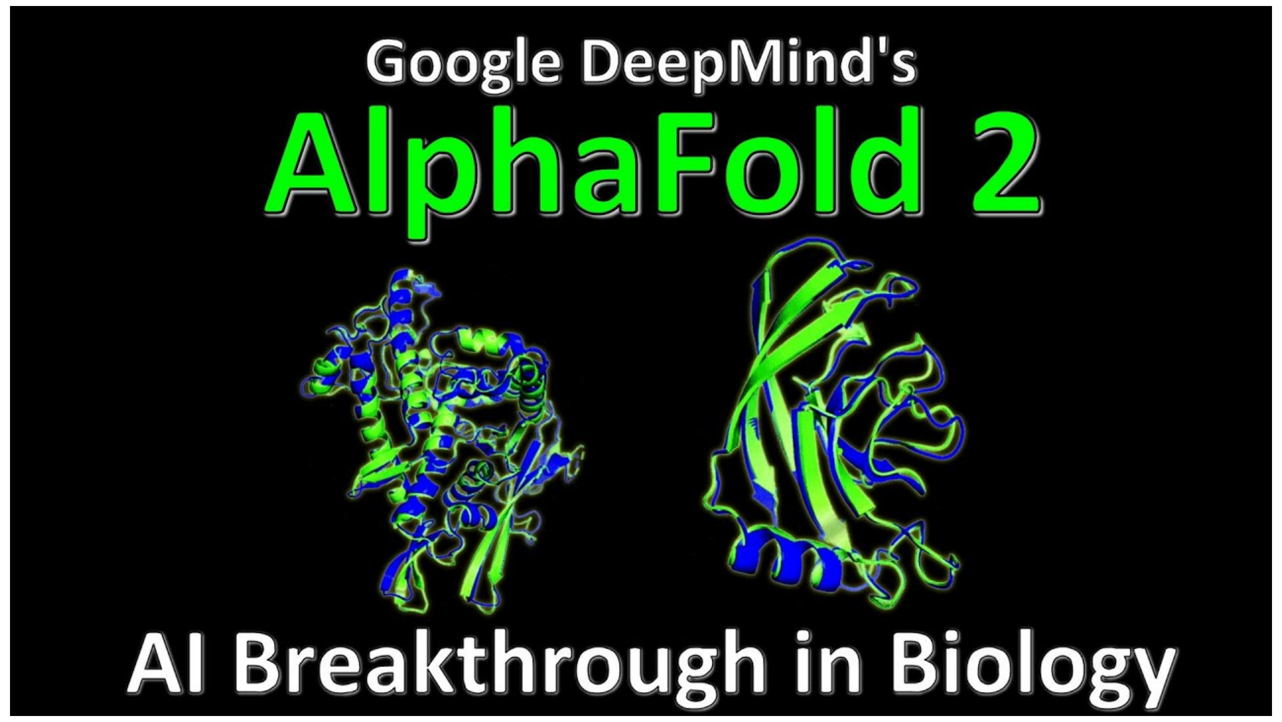
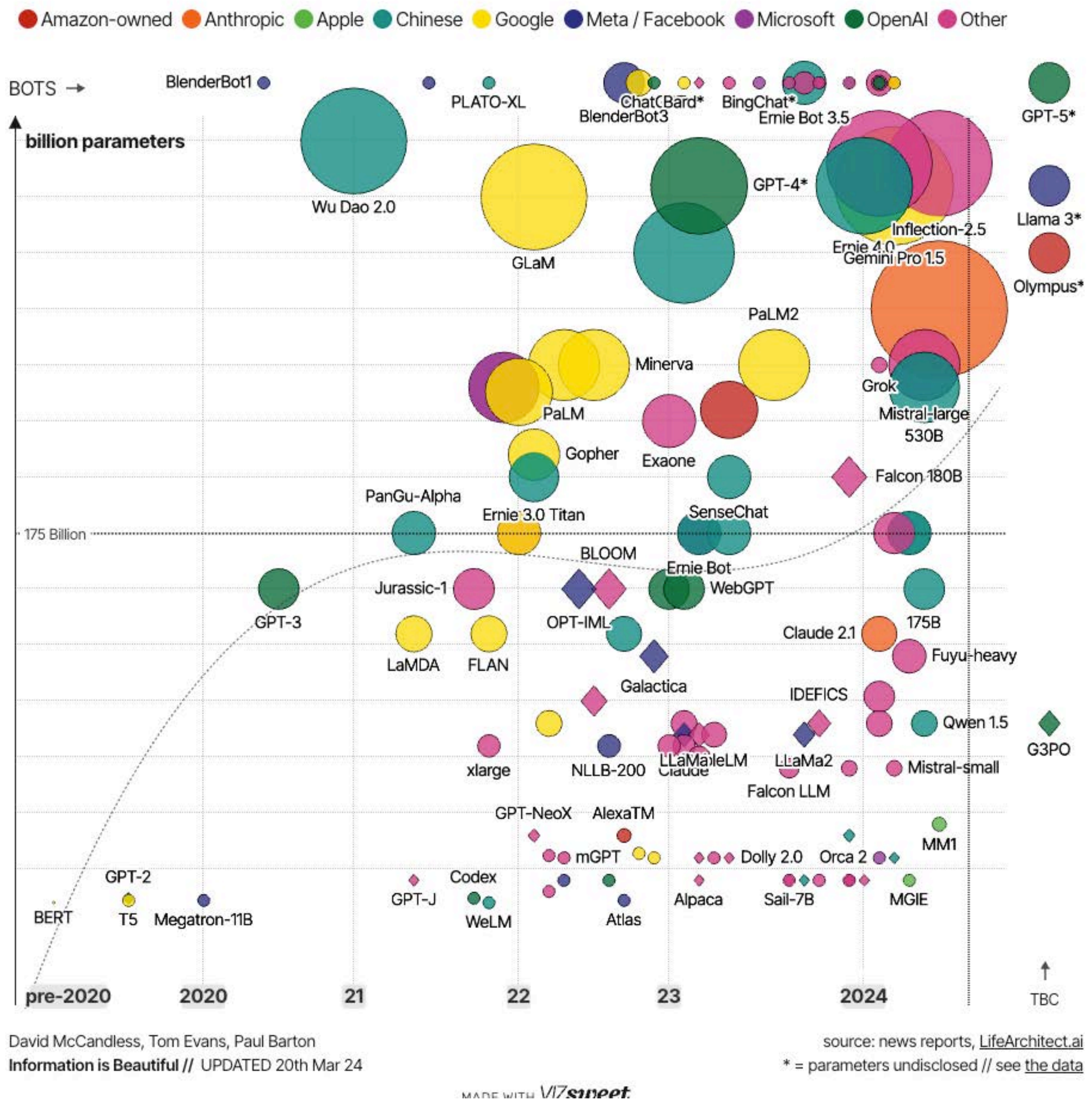
Responsible AI Computing

