POLITECNICO MILANO 1863

On the challenges of sustainable data preparation

Barbara Pernici

COOPIS25, Marbella
October 20, 2025

POLITECNICO MILANO 1863 | RAISE

Who am I: Barbara Pernici

Professor at Department of Electronics, Information and Bioengineering, RAISE (Research on Advanced Information Systems Engineering) group Politecnico di Milano https://isgroup-polimi.github.io/

Background: Electrical engineering + Computer Science

Discount quality

NextGenerationEU – Italian project

Discount quality for responsible data science: Human-in-the-Loop for quality data

https://www.discountquality.polimi.it/

Inspired by Nielsen's Discount Usability Engineering approach

Focus of the talk

- Data preparation challenges
- Sustainable computing
- Measuring indicators
- Green AI and data preparation
- A circular-economy-based framework for data preparation
- Data preparation challenges and some highlights from ongoing research

Data Preparation

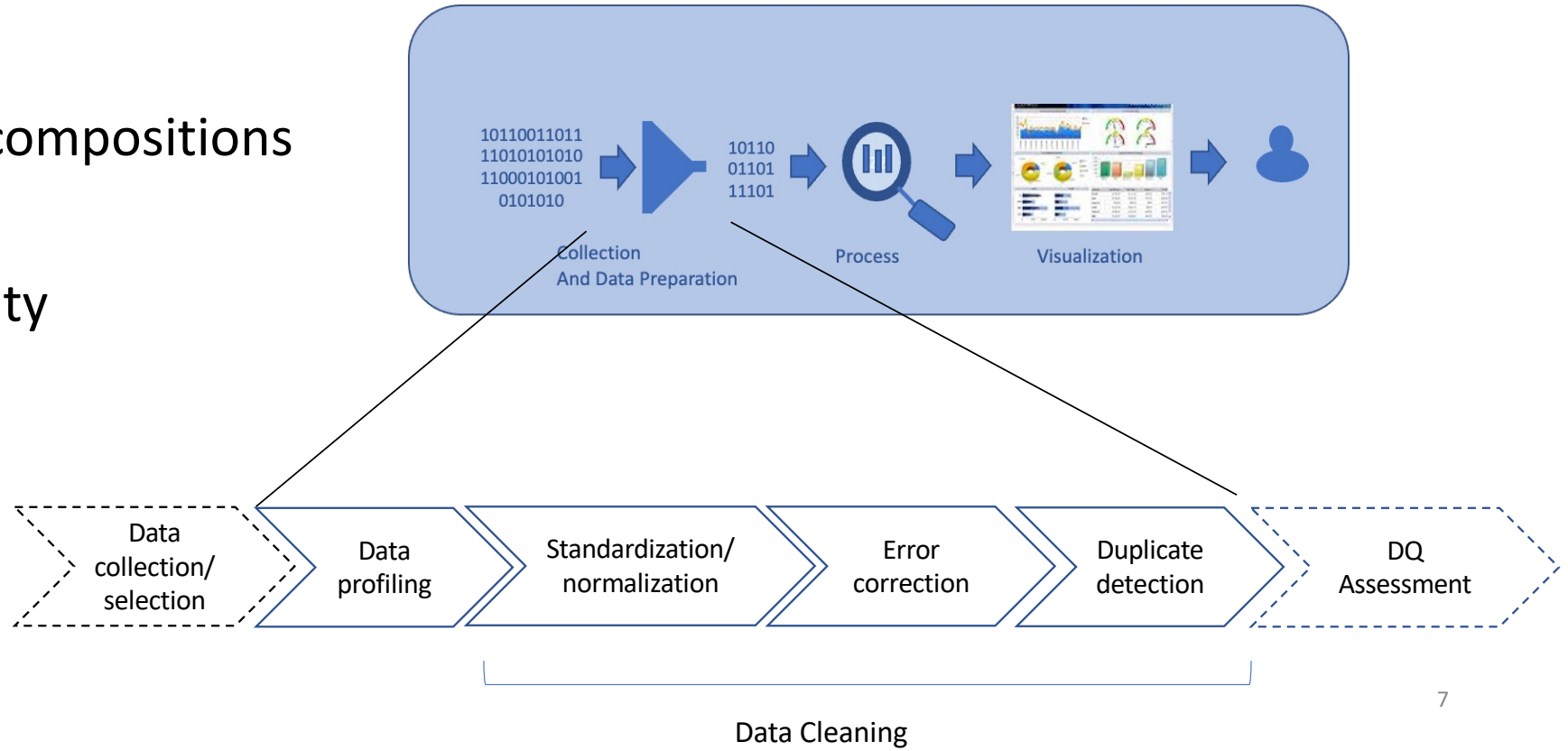
01

Data preparation – why?

- Organizations store a lot of heterogeneous data
  - not cleaned, not “prepared” for the analysis
  - E.g., Data Lake + ELT approach
- Even if the single data source is well curated
  - later it might be integrated with others - need preparation
  - too expensive to keep a unified view
- Only a fraction of the data will actually be used

Pipelines for data-driven decision making

- Goal: reduce needed resources
- Design compositions
- Adaptivity



Importance of data quality in data preparation

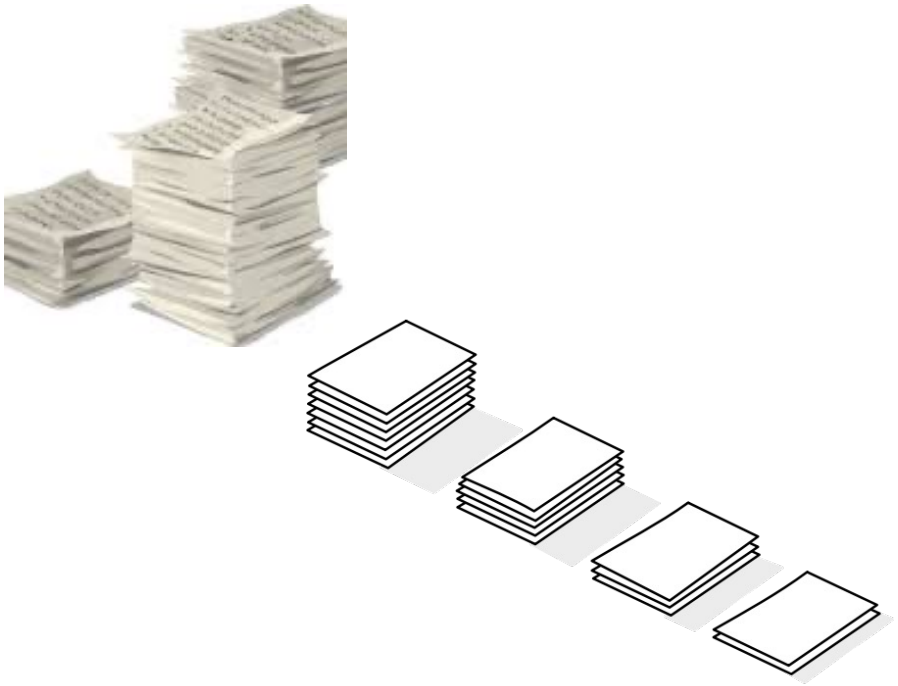
- Multiple sources
  - integration, deduplication, cleaning
- Economic impact of poor DQ for companies
- Significance of DQ dramatically increasing in AI and Model training
  - influences prediction accuracy
  - increasing focus on data

Data-centric AI (DCAI)

systematically engineering the data used to build an AI system

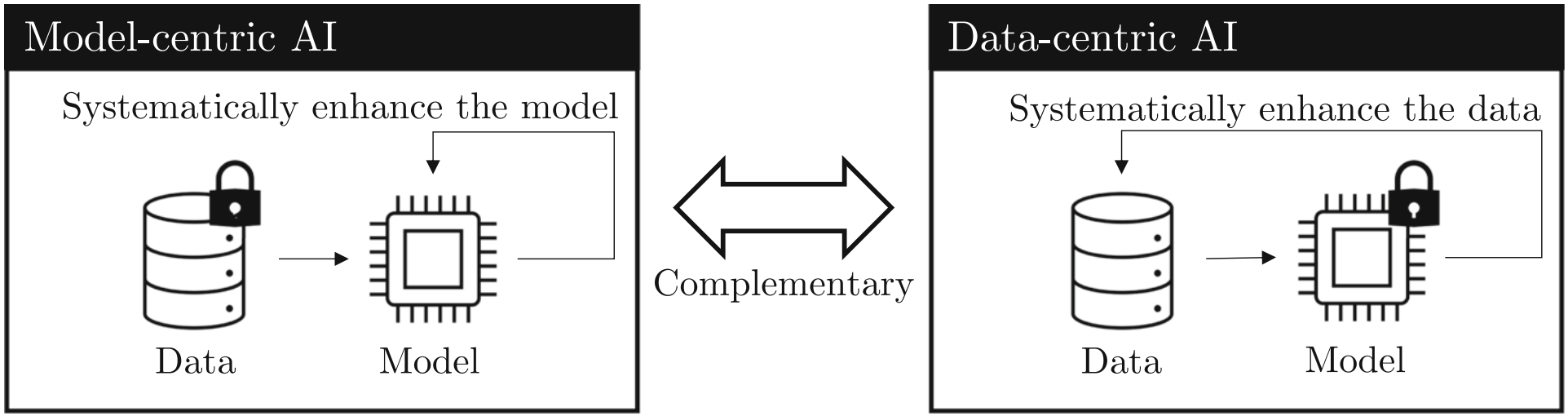
- Carefully selecting and preprocessing data
- Reducing data redundancy
- Avoiding overfitting

- Sample and size of samples
  - Feature selection
  - Reducing the number of data points



Verdecchia, Roberto, et al., ICT4S, IEEE, 2022.

Data-centric artificial intelligence



J Jakubik et al, BISE, 2024

Challenges

- Very large amounts of data
- Importance of data quality
- Processing time
- Environmental sustainability

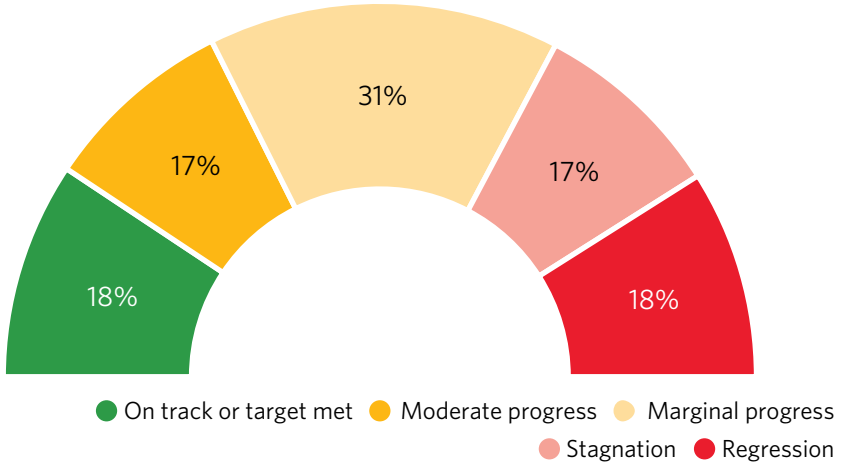
Sustainability and IT

02

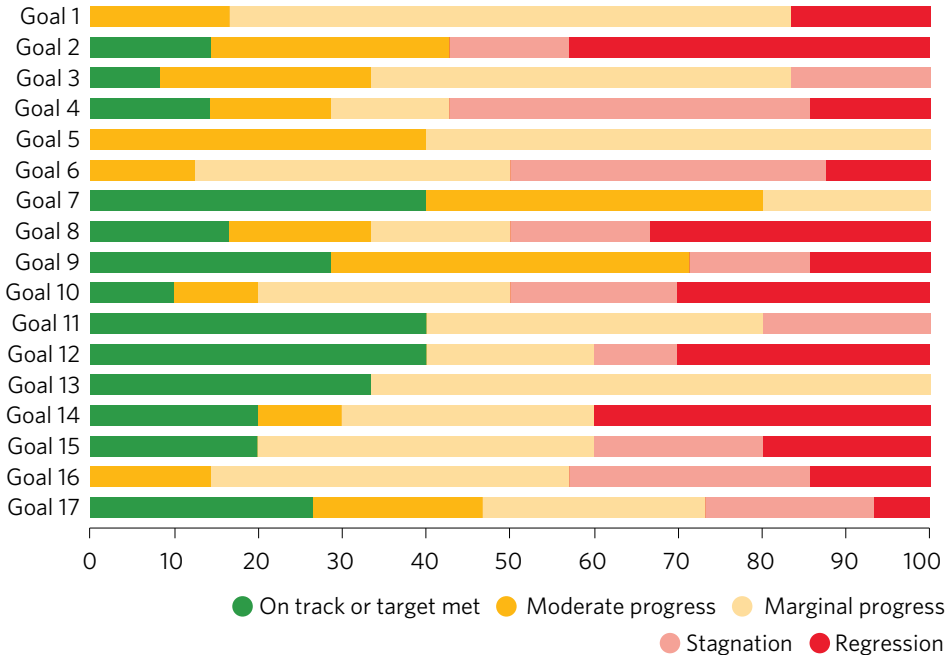
UN Sustainable Development Goals

Progress towards sustainable development goals

Overall progress across targets based on 2015–2025 global aggregate data



Progress assessment for the 17 Goals based on assessed targets, by Goal (percentage)



