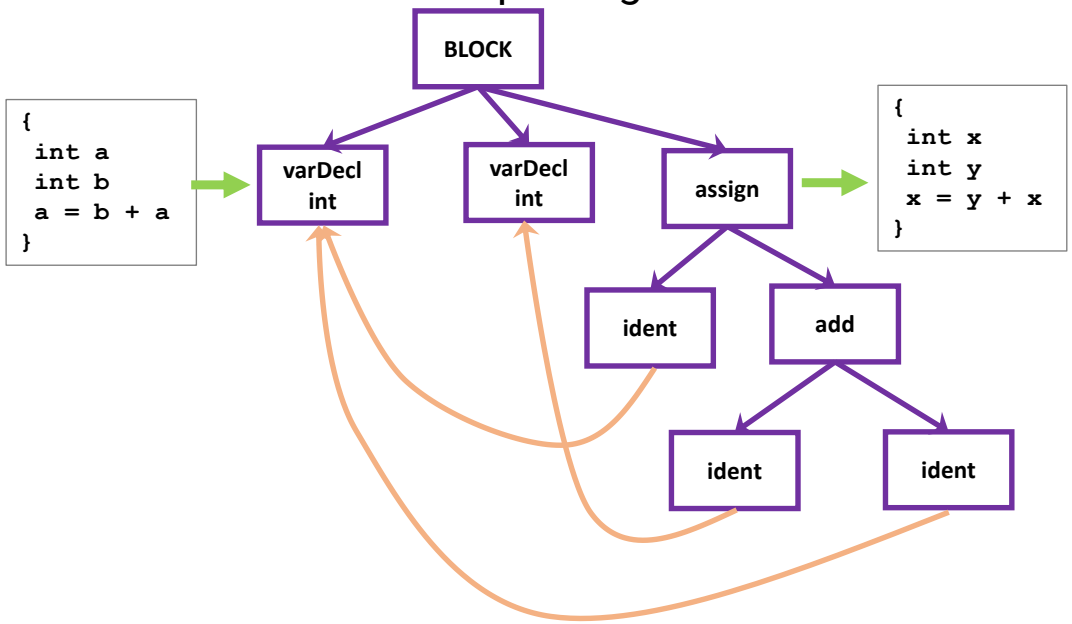
Multi-Agent Network in Citizen Science



AI in Cyber Security

Graph Neural Networks for Softw. Vulnerability Detection

- Input: Software code
- Output: Is the code vulnerable?
  - Future work: "Automatic" patch generation



Adding edges of name dependence

AI in Cyber Security

Graph Neural Networks for Softw. Vulnerability Detection

3-properties encoding per AST node (with variable names)

Construct	Class	Name	Type
int vname	varDecl	vname	int
char[6] cname	varDecl	cname	char[6]
if	control	if	-
0.05	literal	0.05	float
*****	mathOp	mul	-
fputs(...)	call	fputs	-
vname	ident	vname	int
stdout	ident	stdout	-
{...}	block	-	-

3-properties encoding per AST node (without variable names)

Construct	Class	Name	Type
int vname	varDecl	-	int
char[6] cname	varDecl	-	char[N]
if	control	if	-
0.05	literal	-	float
*****	mathOp	mul	-
fputs(...)	call	fputs	-
vname	ident	var	int
stdout	ident	stdout	-
{...}	block	-	-

AI in Cyber Security

Graph Neural Networks for Softw. Vulnerability Detection

- Evaluation
  - Improving accuracy

			Chromium+Debian		FFmpeg+Quemu		VDISC	
Model	Graph	Encoding	Acc	F1	Acc	F1	Acc	F1
codeAST	AST	code	92.01	30.20	55.36	57.01	77.82	75.57
3propASG	ASG	3-Prop.	**92.34**	**44.97**	**60.35**	62.30	**81.27**	**79.86**
codeAST+	AST+	code	90.89	25.86	58.38	46.66	75.67	74.49
3propASG+	ASG+	3-Prop.	**92.34**	44.59	57.04	**62.99**	80.94	79.63

  - Improving memory footprint

#nodes	#tokens	memory code-based	memory 3-prop.	code-based/3-prop.
4,409	33,659	59G	**5.3M**	11,220
7,012	54,157	152G	**8.4M**	18,052
12,077	96,805	468G	**14.5M**	32,268

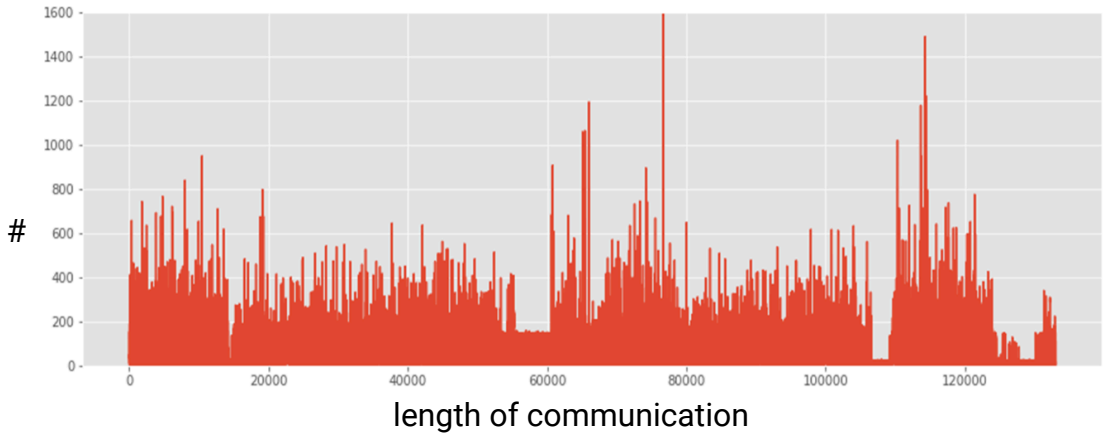
  - Reducing vocabularity explosion

Dataset	Code-Based	3-prop.	3-prop./Code-Based
Chromium+Debian	57,027	35,416	62.10%
FFmpeg+Qume	66,791	45,795	68.56%
VDSIC	449,148	312,948	69.68%

AI in Customer Communication

(Project with ZVO/Germany)

- Real-world data is heterogeneous



#classes	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
#samples	119655	12031	1079	199	31	16	6	7	2	3	5	3	2	2	4

Class	C0	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12	C13	C14	C15	C16	C17
#Comm	9770	21062	14245	235	1307	251	4009	3610	9410	23533	20676	43	2866	9617	8144	12126	4748	2700

AI in Customer Communication

(Project with ZVO/Germany)

- Performance of various approaches

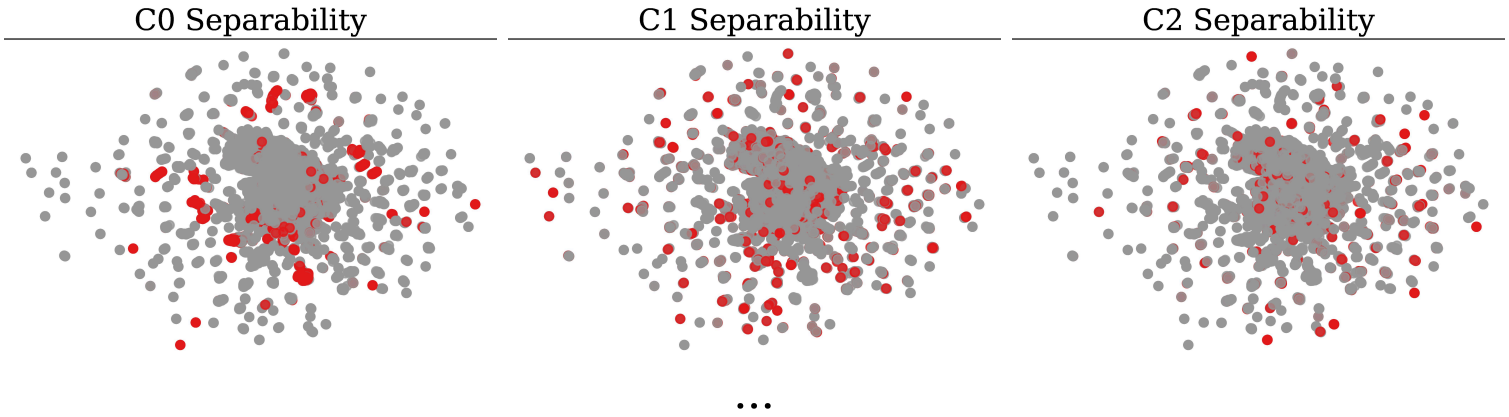
Model	MLP-TF	MLP-TFIDF	CNN-sl	CNN-G	CNN-G-u	CNN-F	CNN-F-u	LSTM-sl	LSTM-G	LSTM-G-u	LSTM-F	LSTM-F-u
Accuracy	55.1	53.2	58.7	50.6	57.5	55.1	56.4	57,4	53.5	61.2	51.2	58.0
Precision	61.6	59.2	66.3	57.2	64.8	62.4	64.4	64.4	60.0	69.0	57.3	65.2
Recall	60.4	58.2	65.8	56.4	64.1	62.1	64.3	62.6	58.3	67.9	55.6	63.7
AUC	79.7	78.7	82.2	77.6	81.5	80.4	81.5	80.6	78.5	83.2	77.2	81.1
Time (s)	961	285	1388	1805	1674	2282	1005	5728	14817	7973	22757	5775