Trustworthiness, Sustainability, and Equity

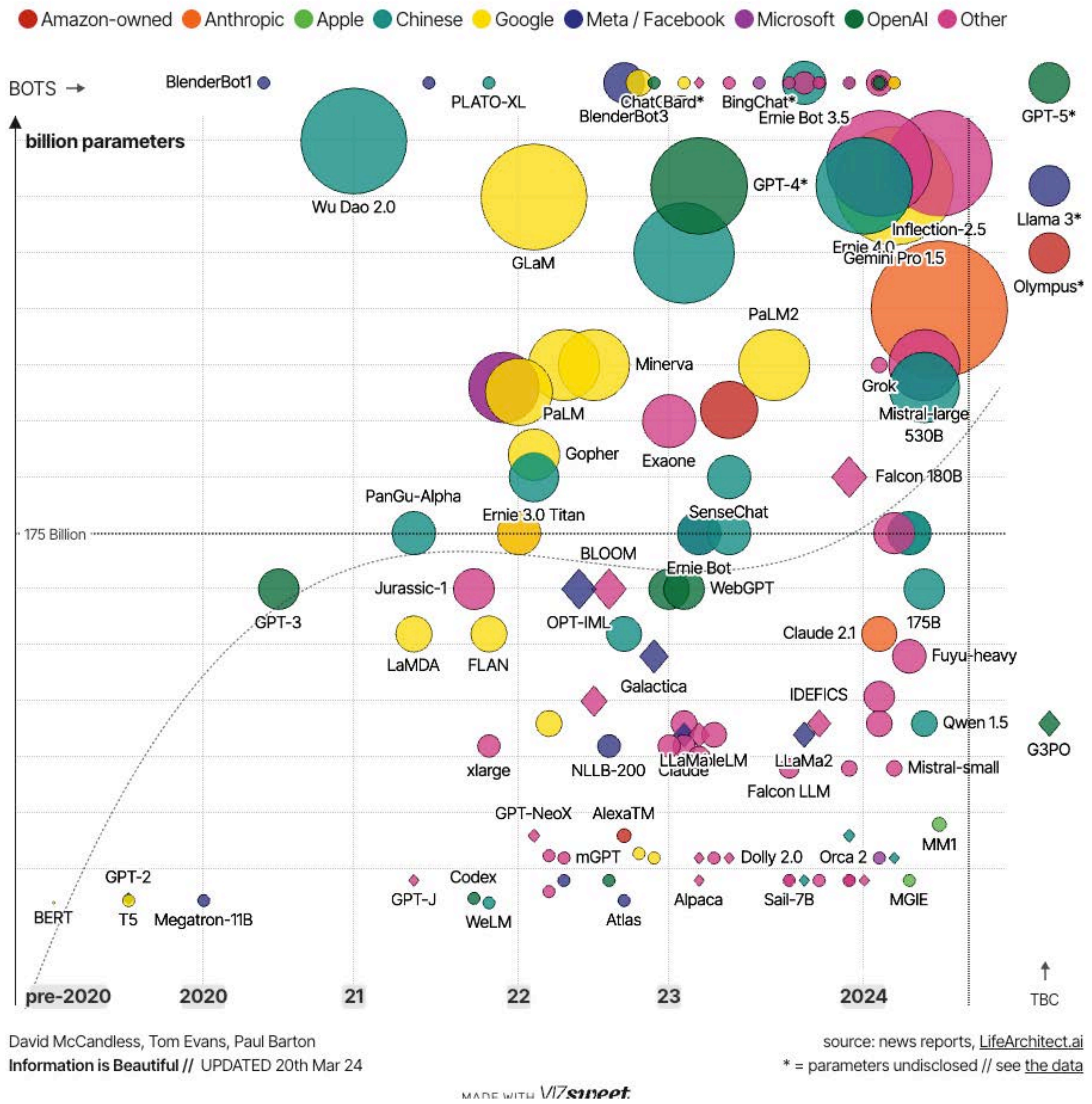
AI is booming…



The scaling law of AI

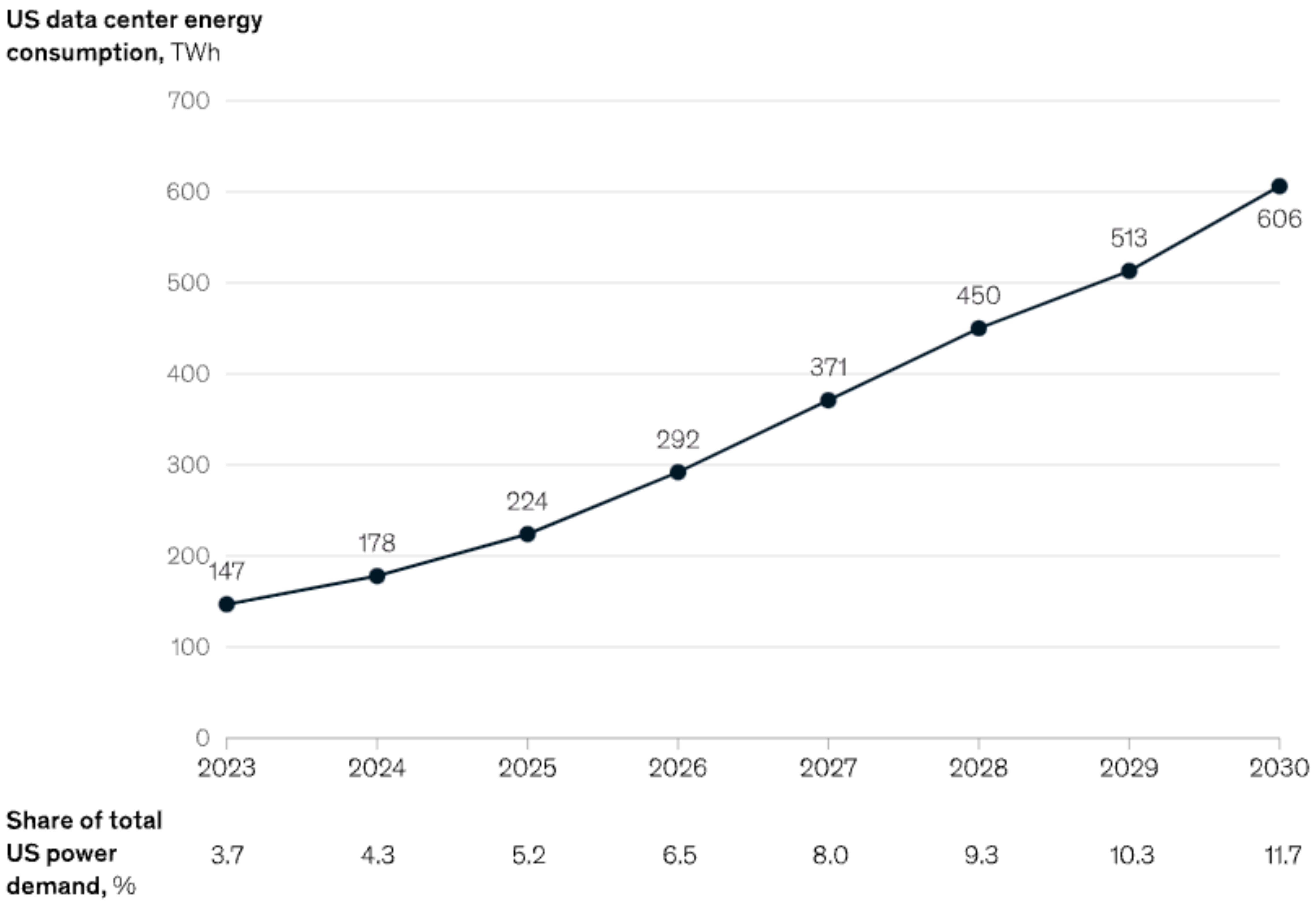


The scaling law of AI



AI's growing appetite for energy

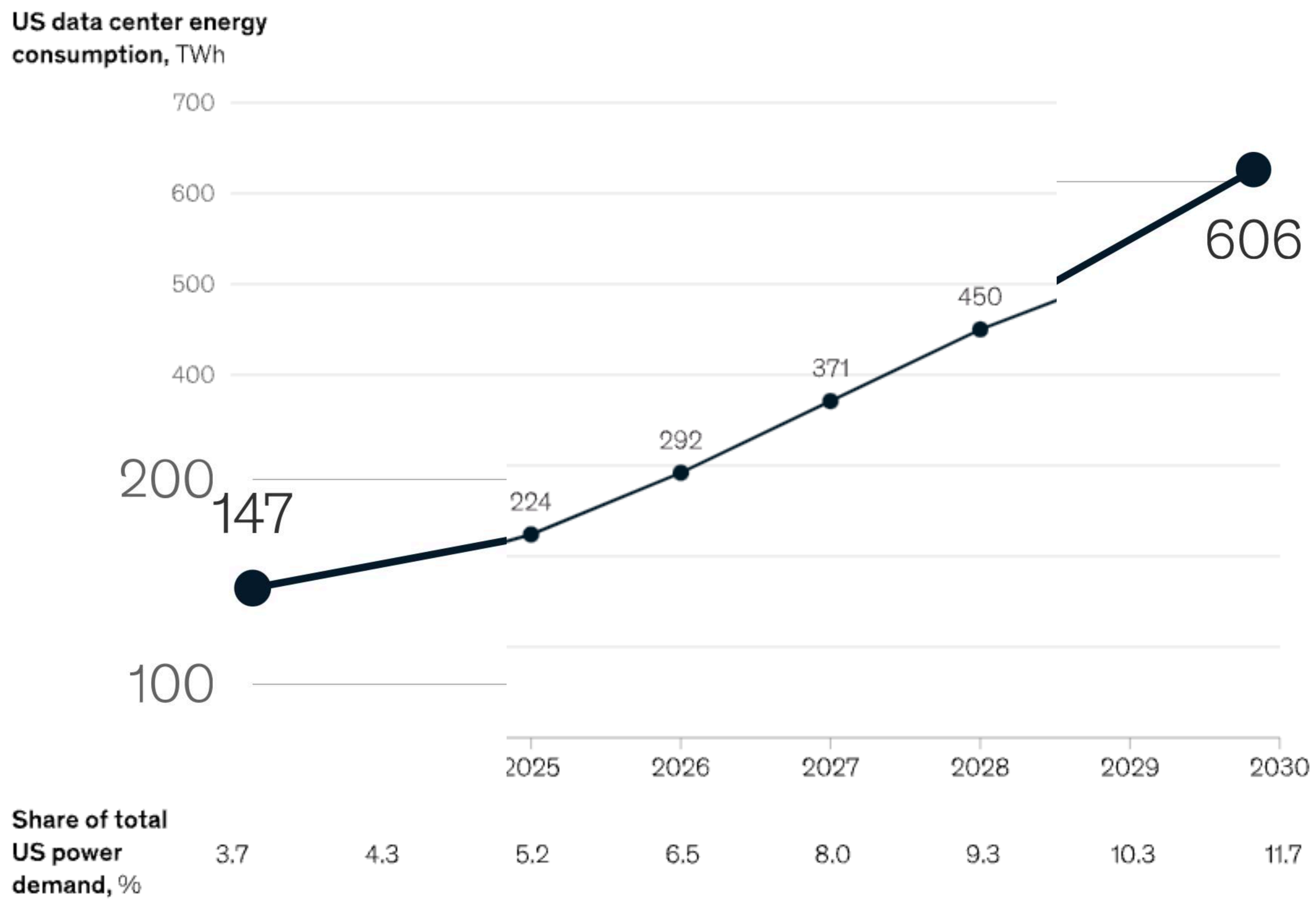
Terawatt-hours (TWh) of electricity demand, medium scenario



[1] McKinsey & Company (2024) How data centers and the energy sector can sate AI’s hunger for power
[2] IEA (2024), Electricity 2024, IEA, Paris https://www.iea.org/reports/electricity-2024