UN Report June 2025

Sustainability and IT

Contribution of IT and in particular AI towards achieving the goals

→ **Computationally intensive, surge in resource usage and energy consumption**

Sustainability and IT

Co ☰ COMMUNICATIONS OF THE ACM July 2025 goals

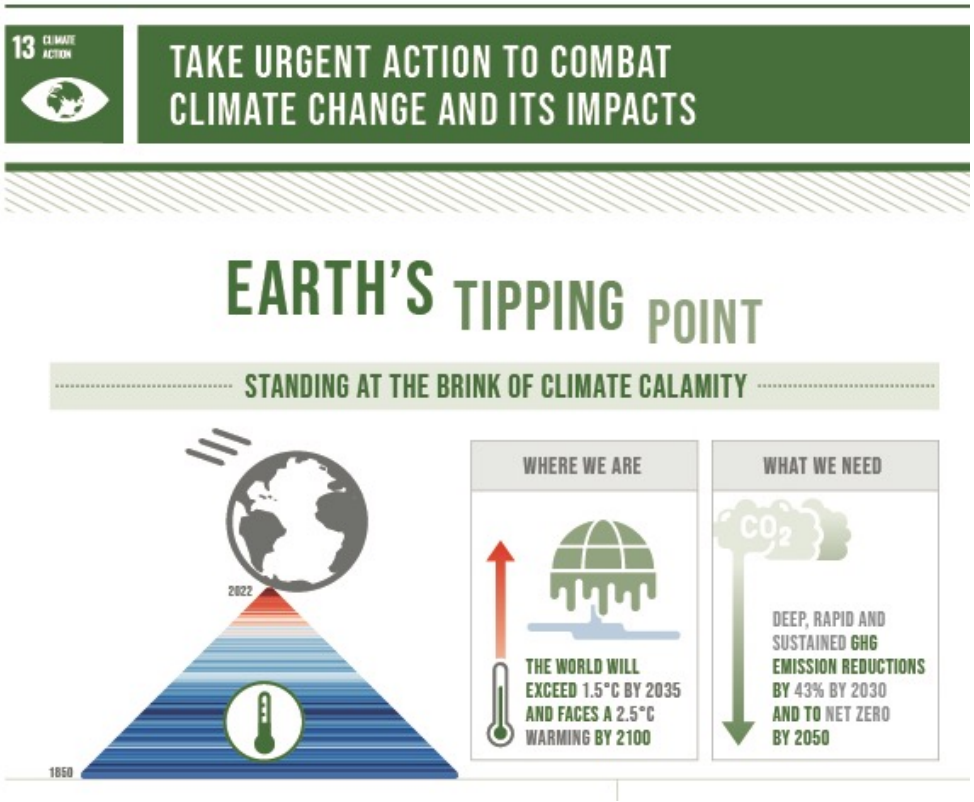
Sustainability and Computing

Compu  onsumption

Computing sustainability / Green IT / Sustainable AI

- SDG13 Climate action
- SDG16 Peace and justice
- SDG9 Gender equality
- ….

SDG13 – Climate action

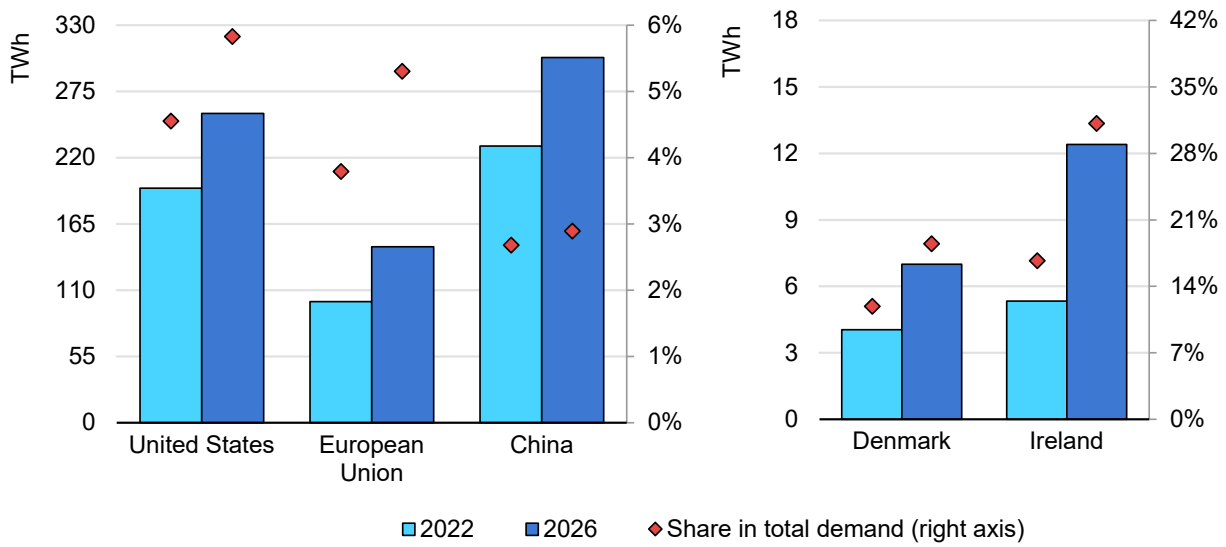


Indicators

- Emission reduction
  - -> CO_2 reduction (CO_2 equivalent)

Energy consumption and data trends

Estimated data centre electricity consumption and its share in total electricity demand in selected regions in 2022 and 2026



Source: IEA. International Energy Agency Website: www.iea.org
Report Electricity 2024

Energy consumption of data centers (US)

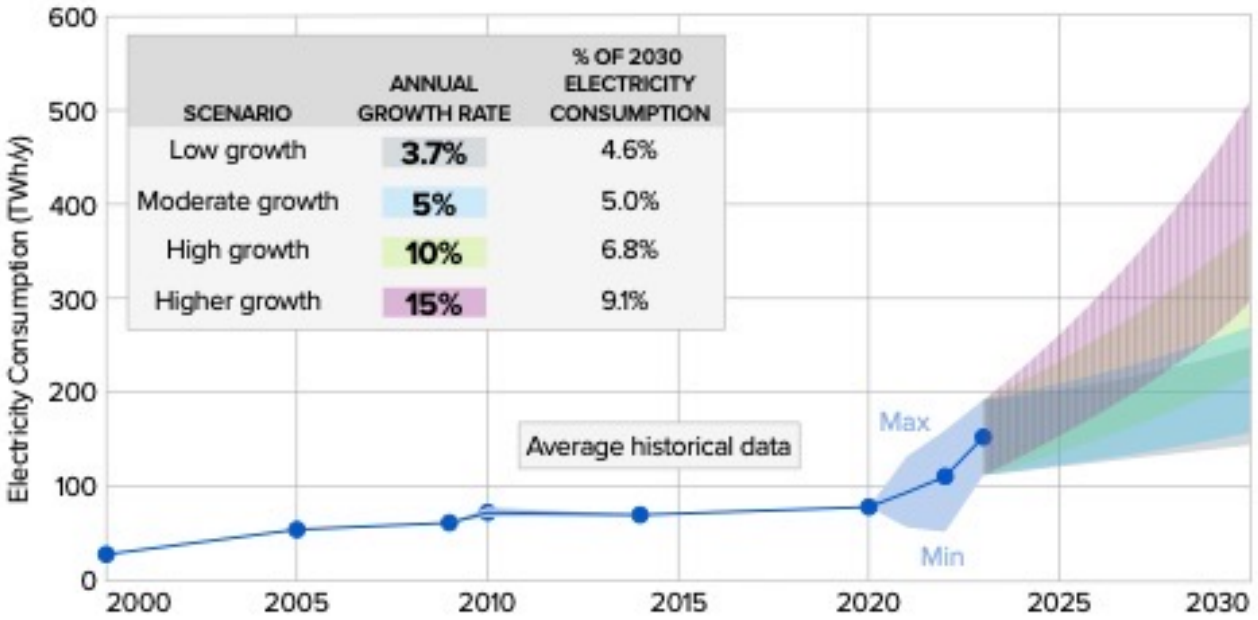


Figure ES-1. *Projections of potential electricity consumption by U.S. data centers: 2023–2030 . % of 2030 electricity consumption projections assume that all other (non-data center) load increases at 1% annually.*

EPRI Electric power research Institute White paper Powering intelligence 2024

Sustainability in Computing

- Efficiency vs sustainability
  - Efficiency: less time, fewer human resources, less storage
  - Energy efficiency:
    - Green IT vs IT for Green
    - outcome/total resources
    - Resources: humans, technical: GPU / CPU, RAM
    - Green IT / Green AI / Sustainable AI

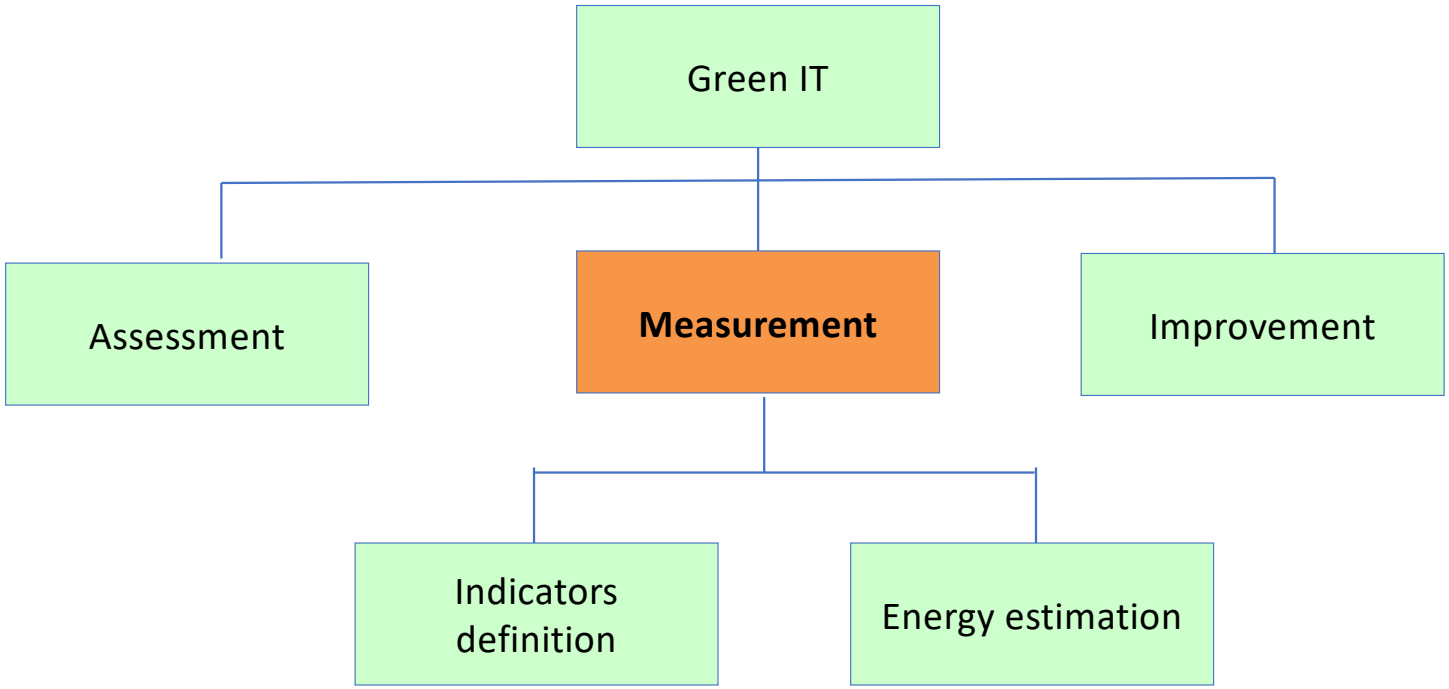
- Sustainability
  - Focus on the life cycle
  - Circular economy reference model

Electricitymaps

Green IT and measuring indicators

03

Measuring indicators



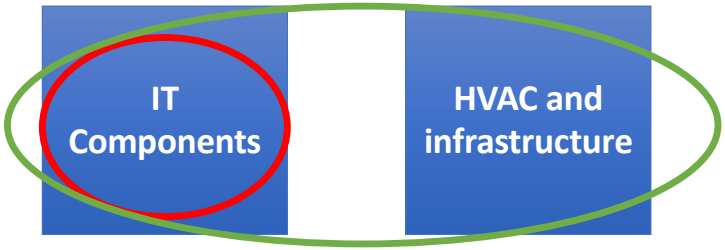
Vitali-Pernici IJCIS 2014

Common Green IT indicators

- PUE
  - Efficient use of energy
- Energy mix
- Resource usage
  - Proxy: Execution time

PUE

- PUE
- **Power Usage Effectiveness**
- The Green Grid, 2007



PUE= Total Facility Power / IT Equipment Power

- how efficiently the electricity is used from the data center control volume to the IT Equipment
- Average around 1.8, best ones (Google, Facebook) claim approx 1.06

Source: M. Zhang, 2024
https://dgtlinfra.com/pue-power-usage-effectiveness/

Energy estimation – energy waste

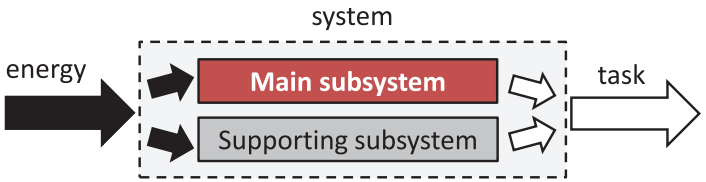


Fig. 2. A system and (sub)systems.