- BERT-transfomer even worse

AI in Customer Communication

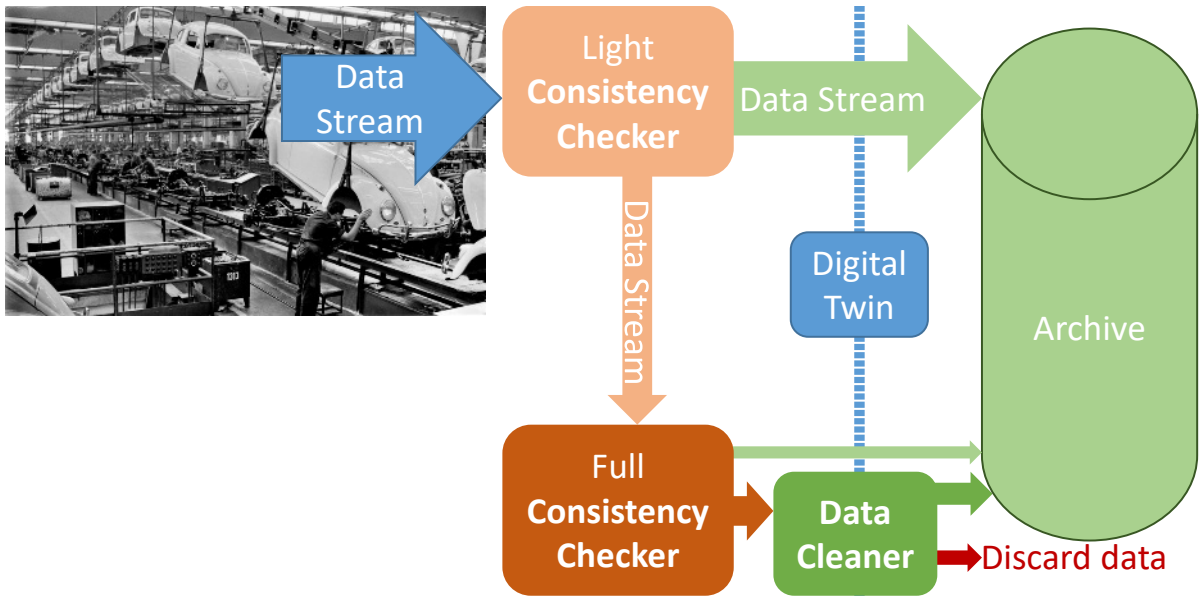
(Project with ZVO/Germany)

- Visualization of all communications
  - Similar communications are close to each other
  - Communications of one class are marked in red
    - not close to each other, difficult to group them
      ⇒ Low separability of classes



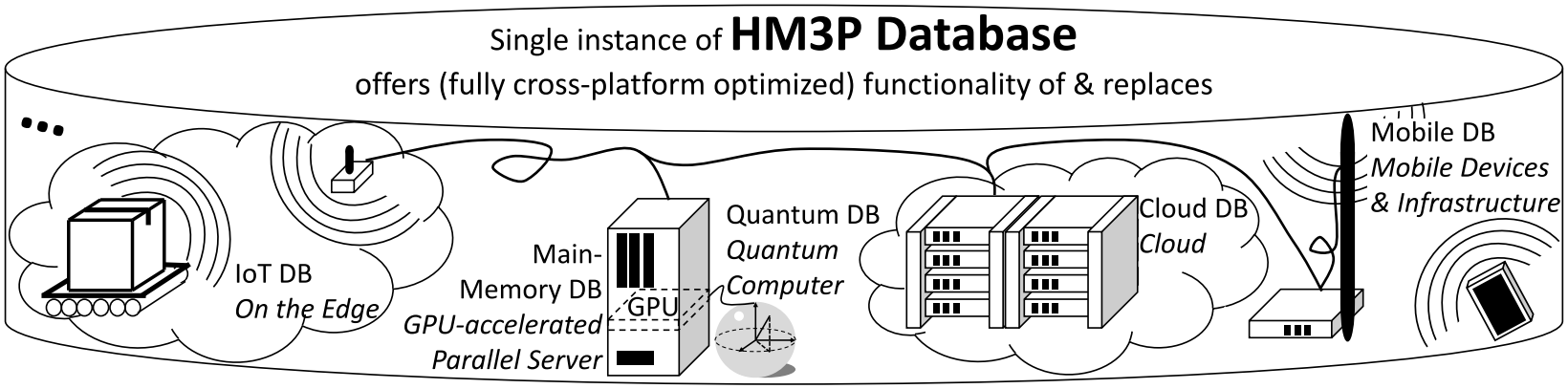
Green Computing in Industry 4.0

(Project with Bosch)



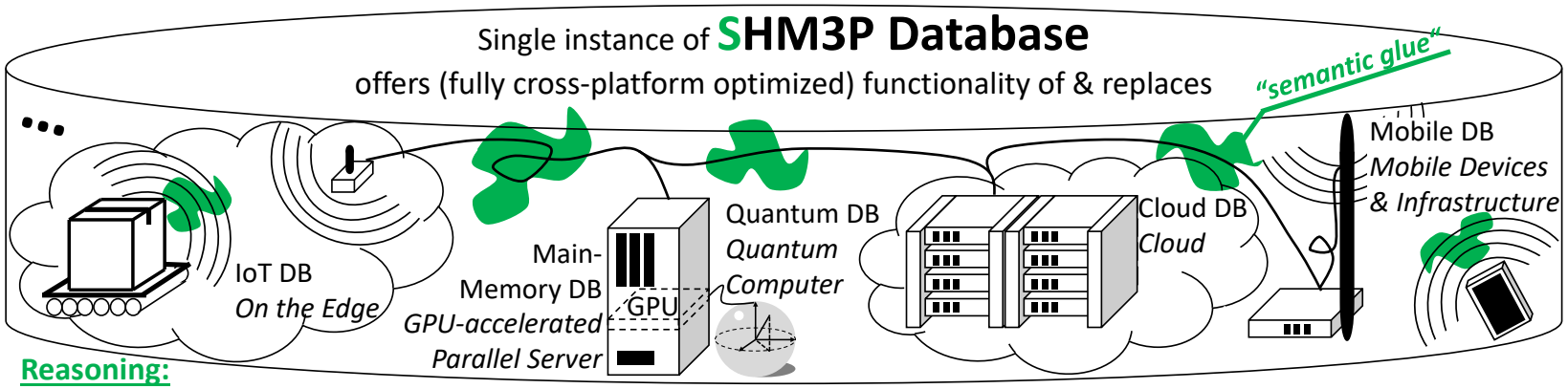
- Energy savings by lightweight components during normal operation and switching on full components for inconsistency handling
  - CO_2e emissions can be reduced by a factor of about 0.6
  - in one year 262 $kgCO_2e$ in EU for a medium-sized plant

Hybrid Multi-Model Multi-Platform (HM3P) Database



- full and uniform data integration at database level
- performance: fully optimized across different data models
- transparent fault-tolerance
- SQL standards: relational ('87), XML ('03), temporal ('11), JSON ('16), Multi-dimensional Arrays ('19), schemaless ('19), streams ('20?), property graphs ('21?)
- **features of different types of databases running on different platforms can be used**

Variant: Semantic HM3P (SHM3P) DB



- Semantic Layer as glue between other models and platforms
- new challenges like integrating different types of reasoners in a transparent global reasoner

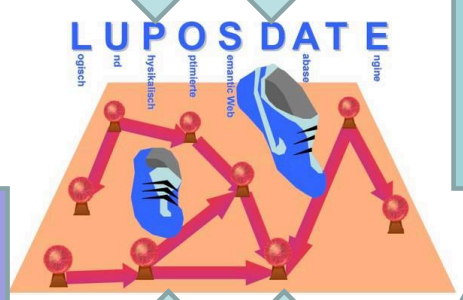
\+ Features of HM3P databases
\+ Easier data integration

\- Performance issues may occur due to semantic layer

Semantic Web DBMS LUPOSDATE

Support of:
- SPARQL Queries
- RIF Rules
- RDF Schema
- OWL (via OWL2RL in RIF)

Indexing:
- Stream Processing
- Main memory for small datasets
- Disk-based for large datasets
  - RDF3X
- Cloud: HBase
- P2P