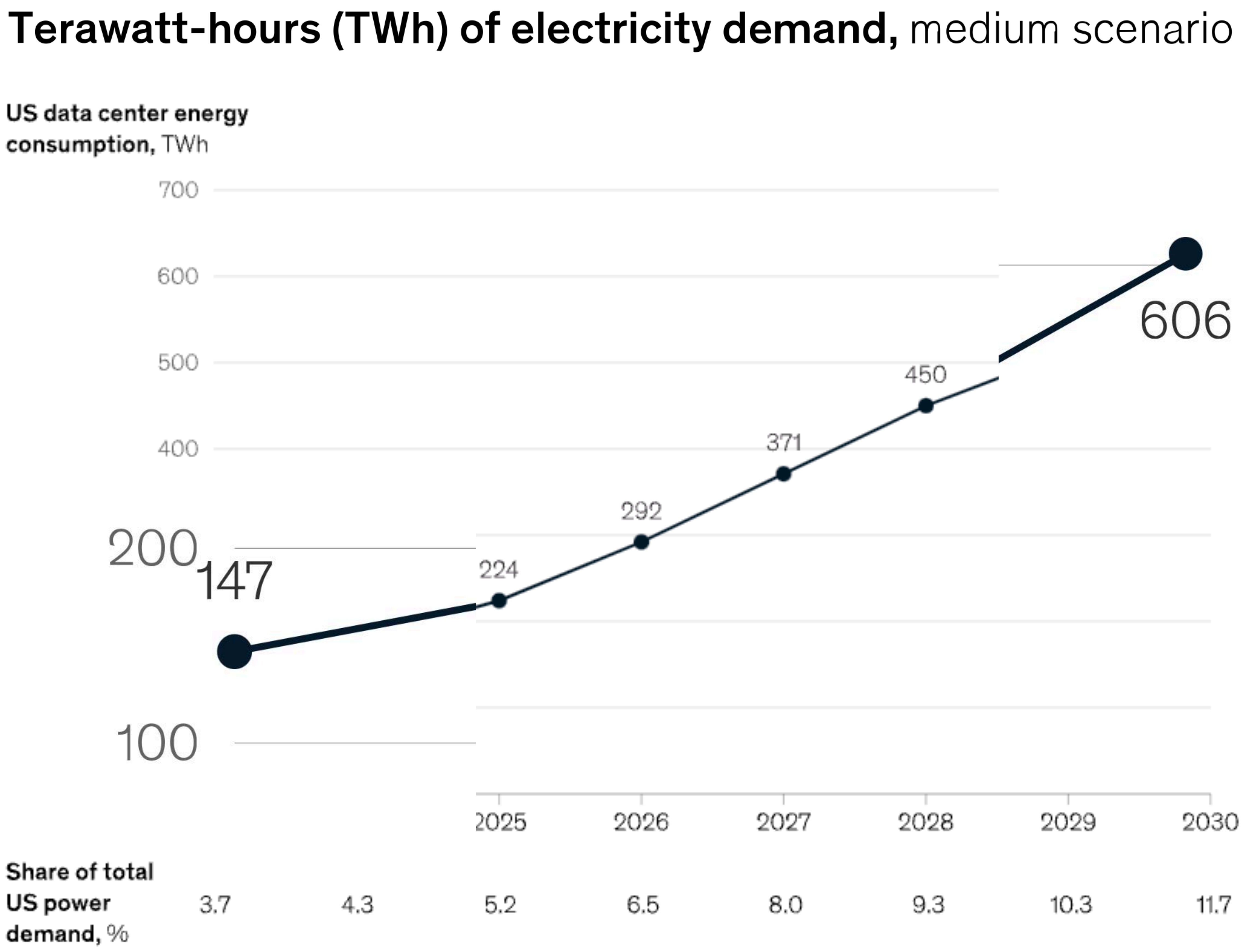
AI's growing appetite for energy

