DIMA/BIFOLD 2025

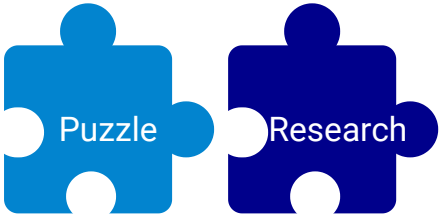
Jigsaw Puzzle of Research for Data Management

13.10.2025

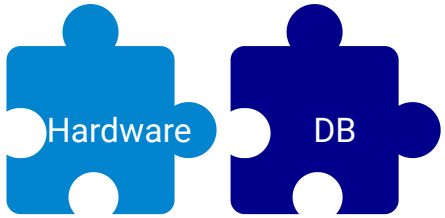
Professor Dr. rer. nat. habil. Sven Groppe
https://www.ifis.uni-luebeck.de/~groppe

Jigsaw Puzzle of Research for Data Management

- "Rules" for this talk
  - Only very short introduction into different research areas
  - Collection of concrete examples of "interdisciplinary" research results
    - 1 slide for each example (exceptional 2-3 slides) besides few slides for introduction of larger topics
    - Focus on result and showing benefits, advantages and potential of combining emergent technologies on emergent hardware architectures
    - Not much details about the how
  - Puzzle pieces for emphasizing what has been combined
  - References are provided for further reading

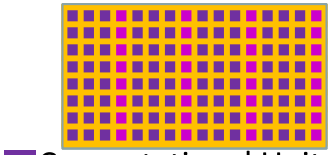


Architectures of Emergent Hardware



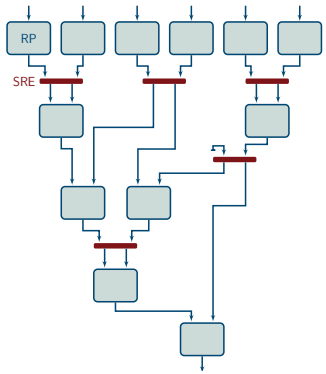
	Multi-Core CPU	Many-Core CPU	Graphics Processing Unit (GPU)	Field Programmable Gate Arrays (FPGA)	Quantum Computer/Annealer
Cores	~10	~100	~1000	~100 000	~100 / ~5000 qubits
Core Complexity	Complex (optimized for single thread performance)			Simple	Complex
Computational Model	MIMD + SIMD		SIMD	Data-Flow	Universal/Adiabatic Quantum computing
Parallelism	Thread and Data Parallel		Data Parallel	Arbitrary	Quantum Logic Gates/ Fluctuation (glob.Min.): Interference
Memory Model	Shared	Distributed			Quantum Superposition
Power	150 W	200 W	250 W	50 W	25 KW
Database Op.	Query Optimization (Enumeration of Plans), Concurrency Control				
	Efficient Processing Of				
	Query Processing				

Legend: Computational Unit, Execution Controller, Interconnection Network, On-Chip Memory



LUPOSDATE on FPGA

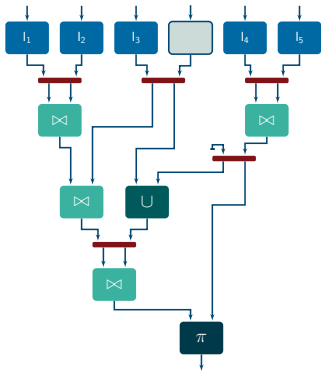
- Design Choice:
  - At system deployment time generated RMs are configured into RPs at system runtime
    ⇝ avoids long synthesis time



SP²B Query 4

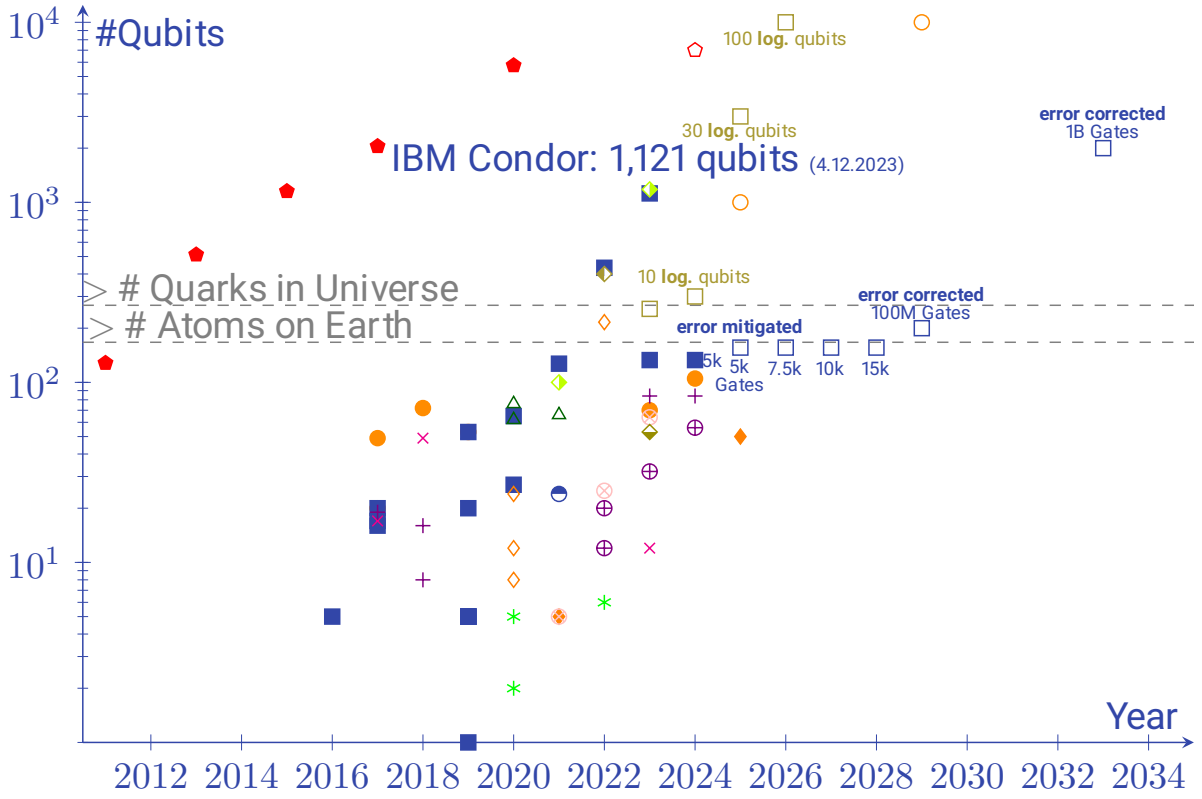
➡

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



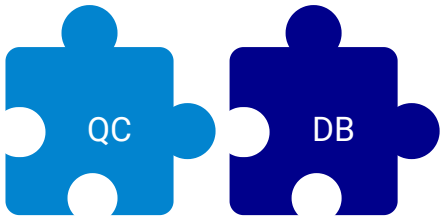
- Benchmark Results
  - Reconfiguration reduced from about half hour to few milliseconds (< 20 ms for all queries) when using semi-static operator graphs
- SP²B Benchmark
  - Dataset sizes from 66 to 262 million triples
  - **Speedups between 4 and 32 times** (dependent on query and dataset size)

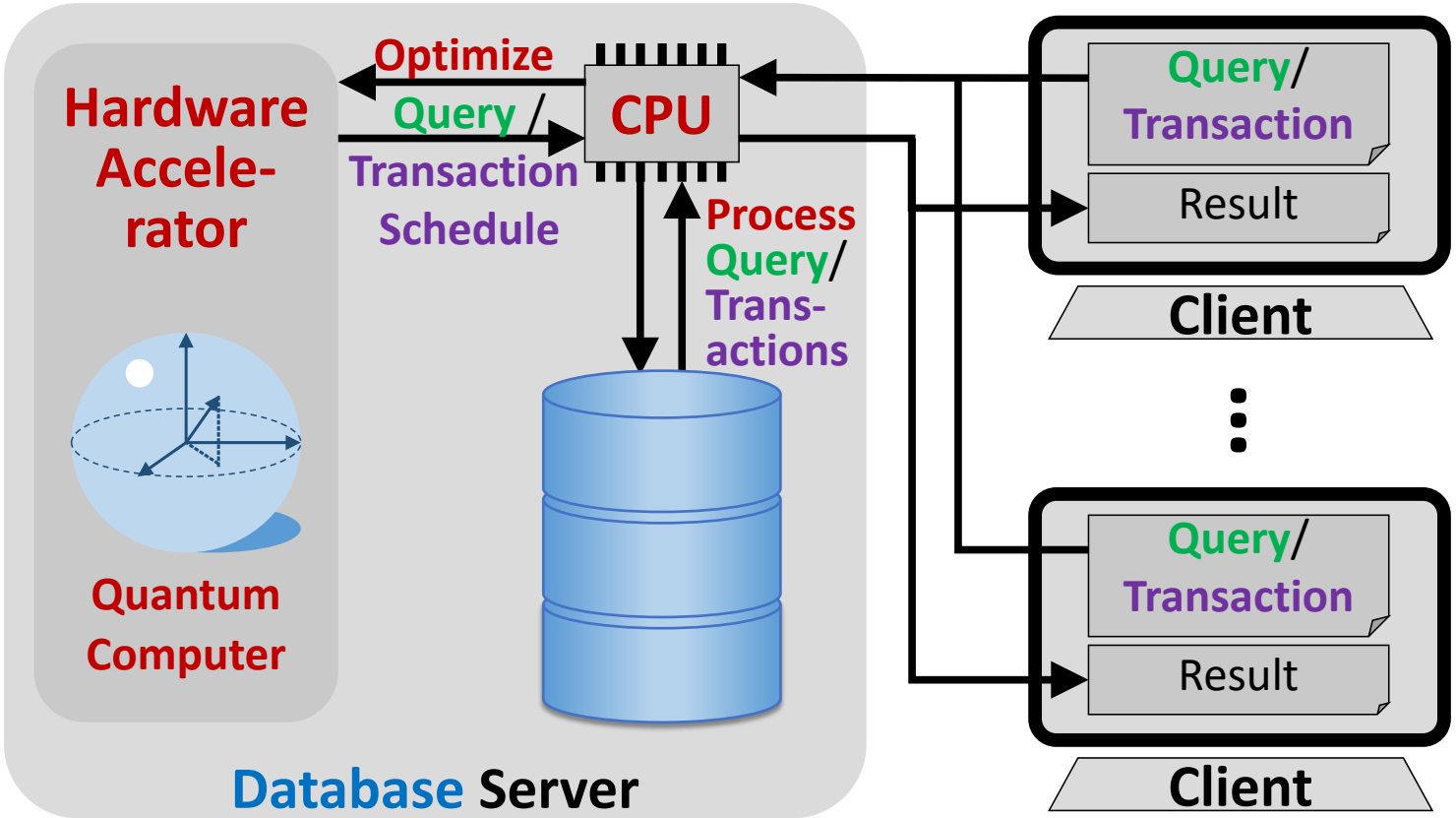
Timeline of Quantum Computers



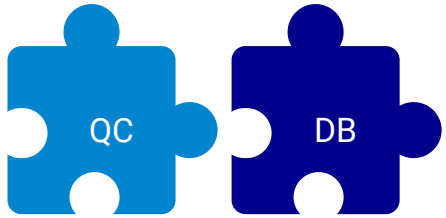
- Google
- IBM
- Intel
- Rigetti
- QuTech
- USTC
- Xanadu
- Quantum Brilliance
- Quantinuum
- Quantware
- RIKEN
- AQT
- Atom Computing
- M Squared Lasers
- |QuEra⟩
- D-Wave *(Quantum Annealing)*

