



Laboratoire d'Informatique et d'Automatique pour les Systèmes

Green Analytics

Ladjel BELLATRECHE

LIAS/ISAE-ENSMA, Poitiers, France

bellatreche@ensma.fr

<http://www.lias-lab.fr/members/bellatreche>



IoTBDs'2026

Benidorm, Spain, 21-24 May 2026



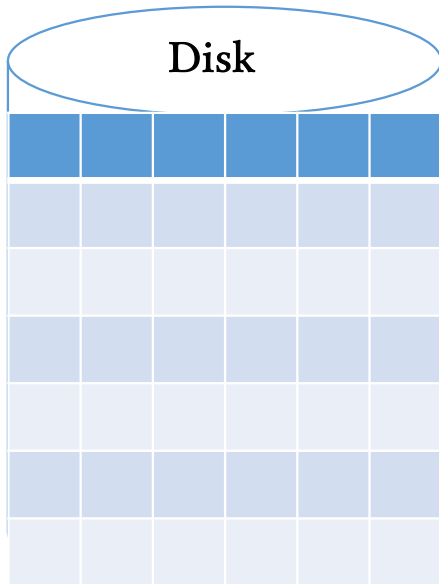
ISAE-ENSMA

- ▶ National Engineering School for **Mechanics** and **Aerotechnics**, Futuroscope, Poitiers
- ▶ Member of ISAE-Group: SUPAERO, ENSMA, SUPMECA, ESTACA, ENAC, Ecole de l'Air et de l'Espace
- ▶ Ranked by Shanghai Ranking (between 151-200)



Preamble

- ▶ Disk-oriented relational databases
- ▶ Research question: reduction of response time of my SQL query



Inputs/Outputs

PostgreSQL 12.4 DBMS

Query 17e from Join Order Benchmark (JOB) workload:

```

SELECT MIN(n.name) AS member_in_charnamed_movie
FROM cast_info AS ci
JOIN title AS t ON ci.movie_id = t.id
JOIN movie_keyword AS mk ON t.id = mk.movie_id
JOIN keyword AS k ON mk.keyword_id = k.id
AND k.keyword = 'character-name-in-title'
JOIN movie_companies AS mc ON ci.movie_id = mc.movie_id
AND t.id = mc.movie_id
JOIN company_name AS cn ON mc.company_id = cn.id
AND cn.country_code = '[us]'
JOIN name AS n ON n.id = ci.person_id;
    
```

6 joins
2 selections

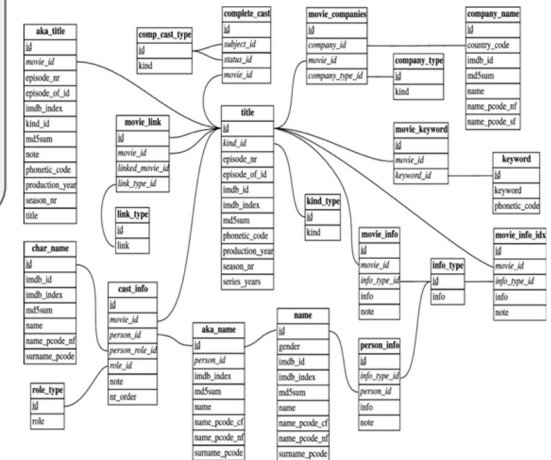
Size of JOB tables (#rows)

- cast_info (36244344)
- title (2528312)
- movie_companies (2609129)
- movie_keyword (4523930)
- name (4167491)
- company_name (234997)
- keyword (134170)

Hardware configuration

Component	Details
OS	Ubuntu 24.04.3 LTS
CPU	AMD Ryzen 5 5600X — 6 cores/12 threads
GPU	NVIDIA RTX 3060
RAM	16 GB

Query processor (QP)



QP tasks

- ▶ Selection of the best query plan:
 - **Join order:** $((ci \bowtie (((mk \bowtie k) \bowtie mc) \bowtie cn) \bowtie t)) \bowtie n)$
 - **Join implementations:** Hash join algorithm
 - **Table scan:** sequential scan

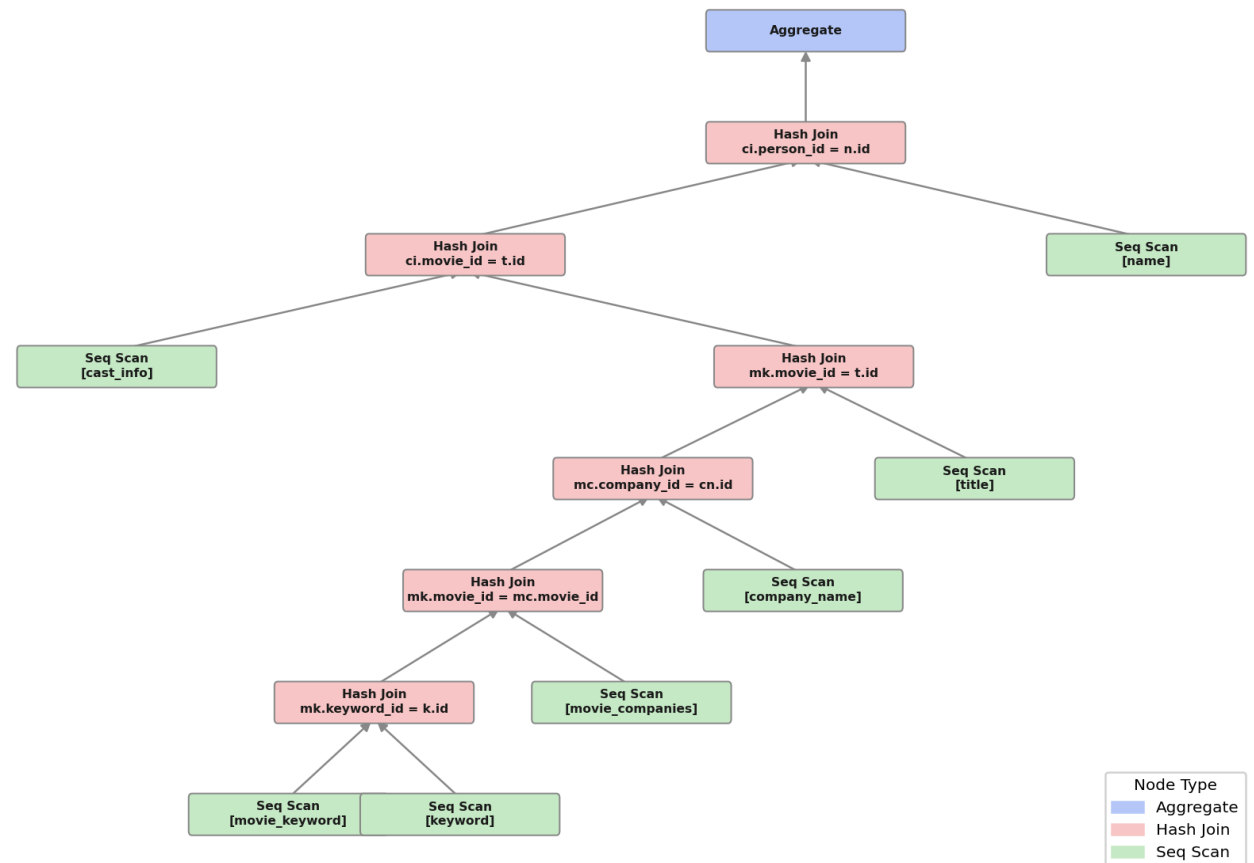
→ Response time: 11,13 s

▶ Thanks to analytical cost models

▶ Example:

- Cost $(mk \bowtie k; \text{Hash join}) = 3 \times (|mk| + |k|)$
- $|mk|$: number of pages of the table mk

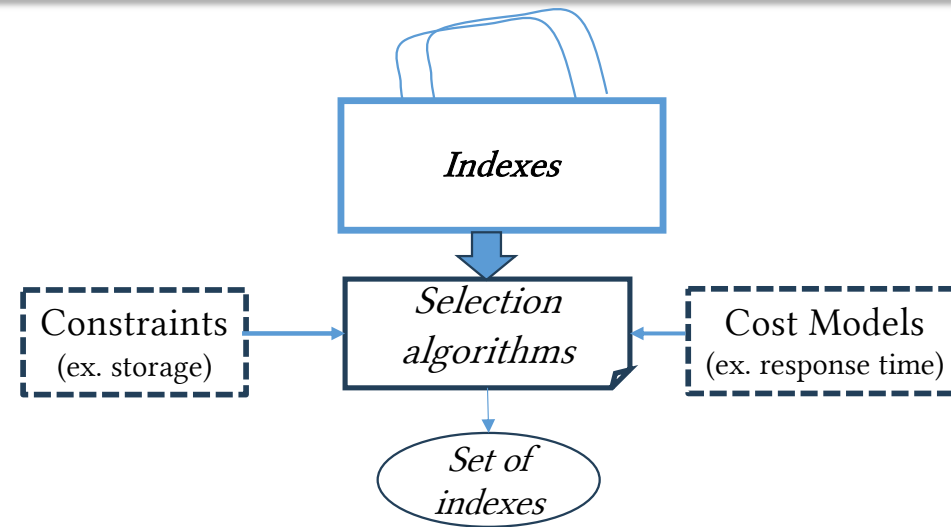
First Generation of QP: Cost-based QP



Cost-based QP

► Usages of an analytical cost model

1. Estimation of:
 - Response time of DB operations (e.g., join, selection)
 - Intermediate results (selectivity factors)
 - Maintenance cost
 - Storage cost
2. Control of optimization structure selection (ex. indexes)

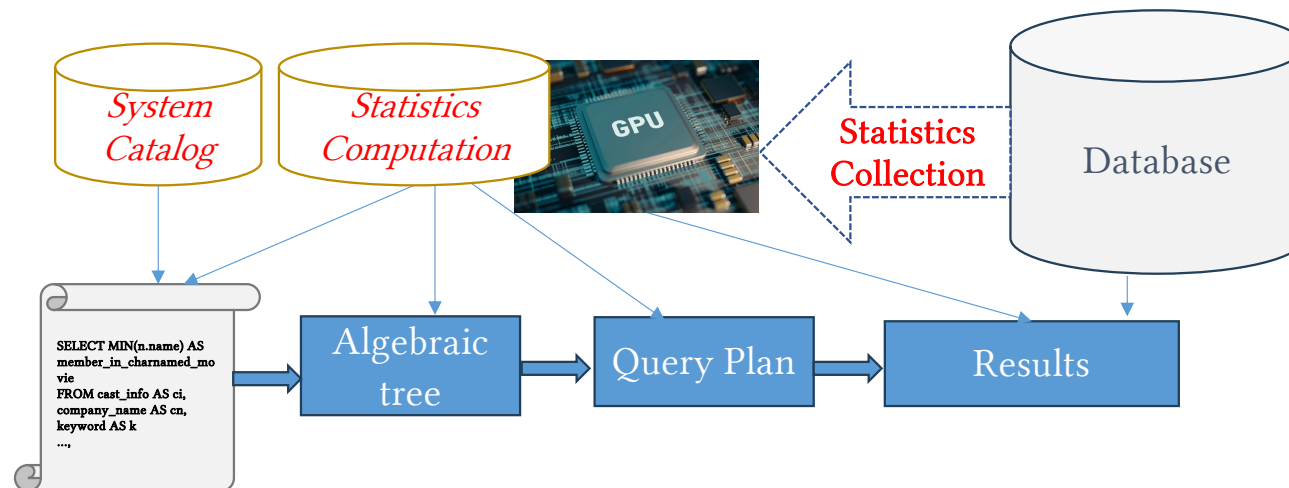


A Software oriented solution to reduce the response time of a query

- L. Bellatreche: *Optimization and tuning in data warehouses*. **Encyclopedia of Database Systems, Springer** (2nd ed.), 2018
- L. Bellatreche, K. Karlapalem, M. Schneider: *On efficient storage space distribution among materialized views and indices in data warehousing environments*. **ACM CIKM**, pp. 397-404, 2000,
- L. Bellatreche, K. Karlapalem, A. Simonet: *Algorithms and support for horizontal class partitioning in object-oriented databases*. **Distributed Parallel Databases**, 8(2): 155-179 (2000)

Hardware accelerators

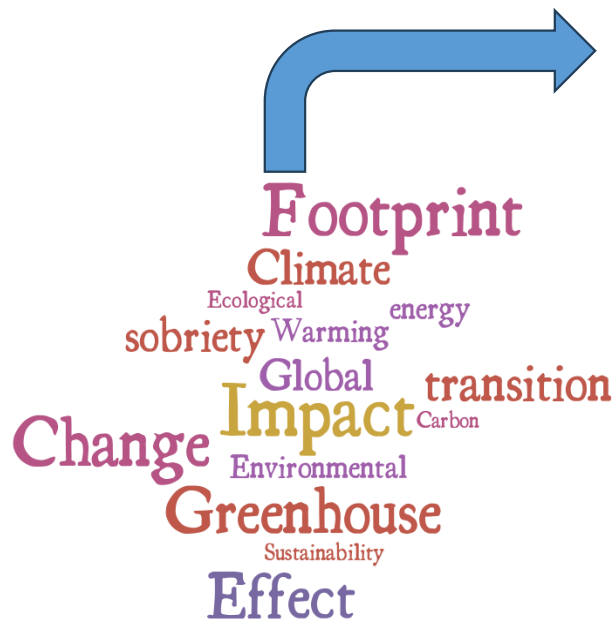
- ▶ Estimation of the size of intermediate results is quite important for query optimization:
 - $((mk \gg k) \gg mc)$ vs. $(mk \gg (k \gg mc))$
 - Bad estimates often lead to the optimizer making wrong decisions.
- ▶ Usage of GPU to run more expensive - but more accurate - estimation algorithm.