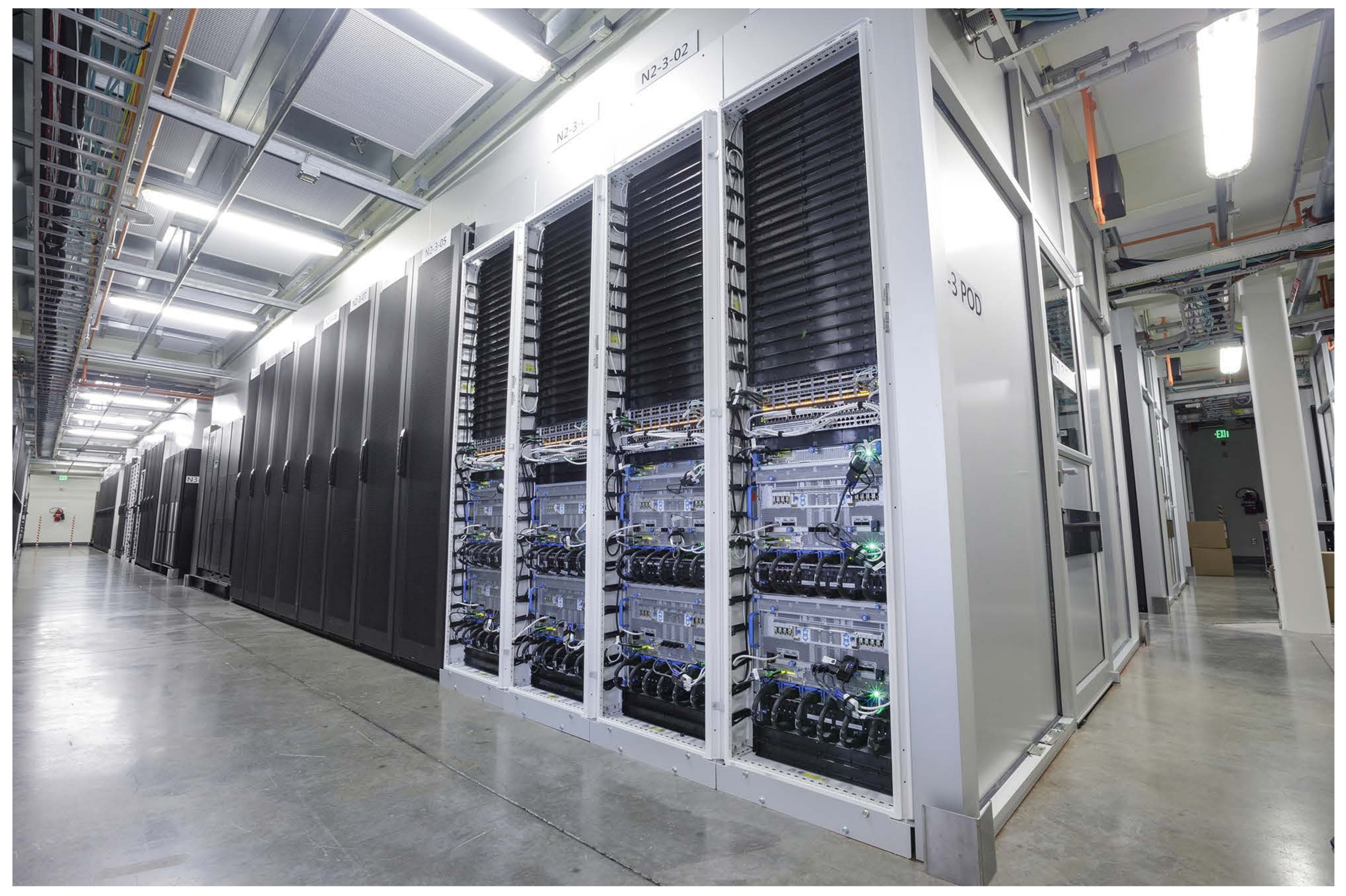
Terawatt-hours (TWh) of electricity demand, medium scenario