Using Hardware Accelerator for optimizing Queries / Transaction Schedules





Approaches for Query/Transaction Schedule Optimization

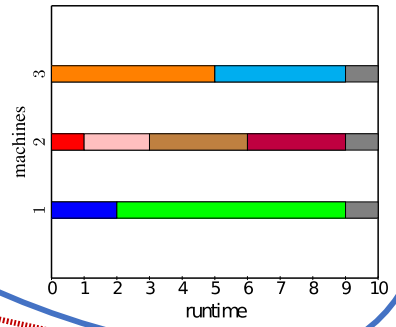


Query Optimization:

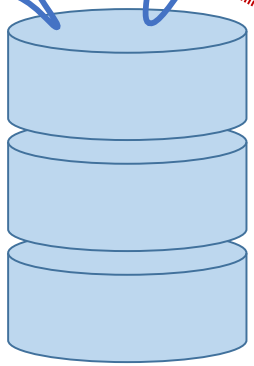
$$\bowtie_{i=1}^{n} R_i \;?\; \begin{array}{c} (R_1 \bowtie R_2) \cdots \bowtie R_n \\ \vdots \\ (R_1 \bowtie R_n) \bowtie (\cdots) \end{array}$$

Transaction Schedule Optimization:

$$\{T_1, \ldots, T_m\} \;?$$

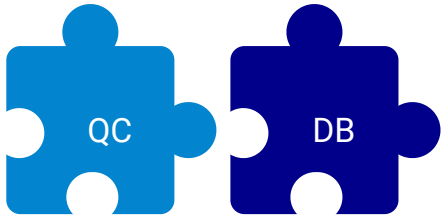


Open Source Relational Database Management System (RDBMS), e.g. PostgreSQL, MySQL

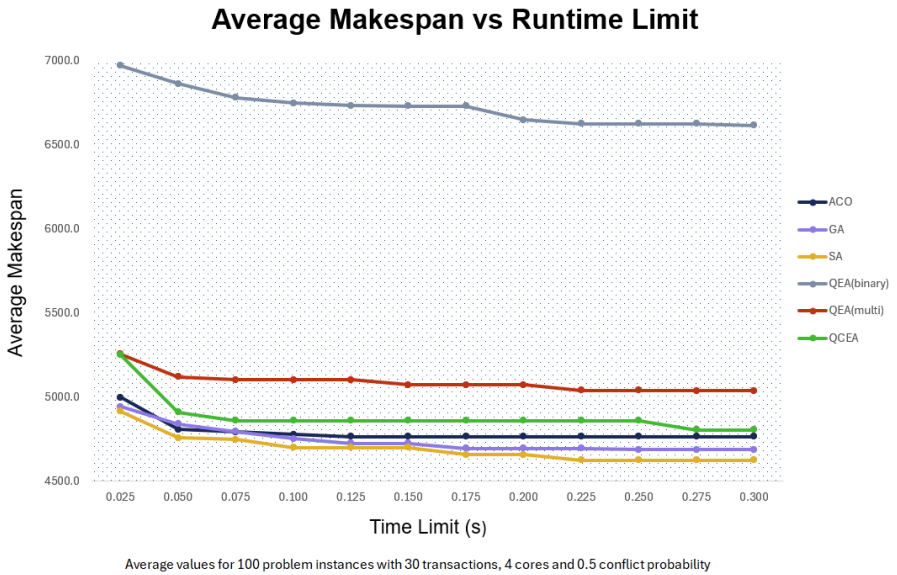


Dynamic Programming
Random Walk
Linear Program-ming
...
Machine Learning
Genetic Algorithm
Simulated Annealing

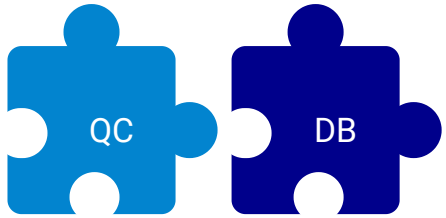
Transaction Scheduling



- 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



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: $O((n \cdot k \cdot R)^2)$

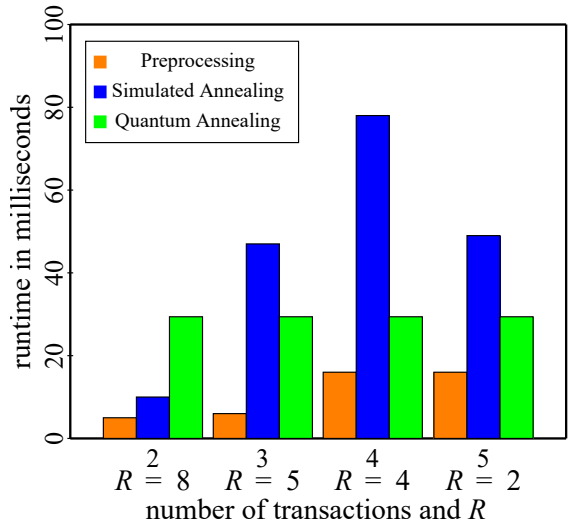
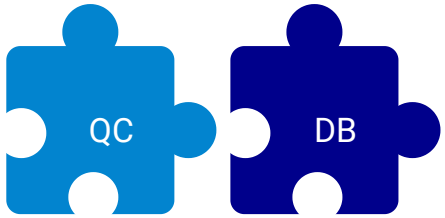


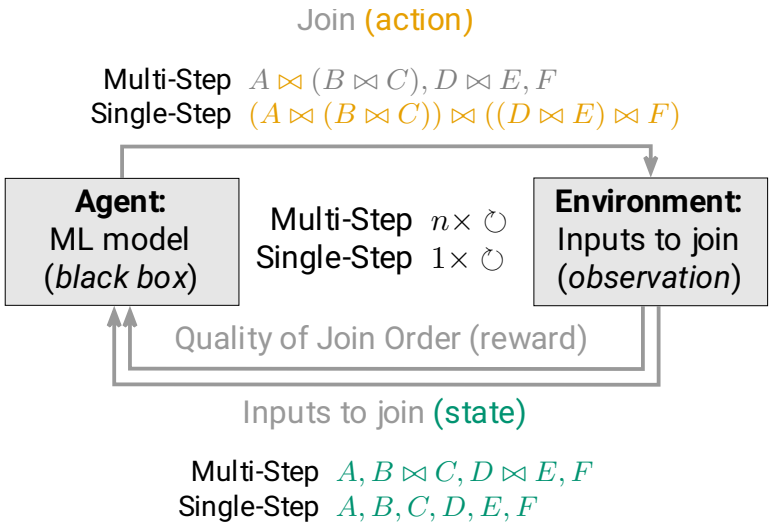
Fig.	k	n	R	O	$l_1, \ldots, l_n$	$r_1, \ldots, r_n$	req. var.
11	2	2	8	$\{\}$	8, 4	0, 4	8
		3	5	$\{(t_1, t_3)\}$	4, 5, 1	1, 0, 4	10
		4	4	$\{(t_2, t_4)\}$	3, 2, 1, 2	1, 2, 3, 2	16
		5	2	$\{(t_1, t_2), (t_4, t_5)\}$	1, 1, 1, 1, 1	1, 1, 1, 1, 1	10

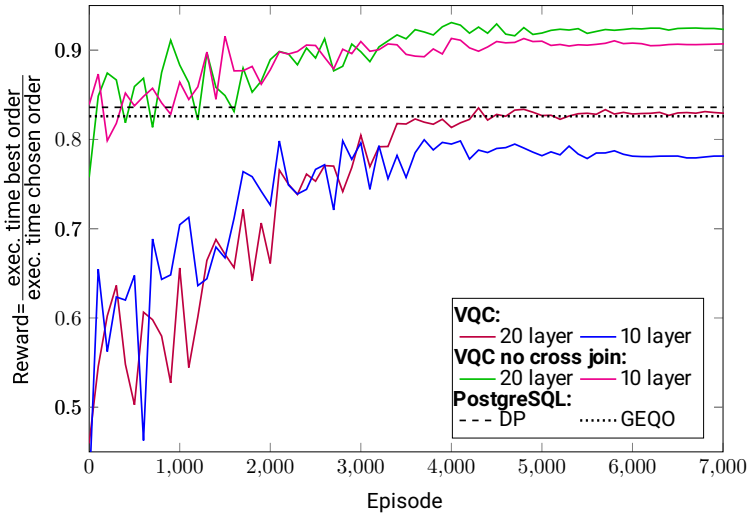
k: #cores, n: #transactions, R: max. makespan, O: conflict matrix, l_x: transaction length, $r_x = R - l_x$

QML for Join Ordering

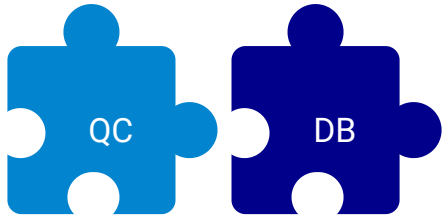


- Design choice: Predict best join order based on real execution times in contrast to estimated costs (EC)
  - can beat exact methods like dynamic programming (DP) based on EC
- Single-Step: Real-world queries (ErgastF1 Benchmark/PostgreSQL)
  - join orders with faster execution than DP (2.8%) and GEQO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

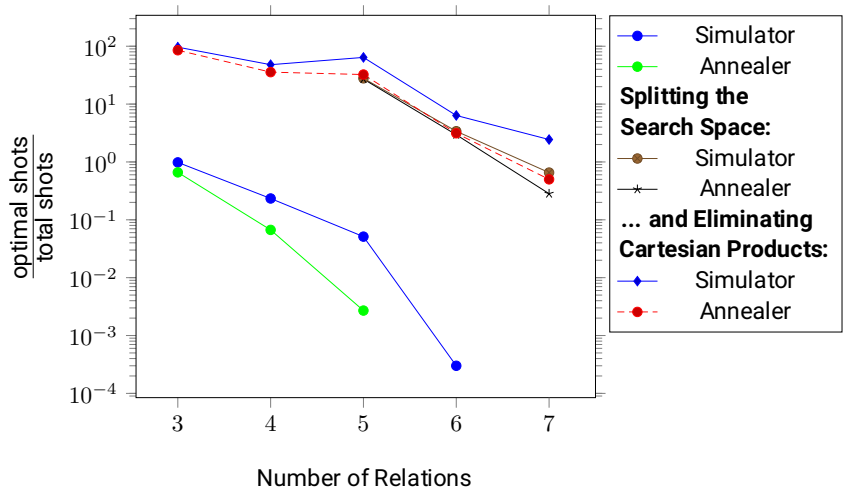


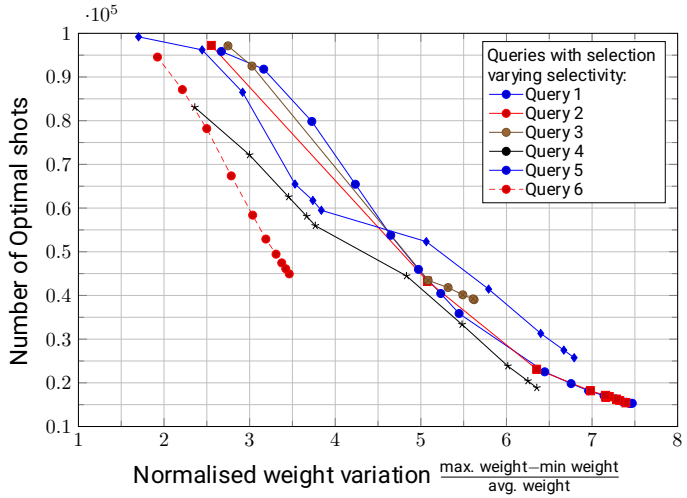


Join Ordering as QUBO Problem

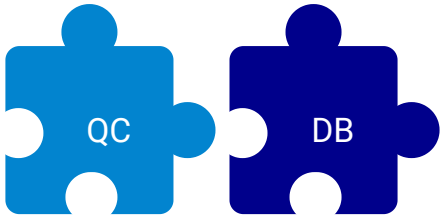


- Design choice: Direct Join Costs in contrast to calculated costs
  - Estimated costs for **each** join $\Rightarrow$ better join order
  - Runtime complexity $O(2^m - m)$ with m relations to join
    - Classical runtime: $O(3^m - 2^{m+1})$ using dynamic programming, best exact: $O(2^m \cdot m^3)$, $O(2^{3m/2}/\sqrt{\epsilon})$ with $(1+\epsilon)$-approx. algo.[SK24]
- Real-world queries (ErgastF1 Benchmark/PostgreSQL)