Agenda

1. Digital Sector Energy Footprint
2. Methodology for studying energy efficiency (DB)
3. Green query processors
4. Summary

2015: COP21 - Paris



Raising awareness on climate change (CC) and health



CC is everyone's business: green initiatives

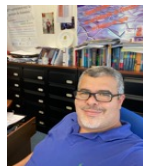
Individuals

Scientists

National Institutions
(Governments, Associations)

International Institutions
(e.g., WHO)

→ Energy efficiency (EE): achievement of the same level of services while consuming less energy

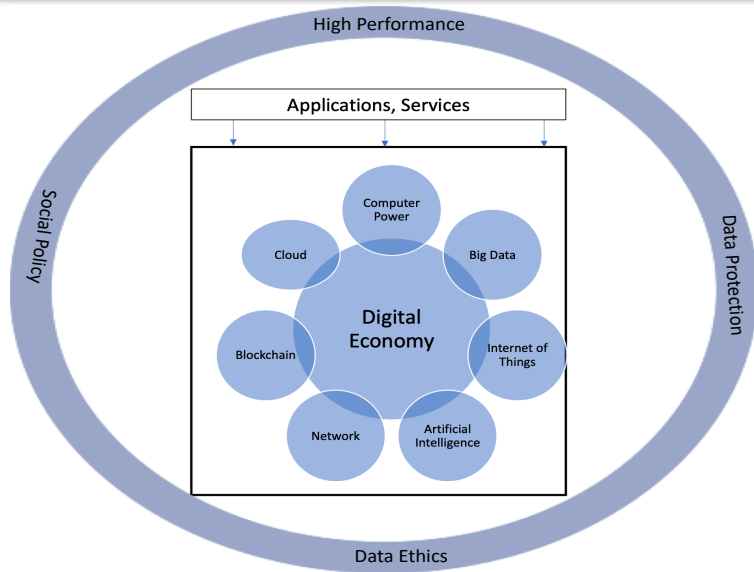


▶ How can I study EE of databases?

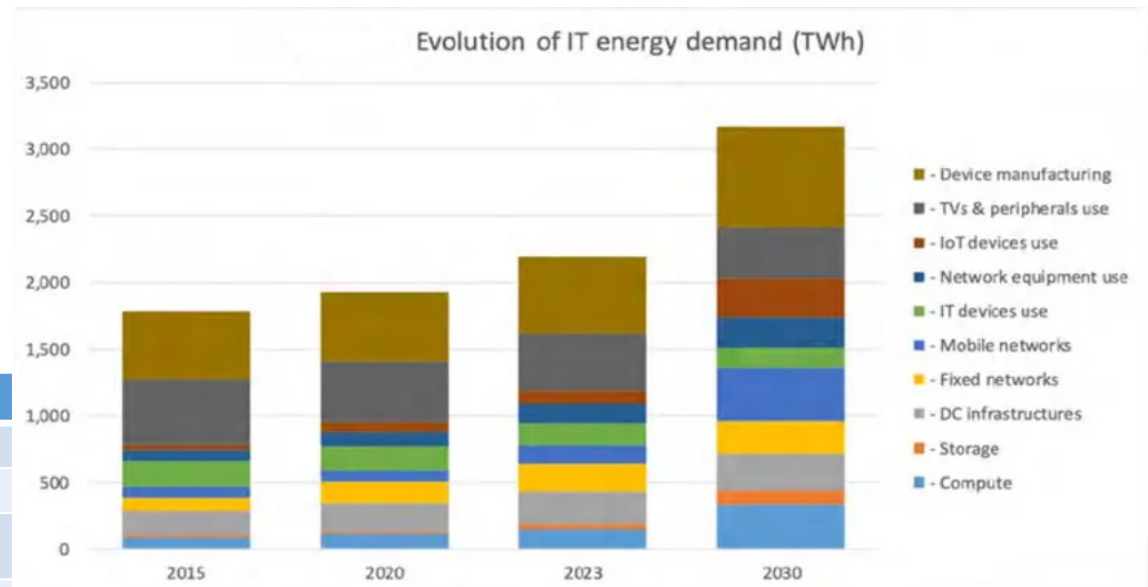
€ Fundings (local, national, and EU)

📄 A niche research topic

Digital sector energy footprint



- ▶ **2030:** IT sector electricity power demand will grow by 50% by 2030
- ▶ **2040:** IT carbon footprint could reach 14%, with data centers contributing to 50% of growth



	2015	2021	Evolution
Internet users	3 billion	4,9 billion	+60%
Internet Traffic	0,6 ZB	3,4 ZB	+440%
Data center workloads	180M	650M	+260%
Data center energy use (excluding crypto)	200 TWh	220-320 TWh	+10-60%
Crypto mining engerg use	4 TWh	100-140 TWh	+2300-3333%
Data transmission network energy use	220	TWh 260-340 TWh	+20-60%

<https://www.i-scoop.eu/sustainability-sustainable-development/it-sector-electricity-demand/>

World's most valuable resource

“Data is the new oil.”

Clive Robert Humby

mathematician, entrepreneur, and Chief Data Scientist, Starcount

“Data is the new currency.”

Antonio Neri, *President Hewlett Packard Enterprise*



“Data is a commodity like gold.”

Matt Shepherd

Head of Data Strategy, BBH London

“At the heart of the digital economy and society is the explosion of insight, intelligence and information – data.”

Data is the lifeblood of the digital economy.

World Economic Forum

A New Paradigm for Business of Data BRIEFING PAPER - JUL Y 2020

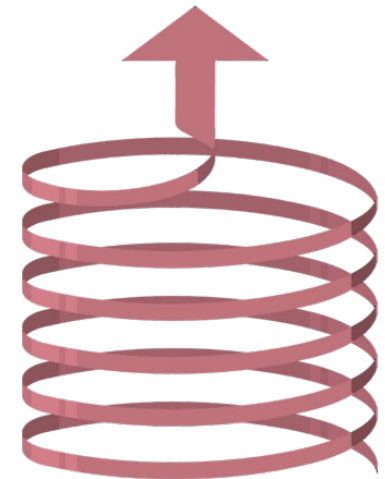
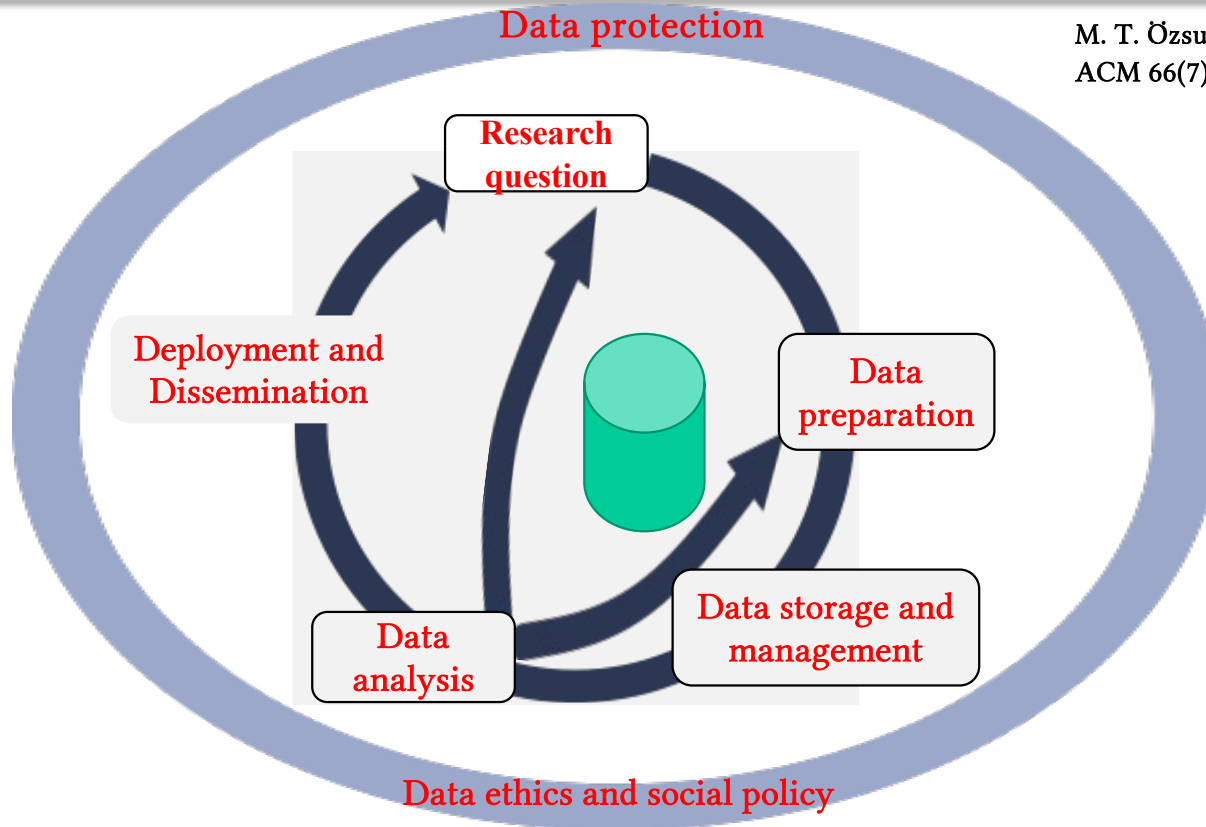
“Data isn’t the new oil, it’s the new CO2