[1] McKinsey & Company (2024) How data centers and the energy sector can sate AI’s hunger for power
[2] IEA (2024), Electricity 2024, IEA, Paris https://www.iea.org/reports/electricity-2024

AI's growing appetite for energy





IEA projects that, by 2026, the global **AI computing** will consume at least 10x the electricity in 2023. [1]

[1] McKinsey & Company (2024) How data centers and the energy sector can sate AI's hunger for power
[2] IEA (2024), Electricity 2024, IEA, Paris https://www.iea.org/reports/electricity-2024

Energy tolls of large language models

Training

Energy tolls of large language models

Training

1,066 MWh
OPT 280B

1,287 MWh
GPT-3 175B [1]

4900 MWh
Llama 3.1 70B

21,588 MWh
Llama 3.1 405B

[1] The energy consumption for GPT-4 is estimated to be at least 7200 MWh in "Preventing the Immense Increase in the Life-Cycle Energy and Carbon Footprints of LLM-Powered Intelligent Chatbots"

Energy tolls of large language models

Training

1,066 MWh
OPT 280B

1,287 MWh
GPT-3 175B [1]

4900 MWh
Llama 3.1 70B

21,588 MWh
Llama 3.1 405B

0.875 MWh
Monthly Electricity

1.97 MWh[2]
Per Vechile

[1] The energy consumption for GPT-4 is estimated to be at least 7200 MWh in "Preventing the Immense Increase in the Life-Cycle Energy and Carbon Footprints of LLM-Powered Intelligent Chatbots"

Energy tolls of large language models

Training

1,066 MWh
OPT 280B

1,287 MWh
GPT-3 175B [1]

4900 MWh
Llama 3.1 70B

21,588 MWh
Llama 3.1 405B

24,672 Homes

10,958 Cars

[1] The energy consumption for GPT-4 is estimated to be at least 7200 MWh in "Preventing the Immense Increase in the Life-Cycle Energy and Carbon Footprints of LLM-Powered Intelligent Chatbots"