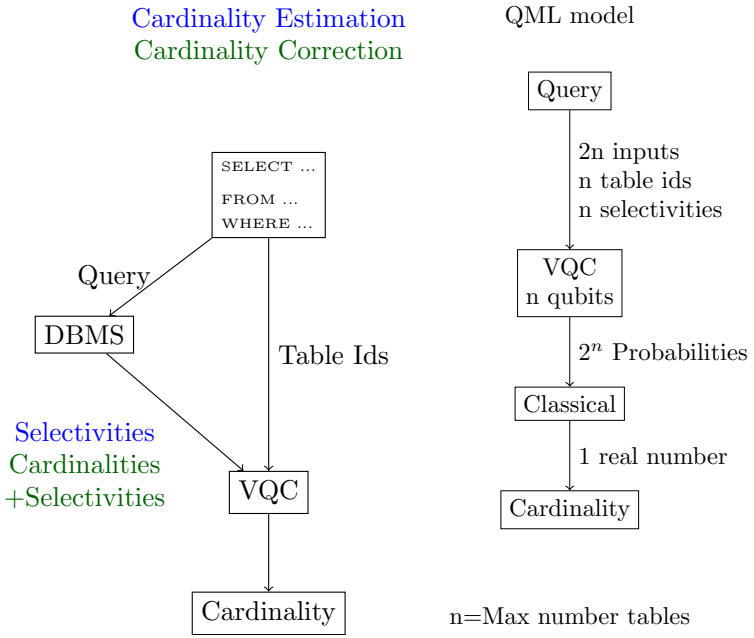


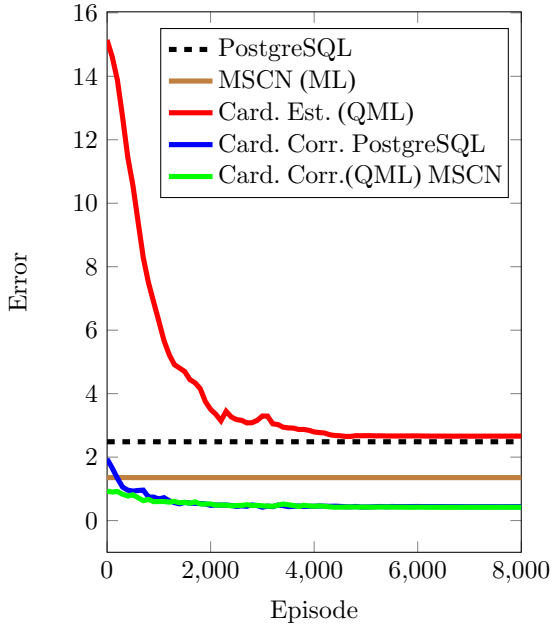


QML for Cardinality Estimation

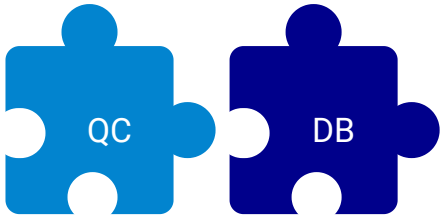


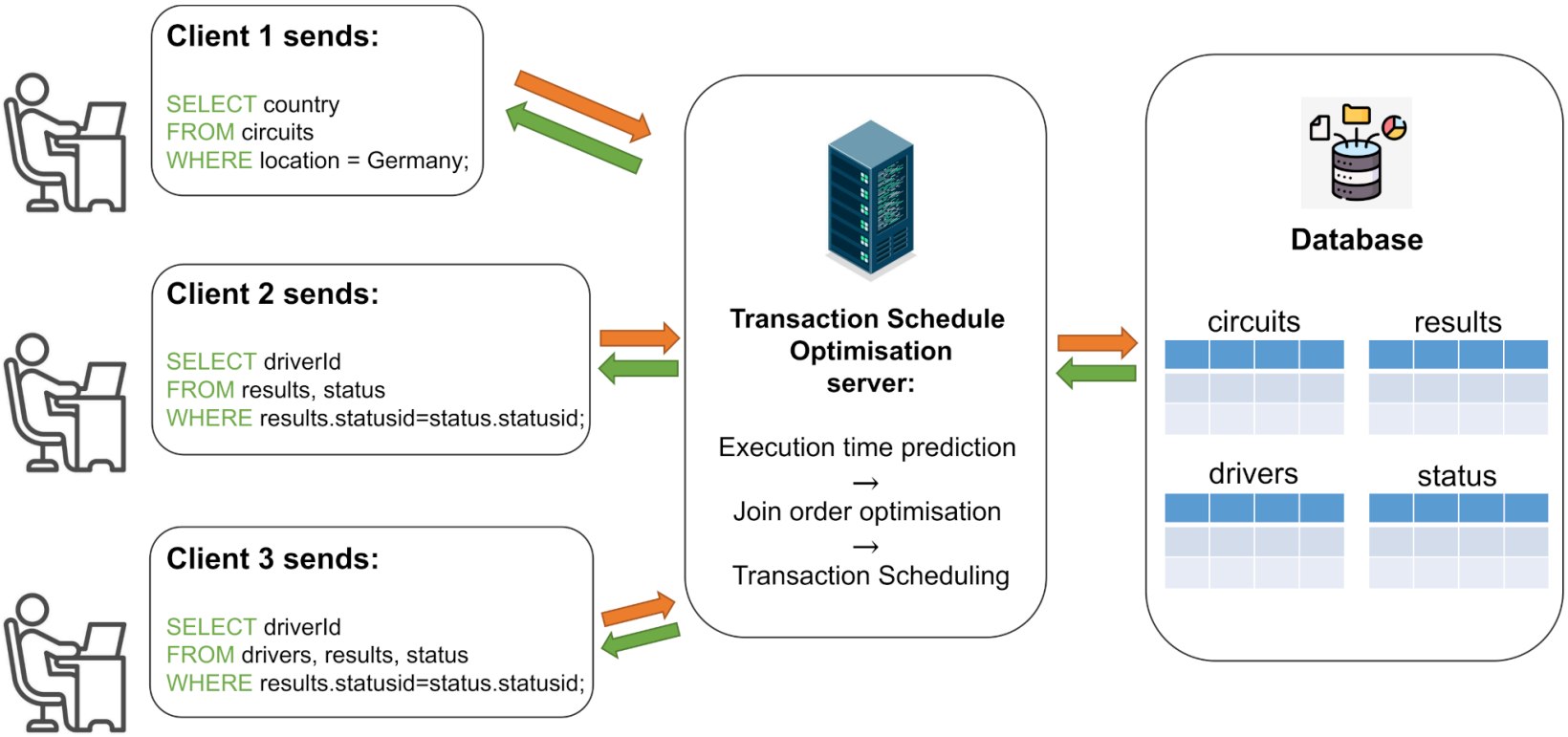
- Design choice: Predict correction factor of a Cardinality Estimator
  - Cardinality Predictor:PostgreSQL|MSCN|Hybrid Quantum Classical Network
  - Cardinality Correction: Hybrid Quantum Classical Network
- Queries of JOB-light Benchmark





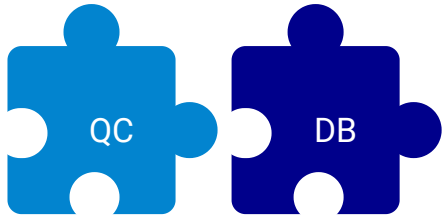
Transaction Scheduling Optimisation Server



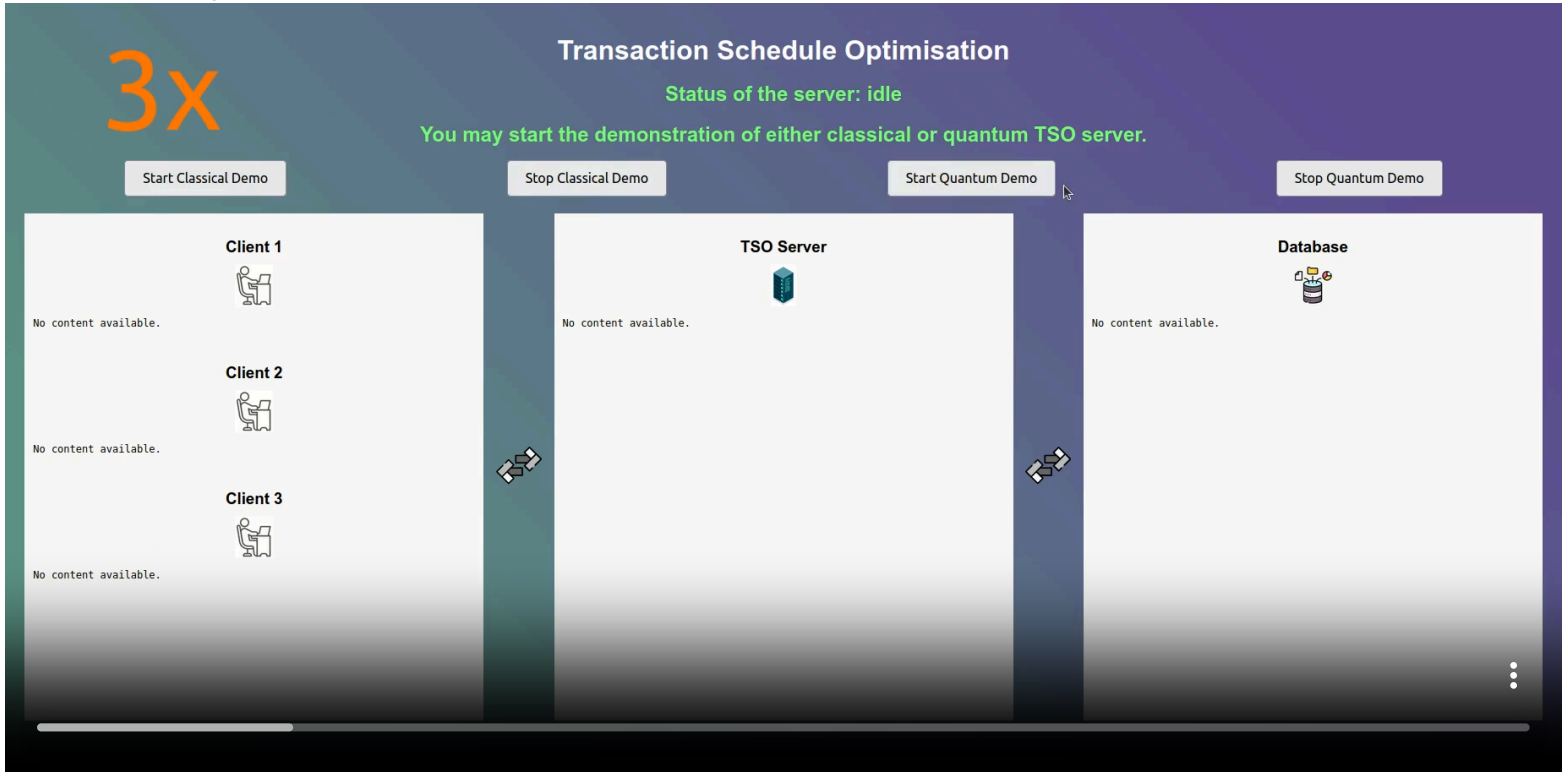


- Demonstrator: fully quantum **versus** fully classical

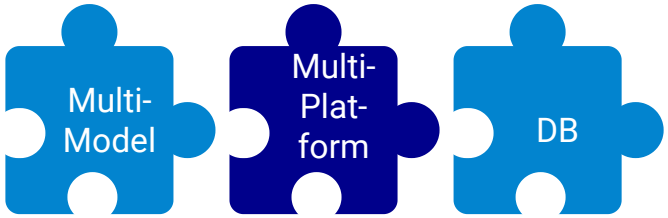
Demo for Query/Transaction Schedule Opt./Exec. Time Prediction