Martin Tisné, Vice President, Luminare Strategic Initiatives

EE of QP

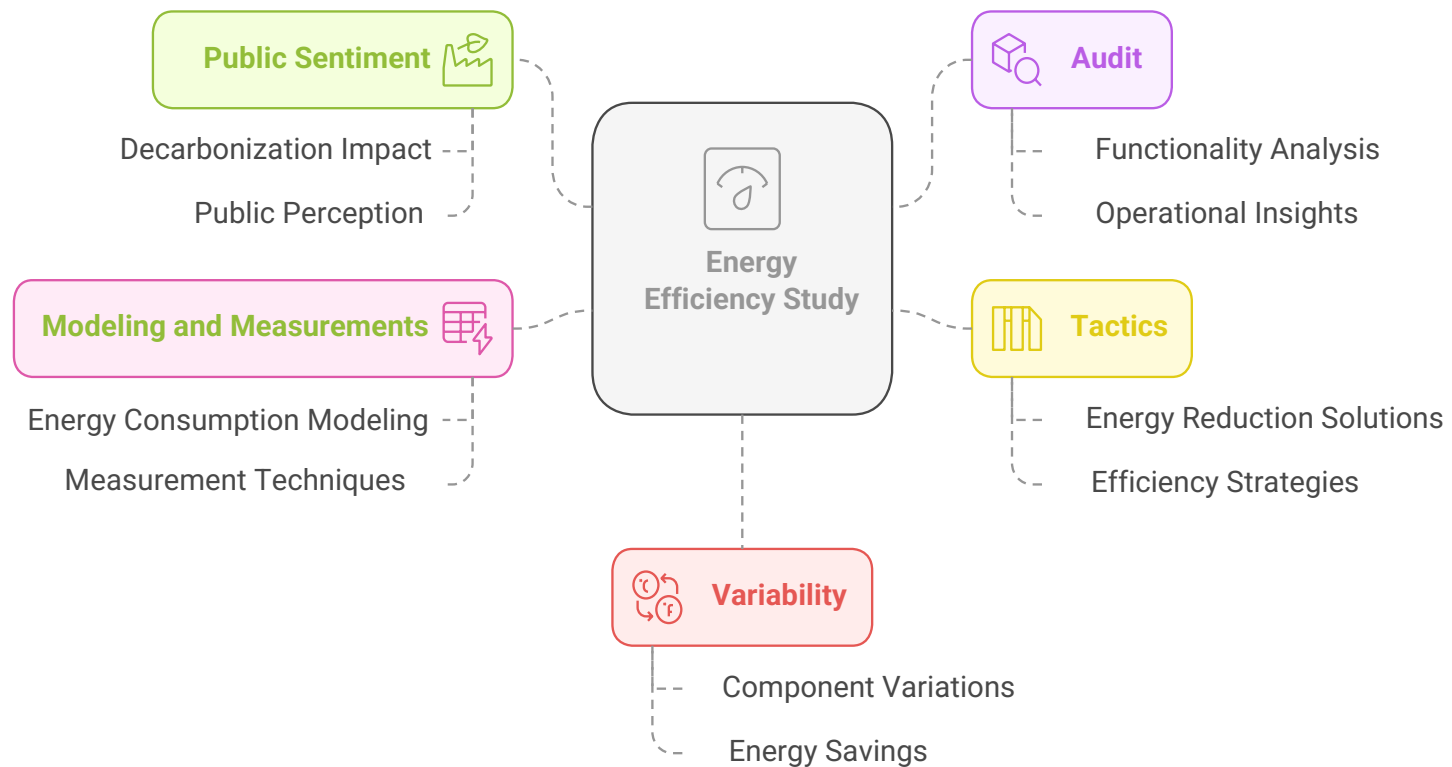
Data Science (DS) era

M. T. Özsu: Data Science - A Systematic Treatment. Commun. ACM 66(7): 106-116 (2023)



- ▶ Developing DS applications demands significant time and resources—**especially energy**.
- ▶ A systematic methodology for studying EE becomes necessary

Our methodology: STAMP-V



1. L. Bellatreche, F. Djellali, W. Macyna, C. Ordonez: *Energy-aware query processing: a case study on join reordering*. **IEEE Big Data**, pp. 3743-3752, 2023
2. S. P. Dembele, L. Bellatreche, C. Ordonez: *Towards green query processing - auditing power before deploying*. **IEEE Big Data**, pp. 2492-2501, 2020

Public Sentiment (1/2)

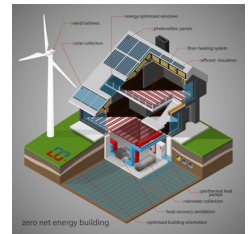
▶ Two controversial questions

1. How can digitalisation be utilized to mitigate the causes and impact of CC?
2. Does the digitalisation sector has a negative impact on the environment and how can it be reduced?



▶ Success stories: DS to the rescue

1. BCOOLER of Google's DeepMind team enhanced the efficiency of cooling systems in data centers by 12.7% (using Reinforcement Learning)
2. Zero Energy Public Buildings: EU IMPROVEMENT
 - Data-driven solutions (ML/DL techniques)
 - o Energy Consumption and Price Prediction



1. J. Tobajas, F. Garcia-Torres, P. Roncero-Sánchez, J. Vázquez, L. Bellatreche, E. Nieto, *Resilience-oriented schedule of microgrids with hybrid energy storage system using model predictive control*, *Applied Energy Journal*, 118092, 306, Part B, 2022
2. F. Chauvet, L. Bellatreche, C. A. S. Silva: *AI approaches for electricity price forecasting in stable/unstable markets: EU improvement project*. *IEEE Big Data*, pp. 4473-4482 , 2022

Public Sentiment (2/2)

“Data is the new oil.”

Clive Robert Humby

mathematician, entrepreneur, and Chief Data Scientist, Starcount



“Data isn’t the new oil, it’s the new CO2

Martin Tisné, Vice President, Luminate Strategic Initiatives

→ AI harming the environment ☹️

- ▶ Cloud data centers (data storage and processing) consumed approximately 205 terawatt-hours (TWh) of electricity in 2020 (*≈ 1% of the world’s total electricity consumption*)
- ▶ **Heating, ventilation,** and **air conditioning** are responsible for a significant percentage of global CO2 emissions
- ▶ Workflow applications are energy intensive: **training, building** and **deploying models**

Model	Number of features	CO2 equivalent emissions
Gopher	280B	322 tons
BLOOM	176B	26 tons
GPT3	175B	502 tons
GPT	175B	70 tons

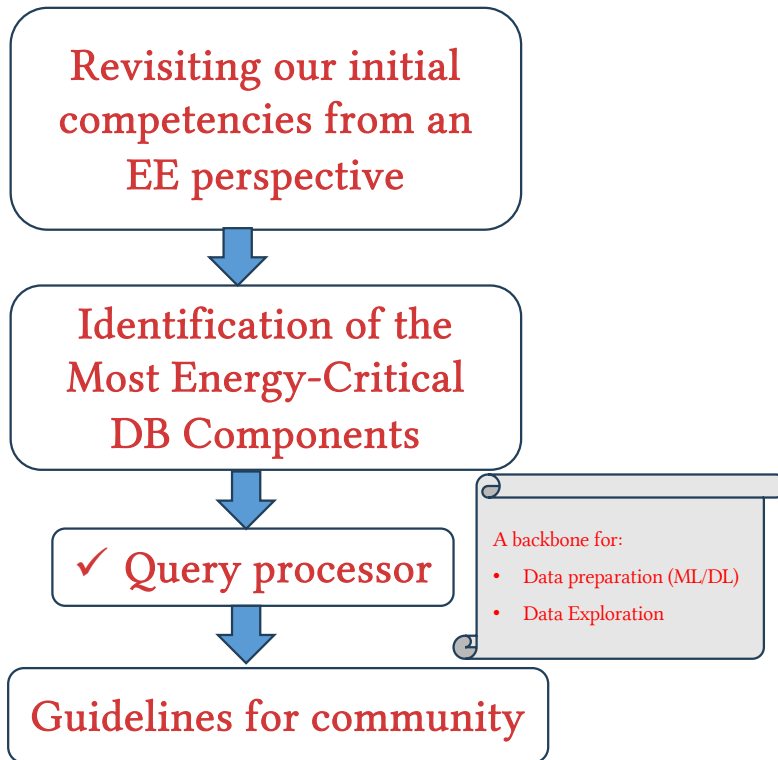
Source: Luccioni et al., 2023

→ *What can we do in this critical situation?*

- *Each scientist has to be energy-aware*

Audit

▶ Our methodology for studying EE of DB

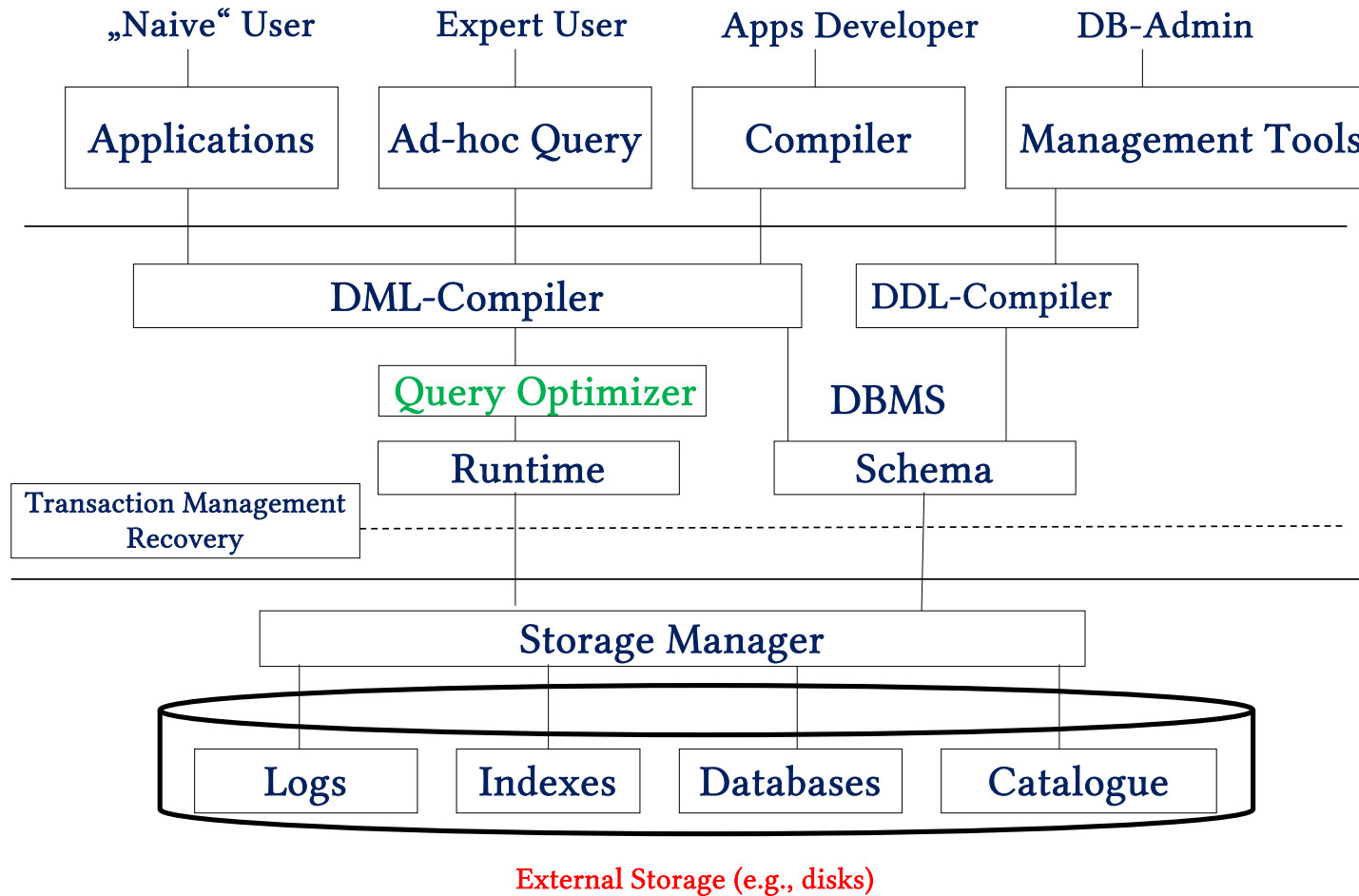


▶ Our guidelines

1. “Audit first then deploy” framework to design green QP
2. Leveraging our initial competencies (QP from a response-time Perspective)
 - a) Quantification of energy consumption of a QP
 - b) Deep understanding of DB ecosystem
 - c) Identify energy-sensitive factors and trade-offs
 - d) Proposal of green QP