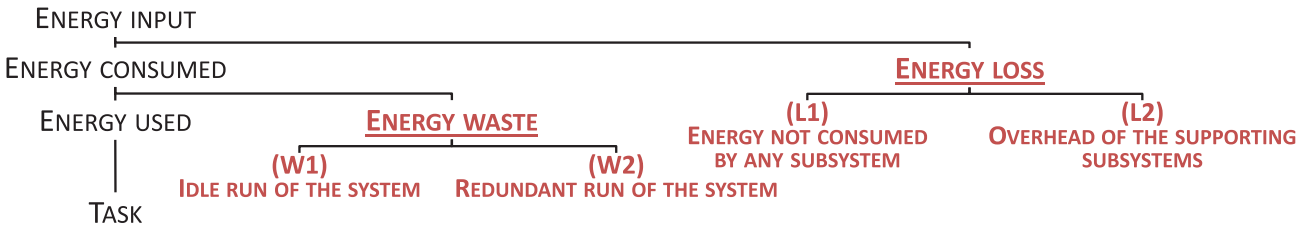
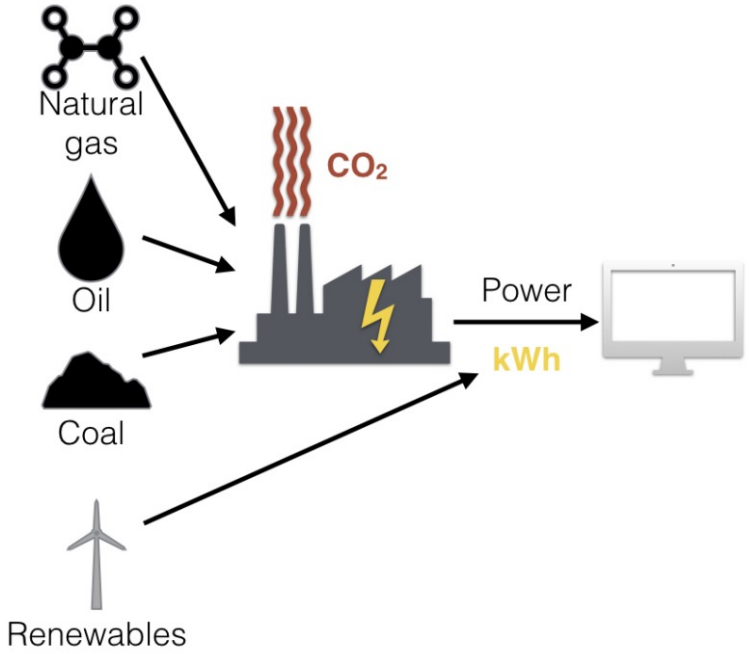


Fig. 3. Critical points within a system where energy is *lost* or *wasted*.

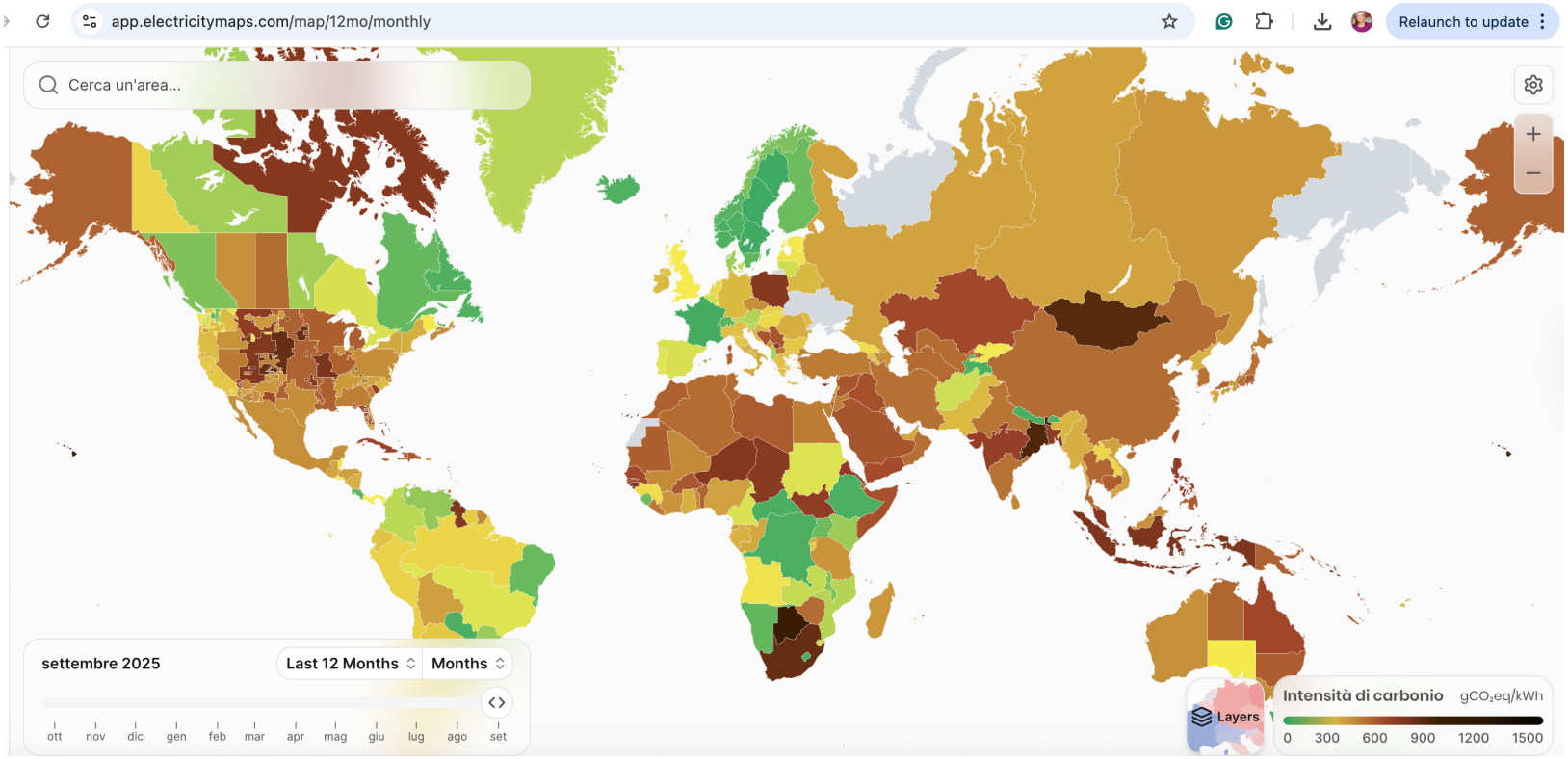
Toni Mastelic, Ariel Oleksiak, Holger Claussen, Ivona Brandic, Jean-Marc Pierson, et al.. Cloud computing: survey on energy efficiency. ACM Computing Surveys, Association for Computing Machinery, 2015, Vol. 47 (n 2), pp. 1-36

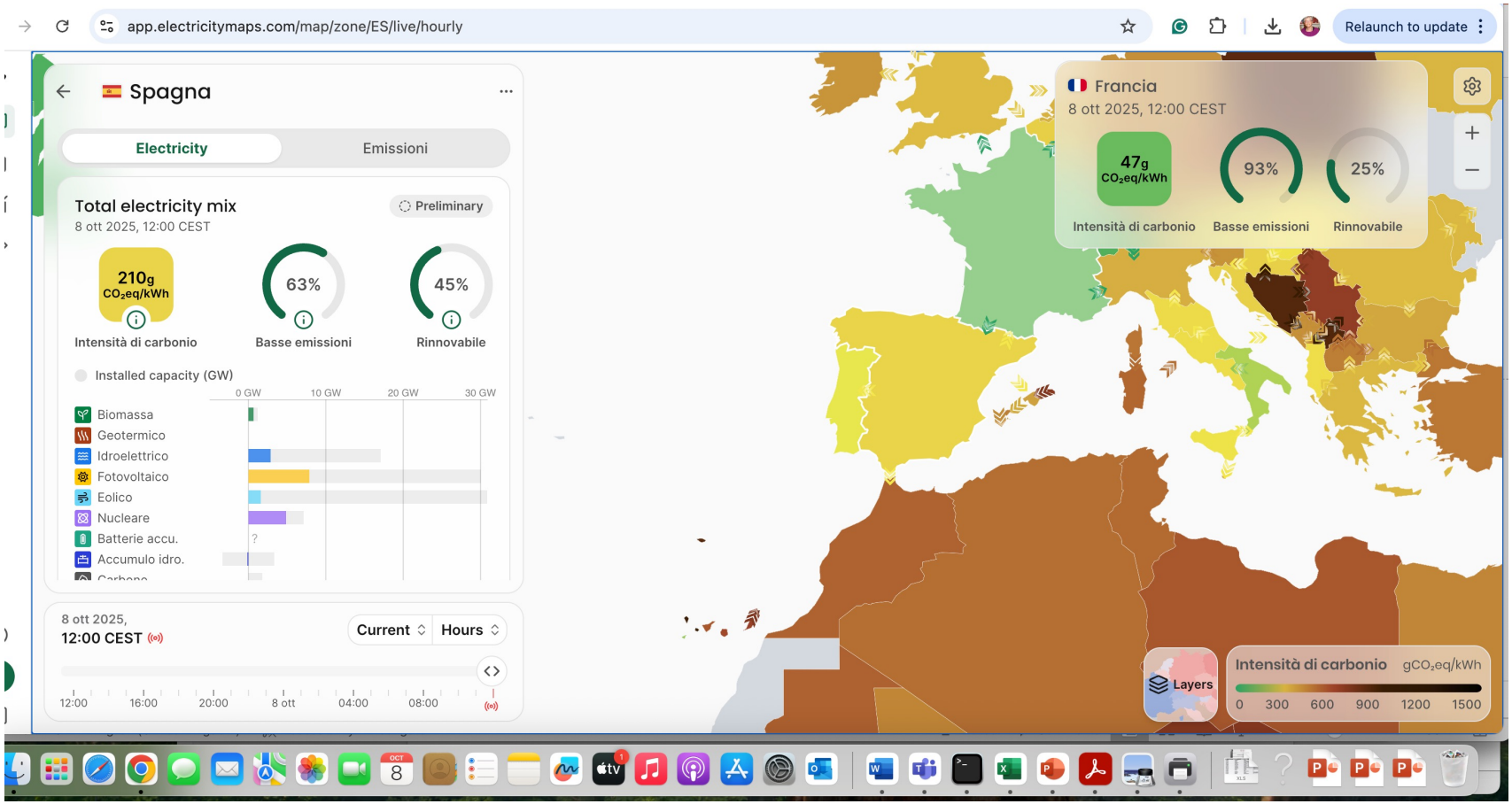
Energy mix



Lottick et al., Energy Usage Reports: Environmental awareness as part of algorithmic accountability Workshop on Tackling Climate Change with Machine Learning @NeurIPS 2019

Carbon intensity – Electricity maps




app.electricitymaps.com/map/zone/ES/live/hourly
Relaunch to update
Spagna
Electricity
Emissioni
Total electricity mix
Preliminary
8 ott 2025, 12:00 CEST
210g CO₂eq/kWh
Intensità di carbonio
63%
Basse emissioni
45%
Rinnovabile
Installed capacity (GW)
0 GW
10 GW
20 GW
30 GW
Biomassa
Geotermico
Idroelettrico
Fotovoltaico
Eolico
Nucleare
Batterie accu.
Accumulo idro.
8 ott 2025,
12:00 CEST
Current
Hours
12:00
16:00
20:00
8 ott
04:00
08:00
Francia
8 ott 2025, 12:00 CEST
47g CO₂eq/kWh
Intensità di carbonio
93%
Basse emissioni
25%
Rinnovabile
Layers
Intensità di carbonio gCO₂eq/kWh
0
300
600
900
1200
1500

Green AI and data preparation

04

Green AI

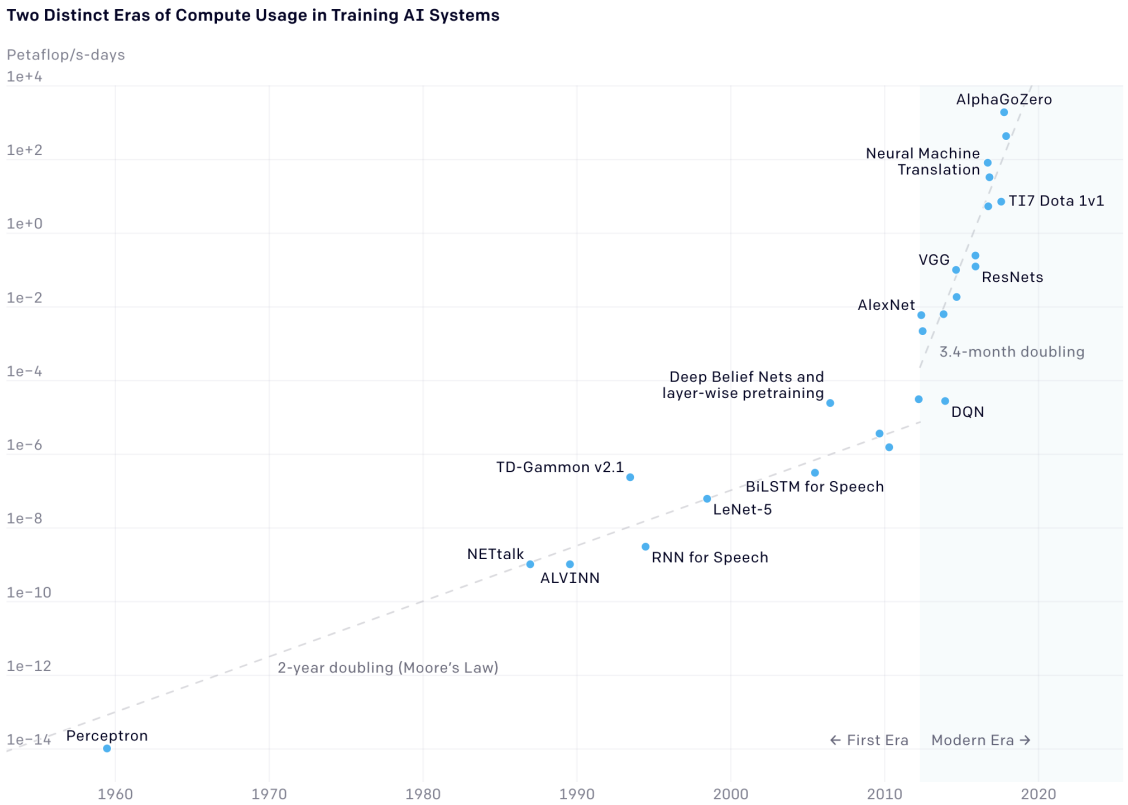
- Phases
  - Data collection and preparation
  - Training
  - Monitoring
  - Inference