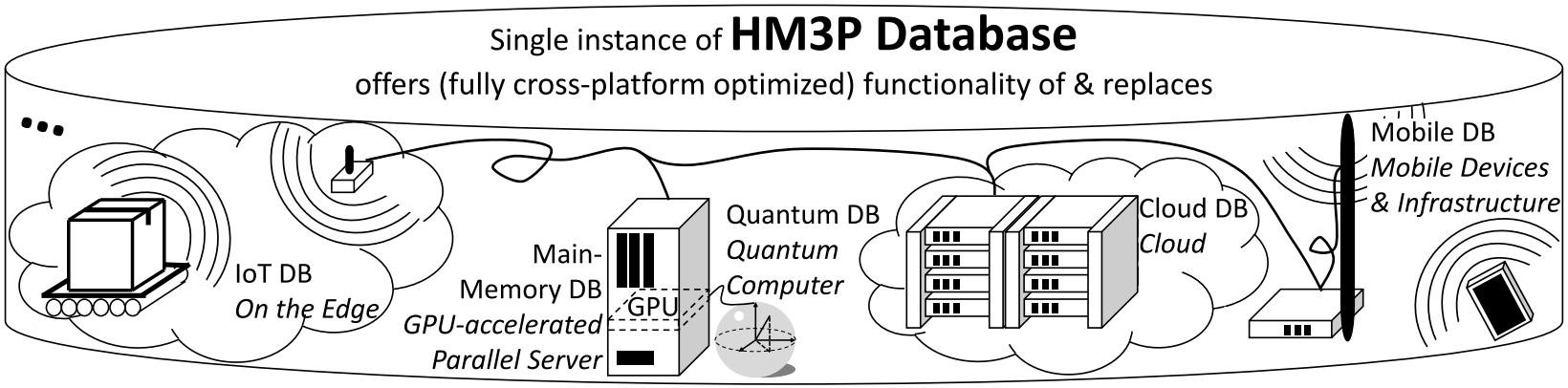


- is publicly available at the Quantum Brilliance web-page...



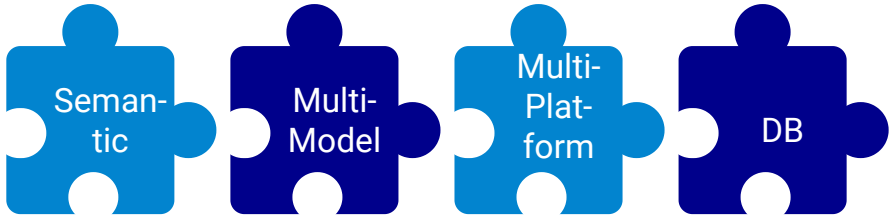
Hybrid Multi-Model Multi-Platform (HM3P) Database

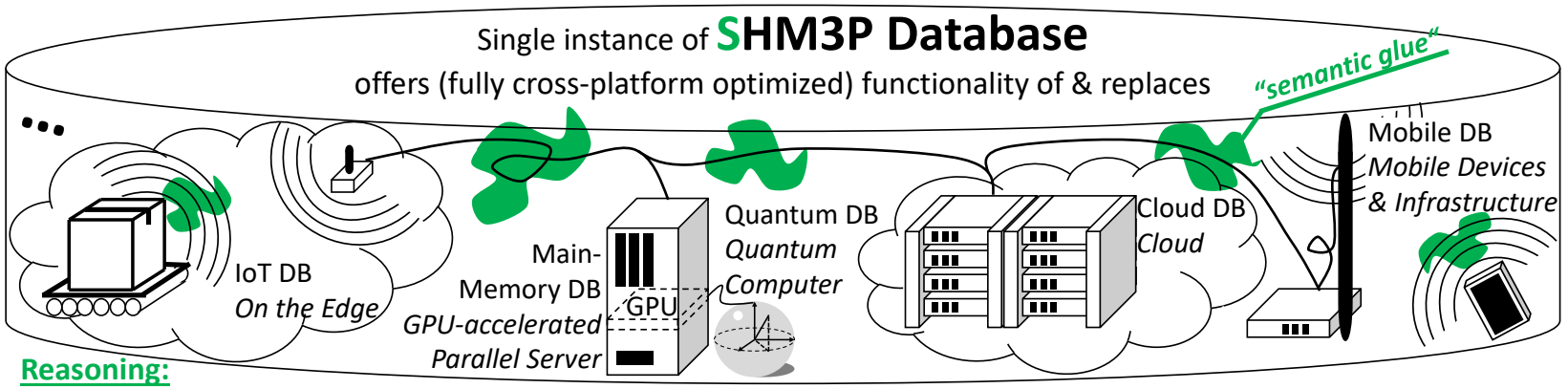




- full and uniform data integration at database level
- performance: fully optimized across different data models
- transparent fault-tolerance
- SQL standards: relational ('87), XML ('03), temporal ('11), JSON ('16), Multi-dimensional Arrays ('19), schemaless ('19), streams ('20?), property graphs ('21?)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

Variant: Semantic HM3P (SHM3P) DB





Reasoning:

Lightweight reasoning on large data sizes of IoT devices

Heavyweight reasoning on moderate data sizes

Heavyweight reasoning on large data sizes

Reasoning on small data sizes of mobile devices

How to integrate the different reasoning capabilities and requirements into one transparent global reasoner?

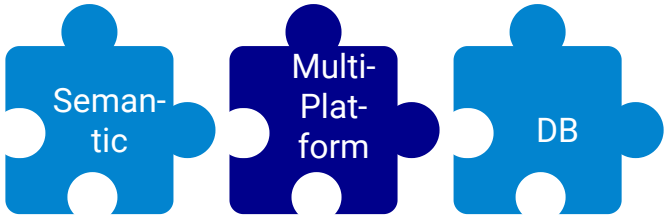
- Semantic Layer as glue between other models and platforms
- new challenges like integrating different types of reasoners in a transparent global reasoner

\+ Features of HM3P databases

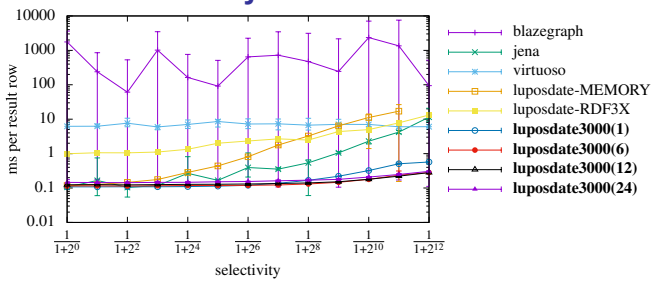
\+ Easier data integration

\- Performance issues may occur due to semantic layer

The Power of Multi-Platform: LUPOSDATE3000

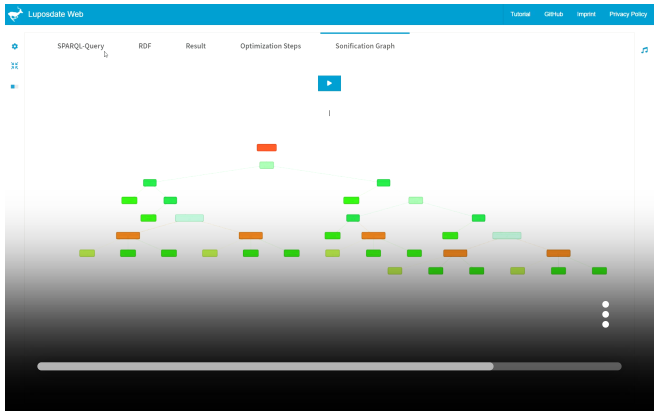


- LUPOSDATE300: Open Source Multi-Platform SW database
  - Design Choice: Implemented in Kotlin for multi-platform support
  - ultra-fast in jvm...



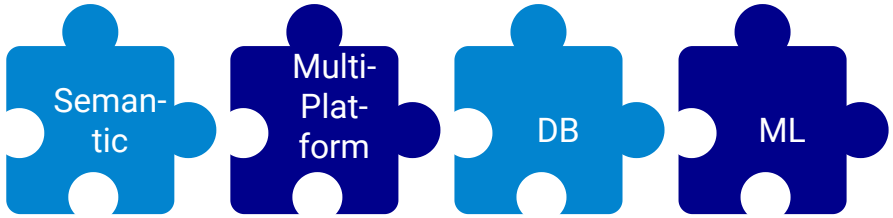
B. Warnke, M.W. Rehan, S. Fischer, S. Groppe: Flexible data partitioning schemes for parallel merge joins in semantic web queries in: BTW'21

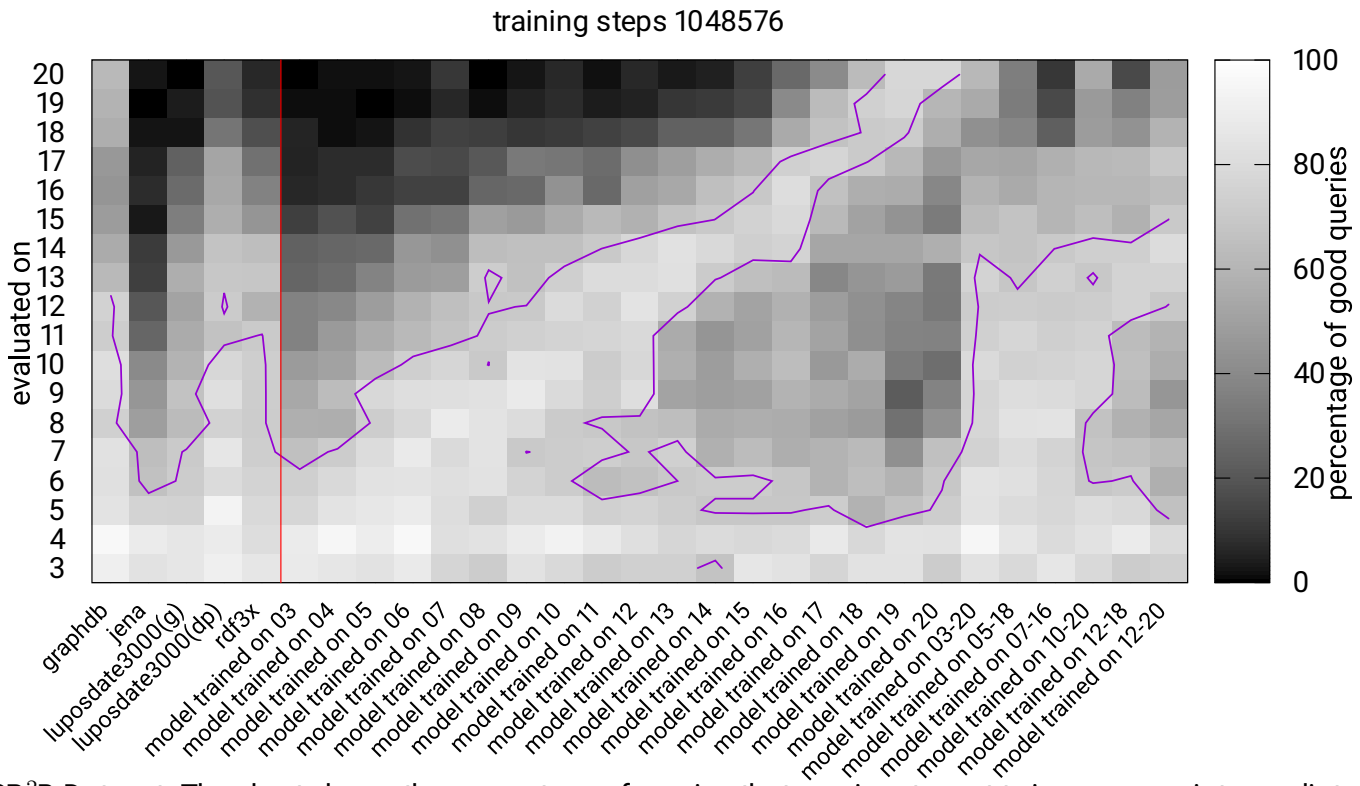
  - ...but also enabling web demos running completely in the browser!