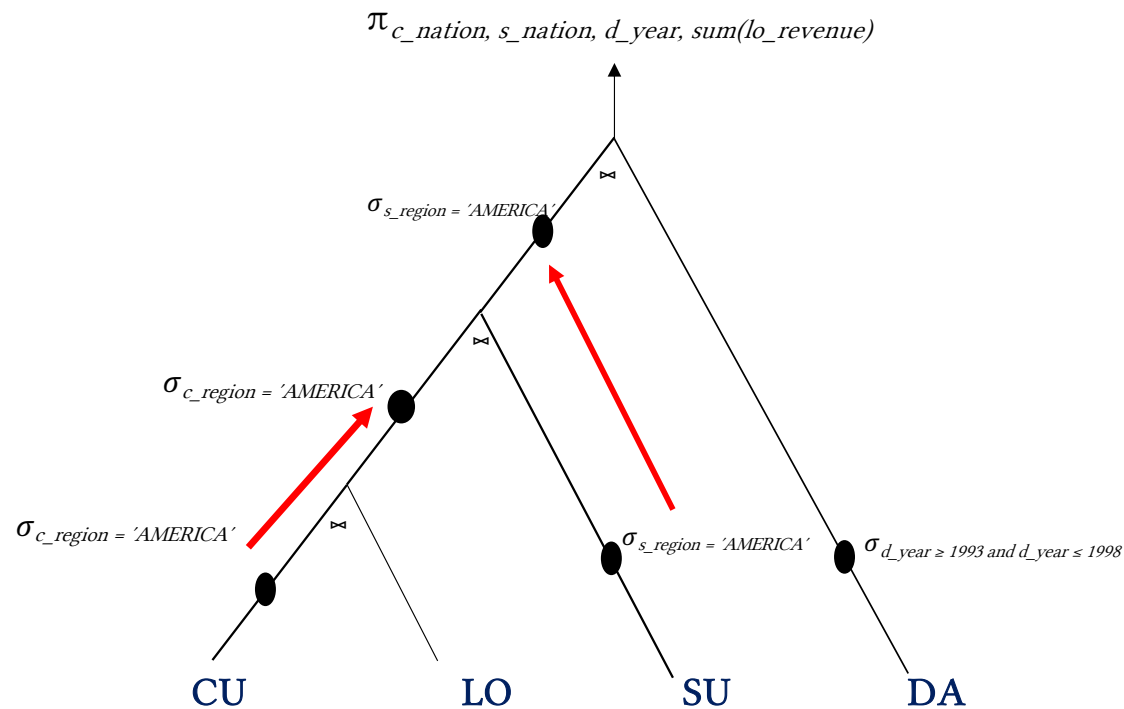


Audit QP



Complexity of query processing

- ▶ Finding an optimal plan is a hard problem
 - Variety of query trees (cascading of selections, **join ordering**, ...)



implementations of a DB operation

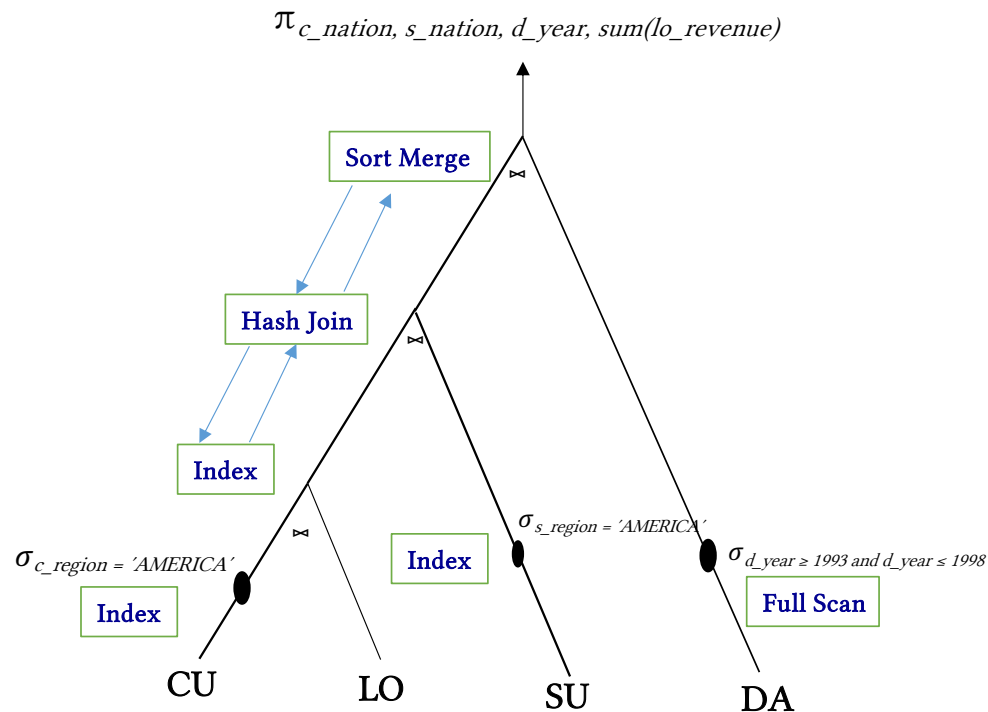
- ▶ A query plan
 - Query tree in which each node has is annotated by its **implementation algorithm**
 - Multiple algorithm per node

I. Unary operations

- Selection
 - Sequential scan
 - Index

II. Binary operations

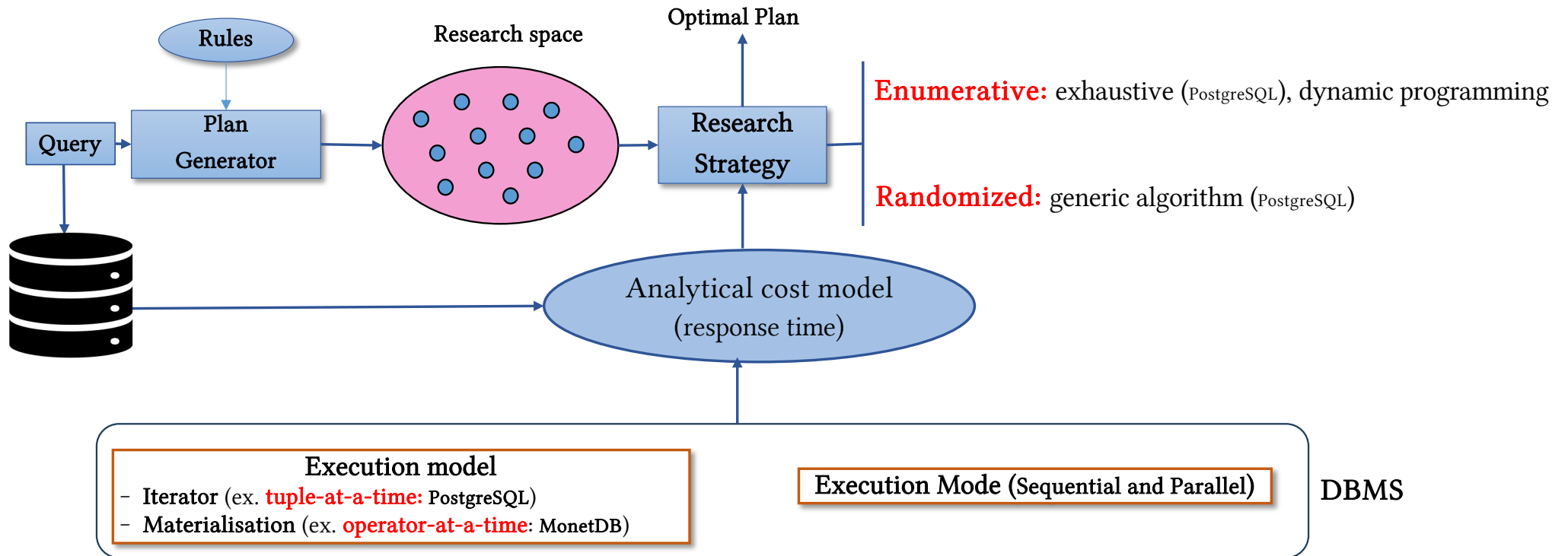
- Join
 - Nested-loop
 - Indexed nested-loop
 - Sort-Merge
 - Hash
 - ...
- Choice based on cost estimate



→ Problem of finding the *best plan*

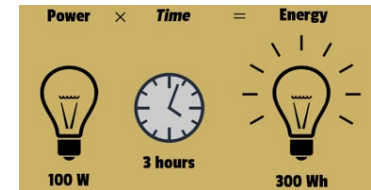
- NP-hard

Cost-based QP



Modeling & Measurements

- ▶ **Energy** (E, joules, watts-hour): the ability to do work.
- ▶ **Power** (P, watts): the rate of energy (E) per a unit of time (T)
- ▶ **Baseline Power:** power consumption when the machine is idle
- ▶ **Active Power:** power consumption due to the execution of the workload
- ▶ **Peak power:** represents the maximum power
- ▶ **Average power:** average power consumed during execution



Existing measurement techniques (1/3)

Energy_Measurement(DB, Workload) → Estimated Energy Cost

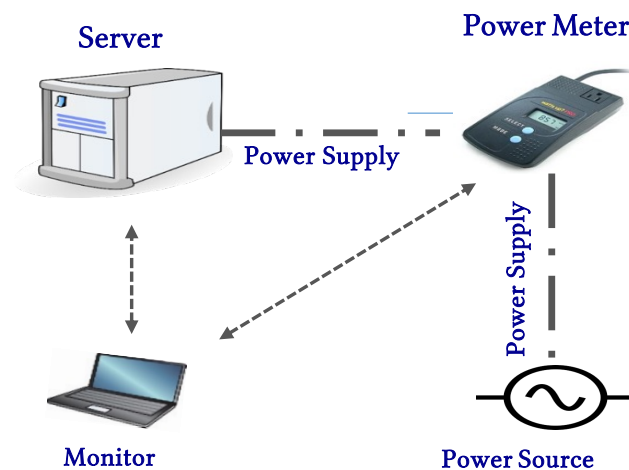
1. Real measurements

- Use of monitor the power consumption of electrical device (ex. yocto-watt)

2. Estimated Measurements: CodeCarbon

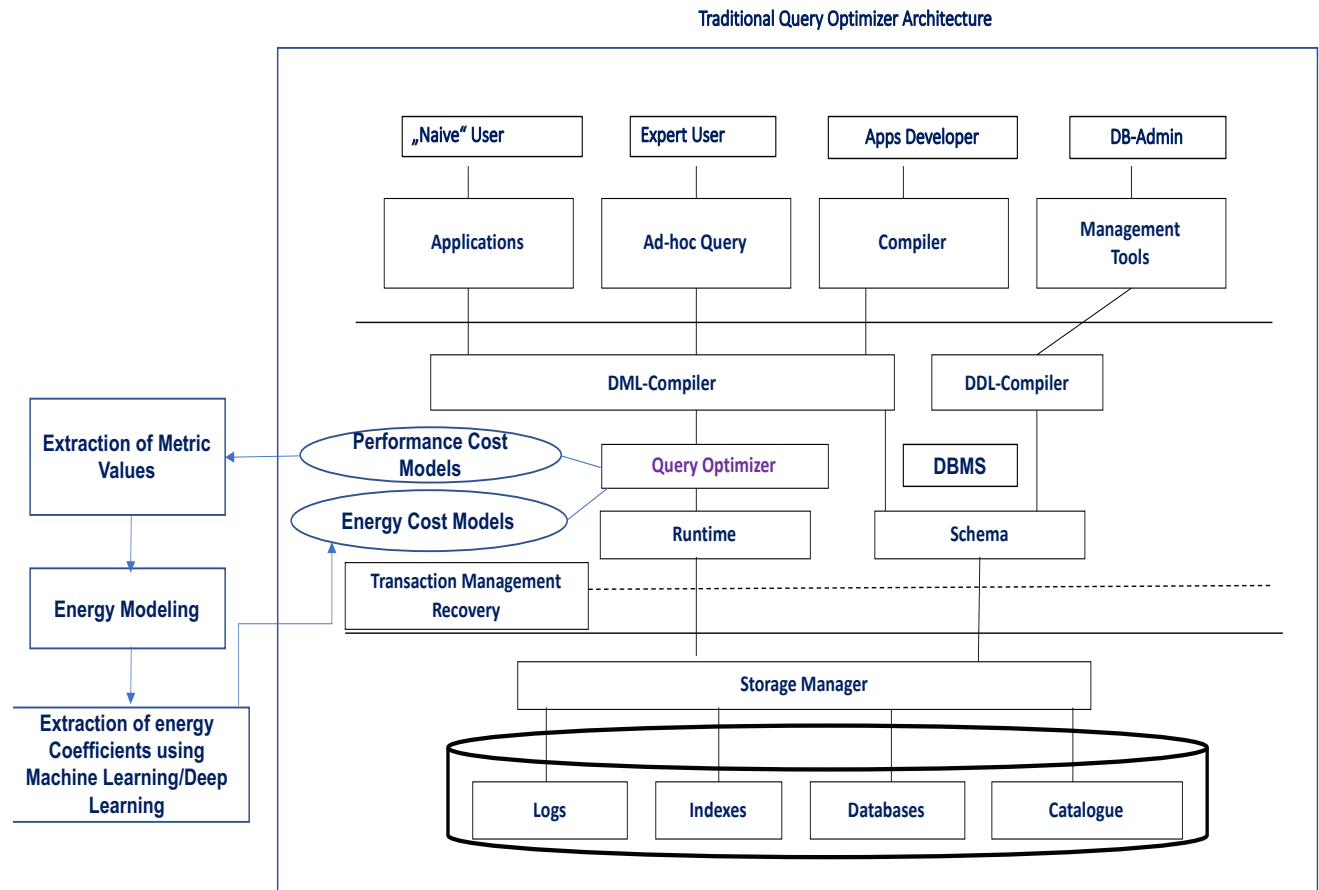


- An open-source tool to estimate carbon emissions from computing
- Developed by Mila, BCG GAMMA, and Haverford College
- Supports Python workflows
- Works with CPU, GPU, and cloud providers
- Monitors (CPU/GPU usage, execution time, location-based carbon intensity)
- Calculates (power consumed, carbon dioxide equivalent (CO₂eq) emitted)
- Outputs (emission logs (.csv), visualization dashboard)