- Incorporating the concept of cost into ML algorithms
  - Report accuracy as a function of computation budget

- Data-centric approaches (efficient data usage)

- *Schwartz, Roy, et al. "Green AI" Communications of the ACM 63.12 (2020): 54-63*

Green AI - training

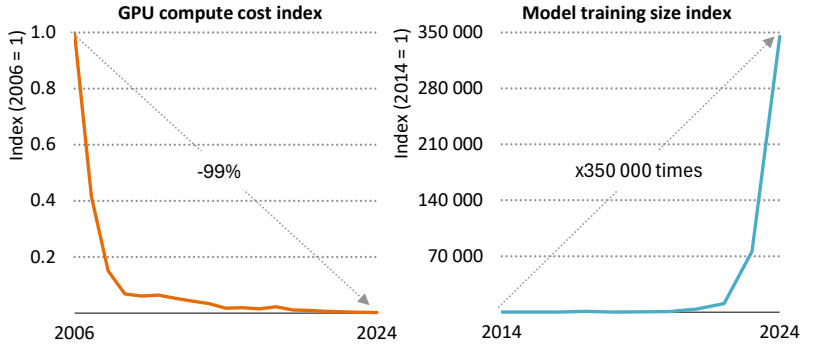


Two phases
Doubling every 3.4 mo

Amodei, D., Hernandez, D., Sastry, G., Clark, J., Brockman, G., & Sutskever, I. (2018). AI and Compute. https://openai.com/index/ai-and-compute/

AI model training

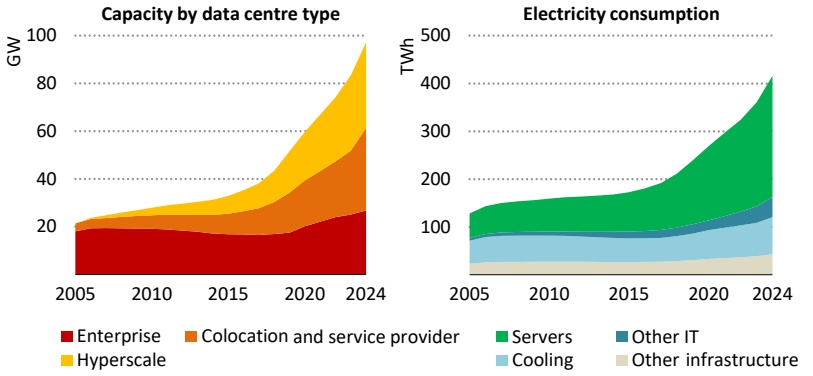
Figure 1.1 ▷ GPU computation cost, 2006-2024, and notable AI model computational training size, 2014-2024



In the past decade, cheaper computing, exponentially more data and research breakthroughs in model design have turbocharged AI model capabilities

Sources: IEA analysis based on data from EpochAI (2024), and Coyle and Hampton (2024).

Figure 2.3 ▷ Total data centre electricity consumption by equipment type and data centre type, 2005-2024



After a decade of limited growth, data centre electricity consumption began to accelerate again after 2015

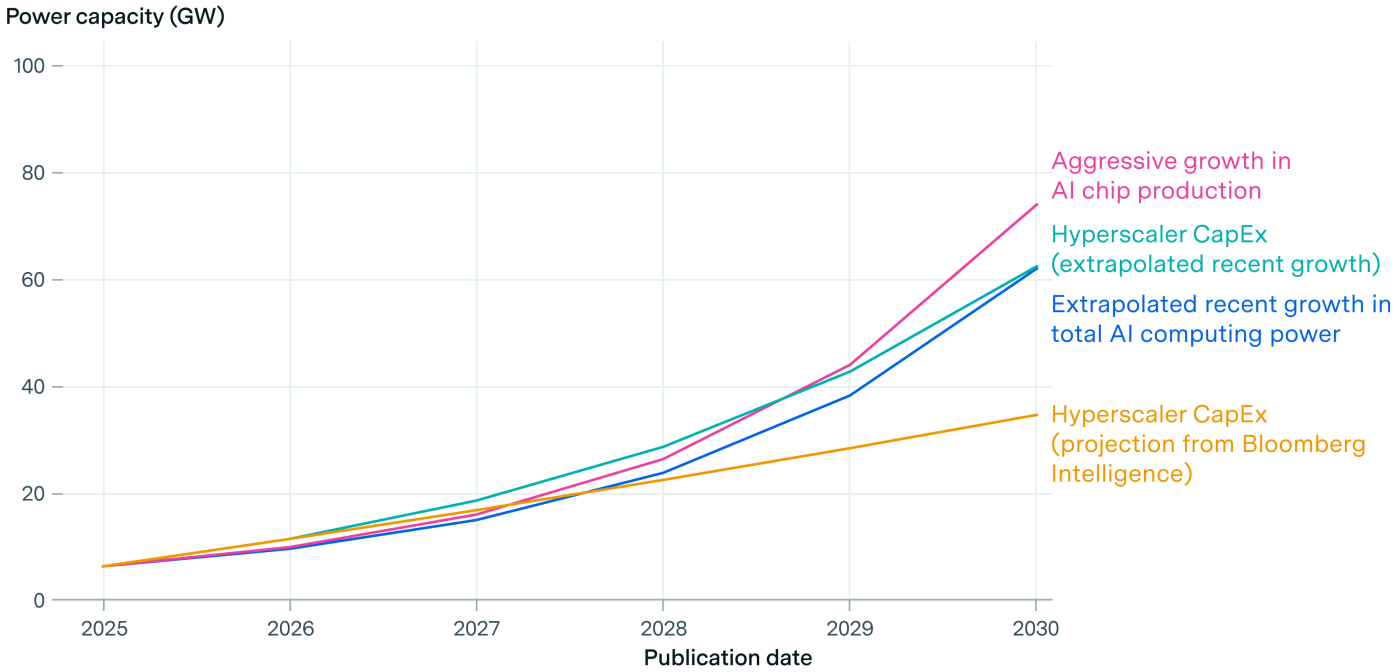
Note: GW = gigawatt; TWh = terawatt hour.

Sources: IEA analysis based on data from IDC (2024a), OMDIA (2025), and SemiAnalysis (2025).

https://www.iea.org/reports/energy-and-ai/energy-demand-from-ai

Modeling AI energy consumption trends

Forecasted total capacity of U.S. AI data centers



EPRI, Epoch AI Report 2025
Scaling Intelligence: The Exponential Growth of AI's Power Needs

Some factors

***4x** per year training compute growth / **1.26x** per year training duration / **1.4x** per year efficiency growth = **2.27x** per year power growth*

EPRI, Epoch AI Report 2025
Scaling Intelligence: The Exponential Growth of AI's Power Needs

Salehi, Shirin, and Anke Schmeink. "Data-Centric Green Artificial Intelligence: A Survey." *IEEE Transactions on Artificial Intelligence* 5.05 (2024): 1973-1989.

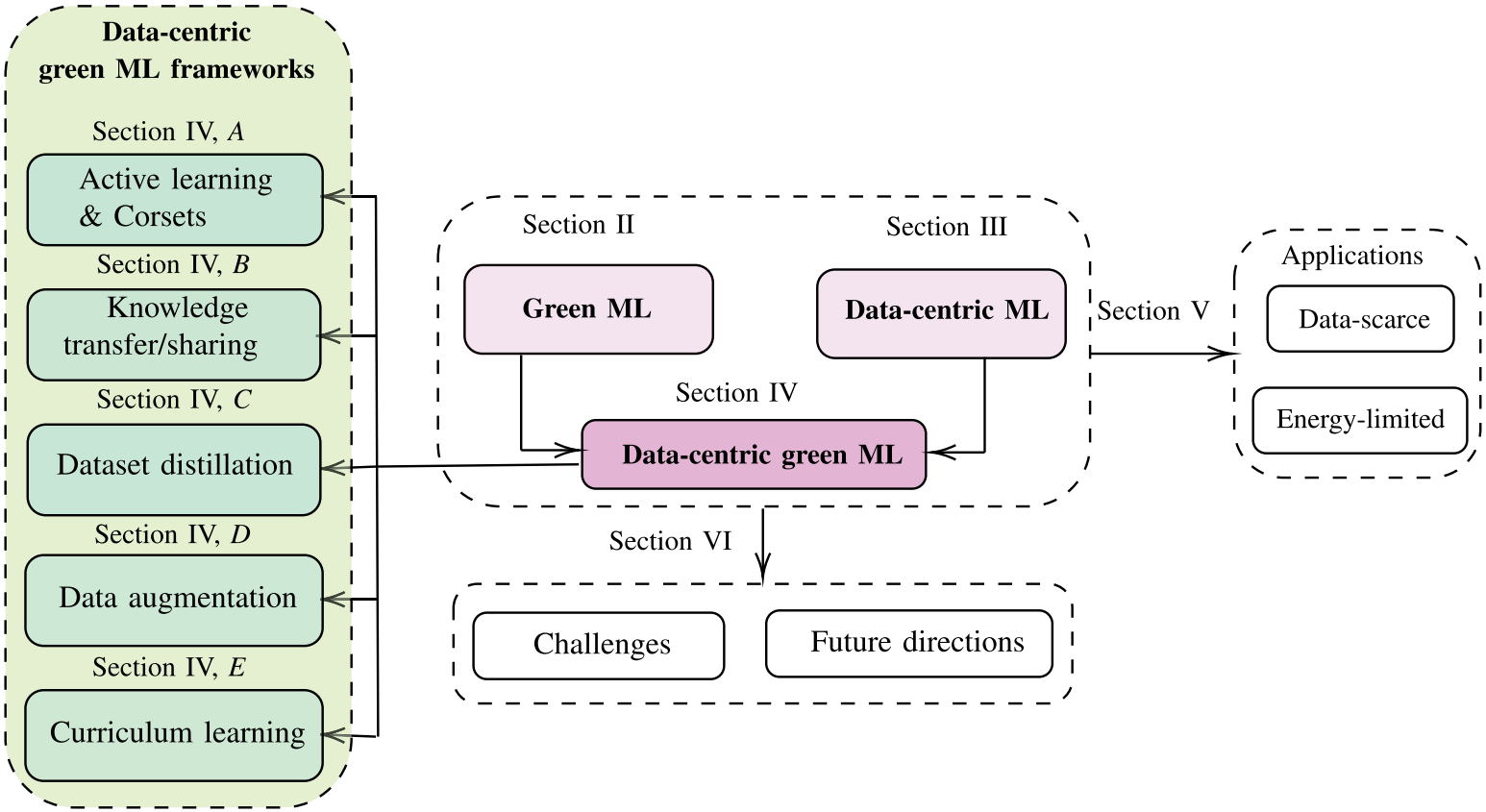
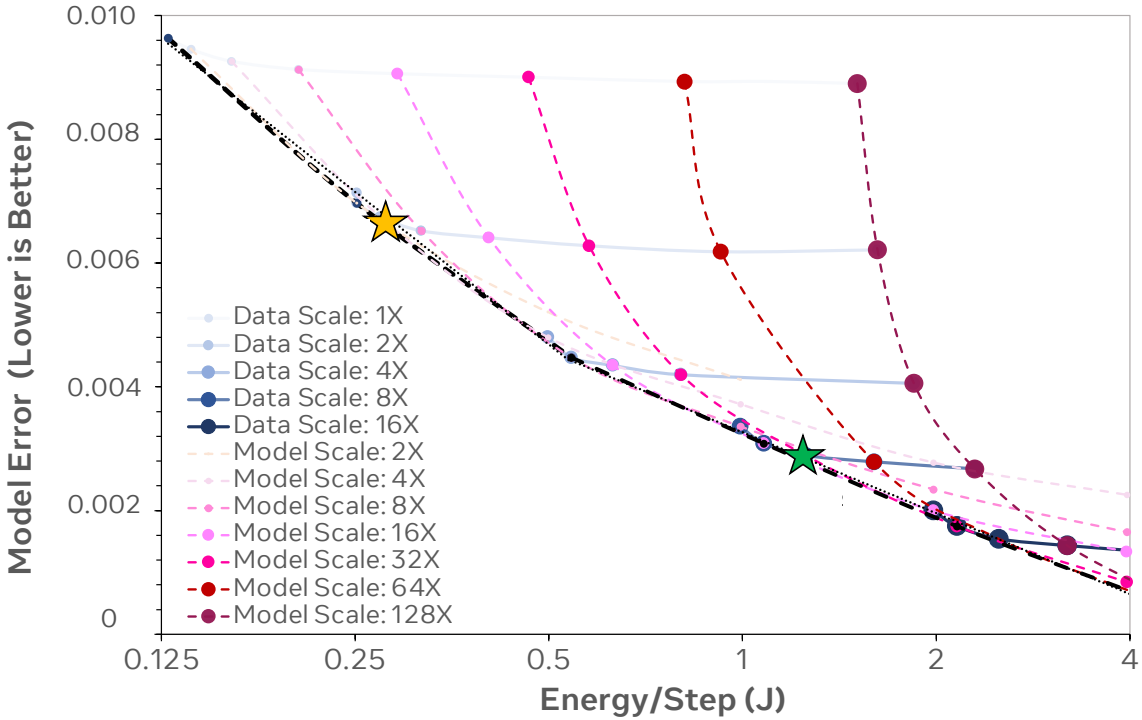


Fig. 1. Schematic structure of this article and the relationship between the adjacent sections.