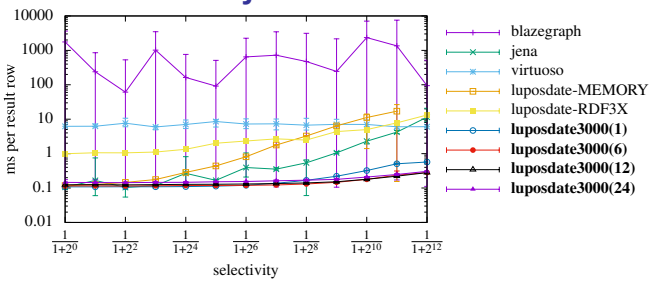


Visualizations:
- Visual Editor
  - Queries (SPARQL)
  - Rules (RIF)
  - Data (RDF) in
    - 2D and
    - 3D
  - Logical Optimization Rules
- Summaries of RDF Data
- Operator graph
- Processing of Queries and Rules
- Optimization Steps

Extra:
- Parallel Processing
- Distributed Processing
- Cloud Computing
- Mobile Computing
- P2P for Internet of Things
- Compression of RDF Data
- Embedding of SW Languages in Programming Languages
- Speeding up by FPGAs

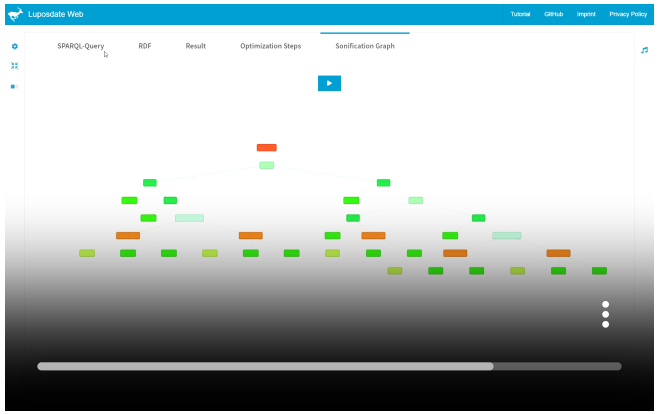
The Power of Multi-Platform: LUPOSDATE3000

- LUPOSDATE300: Open Source Multi-Platform SW database
  - Design Choice: Implemented in Kotlin for multi-platform support
  - ultra-fast in jvm...



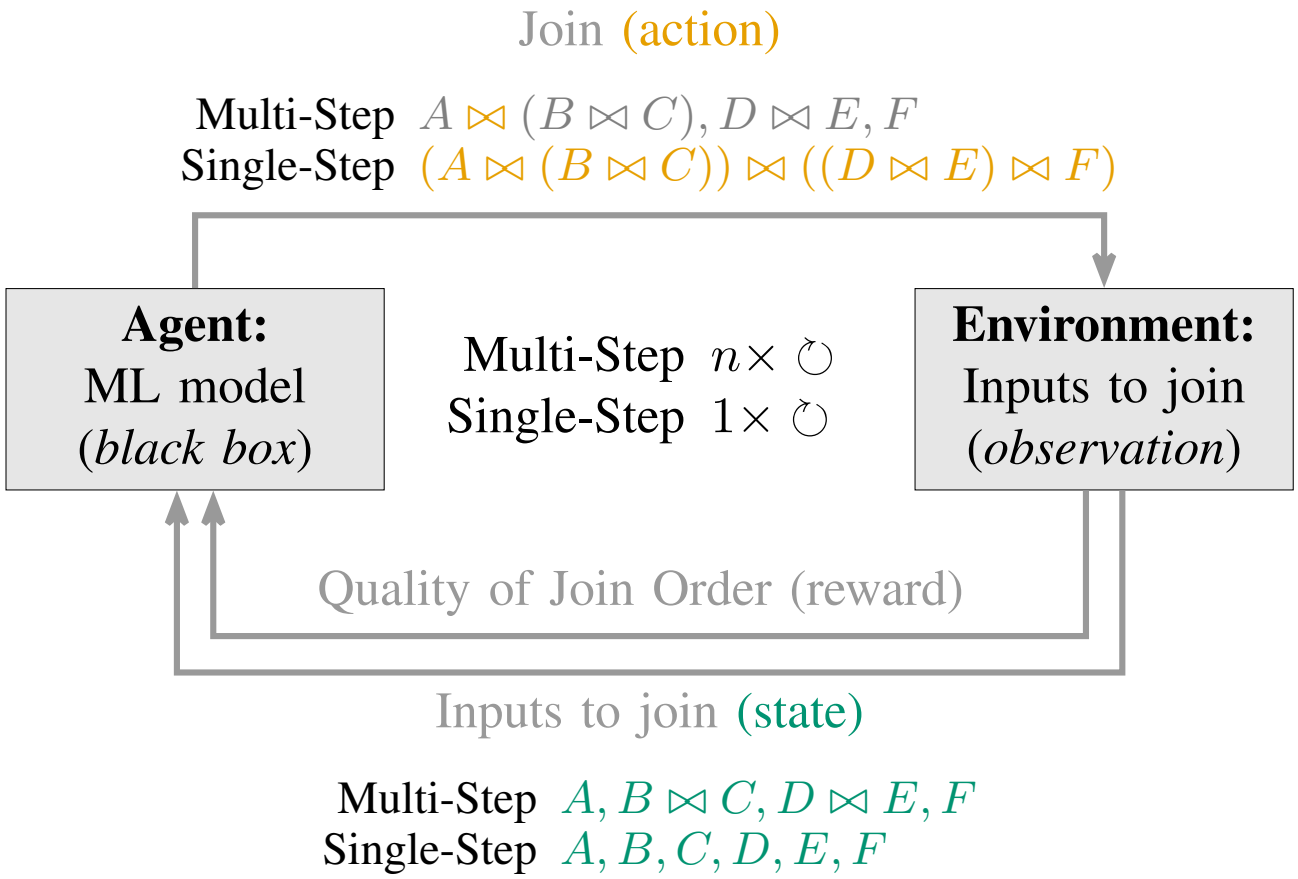
B. Warnke, M.W. Rehan, S. Fischer, S. Groppe: Flexible data partitioning schemes for parallel merge joins in semantic web queries in: BTW'21

  - ...but also enabling web demos running completely in the browser!

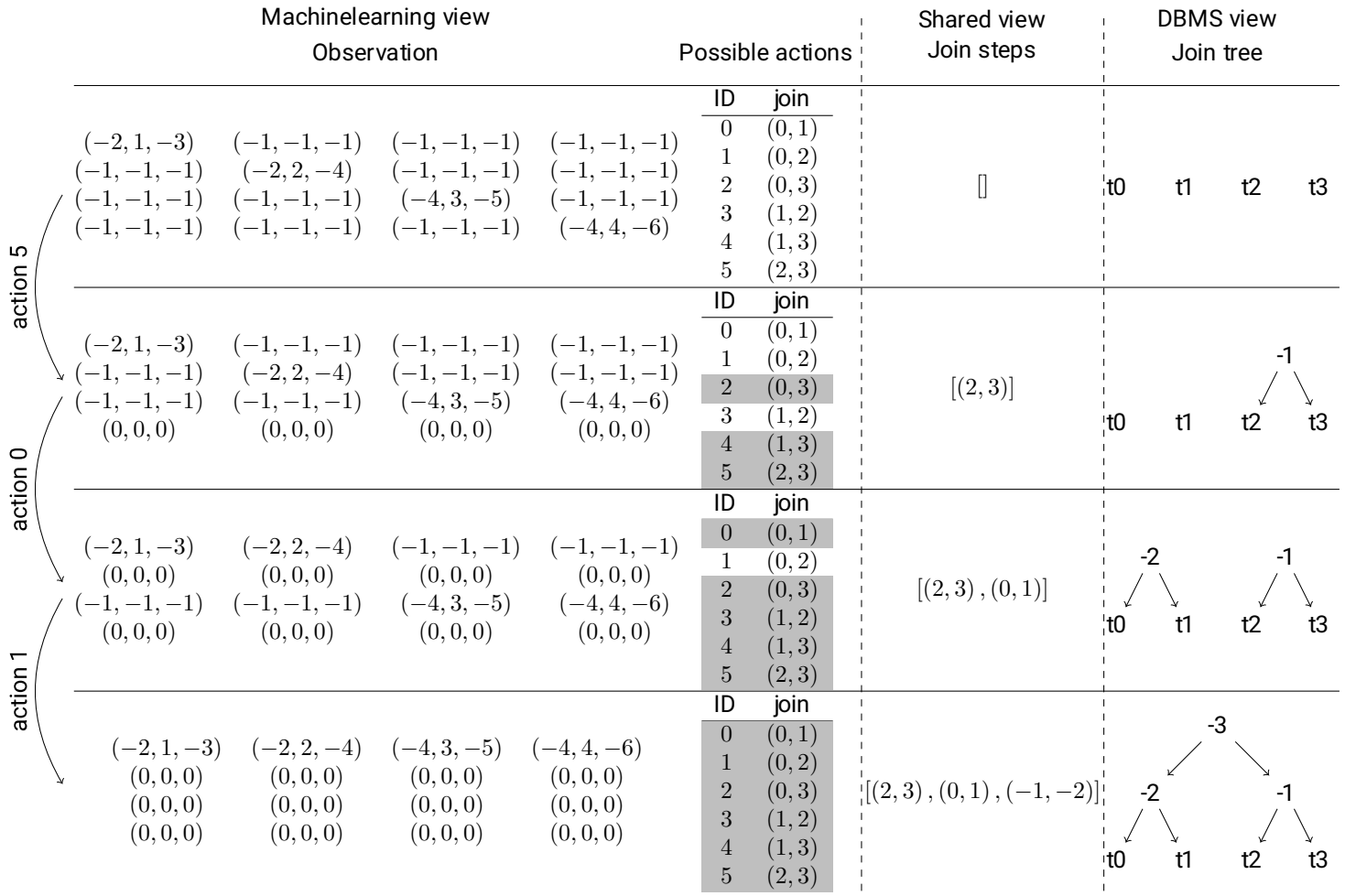


S. Groppe, R. Klinckenberg, B. Warnke. Sound of Databases: Sonification of a Semantic Web Database Engine. PVLDB, 14(12), 2021