Existing measurement techniques (3/3)

4. AI-Driven Approaches



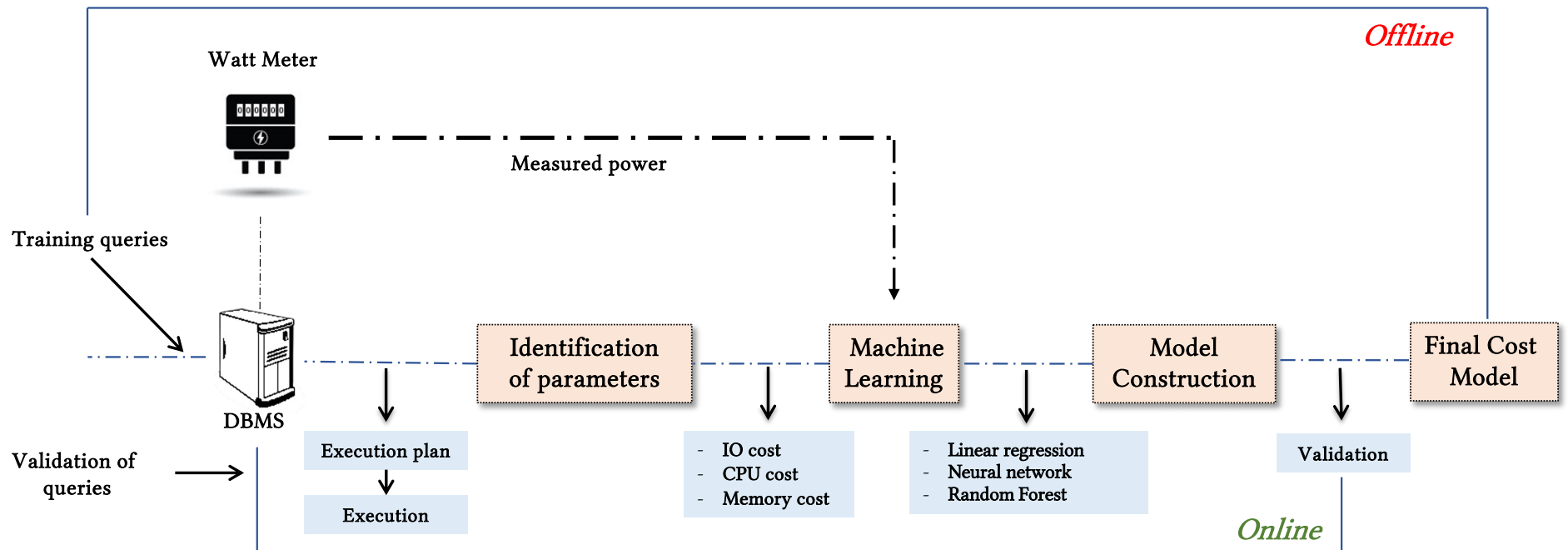
AI-driven techniques

Inputs

- DB schema
- Training queries
- DBMS
- Platform

Used DBMSs

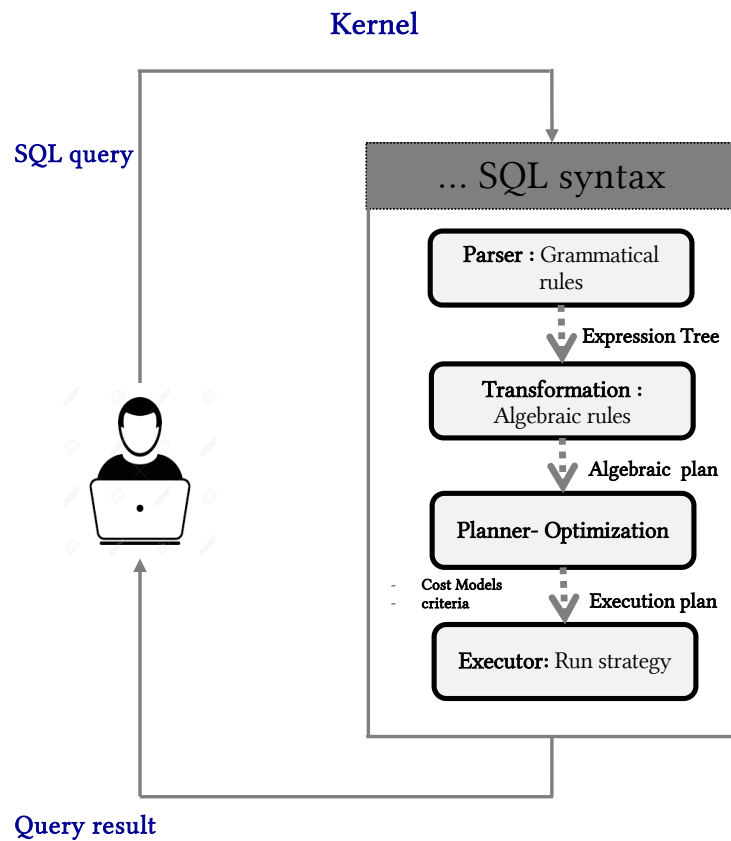
- PostgreSQL : Explain, discard all, ...
- Watt Meter



Case of PostgreSQL DBMS



Widely used in Research and Teaching (OLTP Applications)



► Major audit information

1. Row-oriented storage layout

CID	CName	Gender	Age
1	Simon	M	36

2. Only SQL queries
3. Parallel query processing (multicore)
4. Modelling level: pipeline
5. Other parameters (indexes, selectivity of predicates, buffer politics, size of tables, pages of table, join implementations, ...)

→ Energy Cost = CPU + IO

ML Model

$$\mathbf{Power}(Q_i) = \frac{\sum_{j=1}^p \mathbf{Power}(PL_j) * \mathbf{Time}(PL_j)}{\mathbf{Time}(Q_i)}$$

$$\mathbf{Power}(PL_j) = W_{cpu} * \sum_{u=1}^n C_{cpu_u} + W_{mem} * \sum_{u=1}^n C_{mem_u} + W_{dio} * \sum_{u=1}^n C_{dio_u}$$

Parameters	Meanings
C_{cpu_u}	Number of instructions executed by CPU
C_{mem_u}	Number of read/write operations accessing memory
C_{dio_u}	Number of read/write operations accessing disk
n	Number of operators in the pipeline

→ all provided by PostgreSQL

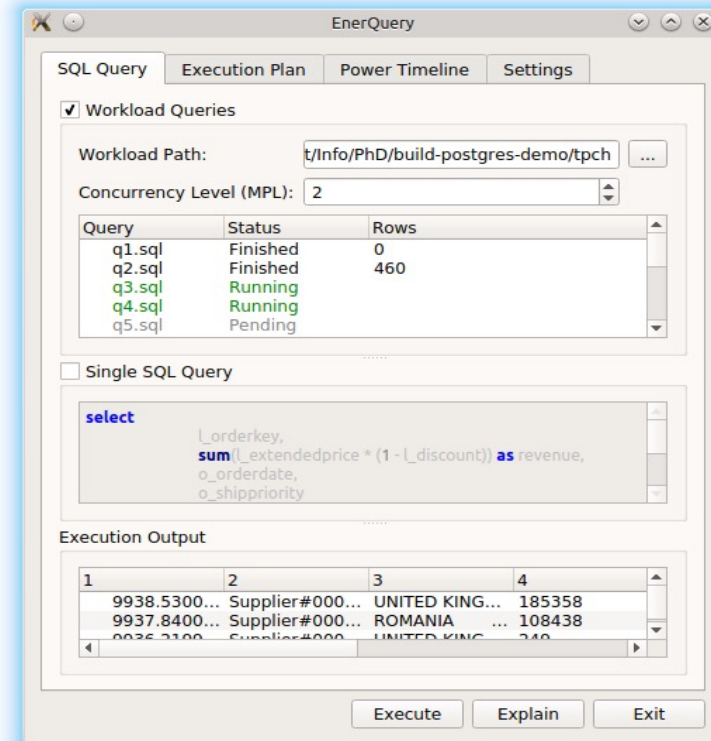
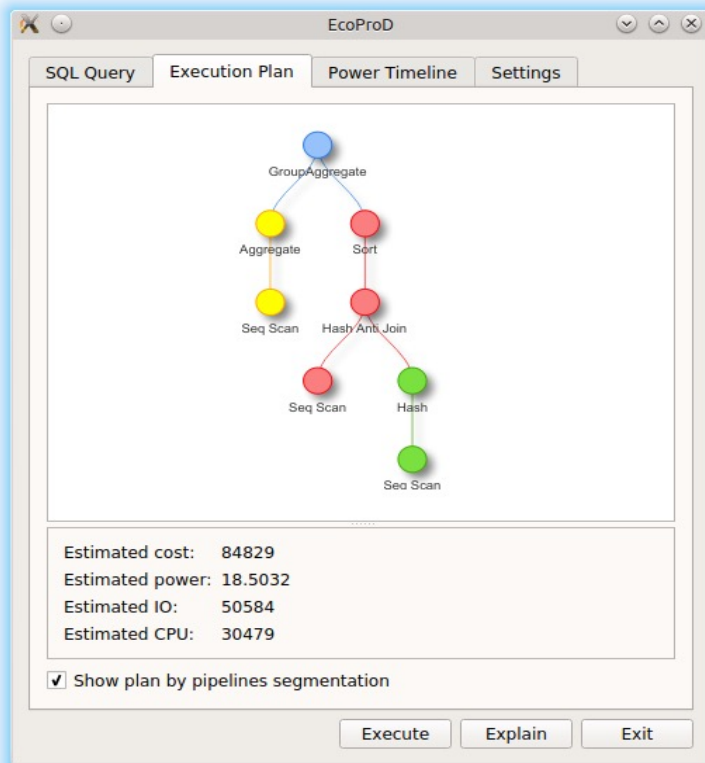
→ ML Technique to calculate β_i : *multiple polynomial regression (degree=2)*

$$P(PL_j^i) = \beta_1 * C_{cpu} + \beta_2 * C_{mem} + \beta_3 * C_{dio} + \beta_4 * C_{cpu} * C_{mem} + \beta_5 * C_{mem} * C_{dio} + \beta_6 * C_{cpu} * C_{dio} + \beta_7 * C_{cpu}^2 + \beta_8 * C_{mem}^2 + \beta_9 * C_{dio}^2 + \beta_0 + \varepsilon$$