Model and data scaling in Recommendation systems



Wu et al., MLSys, 2022

However, when designed well, ***data scaling, sampling and selection strategies*** *can improve the competitive analysis for ML algorithms, reducing the environmental footprint of the process*

Environmental equity

- Limited power-grid capacity
- Impact on other resources (e.g. water)

- Types of **flexibility for AI workloads**:

- Spatial
- Temporal
- Performance (tradeoffs between accuracy and resource consumption)

Hajiesmaili, et al., 2025

ICT eco-feedback

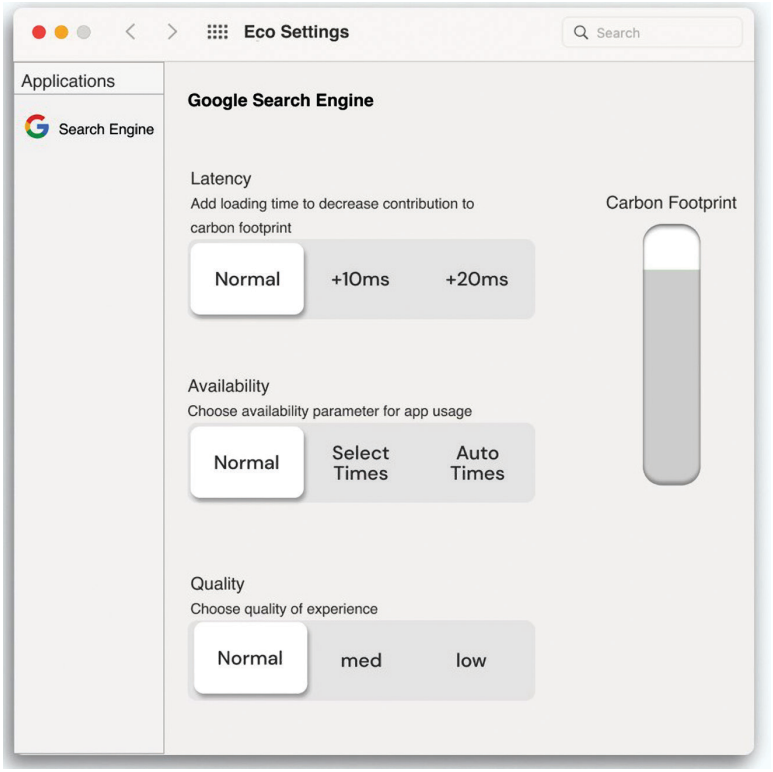
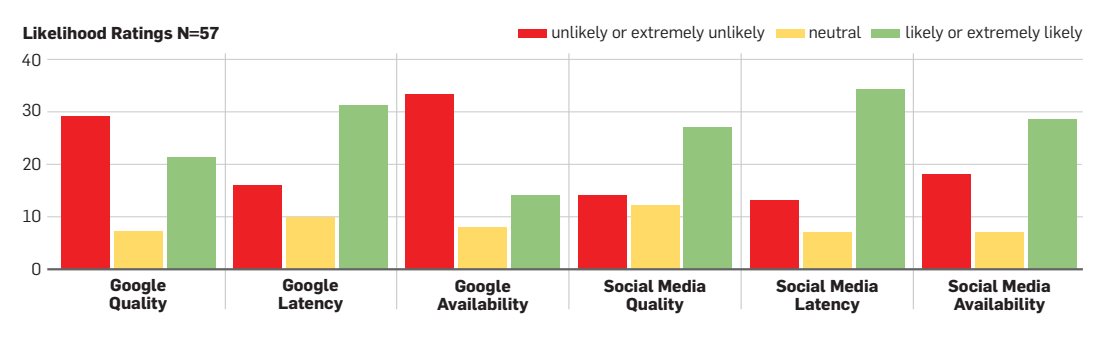


Table 1. Dimensions of user experience.

Dimension	Description	Example
Quality	A measure of the condition of data based on factors such as accuracy, completeness, consistency, reliability, and timeliness.	"You may be required to attempt multiple searches or explore multiple pages of results to find the best result for your query."
Latency	The time it takes for a data packet to travel from one designated point to another.	"Results take a few seconds longer to load than normal."
Availability	The percentage of time that the infrastructure, system, or solution remains operational under normal circumstances.	"Google is only available for set time periods each day. How likely is it that you would choose this option?"

Figure 3. User sentiments: summary of Likert survey items.



Young, Sydney, Udit Gupta, and Josiah Hester. "Empowering Users to Make Sustainability-Forward Decisions for Computing Services." Communications of the ACM 68.7 (2025)

Sustainability in Data preparation

05

Data preparation

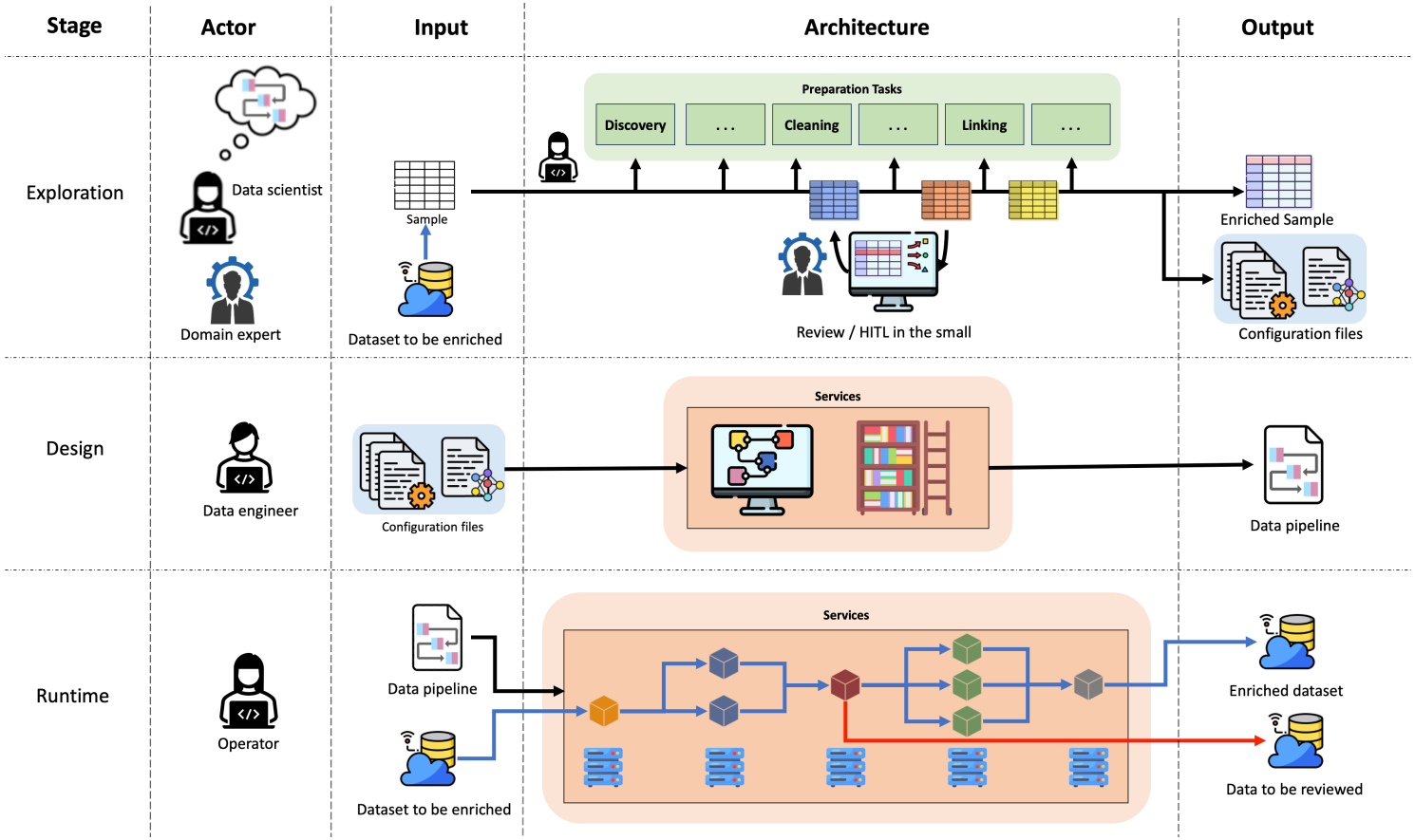
Context and focus:

- Data ecosystems
- Data preparation (also for ML models training)
- Data quality

Sustainability of the data preparation process (“discount” data quality)

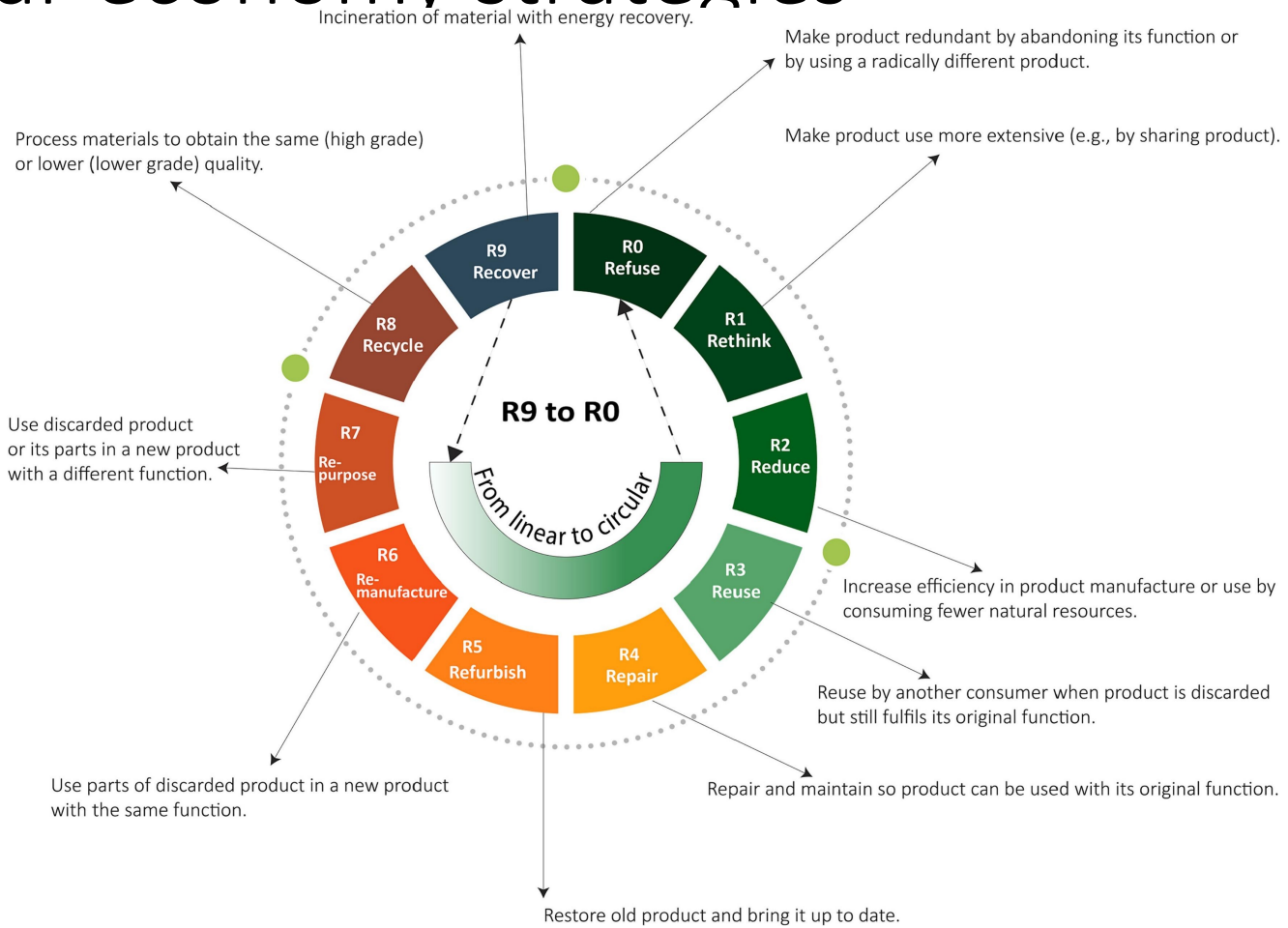
- reducing the computational effort needed to analyze data
- introducing HITL in a sustainable way, to make human contributions effective, keeping them limited in time and size

Data preparation Architecture and phases



Discount Quality project team, SEBD 2024

Circular economy strategies



Muñoz et al., 2024

Strategies for data preparation
Design phase

Smarter product use and manufacture

R1 Rethink	Make product use more intensive via new concepts or business models.
R2 Reduce	Increase efficiency by consuming fewer resources in manufacture or use.