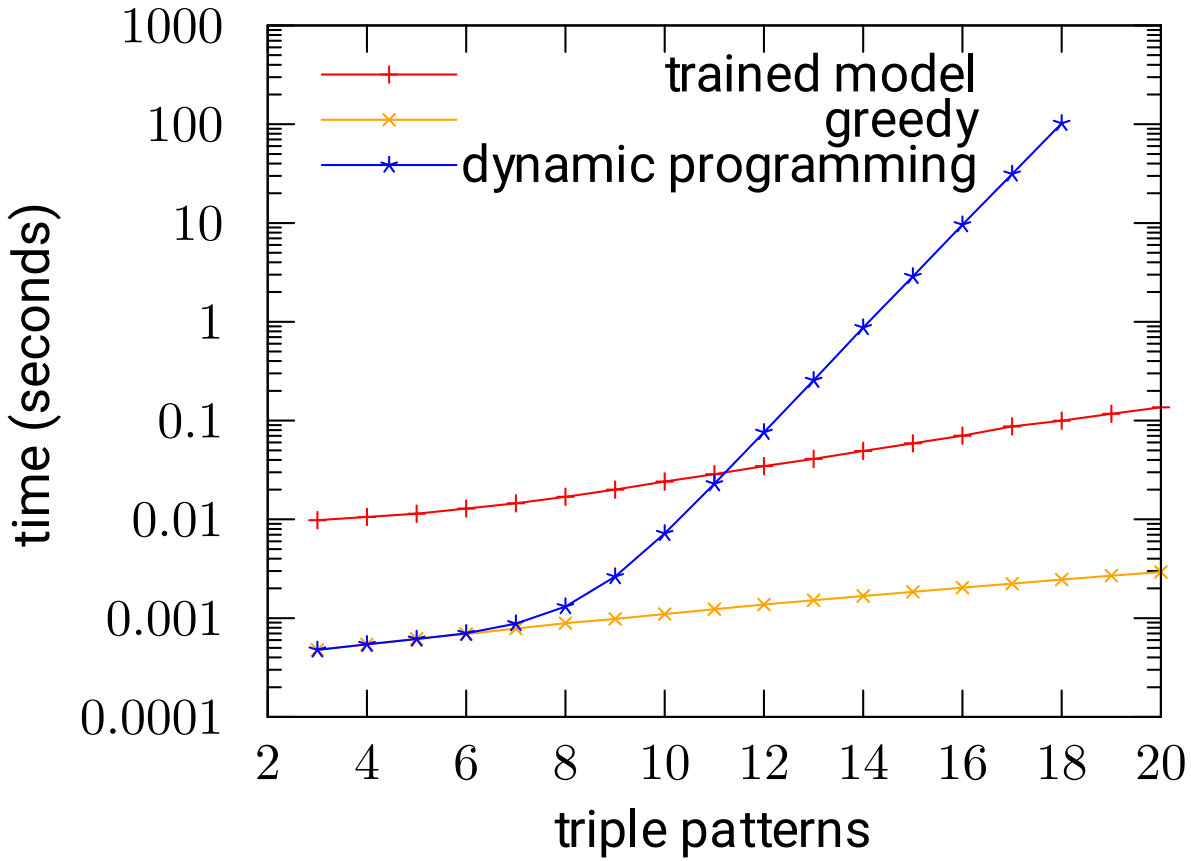
Reinforcement Learning for Query Optimization



Reinforcement Learning for Query Opt.

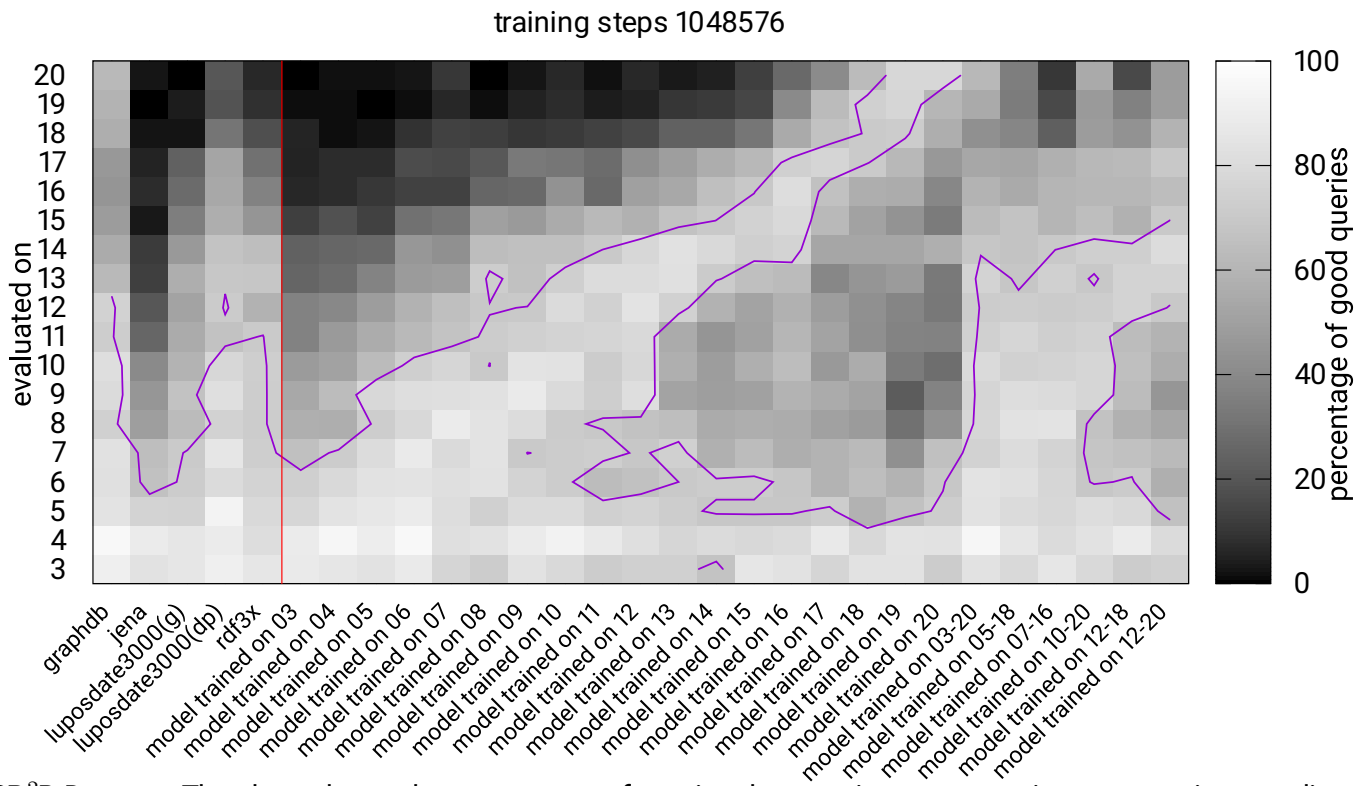


Execution Times of Query Optimization



ML model: maskable Proximal Policy Optimization (PPO) model(25) of stable baselines3 1.5.0, Python 3.9.7

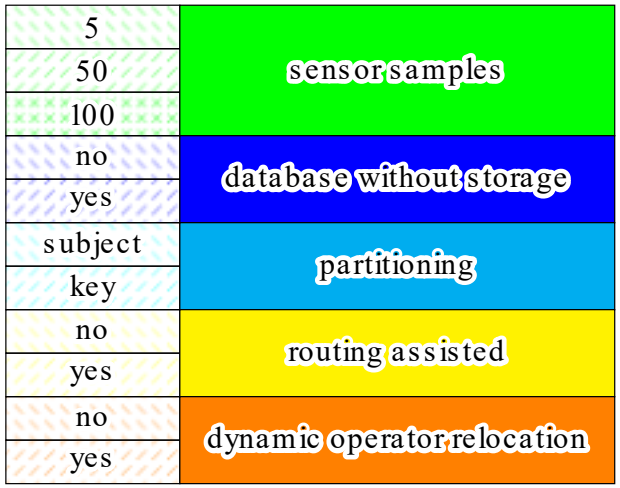
Query Opt. via Reinforcement Learning

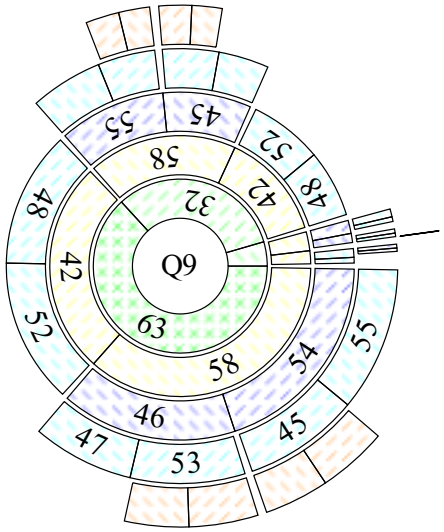


SP^{2}B-Dataset. The chart shows the percentage of queries that require at most twice as many intermediate results as the known best case. The magenta line surrounds the area, where 70% of the queries receive good join orders.