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

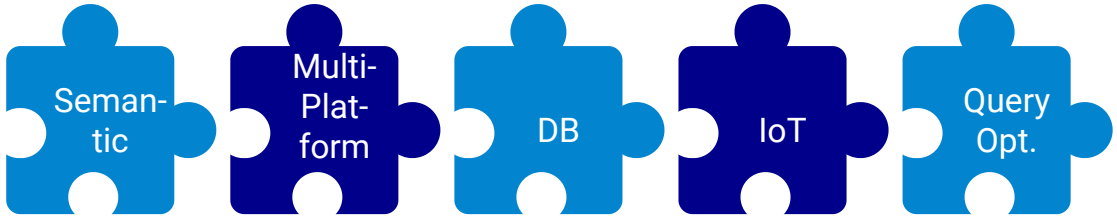
Query Opt. via Rein-forcement Learning



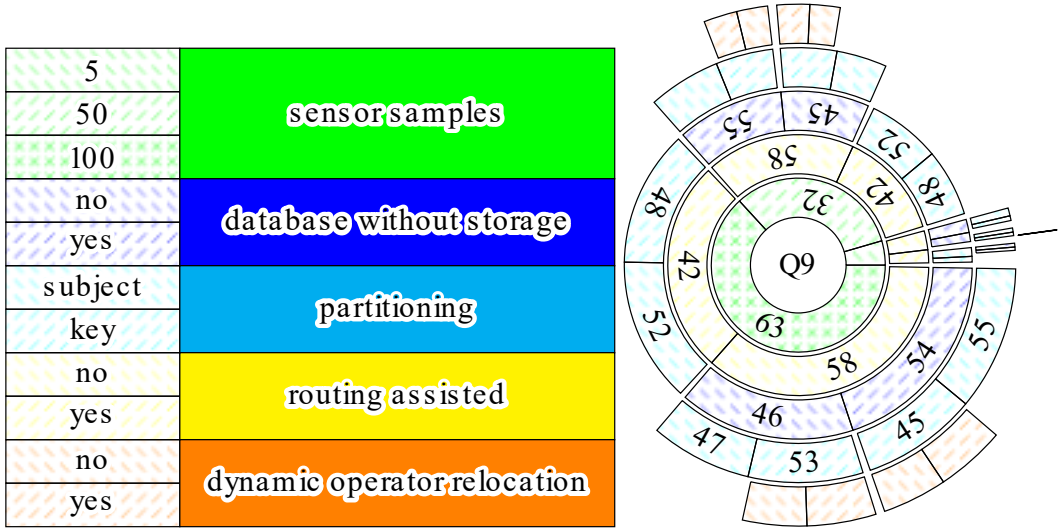


SP^{2}B-Dataset. The chart shows the percentage of queries that require at most twice as many intermediate results as the known best case. The magenta line surrounds the area, where 70% of the queries receive good join orders.

Query Opt. & Routing in IoT

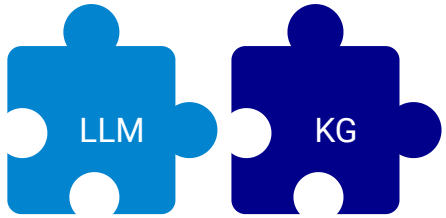


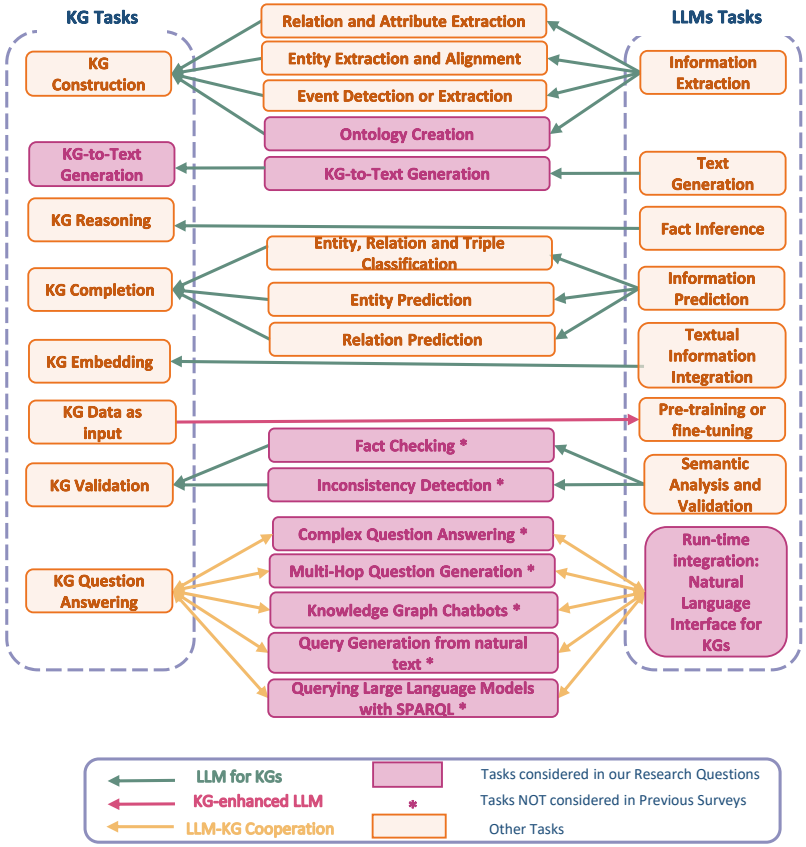
- Design Choice: Combining Routing and Query Optimization in IoT
  - Q9 with 11 triple patterns: Network traffic reduces by
    - routing-assisted join order (instead of centralized query optimization)
    - dynamic relocation of operators whenever the result of a join is more significant than its inputs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

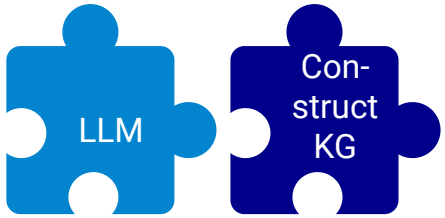
Interplay of LLMs and KGs

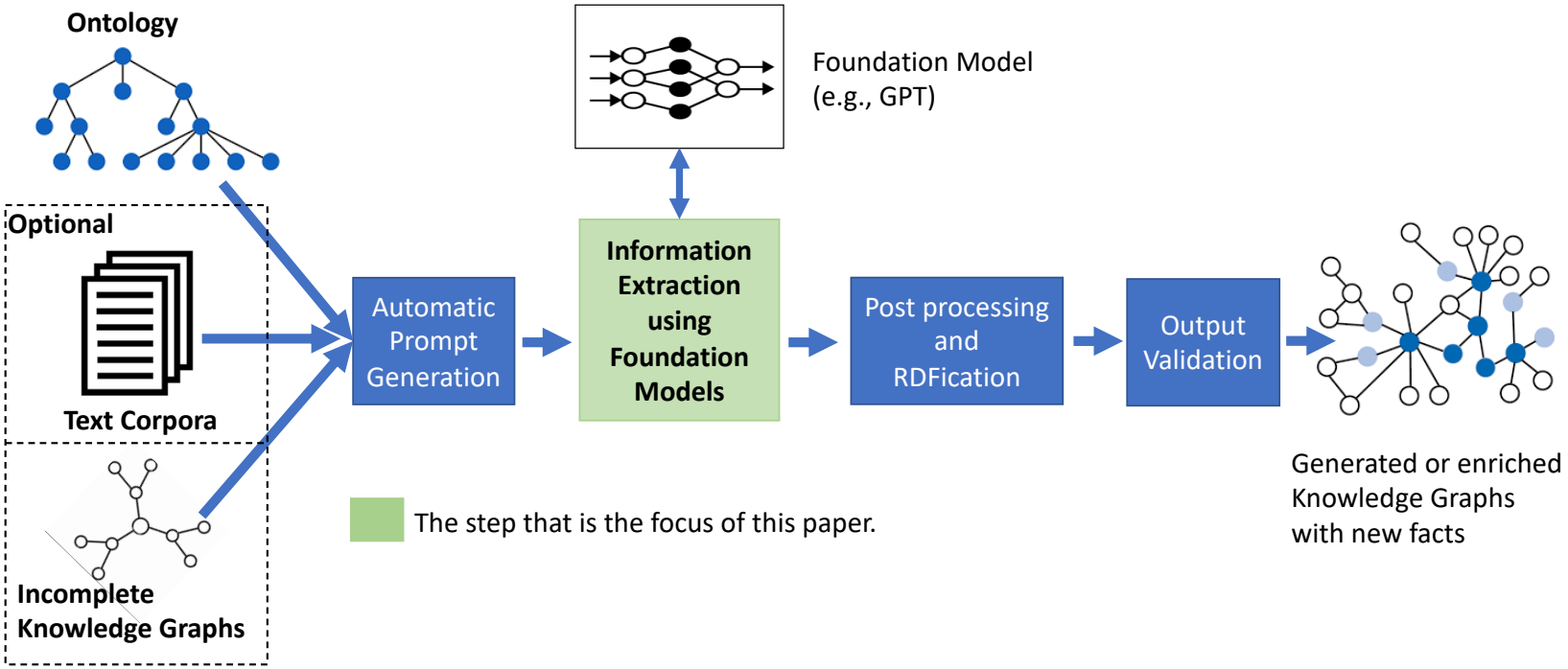




KG Construction with LLMs

(Project with Uni Paris/Toulouse)



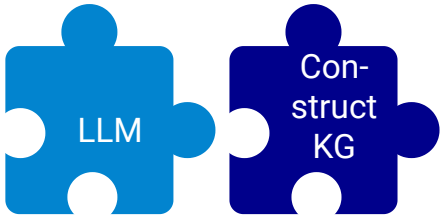


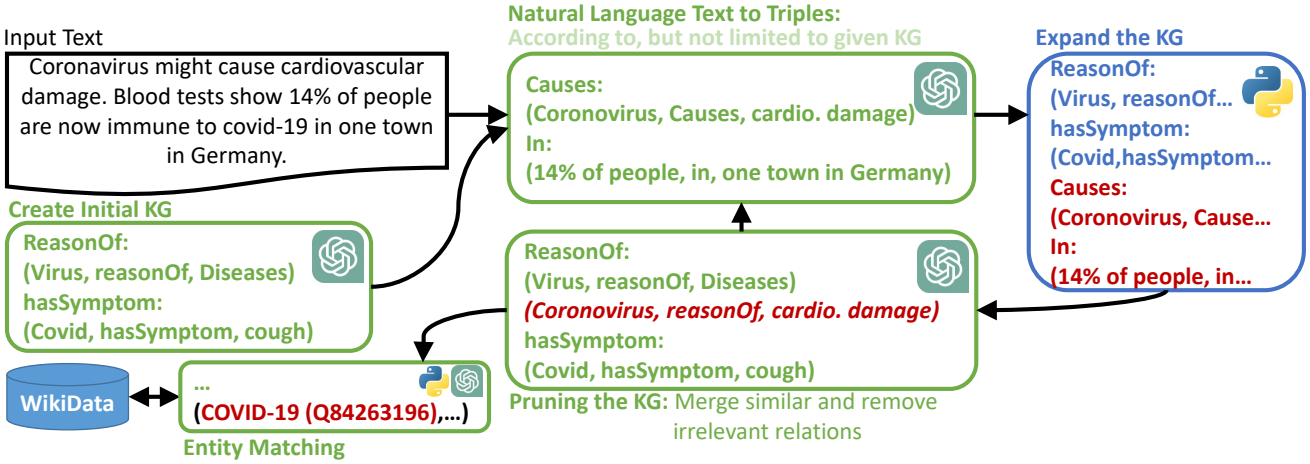
- Knowledge graph (KG) construction, completion, fact validation, extraction facts from unseen text* (limitations due to hallucination)