Data spaces

Models to perform multiple tasks - LLMs

On demand
Selecting relevant tables
Designing efficient pipelines

Strategies for data preparation
Consumption phase

Extend the lifespan of the product and its parts

R3 Reuse	Reuse by another consumer of a discarded product that still fulfills its original function.	Consider context Enrich metadata
R4 Repair	Repair and maintenance of a defective product so it can be used with its original function.	Inspection and pipeline design Considering additional sources
R5 Refurbish	Restore an old product and bring it up to date.	Extend lifecycle of old datasets

Evaluation framework

Data quality evaluation

DQ dimensions:

- ***Conventional DQ*** *(accuracy, completeness, consistency, timeliness, uniqueness, …)*
- ***Data-centric-AI DQ***
  - *Bias, dimensionality, coverage/density, overlap, …*
- ***Human interaction DQ***
  - *Granularity, latency, user control, mental effort, success rate of tasks, …*

During data preparation quality can

improve

decrease

Evaluation framework (2/2)

Cost evaluation

Considering both computational and human resources

- *Assessment cost*
- *Improvement cost*

Waste evaluation

- *Based on the percentage of data being prepared that will actually be used and*
- *The success rate of the technique in solving a given quality issue in a given context.*

Strategies for data preparation
Design phase

Smarter product use and manufacture

R1 Rethink	Make product use more intensive via new concepts or business models.	Data spaces Models to perform multiple tasks - LLMs	DQImprovement Cost
R2 Reduce	Increase efficiency by consuming fewer resources in manufacture or use.	On demand Selecting relevant tables Designing efficient pipelines	DQImprovement Cost DQWaste

Strategies for data preparation
Consumption phase

Extend the lifespan of the product and its parts

R3 Reuse	Reuse by another consumer of a discarded product that still fulfills its original function.	Consider context Enrich metadata	DQImprovement Cost
R4 Repair	Repair and maintenance of a defective product so it can be used with its original function.	Inspection and pipeline design Considering additional sources	DQImprovement and Assessment Cost DQWaste
R5 Refurbish	Restore an old product and bring it up to date.	Extend lifecycle of old datasets	DQImprovement Cost

Challenges, strategies, and methods

- **Pipeline** design
- **On demand** data preparation

Focus on reduce and reuse

- Progressive **visualization**
- **Data enrichment**

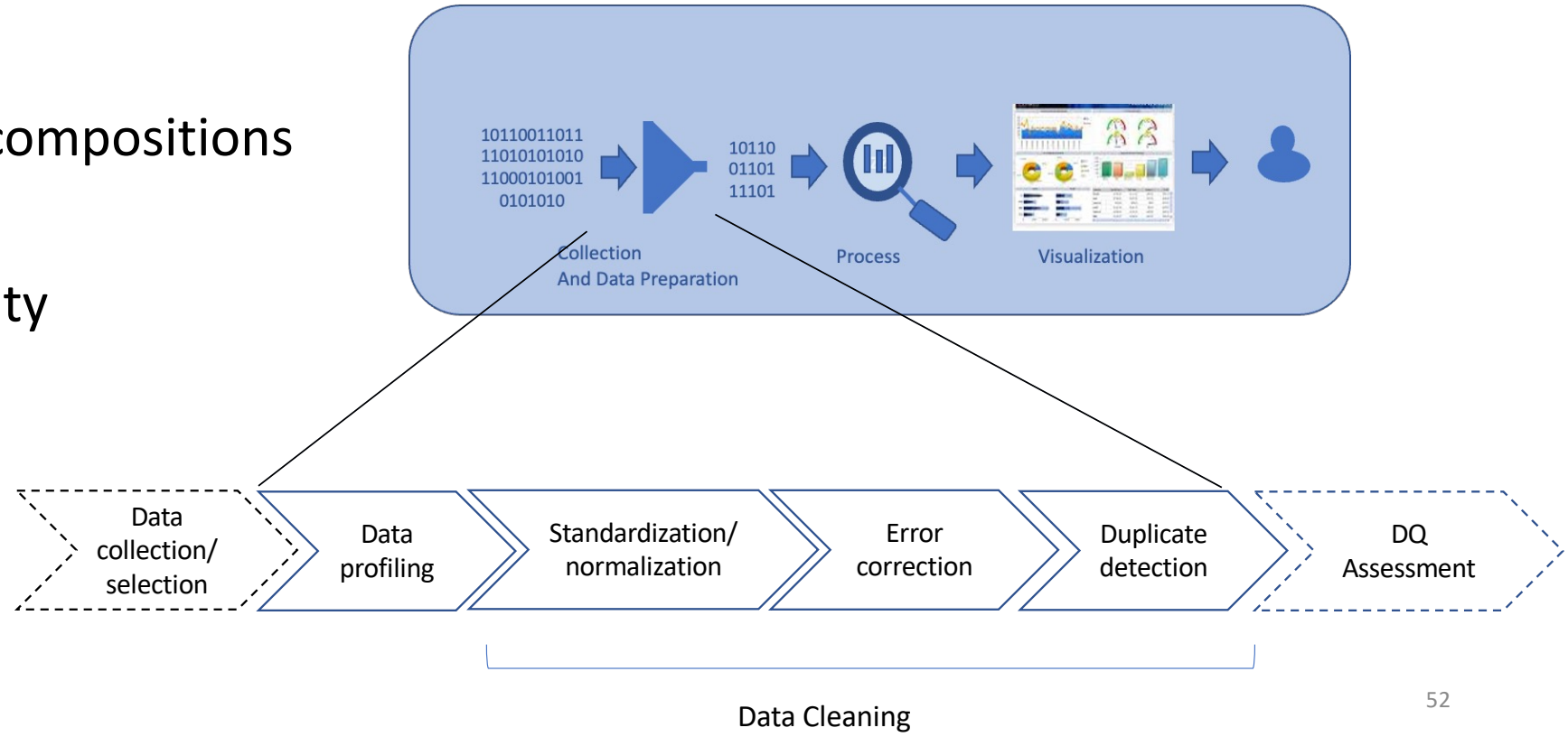
B. Pernici et al., JDIQ, 2025

Pipeline design

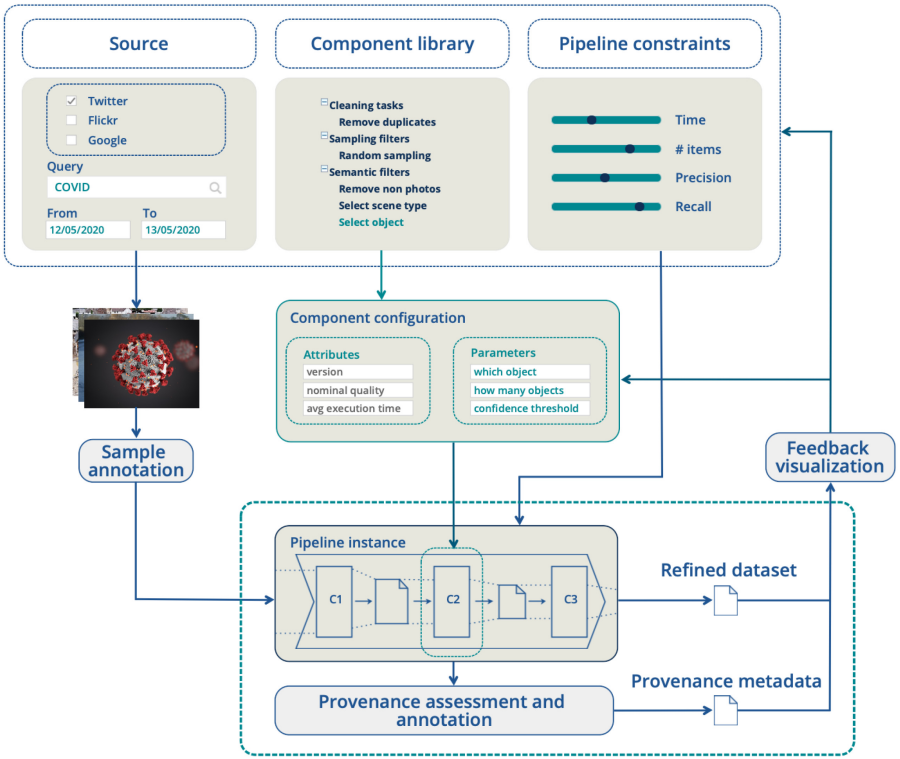
5.1

Pipelines for data-driven decision making

- Goal: reduce needed resources
- Design compositions
- Adaptivity



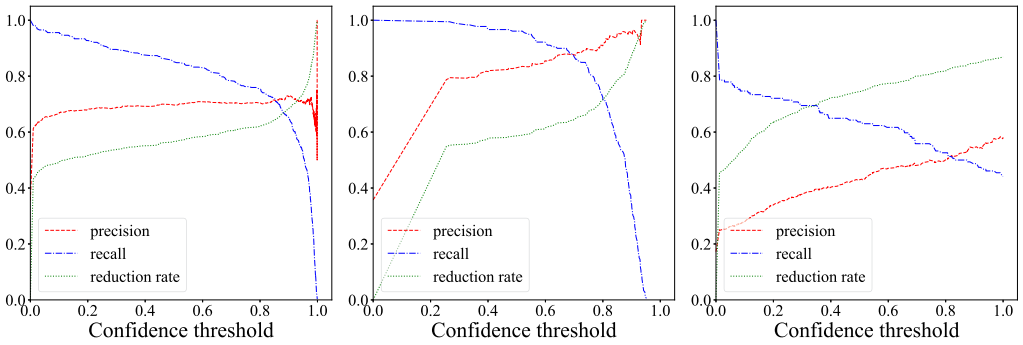
Reducing needed resources by design



Filtering pipelines

Table 2. Component Execution Time Statistics, Estimated on 10 Runs, for One Item

Measure	*Photo*	*TwoPersons*	*PublicPlace*
Mean	45 ms	24 ms	37 ms
StdDev	2 ms	3 ms	2 ms



(a) Photo (b) TwoPersons (c) PublicPlace

Fig. 5. Precision, recall, and reduction rate responses to confidence threshold.

C. Bono et al, Pipeline Design for Data Preparation for Social Media Analysis, IJDIQ, 2024

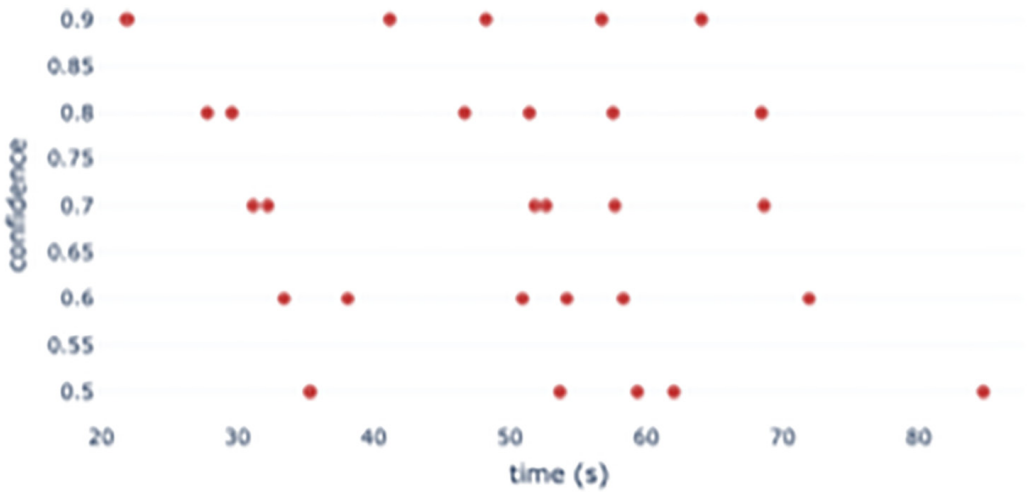


Fig. 7. Execution times for the same pipeline while varying component ordering. Horizontal lines represent different confidence thresholds.

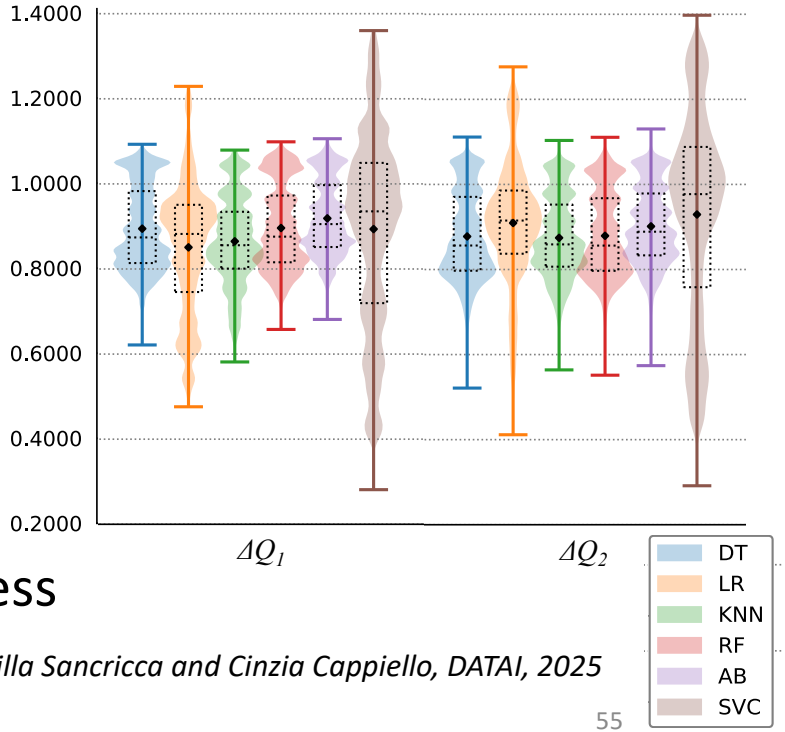
best configuration
[C2: TwoPersons, C1: Photo, C3: PublicPlace]

worst
[C3: PublicPlace, C1: Photo, C2: TwoPersons]