Energy tolls of large language models

Inference: Configurations matters

Energy tolls of large language models

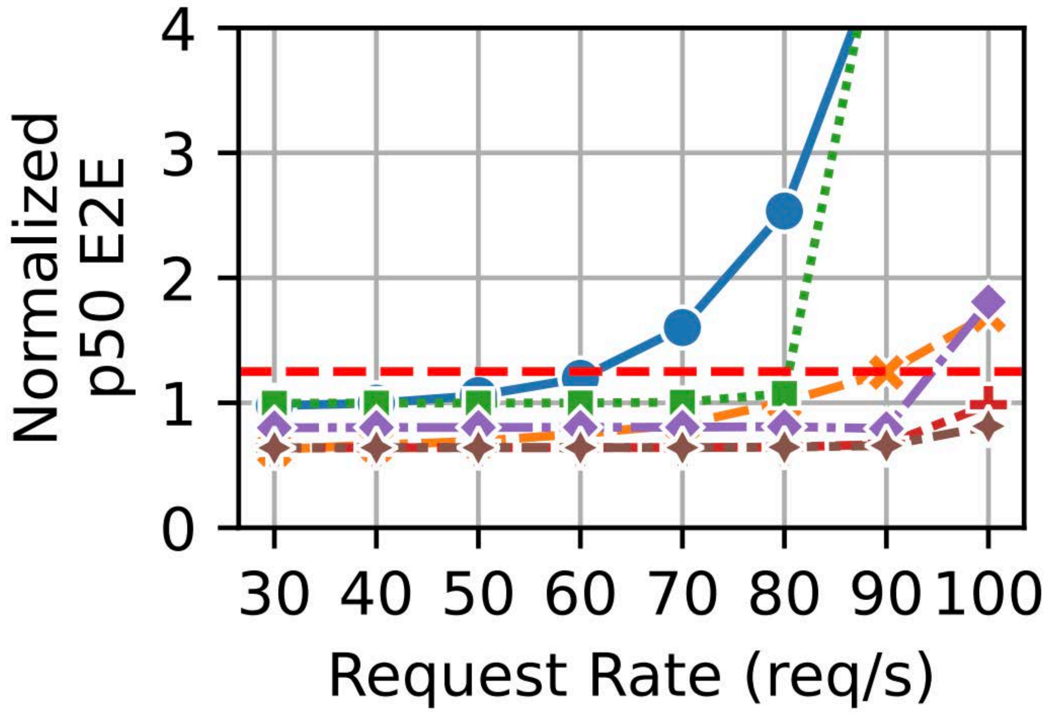
Inference: Configurations matters

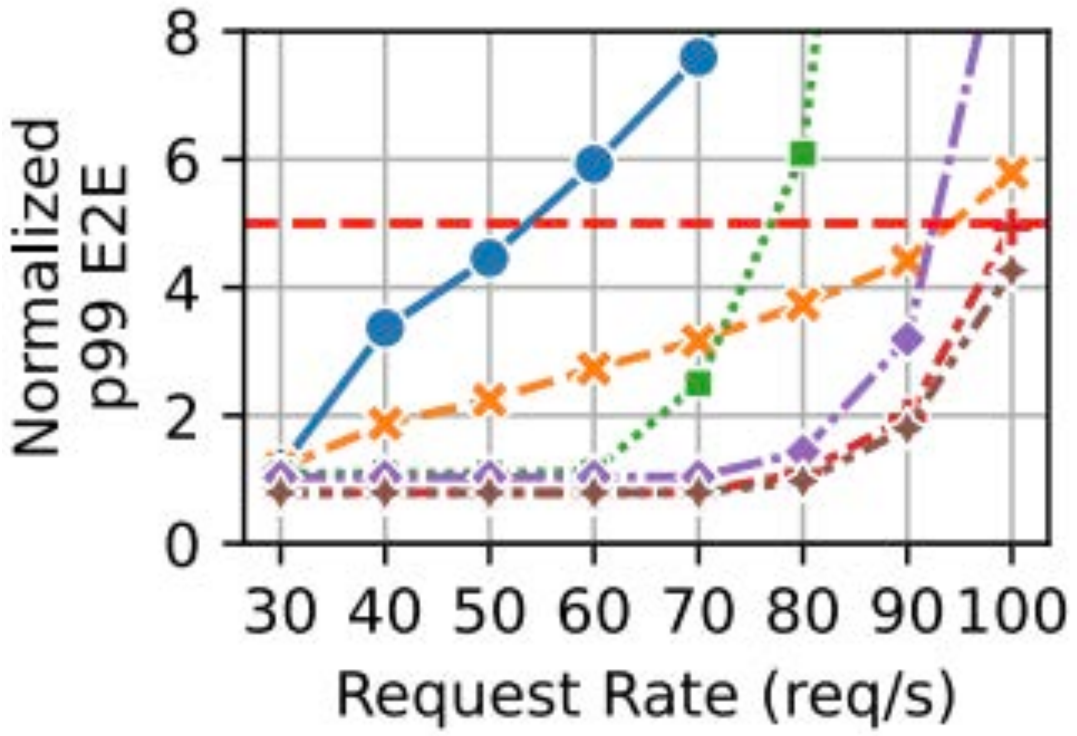
Tensor Parallelism		TP2				TP4				TP8			
GPU Frequency (GHz)		0.8	1.2	1.6	2.0	0.8	1.2	1.6	2.0	0.8	1.2	1.6	2.0
Input	Output												
Short	Short		**0.77**	0.97	1.03	0.94	0.79	0.91	1.01	1.35	1.19	1.29	1.49
Short	Medium		**2.78**	3.45	3.68	3.39	2.82	3.37	3.81	4.55	4.15	4.43	4.74
Short	Long					4.84	**4.17**	4.97	5.52	6.37	5.62	5.59	6.95
Medium	Short			**1.02**	1.09		1.08	1.07	1.20	1.51	1.29	1.34	1.73
Medium	Medium						4.23	**3.91**	4.08	5.34	4.39	4.56	5.44
Medium	Long						4.99	4.66	**4.53**	6.86	5.79	6.52	7.12
Long	Short						**1.51**	1.64	1.76	2.55	2.53	2.83	2.94
Long	Medium										**7.71**	8.81	9.17
Long	Long										12.99	**11.89**	13.21

Energy tolls of large language models

Inference: Configurations matters

Tensor Parallelism		TP2				TP4				TP8			
GPU Frequency (GHz)		0.8	1.2	1.6	2.0	0.8	1.2	1.6	2.0	0.8	1.2	1.6	2.0
Input	Output												
Short	Short		**0.77**	0.97	1.03	0.94	0.79	0.91	1.01	1.35	1.19	1.29	1.49
Short	Medium		**2.78**	3.45	3.68	3.39	2.82	3.37	3.81	4.55	4.15	4.43	4.74
Short	Long					4.84	**4.17**	4.97	5.52	6.37	5.62	5.59	6.95
Medium	Short			**1.02**	1.09		1.08	1.07	1.20	1.51	1.29	1.34	1.73
Medium	Medium						4.23	**3.91**	4.08	5.34	4.39	4.56	5.44
Medium	Long						4.99	4.66	**4.53**	6.86	5.79	6.52	7.12
Long	Short						**1.51**	1.64	1.76	2.55	2.53	2.83	2.94
Long	Medium										**7.71**	8.81	9.17
Long	Long										12.99	**11.89**	13.21