* 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)





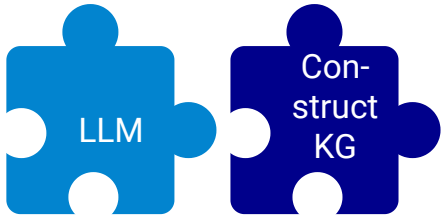
- Result (comparison to manual labeling)

	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)

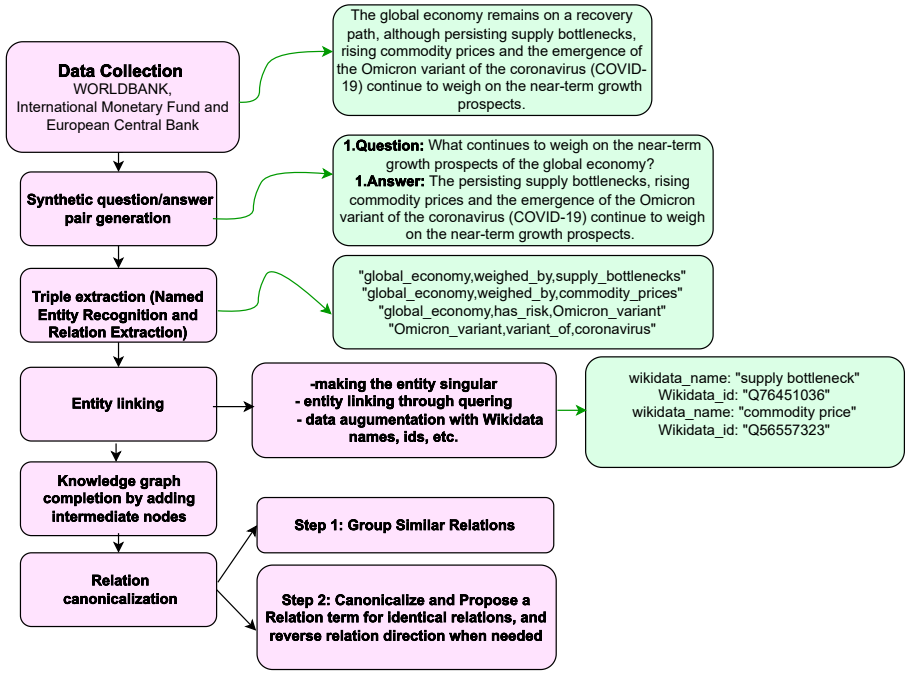
KG Construction with LLMs

(Project with Uni Paris/Toulouse)

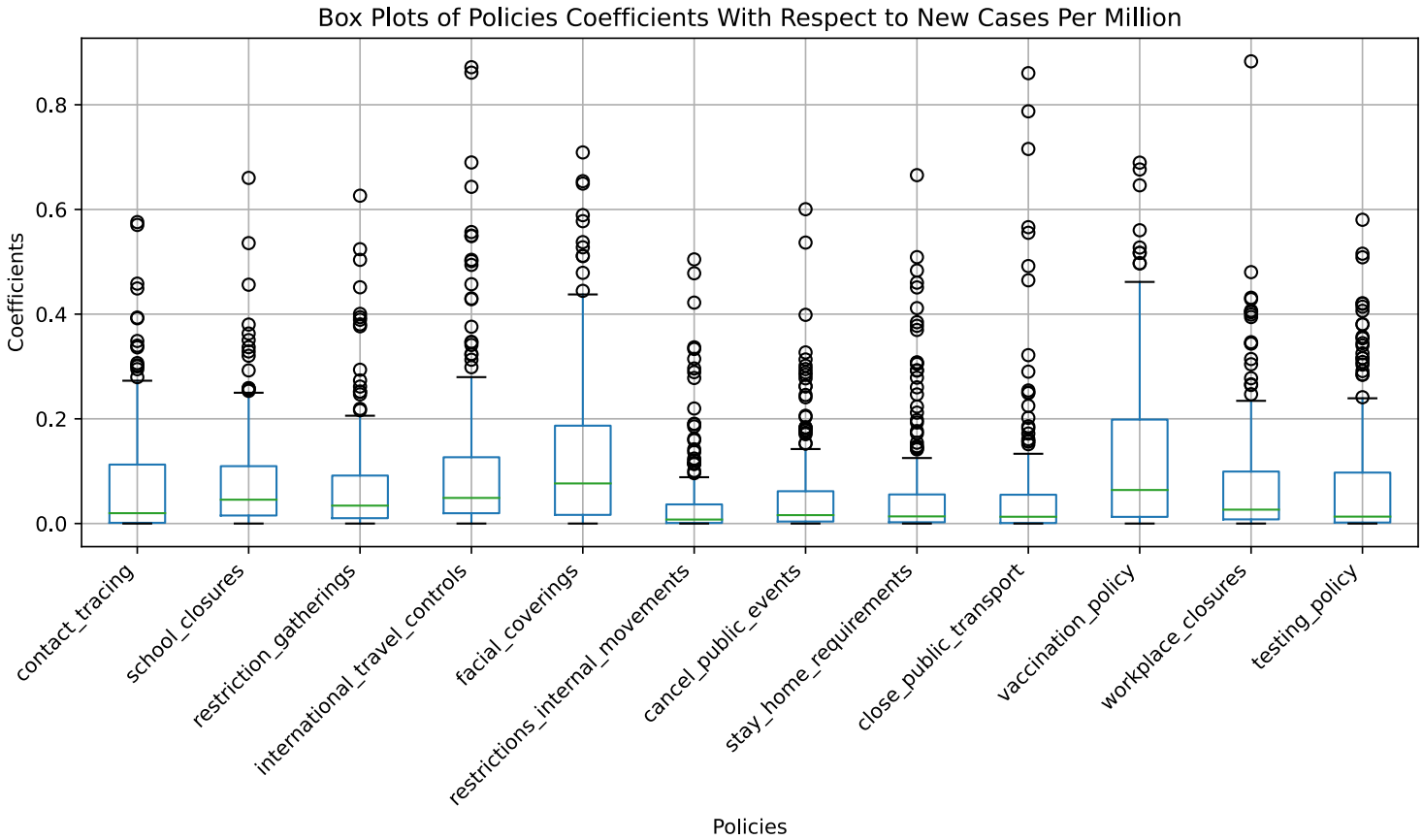


- Khorashadizadeh et al. Construction and Canonicalization of Economic Knowledge Graphs with LLMs, KGSWC 2024.
  - Approach to **Open Information Extraction**
    - extracting structured information directly from unstructured text **without a predefined schema**
  - **Presentation:** Friday, 13.12.2024, 13:15 – 13:40

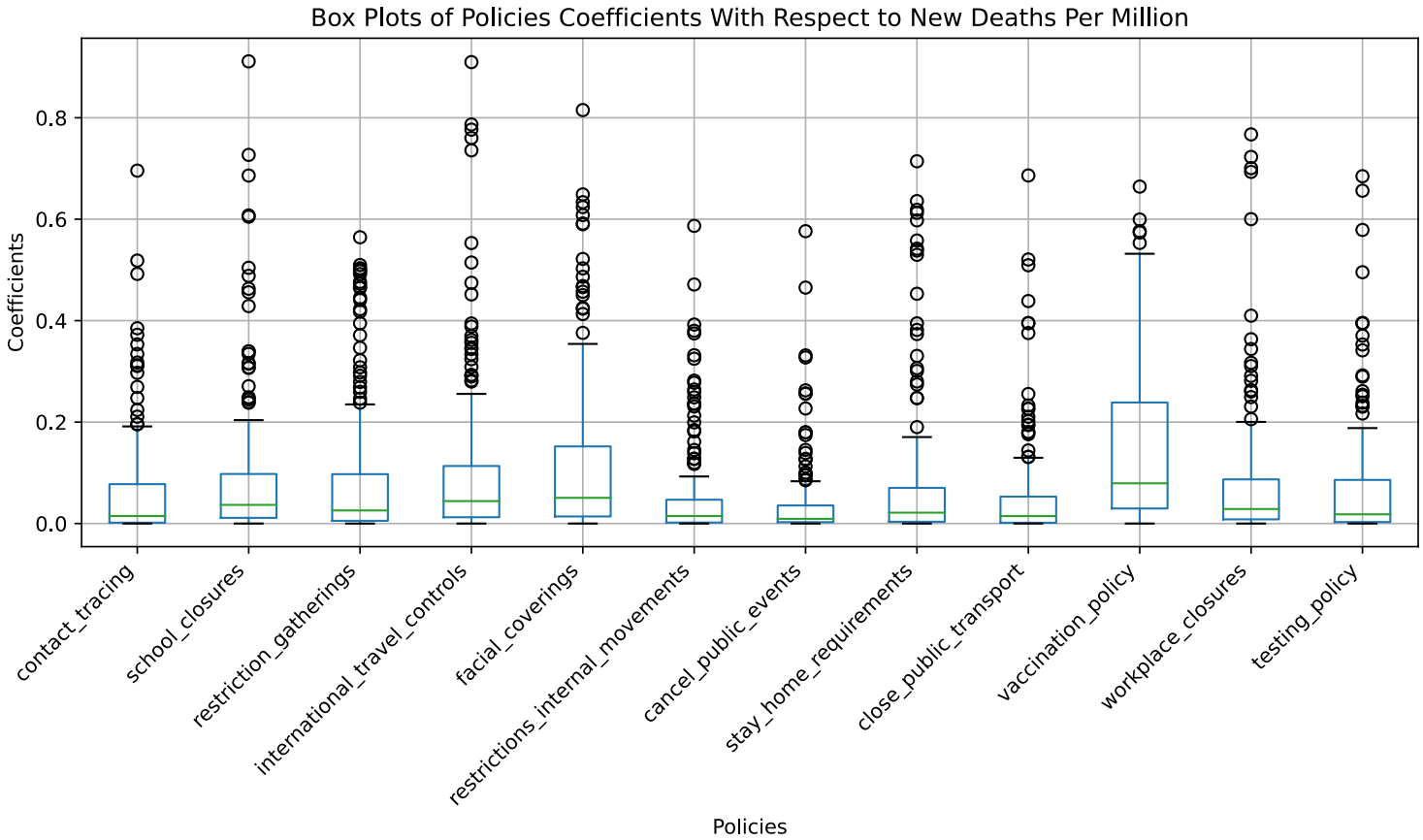
Pipeline:



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

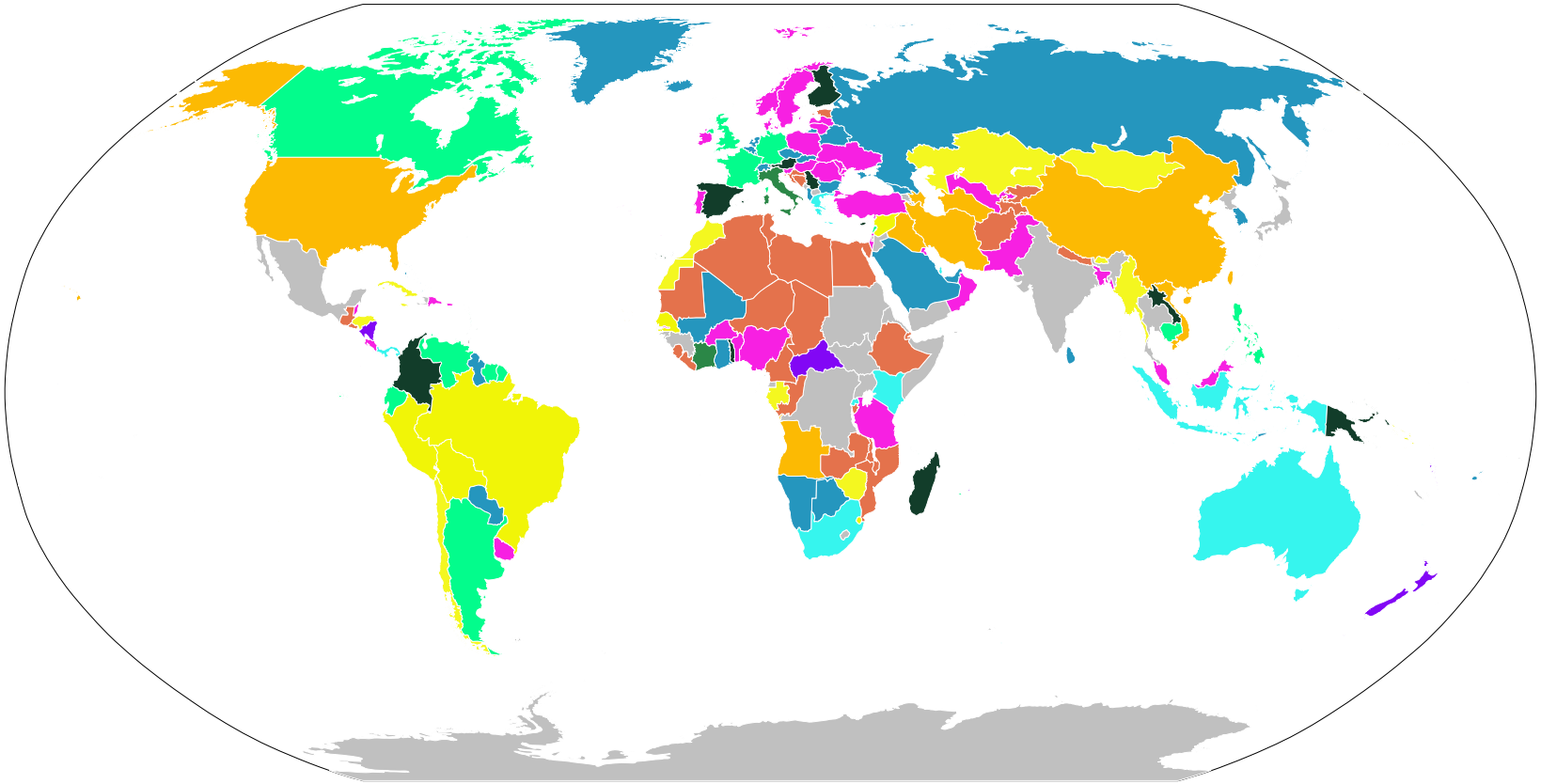


KG for Analysis of COVID-19 Policies (Policy Importance in 183 Nations)

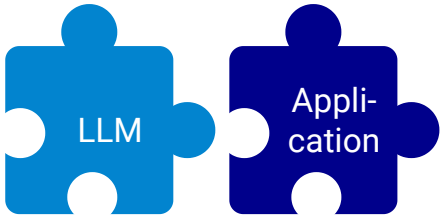


Nations clustered by COVID-19 Policies Importance

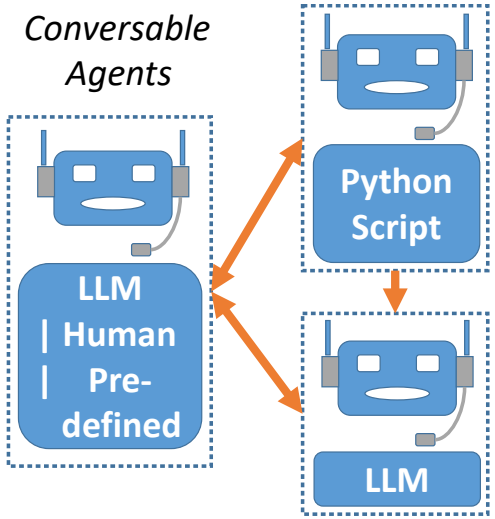
(Policy Importance in 183 Nations)



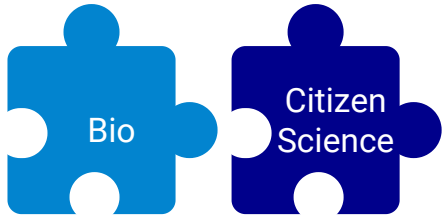
LLM Applications



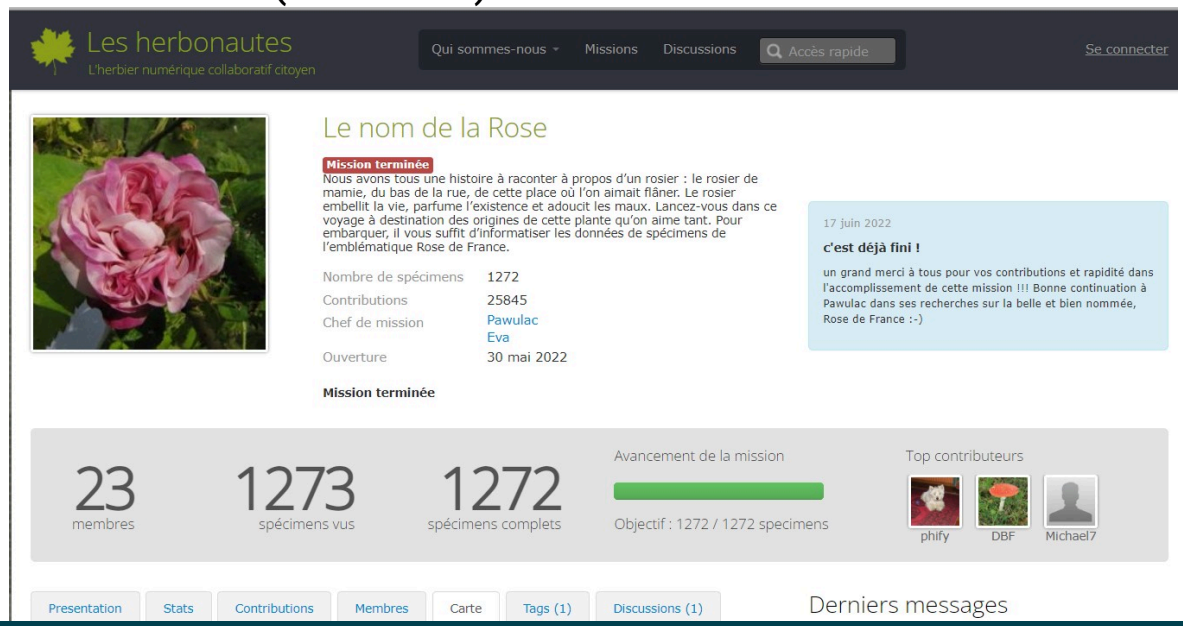
- New kind of application in form of a network of (LLM, human, code) agents
  - Applying LLMs as new paradigm for software development of applications close to natural language/human users
  - Development of new LLM applications like
    - modern chatbots
    - extraction from natural languages text (e.g., for knowledge graph construction)
    - new types of applications with reasoning capabilities (self-reflection etc. for human-like generation of content)



Citizen Science Platforms for Biodiversity

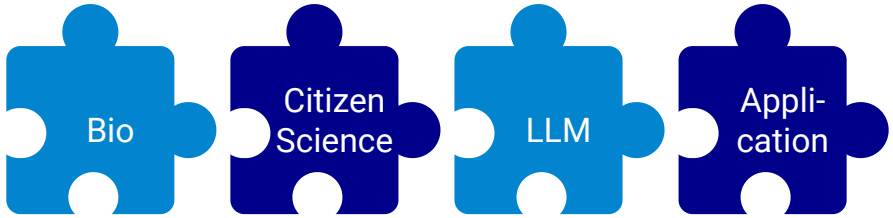


- launched by Muséum national d'Histoire naturelle (MNHN)/Paris
- collaborative way of contributing to the creation of a scientific database based on the millions of photos of plants in the national network of naturalist collections (Récolnat)

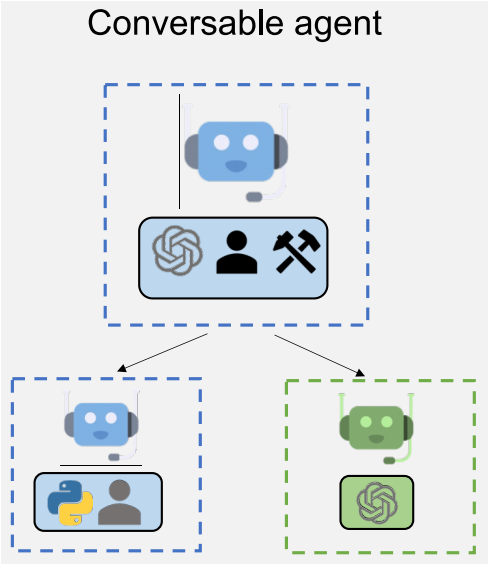


LLMs4Citizen Science

(Les Herbonautes, project with MNHN/Uni Paris)

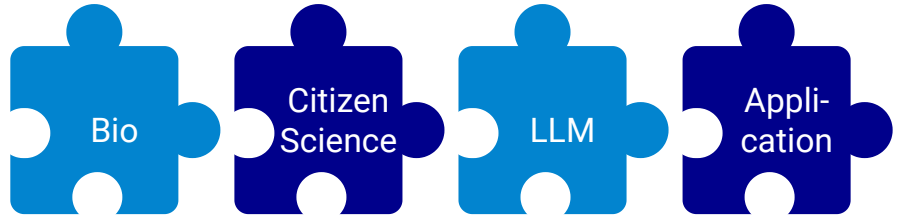


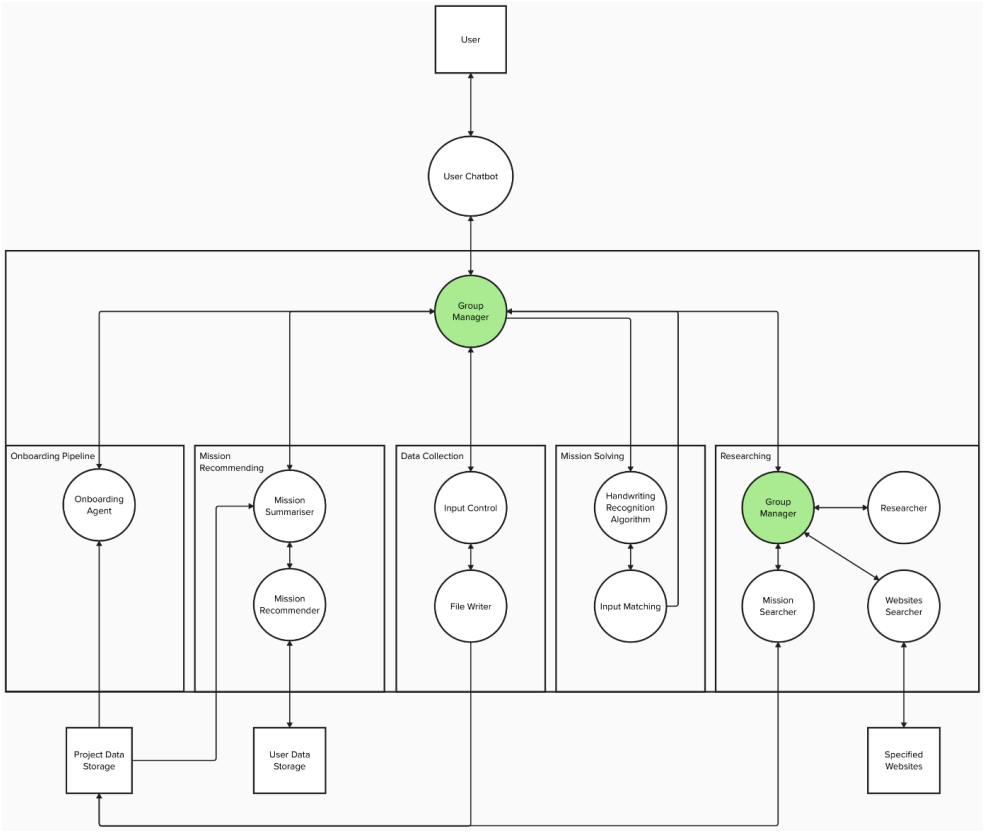
- LLM in form of a network of (LLM, human, code) agents as basis for an easy-to-use chatbot
  - for guiding users through the data validation process and with which users can retrieve important information by posing related questions for adding missing information (as some kind of easy-to-use tool for 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