C. Bono et al, Pipeline Design for Data Preparation for Social Media Analysis, IJDIQ, 2024

Lightweight pipelines

- "Good Enough is Sometimes Better"
- Improve only the **first ranked DQ dimension**
  - **Based on a knowledge base on DQ dimensions and improvement techniques**
- Tests vs improving all dimensions
- Assessing the impact on accuracy
- Considered dimensions: accuracy and completeness



Camilla Sancricca and Cinzia Cappiello, DATAI, 2025

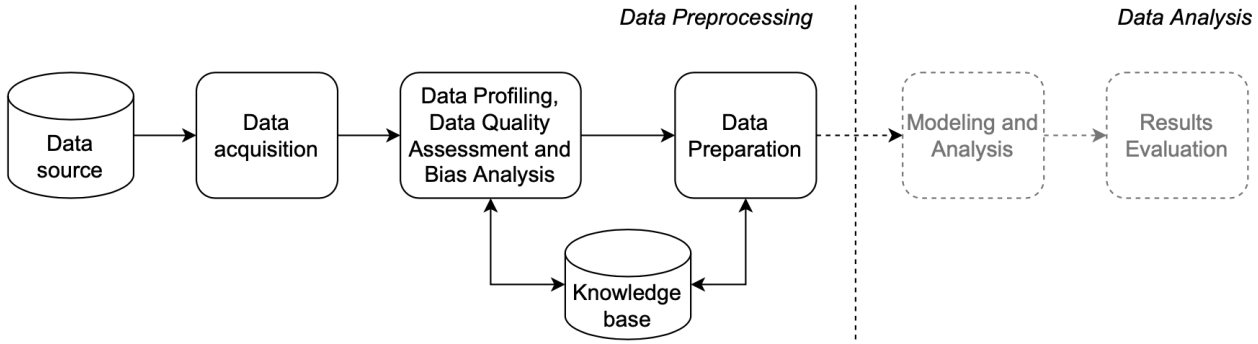
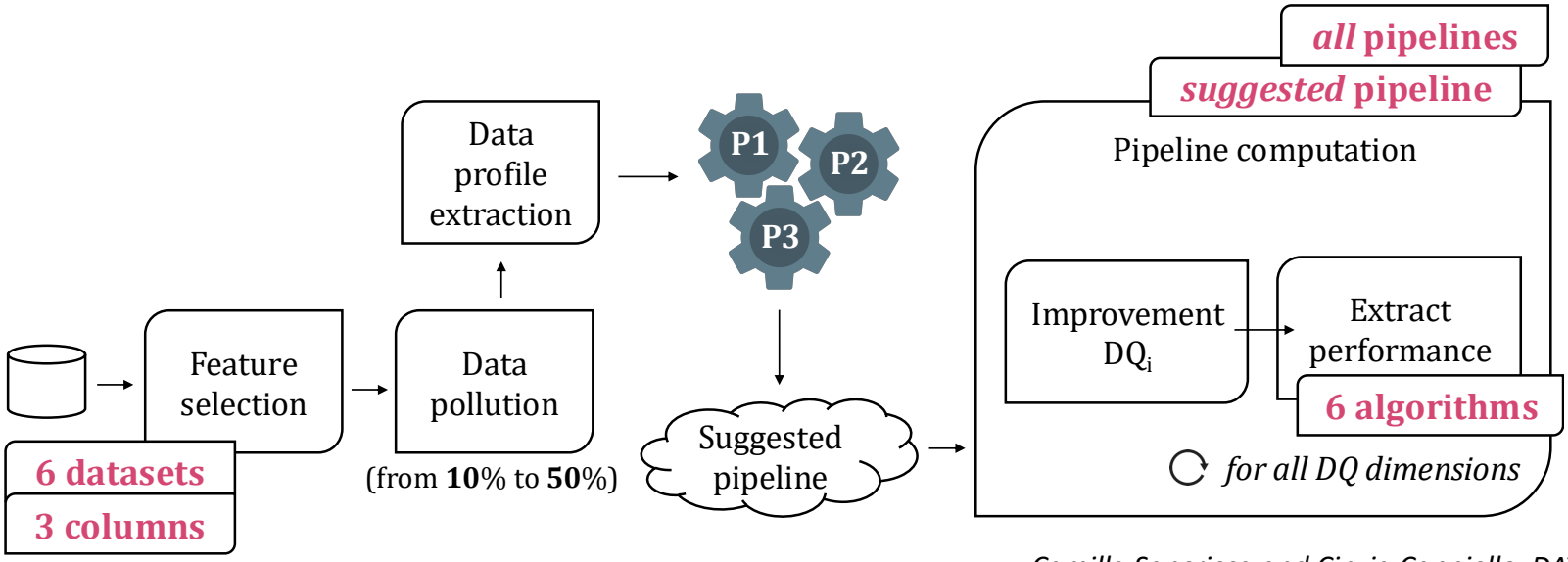


Figure 1: The Data Preparation framework: the high level architecture



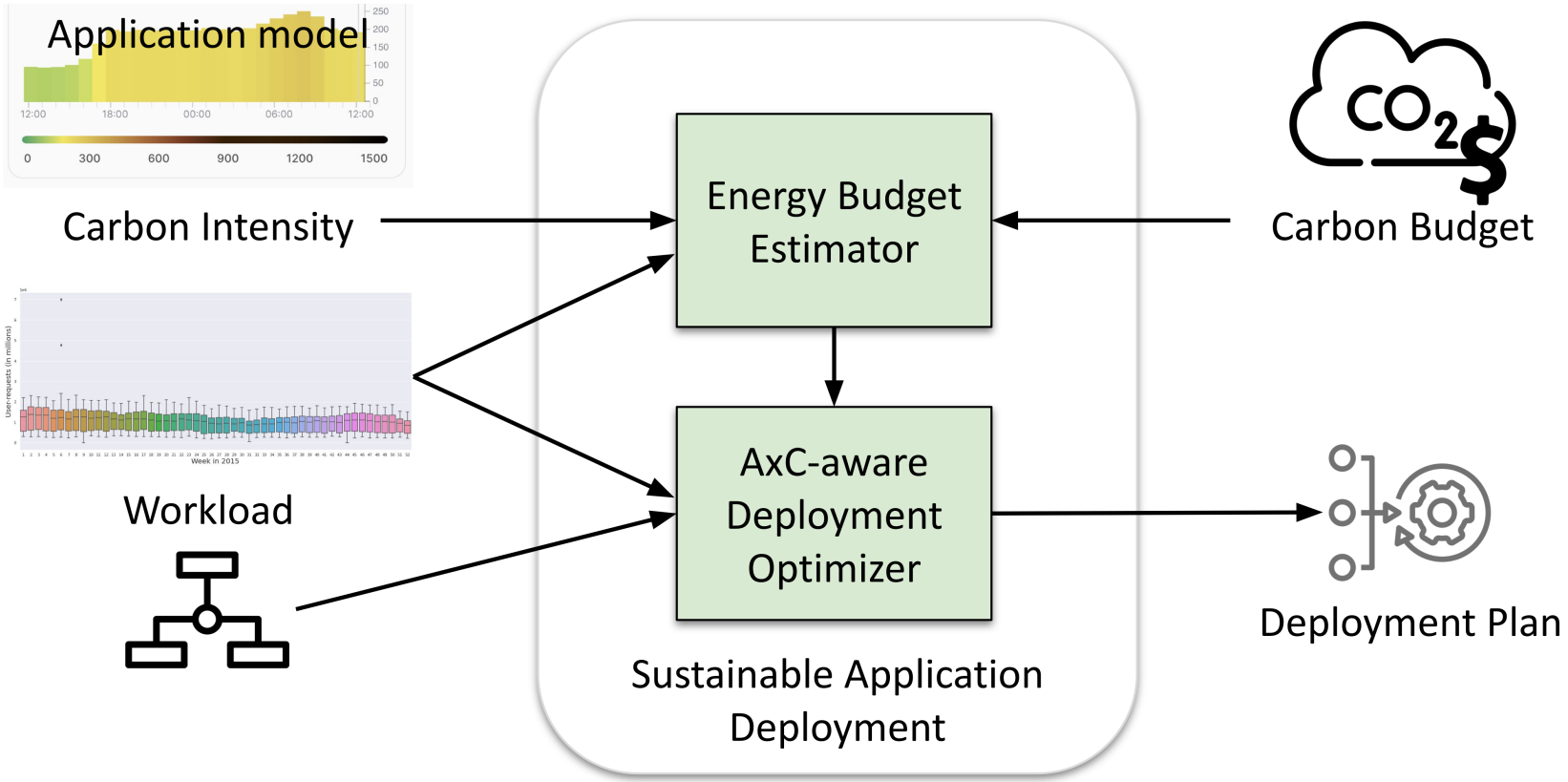
Camilla Sancricca and Cinzia Cappiello, DATAI, 2025

Considering a limited budget for workflows

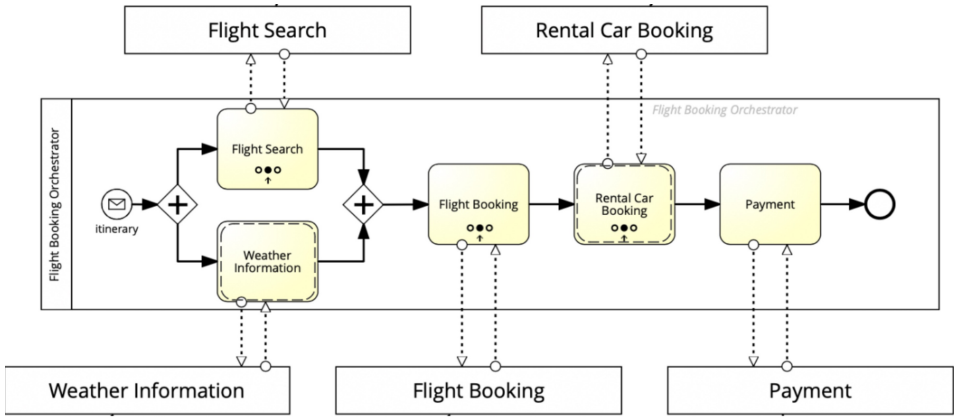
- Organization need to minimize the environment impact of their operations
- Approximation techniques: dynamically adjust workflows
- Balance
  - Sustainability
  - User experience
  - Revenue

Vitali et al, FGCS, 2025

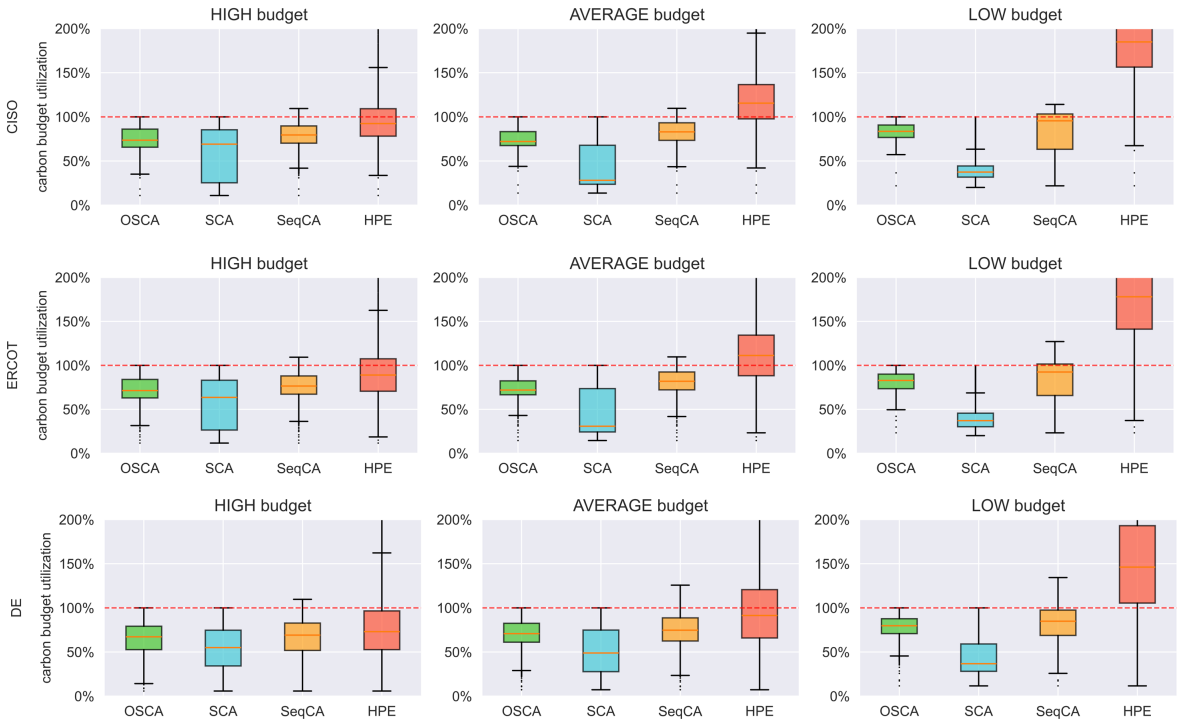
Adaptive workflows



Vitali et al, FGCS, 2025



Component	Versions	Mandatory?
Flight Search	Low Power Normal High Performance	Mandatory
Weather Information	Off Normal	Optional
Flight Booking	Low Power Normal	Mandatory
Rental Car Booking	Off Normal High Performance	Optional
Payment	Normal	Mandatory



Strategies

- Optimised Selection Carbon-Aware algorithm (OSCA)
- High Performance (HPE) “do-nothing”
- Simple Carbon-Aware (SCA): Three fixed configurations: High Performance, Normal Performance, and Low Performance. Selected wrt carbon budget compliance
- Sequential Carbon-Aware (SeqCA), adaptive to emissions

Vitali et al, FGCS, 2025

On demand

5.2

Considerations

- 85% of data not used during information access
- Costly data preparation operations

Integrate cleaning operations in queries

Extract clean samples from dirty data

Select most appropriate table to merge in a data lake

…

ON DEMAND

Entity resolution

Entity Resolution (ER) is the task of identifying records that pertain to the same real-world object.