Query Opt. & Routing in IoT

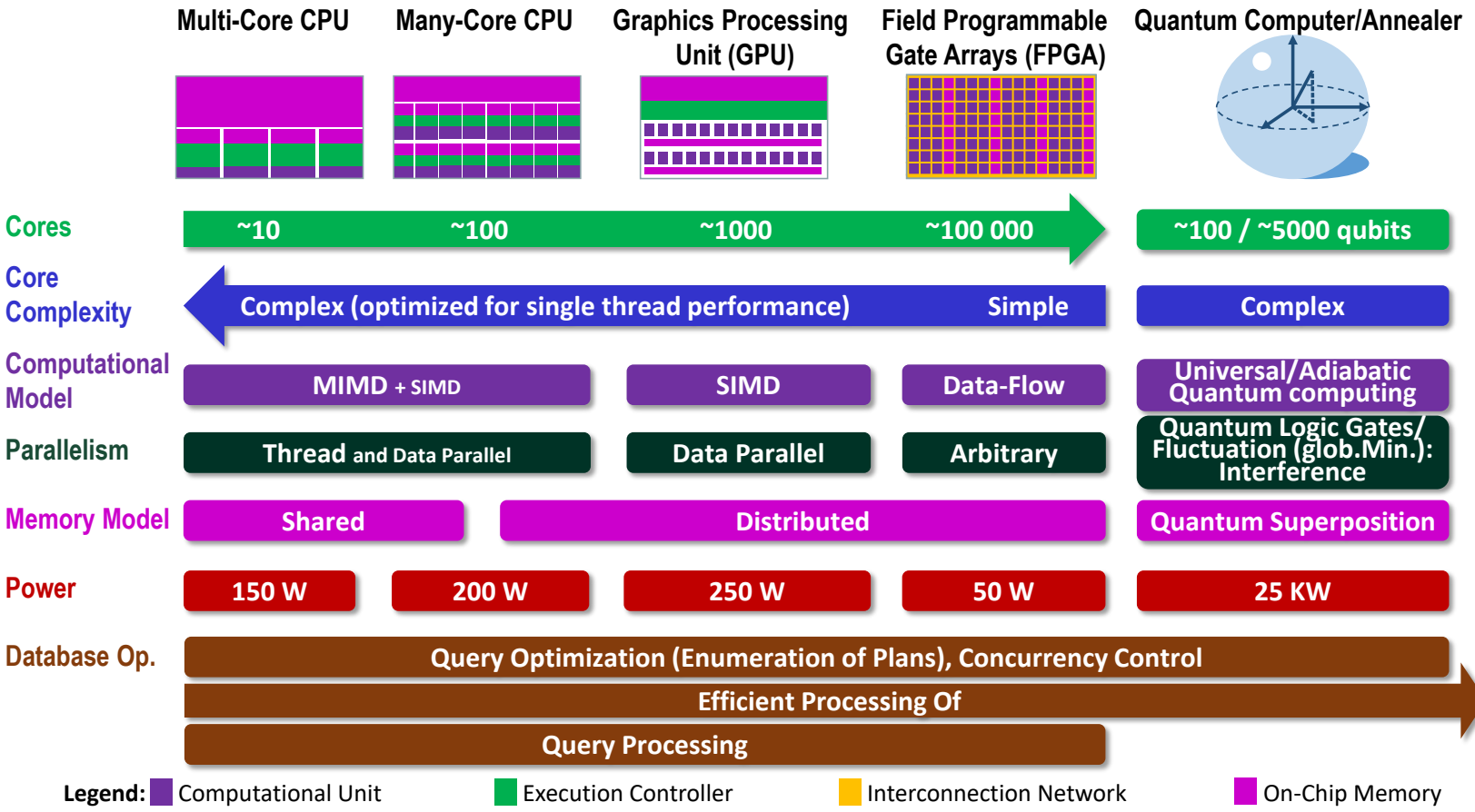
- Design Choice: Combining Routing and Query Optimization in IoT
  - Q9 with 11 triple patterns: Network traffic reduces by
    - routing-assisted join order (instead of centralized query optimization)
    - dynamic relocation of operators whenever the result of a join is more significant than its inputs (4% for large and 23% for small data scales)

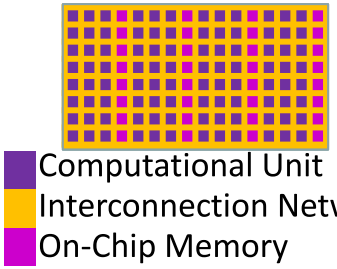




Transferred data for benchmark queries in percent (smaller ⇝ better) dependent on option. More significant effects are in the center of the circle.

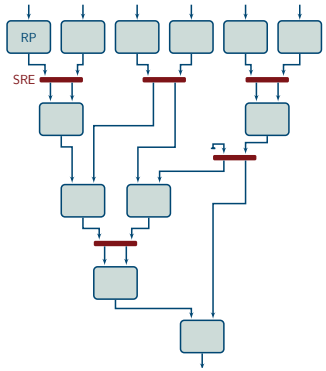
Architectures of Emergent Hardware





LUPOSDATE on FPGA

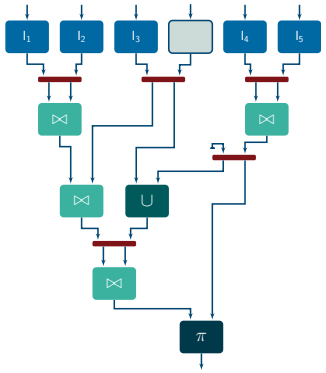
- Design Choice:
  - At system deployment time generated RMs are configured into RPs at system runtime
    ⇝ avoids long synthesis time



SP²B Query 4

➡

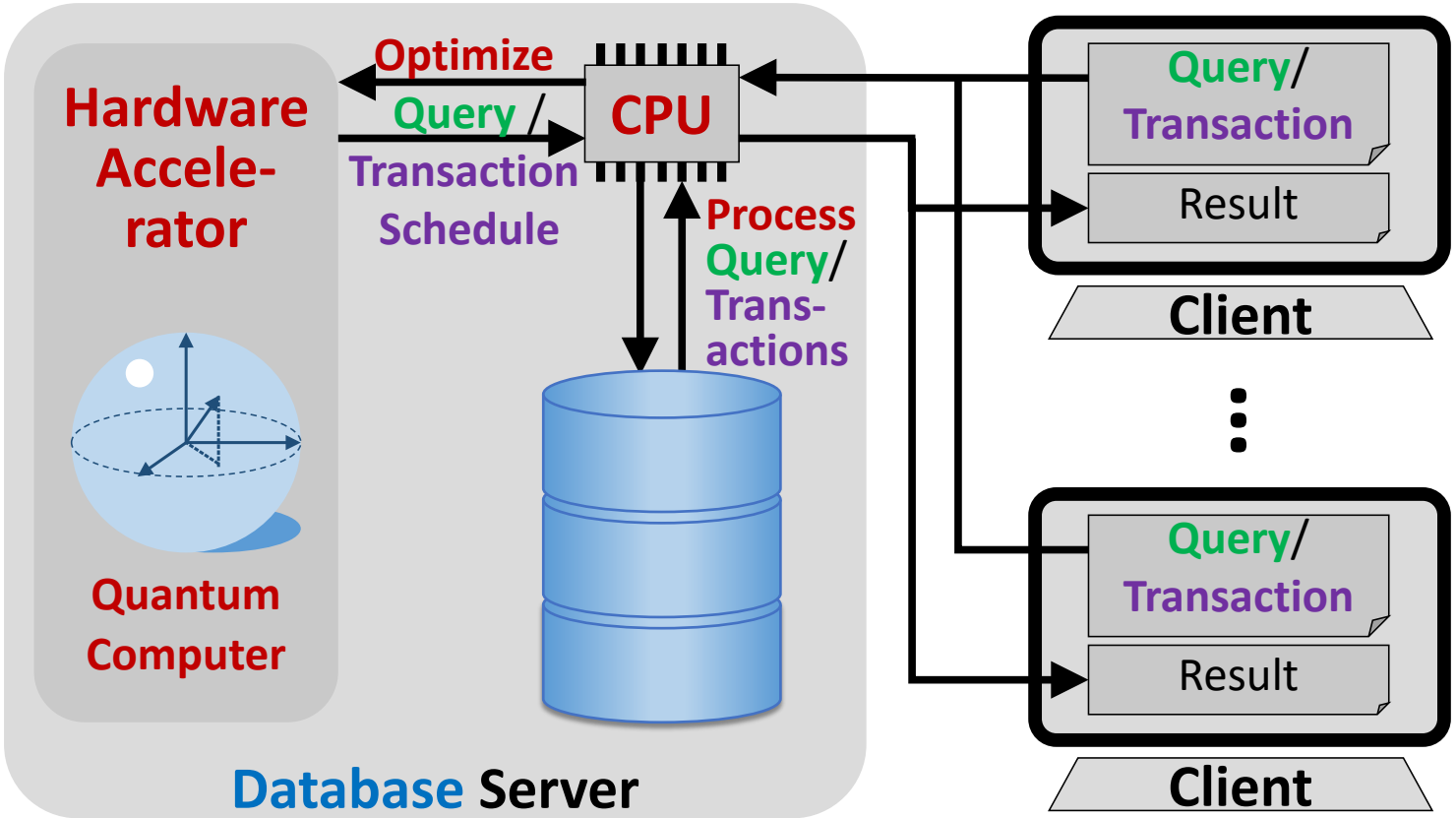
RM: Reconfigurable Module
RP: Reconfigurable Partition
SRE: Semi-static Routing Element