- β_i : Regression coefficients
- ε : Measurement error

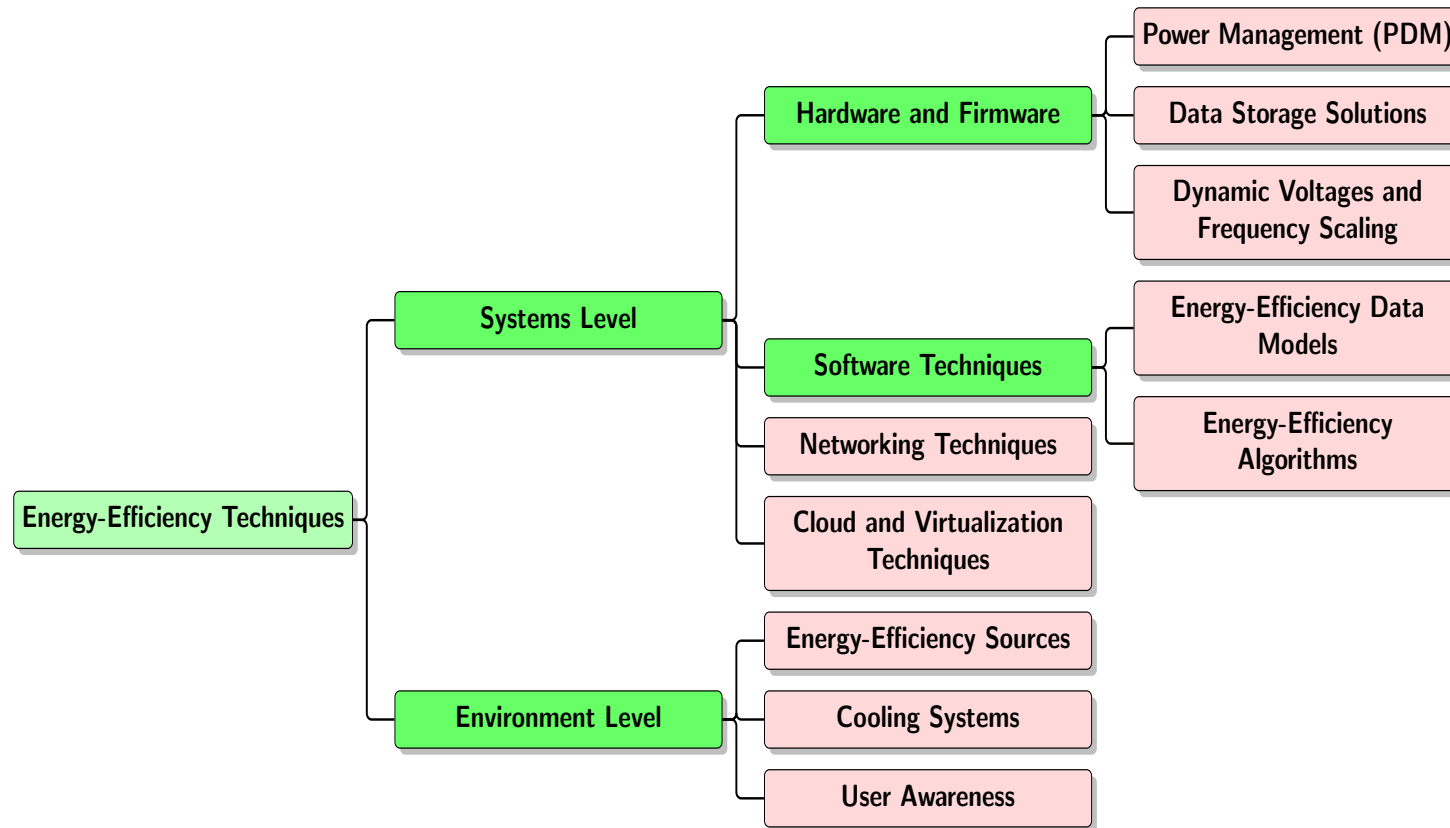
EnerQuery Tool

- ▶ <https://youtu.be/cWK5rf4MBNq>
- ▶ Available at: <https://github.com/lias-laboratory/enerquery>



A. Roukh, L. Bellatreche, C. Ordonez: EnerQuery: *Energy-aware query processing*. ACM CIKM 2016: 2465-2468

Tactics



Hardware Tactics

- ▶ Dynamic Power Management
 - Dynamic component deactivation
 - Dynamic performance scaling
- ▶ Hardware accelerators
 - Graphical Processing Unit (GPU)
 - Field-Programmable Gate Array (FPGA)
 - Tensor Processing Units (TPU)
- ▶ ...

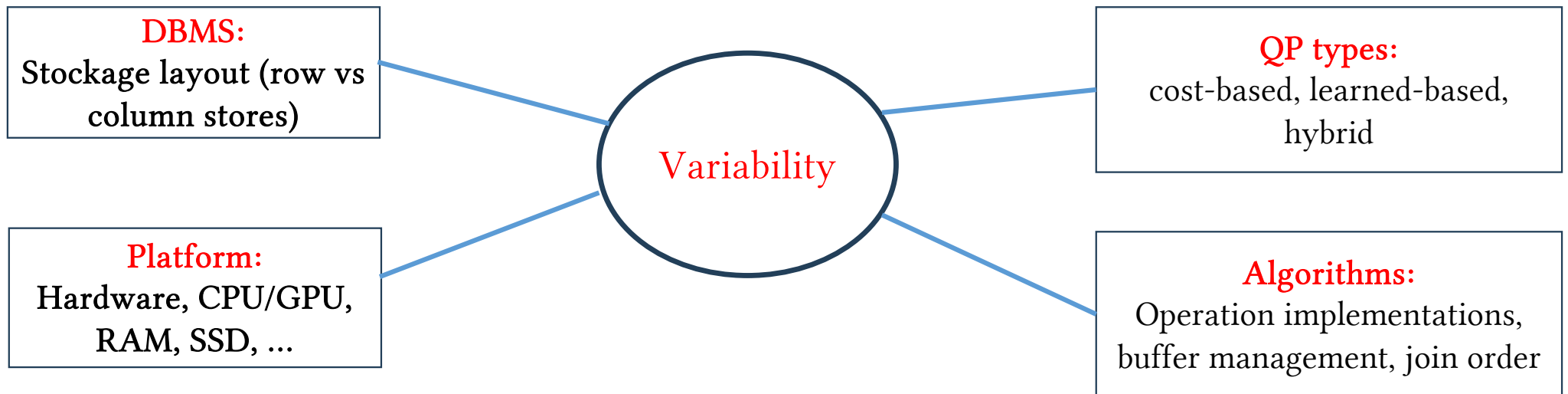
Hardware Tactics

- ▶ **Dynamic Voltage and Frequency Scaling (DVFS) – one of the most popular techniques in DynPS**
- ▶ GOAL: It allows a system to adjust the frequency and supply voltage to particular components within the infrastructure computation
- ▶ DVFS on a multi-core CPU
 - System with only one global voltage for all cores (global DVFS) is energy-inefficient.
 - Global DVFS and per-core DVFS architectures with multiple Voltage Frequency Islands (VFIs) have been proposed.
 - The cores in an island share the same voltage and frequency, but different islands can be executed at various voltages and frequencies

Software Tactics

- ▶ Efficient algorithms and data structures
- ▶ Workload categorization and prediction
- ▶ Query scheduling
- ▶ Buffer management
- ▶ Virtualisation and containers

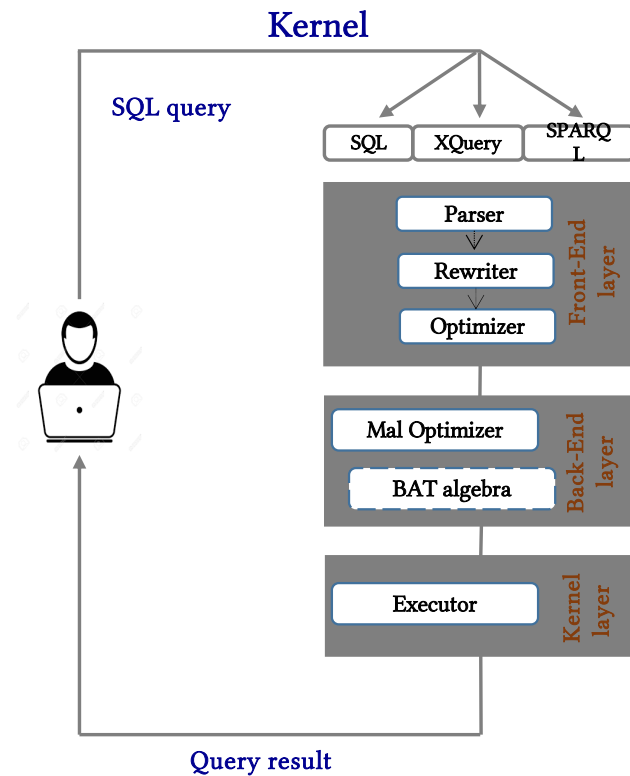
Variability



Variation of DBMS



Analytical Queries



Major audit information

1. Column-oriented storage layout

CID	CName	Gender	Age
1	Ahmed	M	36

2. Compression
3. Multi query languages
4. Parallel query processing (**multicore**)
5. Modelling level: operation
6. ...

$$\rightarrow \text{Energy Cost} = \text{CPU} + \text{IO}$$

ML Model

$$Power(Q_i) = \frac{\sum_{j=1}^n Power(OP_j^i) * Time(OP_j^i)}{Time(Q_i)}$$

$$Power(OP_j^i) = W_{cpu} * C_{cpu_j} + W_{mem} * C_{mem_j} + W_{dio} * C_{dio_j}$$

Parameters	Meanings	
C_{cpu_u}	Number of instructions executed by CPU	} → provided by MonetDB
C_{mem_u}	Number of read/write operations accessing memory	
C_{dio_u}	Number of read/write operations accessing disk	

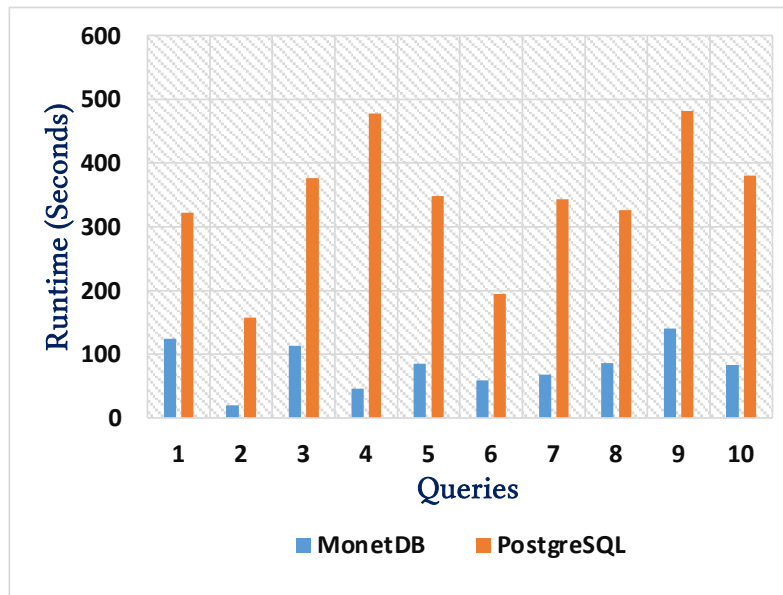
→ Provided by Processor Counter Monitor

→ Machine Learning Technique to calculate β_i : *Multiple Polynomial Regression (degree =3)*