Data Source A

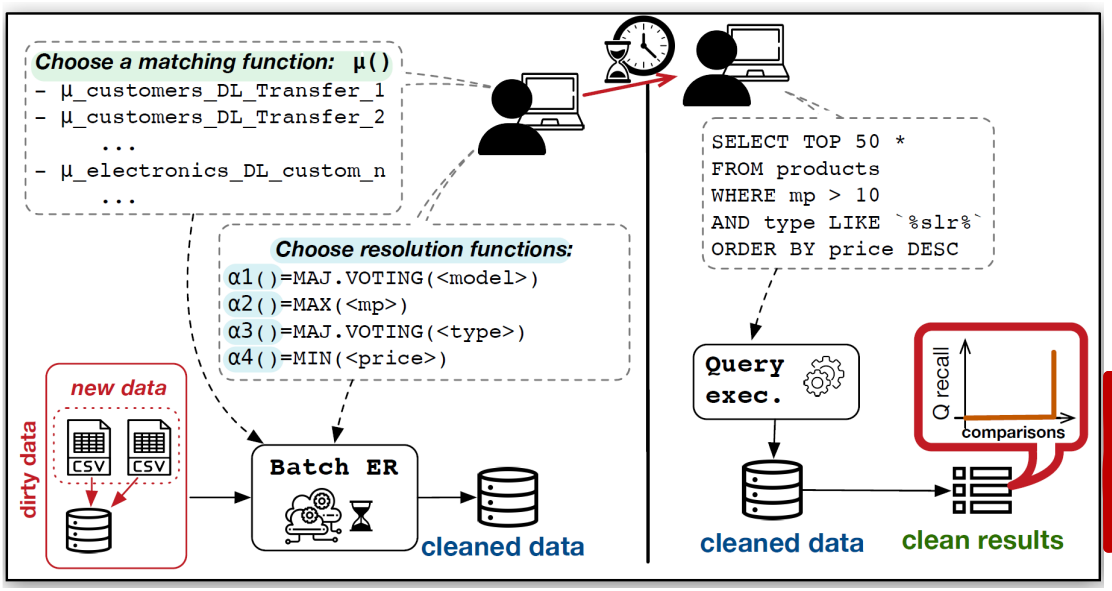
Title	Price	Manifacturer	Description
Macbook	1000	Apple	Laptop, model Air
Pavillion	1500	HP	laptop

Data Source B

Code	Name	Model	Descr
100	HP Pavillion	N-4000	Laptop, weight: 2 kg, Intel i7, RAM 16 Gb, price: 1200€
2001	Apple Macbook	air	The new Apple light weight laptop

DBGroup @ UniMoRe

Towards on demand

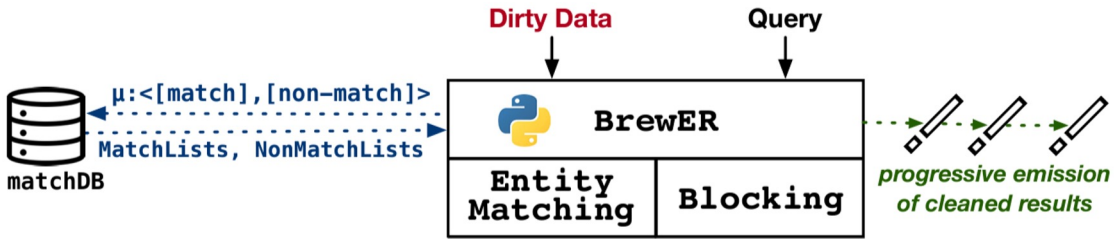


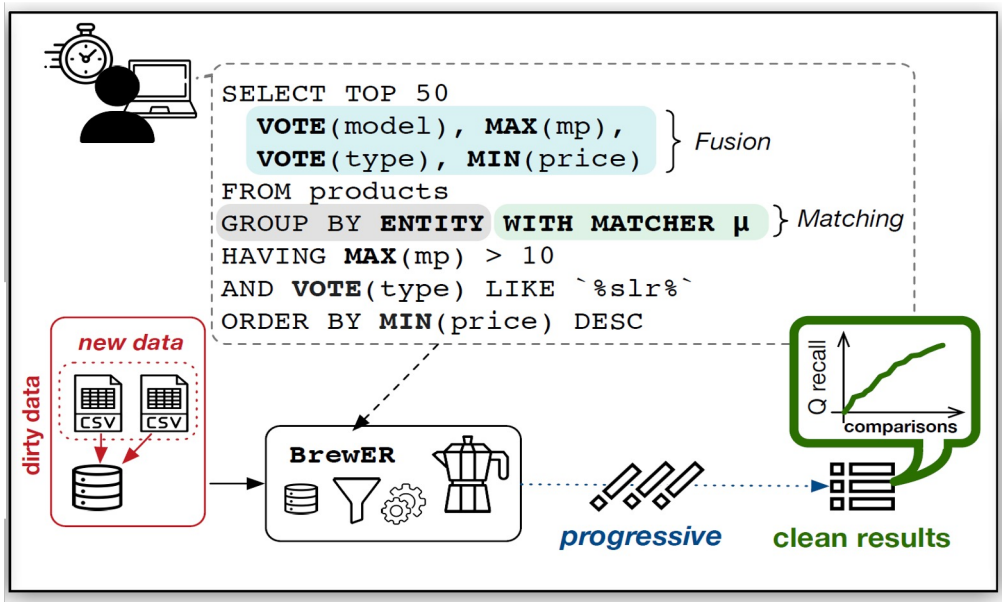
Traditional approach

First correct entity after minutes/hours

On Demand

First correct entity after seconds

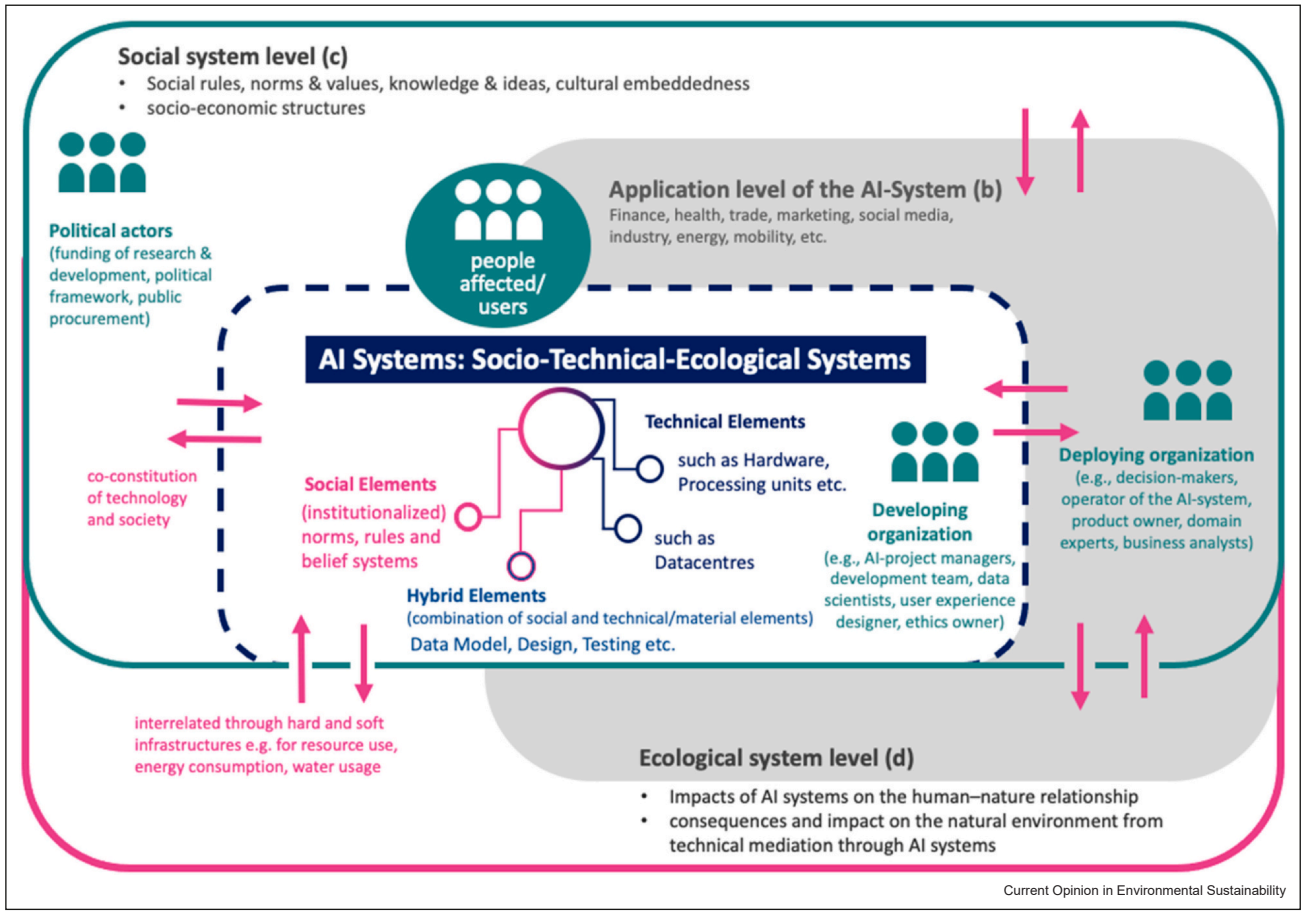




Concluding remarks

- Measuring is complex
- Circular economy strategies provide an analysis framework
- Comparing alternative approaches focusing on specific problems helps
- Much future work is needed
  - Balance time and space requirements and constraints
  - Assess human experience and involvement
  - Define “good enough”…
- Guidelines: EU AI Act encourages Code of Conducts for environmental sustainability (art. 95) “*assessing and minimising the impact of AI systems on environmental sustainability, including as regards energy-efficient programming and techniques for the efficient design, training and use of AI*”

Many different perspectives



AI systems as socio-technical–ecological systems, with impact levels and relevant actors for sustainable AI.

Rohde et al, 2024

THANK YOU!

References

- Barbara Pernici, Cinzia Cappiello, Carlo Bono, Camilla Sancricca, Tiziana Catarci, Marco Angelini, Matteo Filosa, Matteo Palmonari, Flavio De Paoli, Sonia Bergamaschi, Giovanni Simonini, Angelo Mozzillo, Luca Zecchini, Sustainable quality in data preparation, ACM Journal on Data and Information Quality, https://doi.org/10.1145/3769120
- Jakubik J, Vössing M, Kühl N, Walk J, Satzger G. Data-centric artificial intelligence. Business & Information Systems Engineering. 2024 Aug;66(4):507-15
- Communications of the ACM 68.7, special section on Sustainability (2025)
- Hajiesmaili, M., Ren, S., Sitaraman, R., & Wierman, A. (2025). Toward Environmentally Equitable AI. *Communications of the ACM*, *68*(7), 70-73.
- Young, Sydney, Udit Gupta, and Josiah Hester. "Empowering Users to Make Sustainability-Forward Decisions for Computing Services." *Communications of the ACM* 68.7 (2025): 80-85
- Verdecchia, Roberto, et al. "Data-centric green AI an exploratory empirical study." 2022 international conference on ICT for sustainability (ICT4S). IEEE, 2022.
- Santiago Muñoz, M. Reza Hosseini, Robert H. Crawford, Towards a holistic assessment of circular economy strategies: The 9R circularity index, Sustainable Production and Consumption, Volume 47, 2024, Pages 400-412
- Mohammed, S., Ehrlinger, L., Harmouch, H., Naumann, F., Srivastava, D. (2025). The five facets of data quality assessment. *ACM SIGMOD Record*, *54*(2), 18-27
- Patterson, D., Gonzalez, J., Le, Q., Liang, C., Munguia, L. M., Rothchild, D., So, D., Texier, M., Dean, J. (2014). Carbon emissions and large neural network training. arXiv 2021. *arXiv preprint arXiv:2104.10350*.

- Salehi, Shirin, and Anke Schmeink. "Data-Centric Green Artificial Intelligence: A Survey." *IEEE Transactions on Artificial Intelligence* 5.05 (2024): 1973-1989
- Camilla Sancricca and Cinzia Cappiello, Lightweight Pipelines: Good Enough is Sometimes Better, DATAI: 2nd International Workshop on Data-Centric AI, VLDB Workshop, 2025
- Vitali M, Wiesner P, Kreutz K, Gandola R. Adaptive green cloud applications: Balancing emissions, revenue, and user experience through approximate computing. Future Generation Computer Systems. 2025 Sep 25:108143.
- Luca Zecchini, Ziawasch Abedjan, Vasilis Efthymiou, and Giovanni Simonini. RadlER: Deduplicated Sampling On-Demand. PVLDB, 18(12):5319 – 5322, 2025
- Luca Zecchini, Giovanni Simonini, Sonia Bergamaschi, Felix Naumann. BrewER: Entity Resolution On-Demand. Proc. VLDB Endow. 16(12): 4026-4029 (2023)
- WU, Carole-Jean, et al. Sustainable AI: Environmental implications, challenges and opportunities. Proceedings of machine learning and systems, 2022, 4: 795-813.
- Rohde F, Wagner J, Meyer A, Reinhard P, Voss M, Petschow U, Mollen A. Broadening the perspective for sustainable artificial intelligence: sustainability criteria and indicators for Artificial Intelligence systems. Current Opinion in Environmental Sustainability. 2024 Feb 1;66:101411.