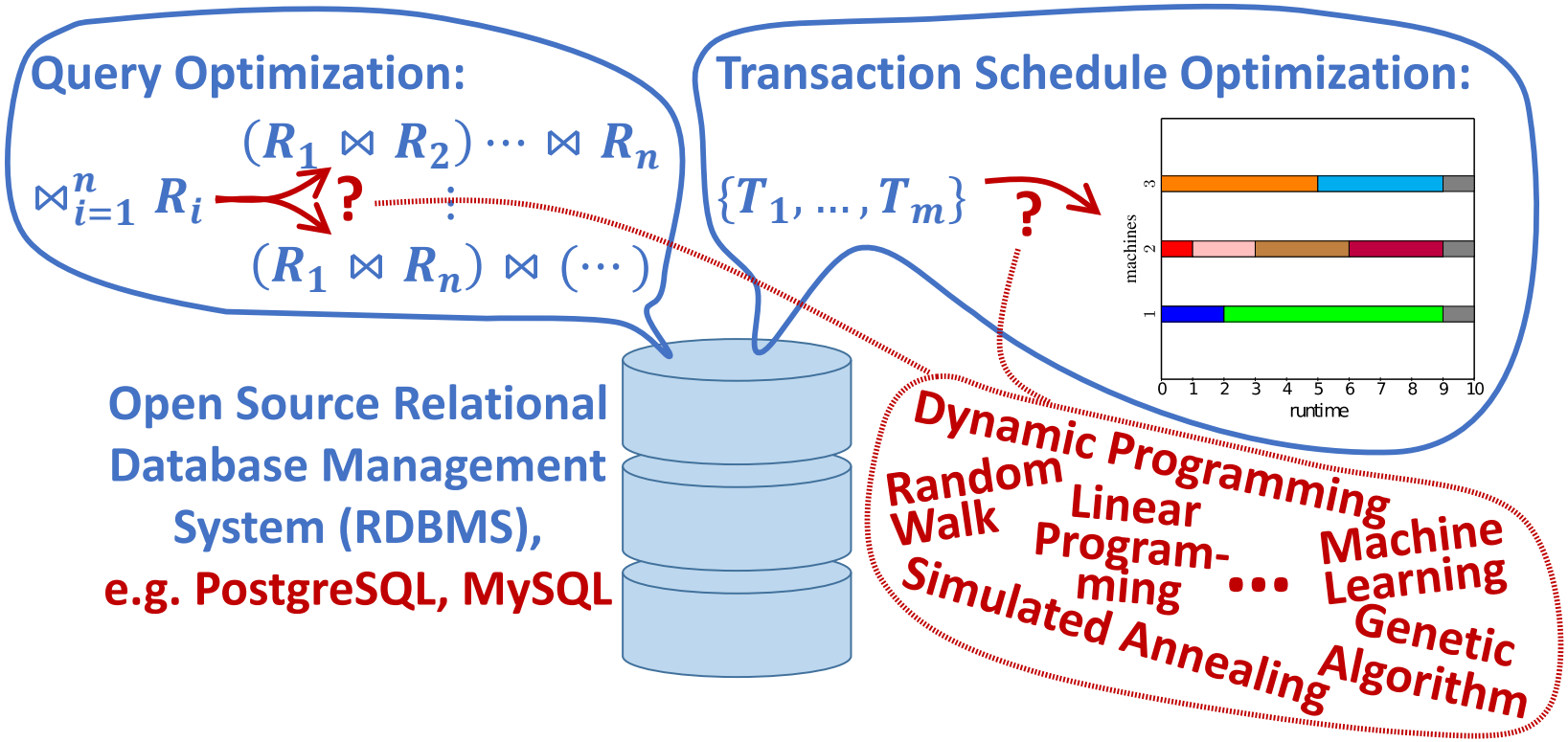
- Benchmark Results
  - Reconfiguration reduced from about half hour to few milliseconds (< 20 ms for all queries) when using semi-static operator graphs
- SP²B Benchmark
  - Dataset sizes from 66 to 262 million triples
  - **Speedups between 4 and 32 times** (dependent on query and dataset size)

Timeline of Quantum Computers

Using Hardware Accelerator for optimizing Queries / Transaction Schedules]

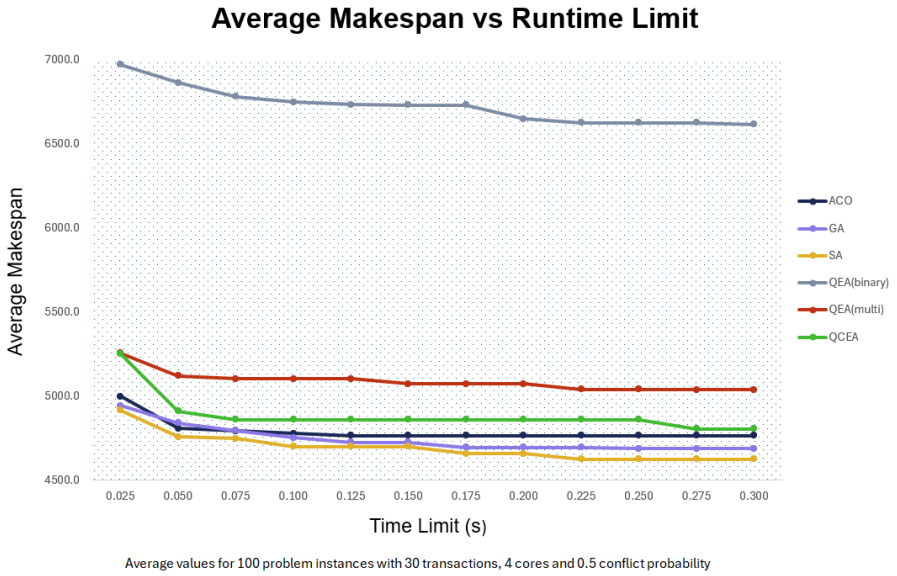


Approaches for Query/Transaction Schedule Optimization



Transaction Scheduling

- Design choice: Fast quantum exact methods and good heuristics for scalable optimization times
- Classical
  - Dynamic Programming | Random Search | Simulated Annealing (SA) | Genetic Algorithm (GA) | Ant Colony Optimization (ACO)
- Quantum
  - Maximum Independent Set | Quantum Evolutionary Algorithms (QEA) (several variants: Binary | (Cultural (QCEA)) Multi-State) | QUBO | Grover



Average values for 100 problem instances with 30 transactions, 4 cores and 0.5 conflict probability

Transaction Scheduling via Quantum Annealing

- Design Choice: Using Quantum Annealing for constant execution time
- Experiments on real Quantum Annealer (D-Wave 2000Q cloud service)
  - first minute free
    (afterwards too much for our budget)
- Versus Simulated Annealing on CPU
- Preprocessing time/Number of QuBits: $O((n \cdot k \cdot R)^2)$