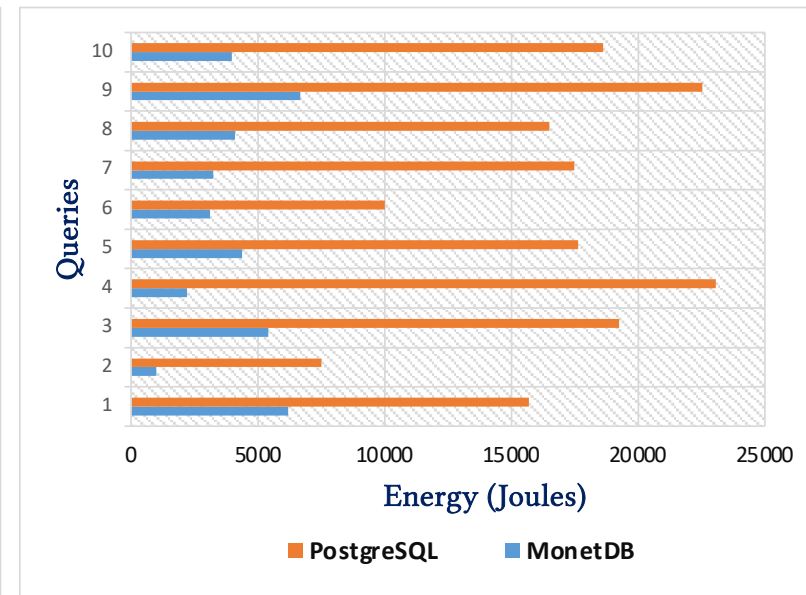
$$P(OP_j^i) = \beta_1 * C_{cpu} + \beta_2 * C_{mem} + \beta_3 * C_{dio} + \beta_4 * C_{cpu} * C_{mem} + \beta_5 * C_{mem} * C_{dio} + \beta_6 * C_{cpu} * C_{dio} + \beta_7 * C_{cpu} * C_{mem} * C_{dio} + \beta_8 * C_{cpu}^2 + \beta_9 * C_{mem}^2 + \beta_{10} * C_{dio}^2 + \beta_{11} * C_{cpu}^2 * C_{mem} + \beta_{12} * C_{mem}^2 * C_{cpu} + \dots + \beta_{17} * C_{cpu}^3 + \beta_{18} * C_{mem}^3 + \beta_{19} * C_{dio}^3 + \beta_0 + \varepsilon$$

- β_i : Regression coefficients
- ε : Measurement error

PostgreSQL vs. MonetDB



Execution time of queries
[30 GB data set of TPC-H benchmark]



Energy consumed of queries queries
[30 GB data set of TPC-H benchmark]

MonetDB consumes less energy than PostgreSQL

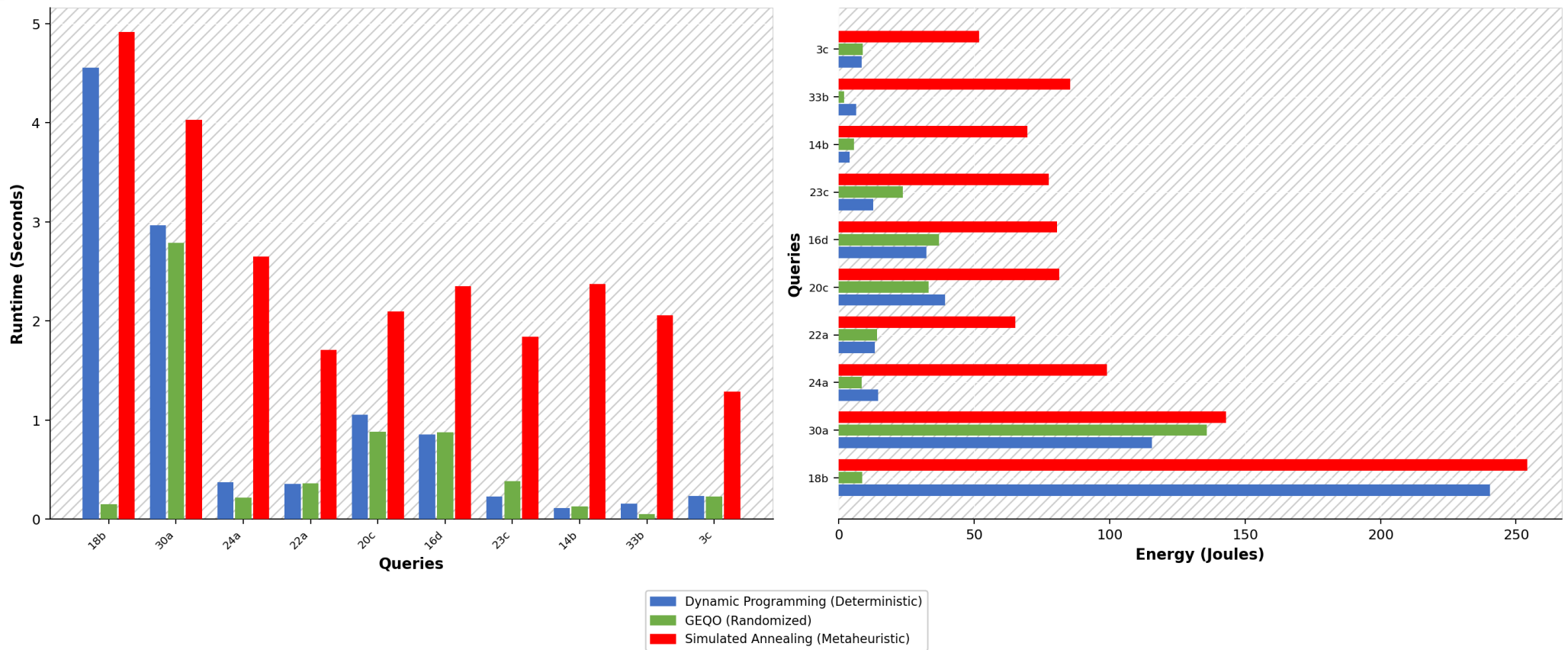
Variability of join ordering

- ▶ The number of possible join trees grows rapidly as the number of join relations increases
- ▶ If a query involves 21 tables, then employing left deep tree, the number of potential join orders is 21!
- ▶ NP-hard problem
- ▶ A QP is responsible for identifying the optimal join strategy and join implementation

1. Exact solutions
2. Randomized Solutions (e.g., genetic algorithm, simulated annealing)
3. Mixed integer linear programming
4. Heuristics (e.g., minimum selectivity, Monte Carlos Tree Search)

Our experimental scenario: the DBMS remains fixed while the join ordering algorithm is varied.

Results of the variation



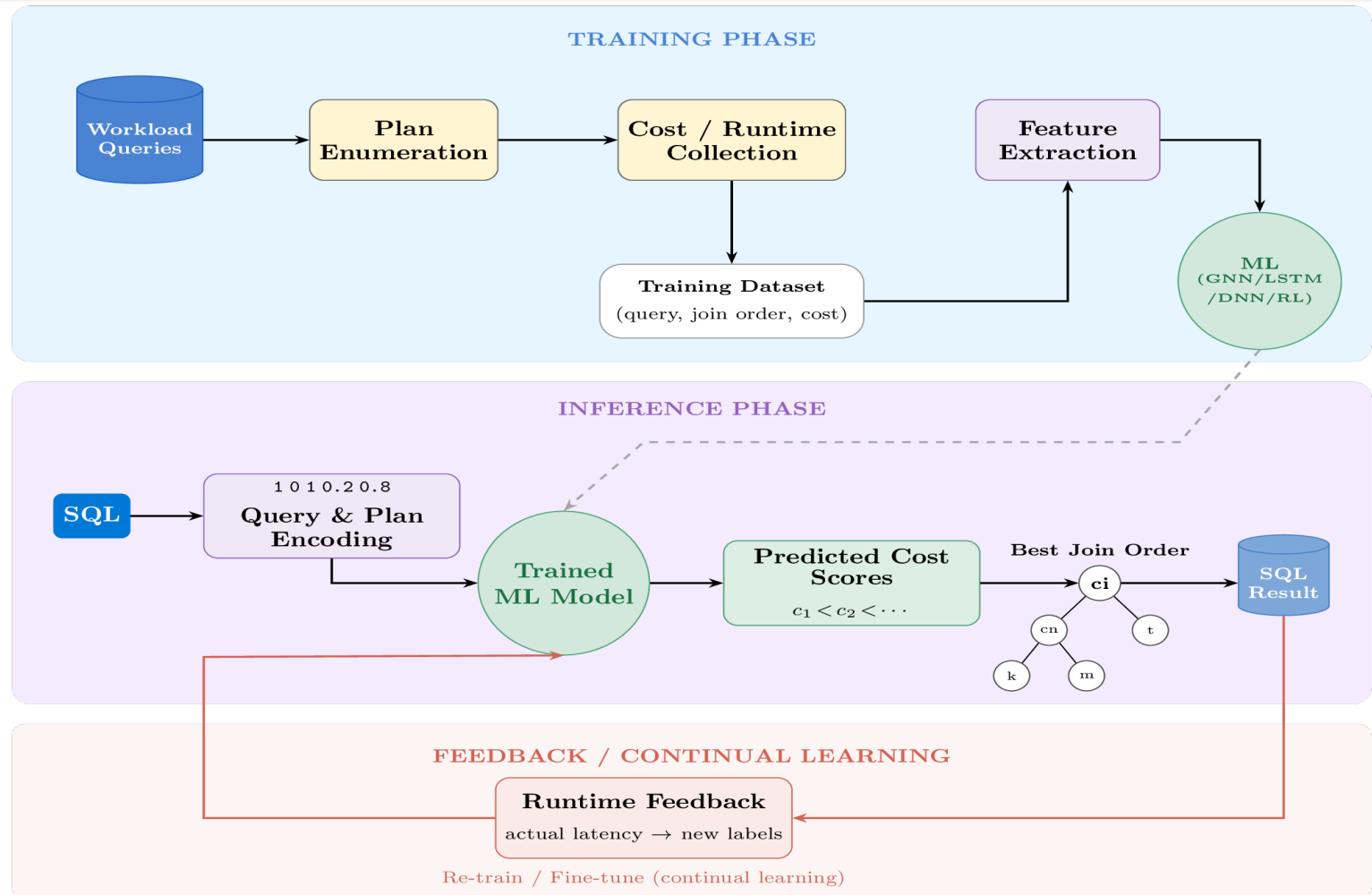
Execution time of queries
[JOB benchmark — IMDB dataset]

Energy consumed of queries
[JOB benchmark — IMDB dataset]

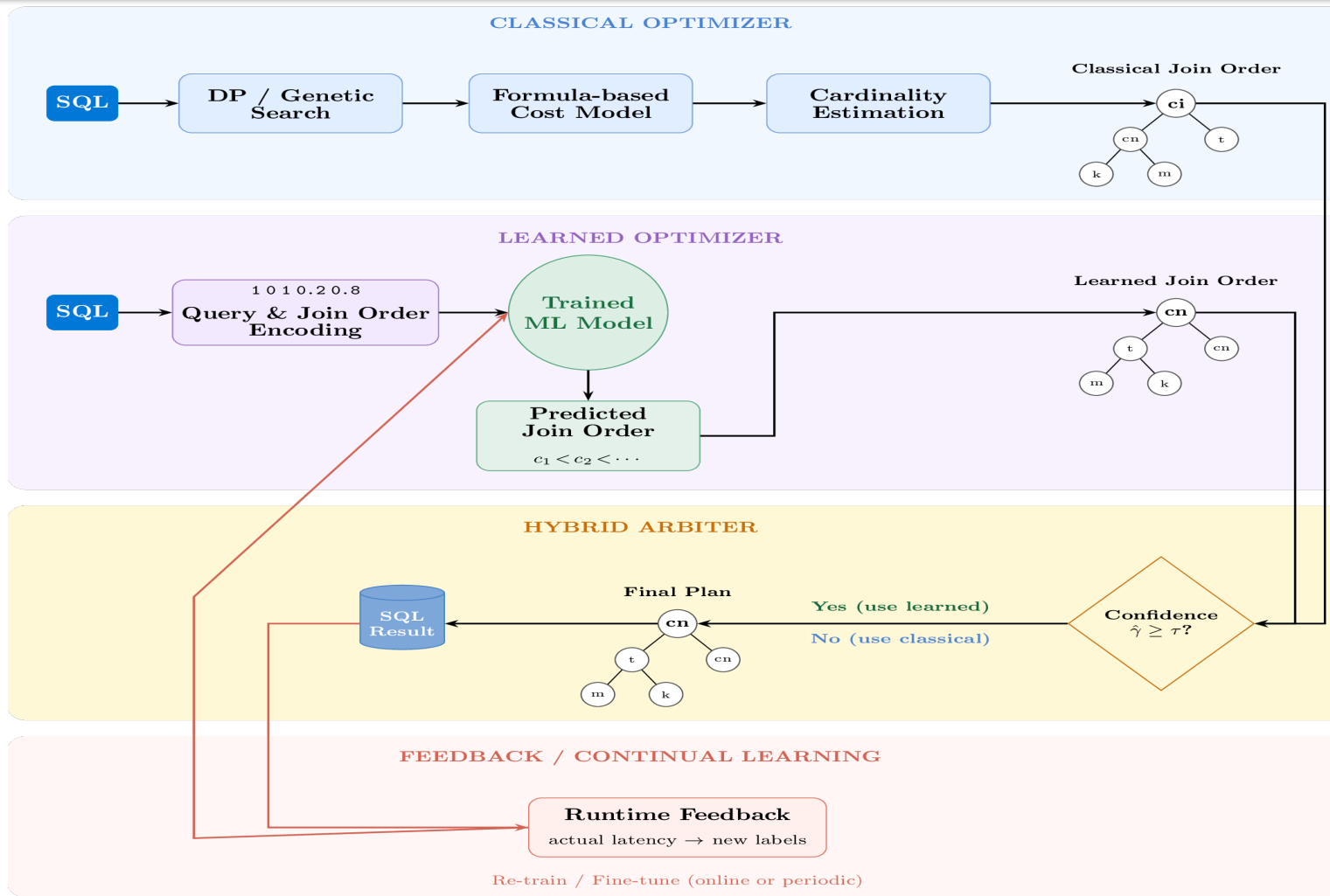
With the same optimizer, join ordering algorithm can drastically change energy use and performance.