LLMs4Citizen Science

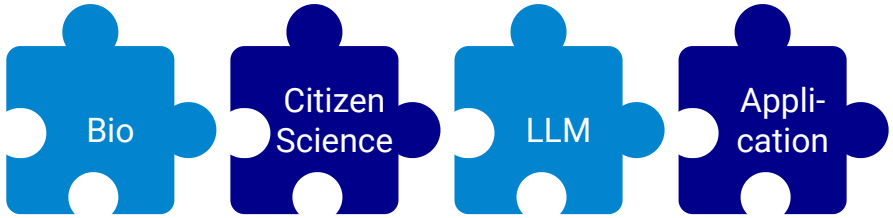
(Les Herbonautes, project with MNHN/Uni Paris)

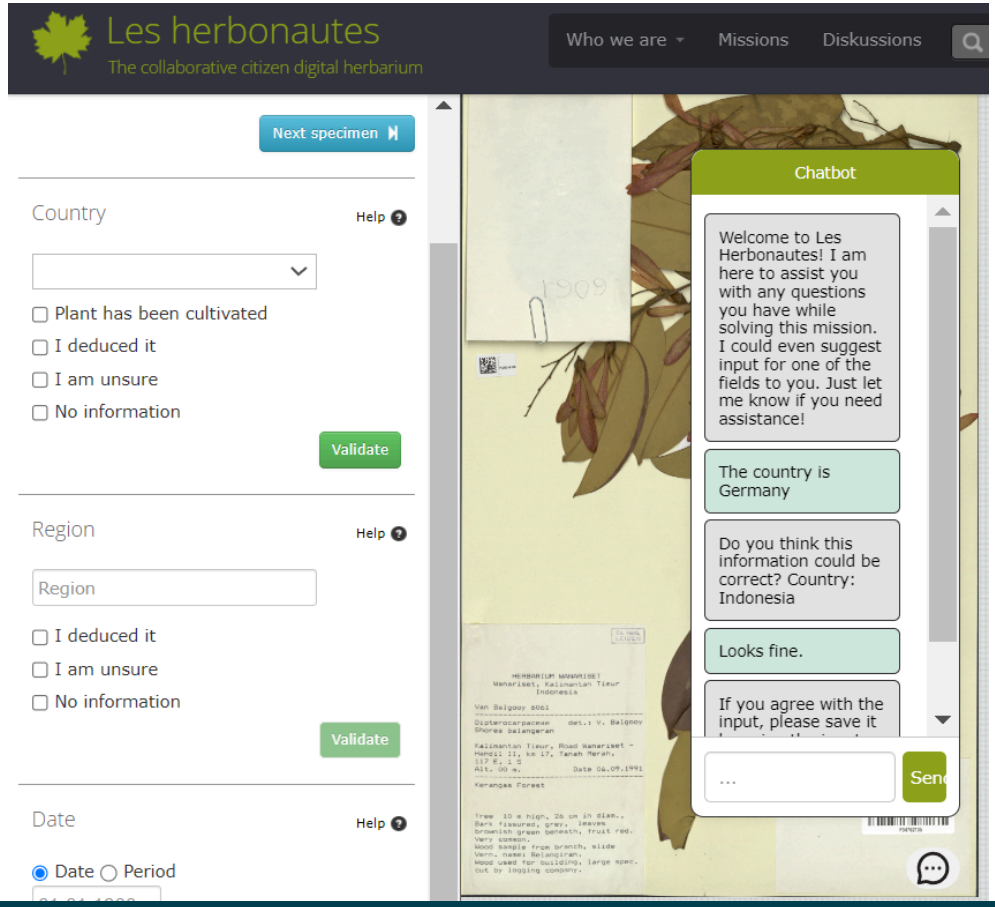




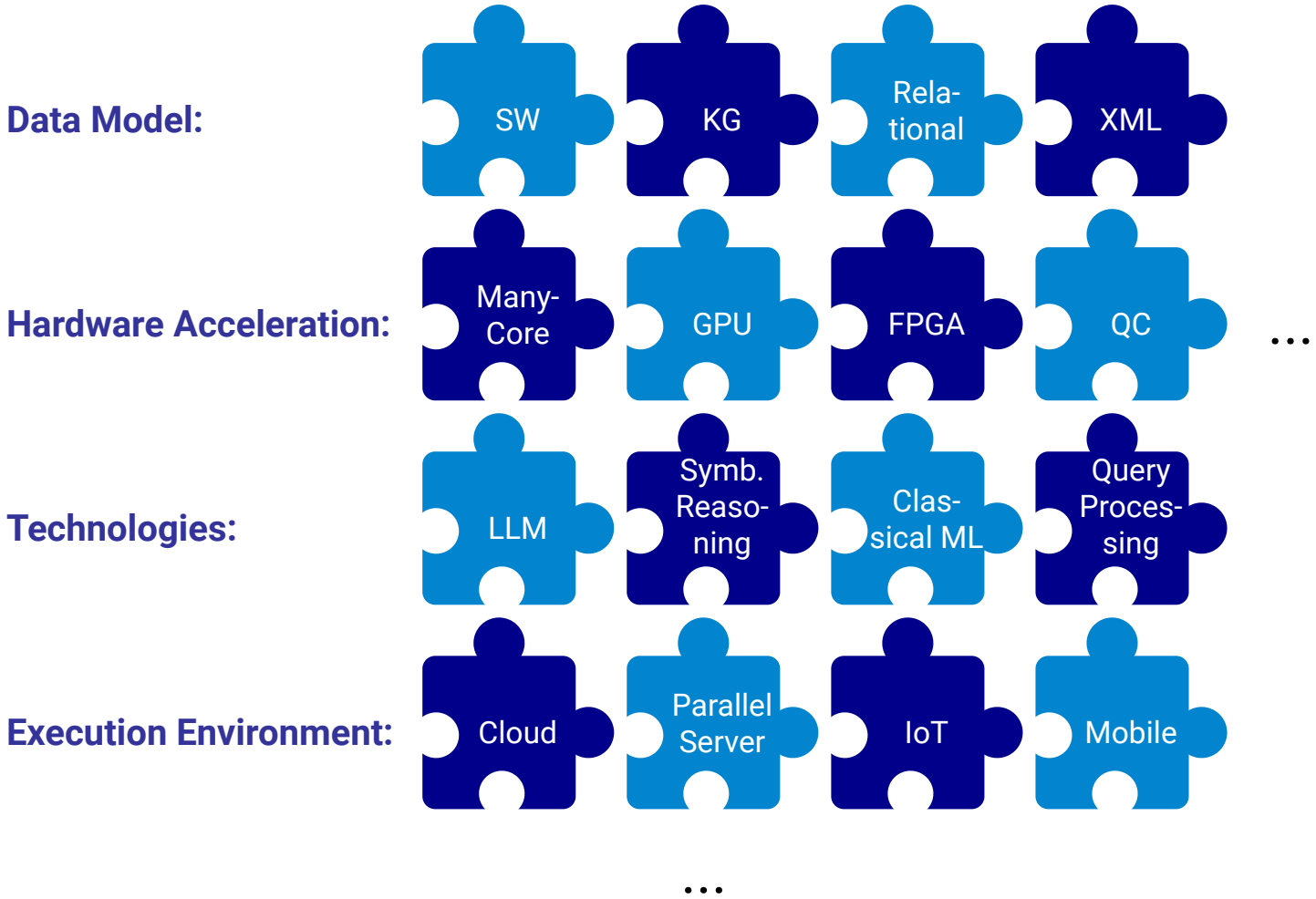
LLMs4Citizen Science

(Les Herbonautes, project with MNHN/Uni Paris)



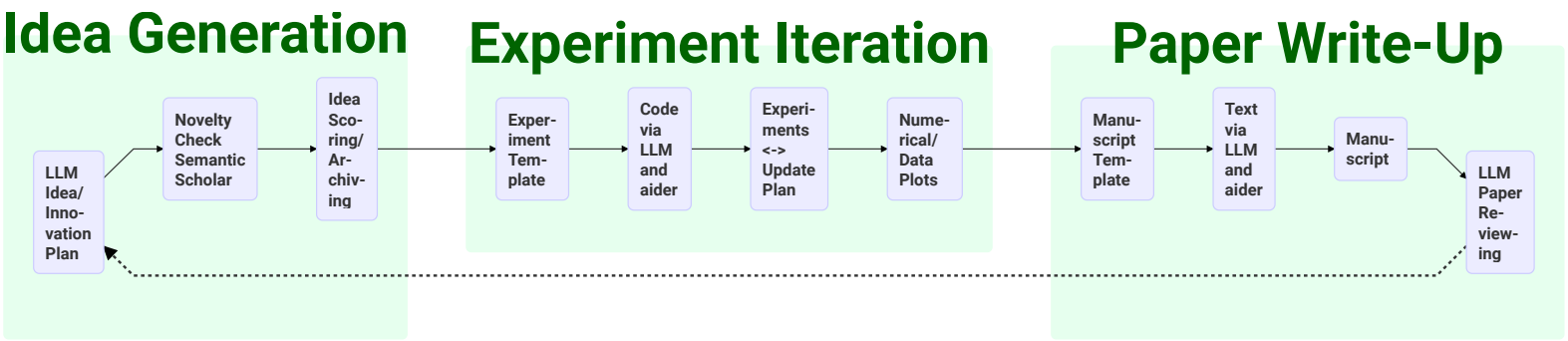


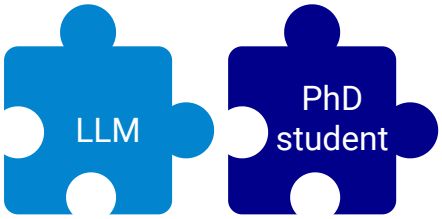
What's next to combine?



Are we going to be replaced? Can the job been done by AI?

- The AI Scientist:





Quantum and Intelligent Computing (**QIC**)

- Head: Prof. Dr. rer. nat. habil. Sven Groppe
- Projects and Research Assistants
  - **QC4DB**: Accelerating Relational Database Management Systems via Quantum Computing (BMBF)
    - Umut Çalikyilmaz
    - Tobias Winker
    - Nitin Nayak
  - **QualityOnt**: High Quality Knowledge Graphs from recent English, French and German Emergent Trends with the example of COVID-19 (DFG/ANR)
    - Hanieh Khorashadizadeh
  - **Semantic Data** Integration and Analysis (Bosch)
    - Simon Paasche (External PhD Student)
  - **Hybrid**^{2}-Index Structures for Main Memory Databasess (DFG)
    - Tobias Groth
  - **BigSIoT**: Big Data Management für das Semantic IoT (DFG)
    - Dr. Benjamin Warnke
  - **SW-FPGA-DB**: Hardwarebeschleunigung von SW DBs durch dynamisch rekonfig. FPGAs (DFG)
    - Dr. Dennis Heinrich
  - **FPGA-Cluster DB**: Beschleunigung rel. DBs mittels laufzeitadaptiver FPGA Cluster (ZIM-BMWi)
    - Dr. Stefan Werner