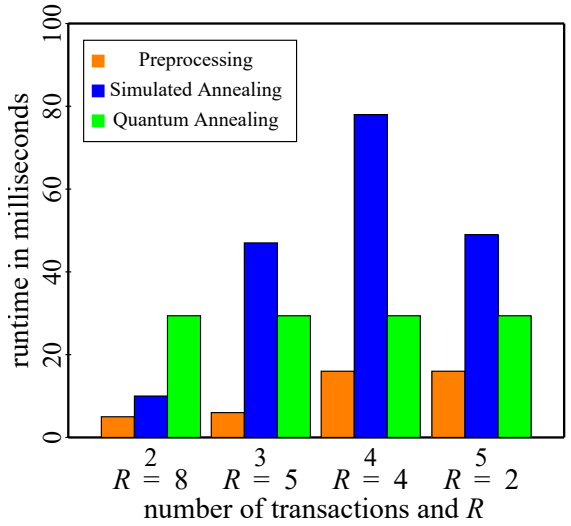
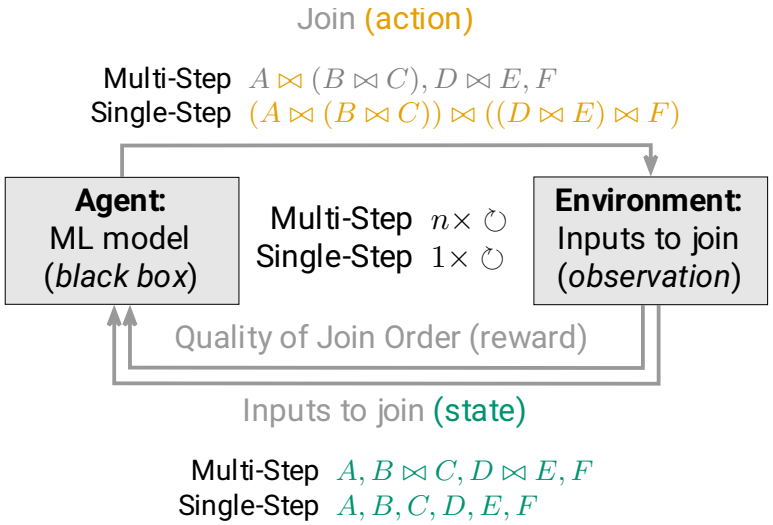


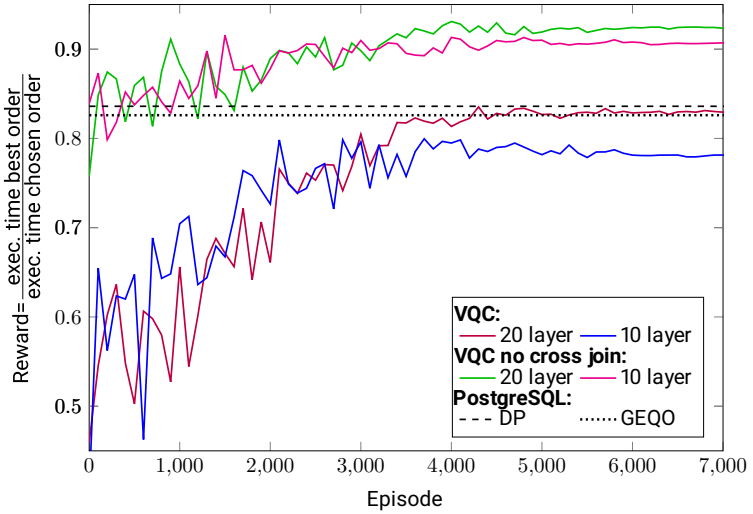
Fig.	k	n	R	O	$l_1, \ldots, l_n$	$r_1, \ldots, r_n$	req. var.
11	2	2	8	$\{\}$	8, 4	0, 4	8
		3	5	$\{(t_1, t_3)\}$	4, 5, 1	1, 0, 4	10
		4	4	$\{(t_2, t_4)\}$	3, 2, 1, 2	1, 2, 3, 2	16
		5	2	$\{(t_1, t_2), (t_4, t_5)\}$	1, 1, 1, 1, 1	1, 1, 1, 1, 1	10

k: #cores, n: #transactions, R: max. makespan, O: conflict matrix, l_x: transaction length, $r_x = R - l_x$

QML for Join Ordering

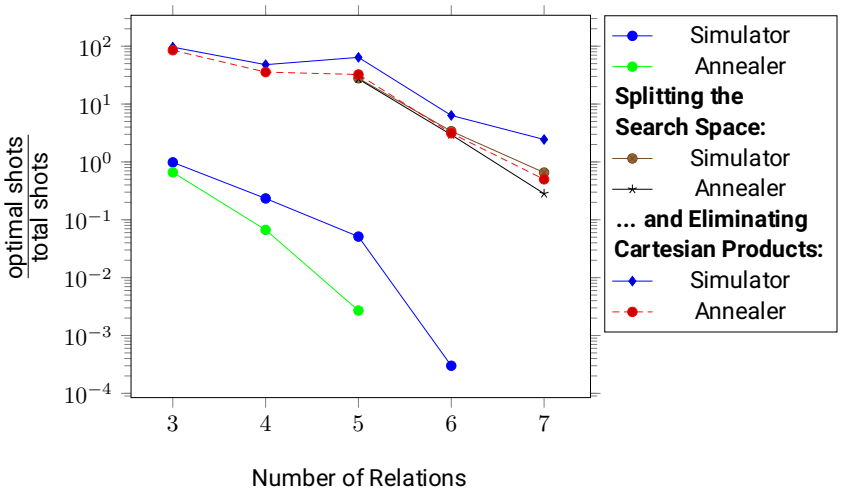
- Design choice: Predict best join order based on real execution times in contrast to estimated costs (EC)
  - can beat exact methods like dynamic programming (DP) based on EC
- Single-Step: Real-world queries (ErgastF1 Benchmark/PostgreSQL)
  - join orders with faster execution than DP (2.8%) and GEQO (11.2%)
  - close to classical ML like RTOS (6%) / best join orders (16.8%)

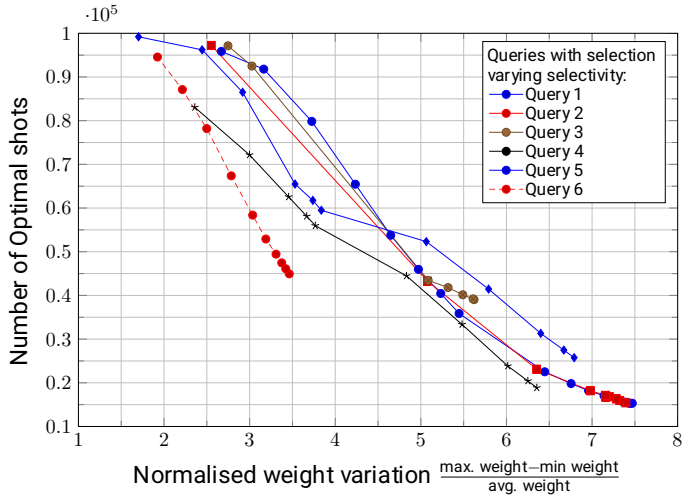




Join Ordering as QUBO Problem

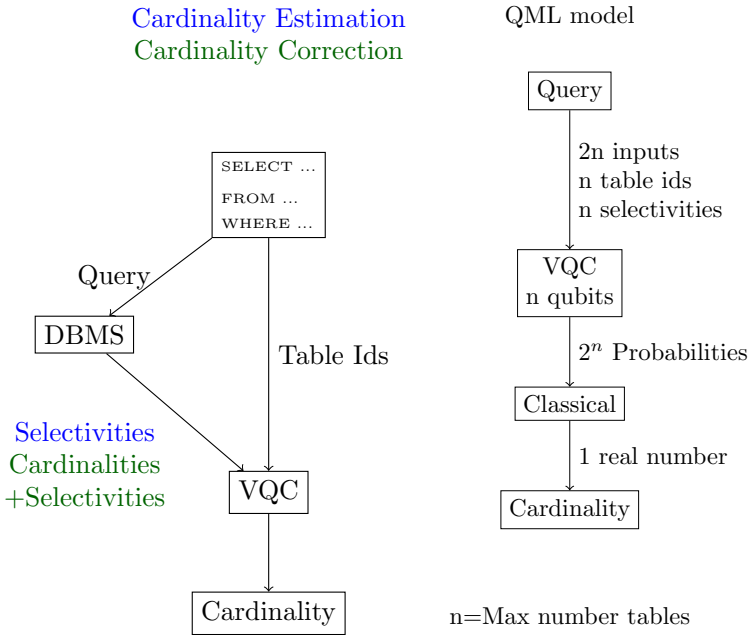
- Design choice: Direct Join Costs in contrast to calculated costs
  - Estimated costs for **each** join $\Rightarrow$ better join order
  - Runtime complexity $O(2^m - m)$ with m relations to join optimal for join ordering based on direct costs
    - Classical runtime: $O(3^m - 2^{m+1})$ using dynamic programming
- Real-world queries (ErgastF1 Benchmark/PostgreSQL)