Variability of types of QP

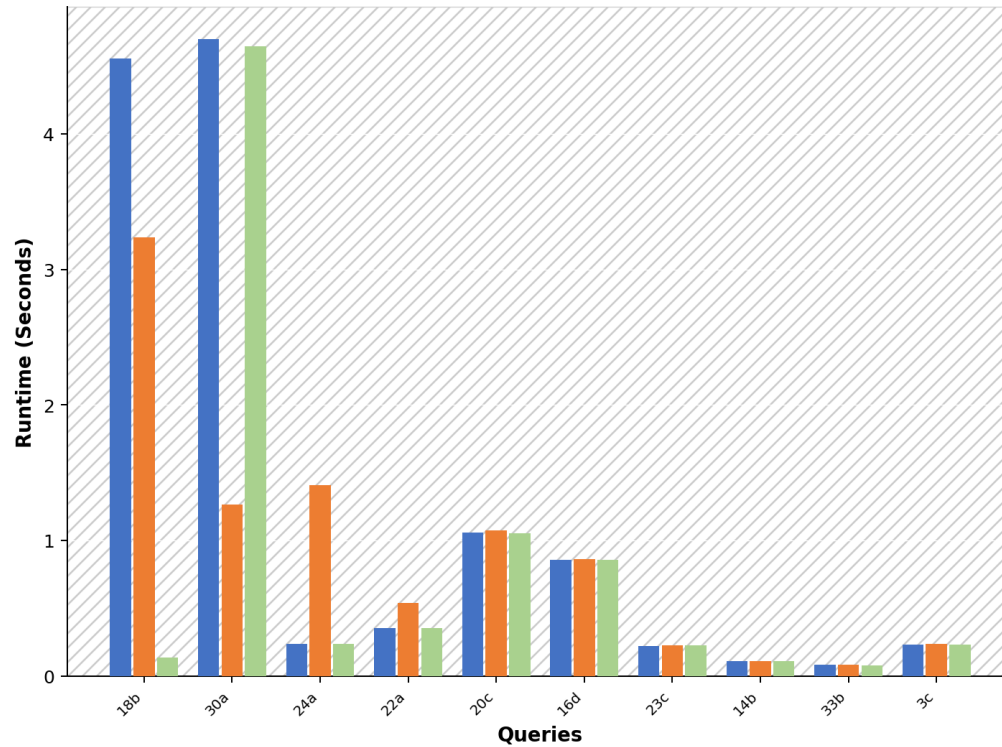
Learned-based QP



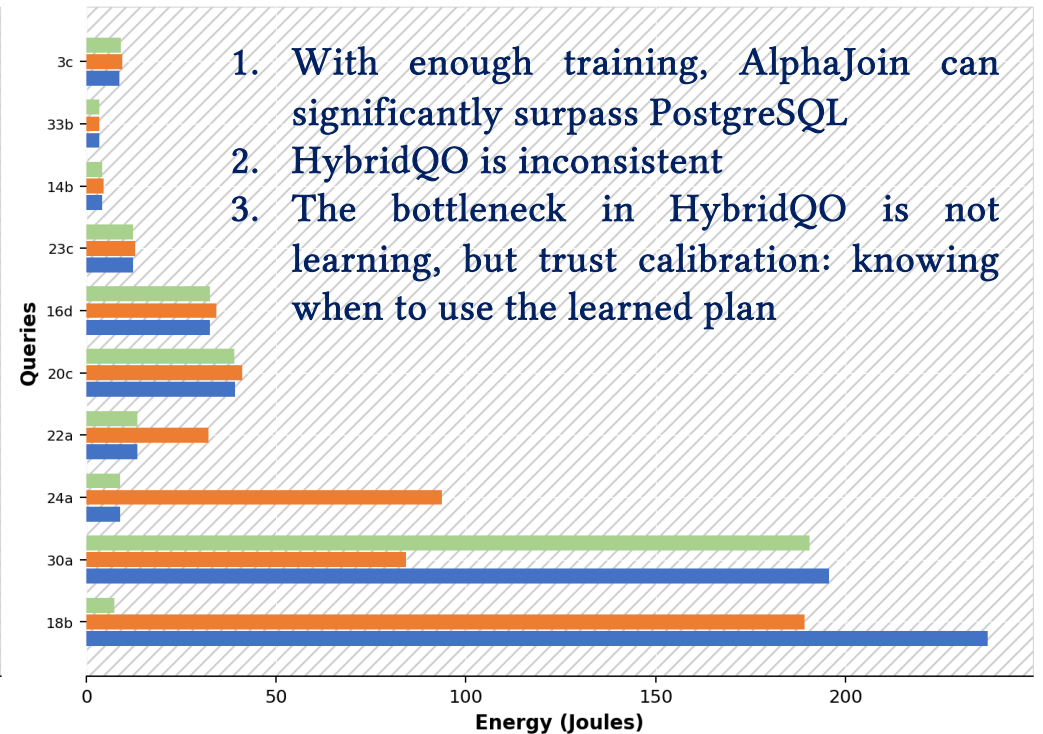
HybridQO



PostgreSQL | AlphaJoin | HybridQO



Execution time of queries
[JOB benchmark — IMDB dataset]



1. With enough training, AlphaJoin can significantly surpass PostgreSQL
2. HybridQO is inconsistent
3. The bottleneck in HybridQO is not learning, but trust calibration: knowing when to use the learned plan

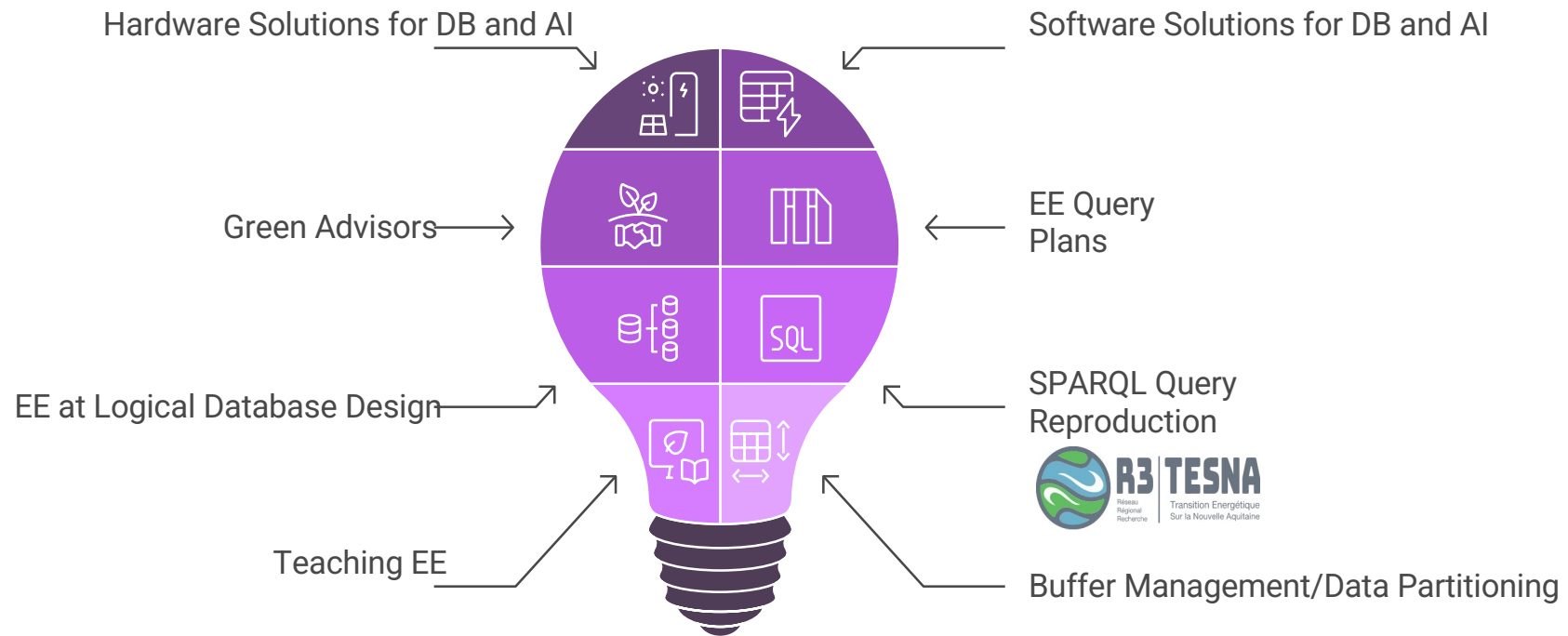
Energy consumed of queries
[JOB benchmark — IMDB dataset]

K. Bennamane, S. Cheikh, L. Bellatreche, How Good are Learned and Hybrid Join Order Strategies from Energy Perspective?, to appear in DaWak'2026

Summary

- ▶ Energy awareness is the **responsibility** of every scientist
- ▶ EE in digitalization and AI is a **niche area**
- ▶ DS's **positive** and **negative** impacts on energy consumption
- ▶ STAMP-V: A comprehensive **framework** for studying EE
 1. Sentiment Analysis
 2. Audit
 3. Tactics
 4. Modeling & Measurements
 5. Variability
- ▶ Application of STAMP-V to QP
- ▶ Studying EE requires a deep understanding of hardware and software

Open issues



Big Thanks



Dr. Amine ROUKH



Dr. Ahcène BOUKORCA



Prof. Carlos ORDONEZ
Univ. of Houston
USA



Prof. Sadok BEN YAHIA
Univ. of Southern Denmark



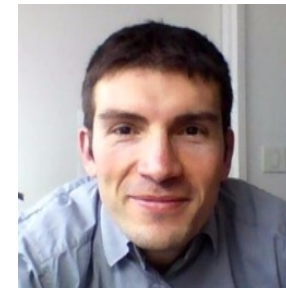
Dr. Simon Pierre DEMBELE



Dr. Abdelkader OUARED



Khadidja BENNAMANE



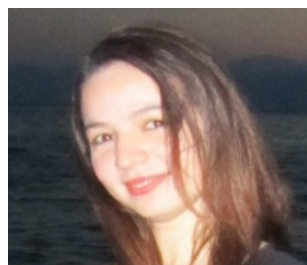
Prof. Stéphane JEAN,
Poitiers Univ.



Dr. Wojciech MACYNA
Wroclaw University
Poland



Dr. Issam GHABRI



Dr. Selma BOUARAR



Dr. Cheikh SALMI
University of Boumerdes
Algeria

Questions? Suggestions? Criticisms?



bellatreche@ensma.fr
www.lias-lab.fr/members/bellatreche