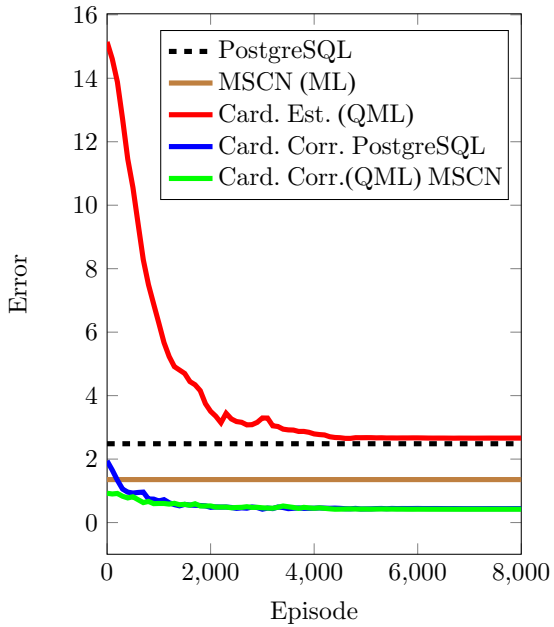




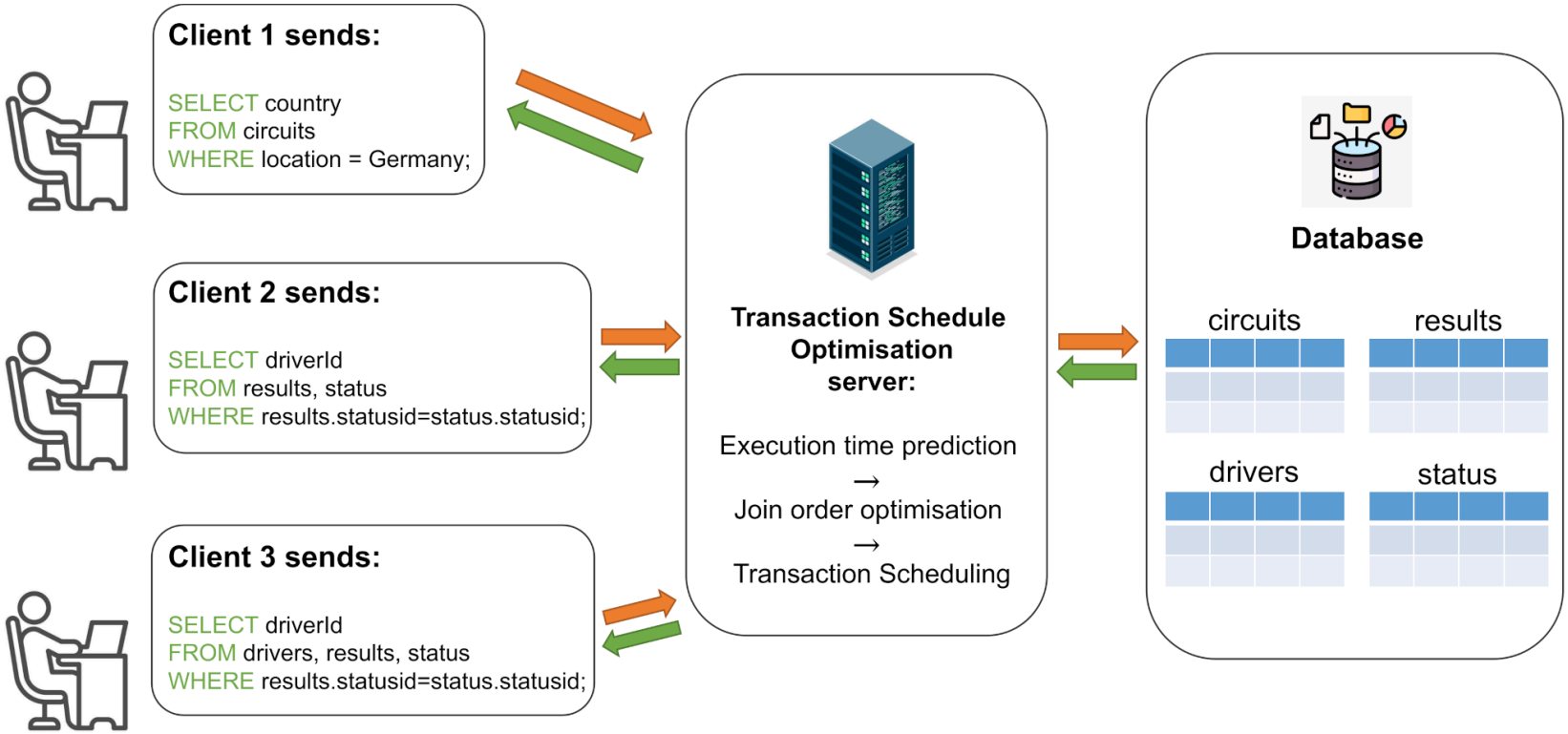
QML for Cardinality Estimation

- Design choice: Predict correction factor of a Cardinality Estimator
  - Cardinality Predictor:PostgreSQL|MSCN|Hybrid Quantum Classical Network
  - Cardinality Correction: Hybrid Quantum Classical Network
- Queries of JOB-light Benchmark





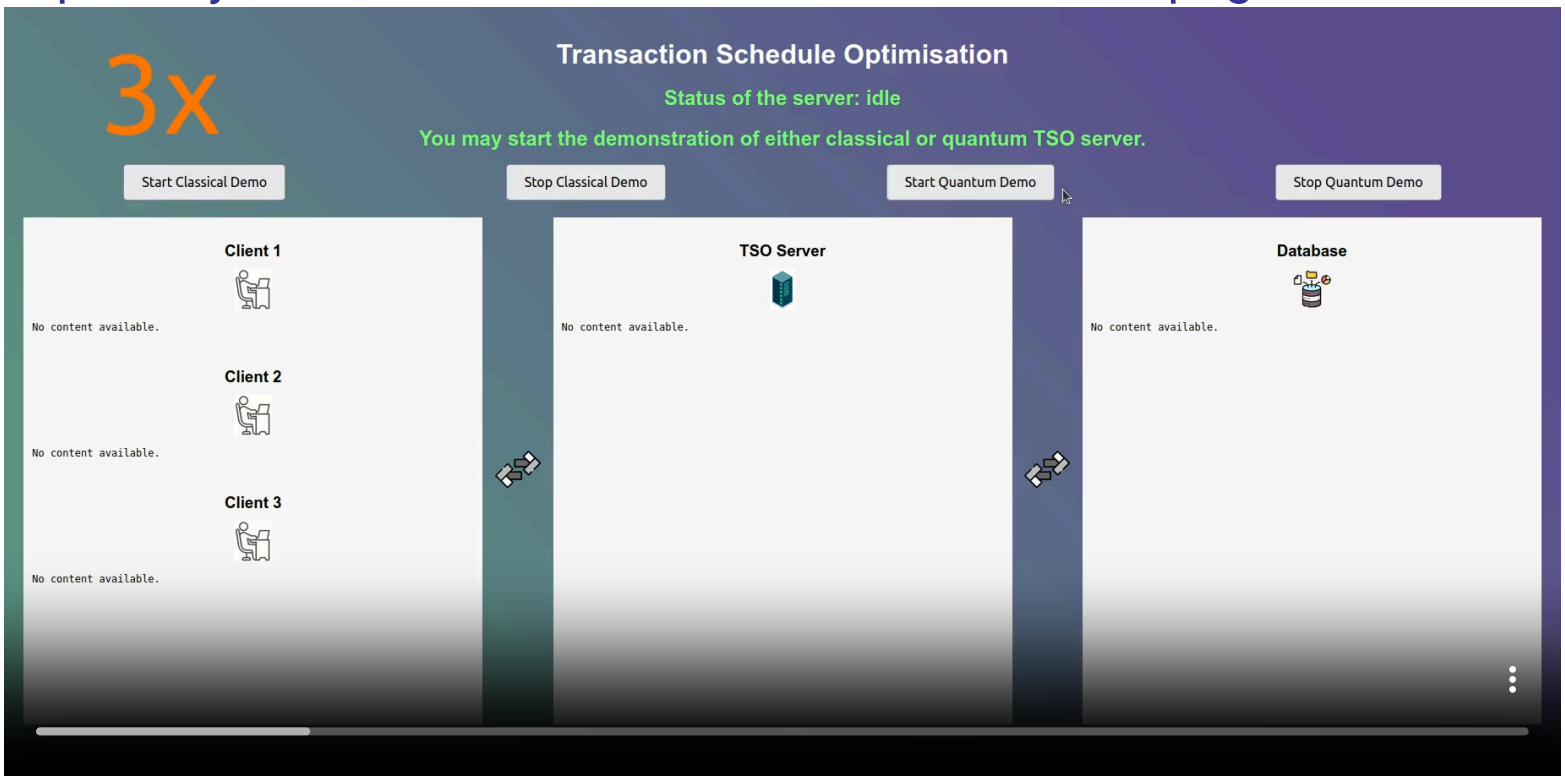
Transaction Scheduling Optimisation Server



- Demonstrator: fully quantum **versus** fully classical

Demo for Query/Transaction Schedule Opt./Exec. Time Prediction

- is publicly available at the Quantum Brilliance web-page...



Summary

- Teaching
  - 50 lectures
  - Easy-to-start tutorials for students for much practical expertise
  - Many supervisions of students at PhD, master and bachelor level
  - Involving students in research contributions (so far 25% of publications)
- Research
  - $\approx$ 2M € project grants
  - $>$ 191 publications
  - Well integrated in the top research community as shown by numerous scientific services
  - Main topics
    - Advanced LLM Applications (KG Construction, Chatbots, biodiversity, ...)
    - Software Vulnerability Prediction
    - Green Computing
    - (Quantum) Machine Learning for Query Optimization
    - Quantum Computing for Data Management
    - Semantic Web