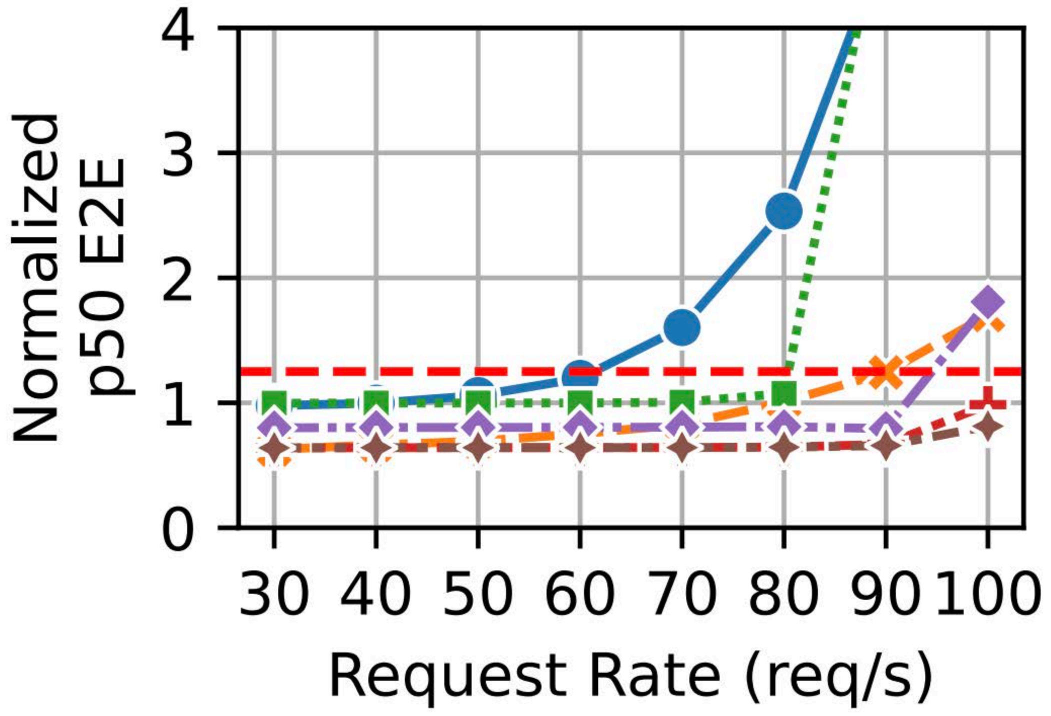


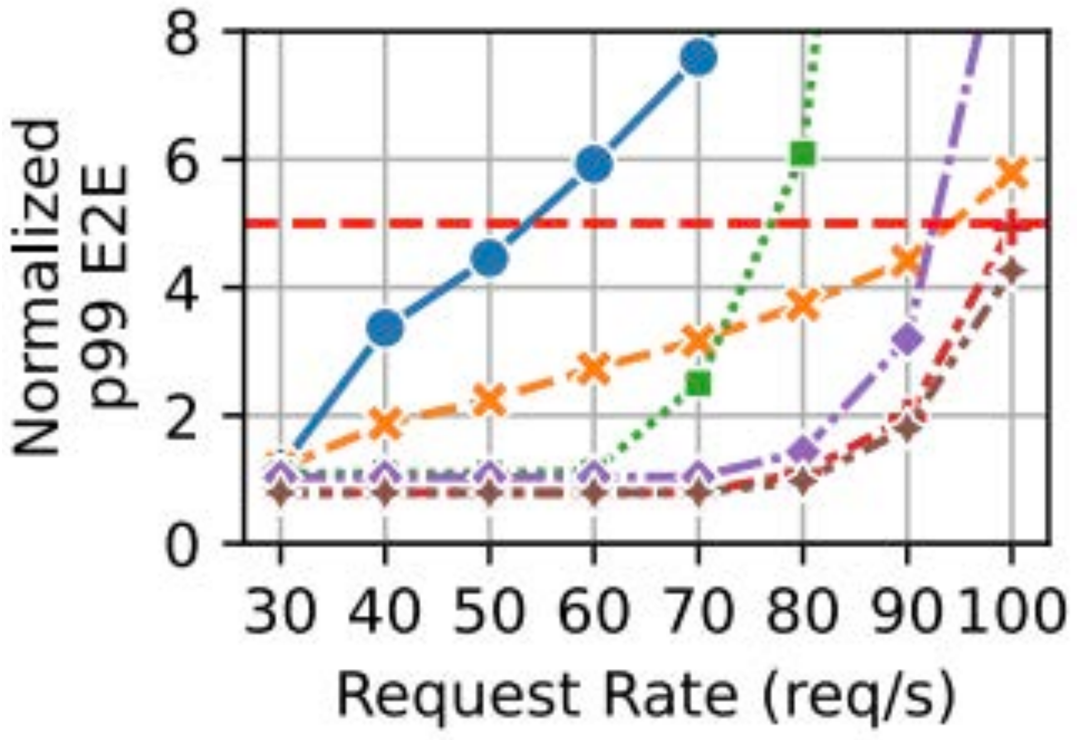


Energy tolls of large language models

Inference: Configurations matters

Tensor Parallelism		TP2				TP4				TP8			
GPU Frequency (GHz)		0.8	1.2	1.6	2.0	0.8	1.2	1.6	2.0	0.8	1.2	1.6	2.0
Input	Output												
Short	Short		**0.77**	0.97	1.03	0.94	0.79	0.91	1.01	1.35	1.19	1.29	1.49
Short	Medium		**2.78**	3.45	3.68	3.39	2.82	3.37	3.81	4.55	4.15	4.43	4.74
Short	Long					4.84	**4.17**	4.97	5.52	6.37	5.62	5.59	6.95
Medium	Short			**1.02**	1.09		1.08	1.07	1.20	1.51	1.29	1.34	1.73
Medium	Medium						4.23	**3.91**	4.08	5.34	4.39	4.56	5.44
Medium	Long						4.99	4.66	**4.53**	6.86	5.79	6.52	7.12
Long	Short						**1.51**	1.64	1.76	2.55	2.53	2.83	2.94
Long	Medium										**7.71**	8.81	9.17
Long	Long										12.99	**11.89**	13.21

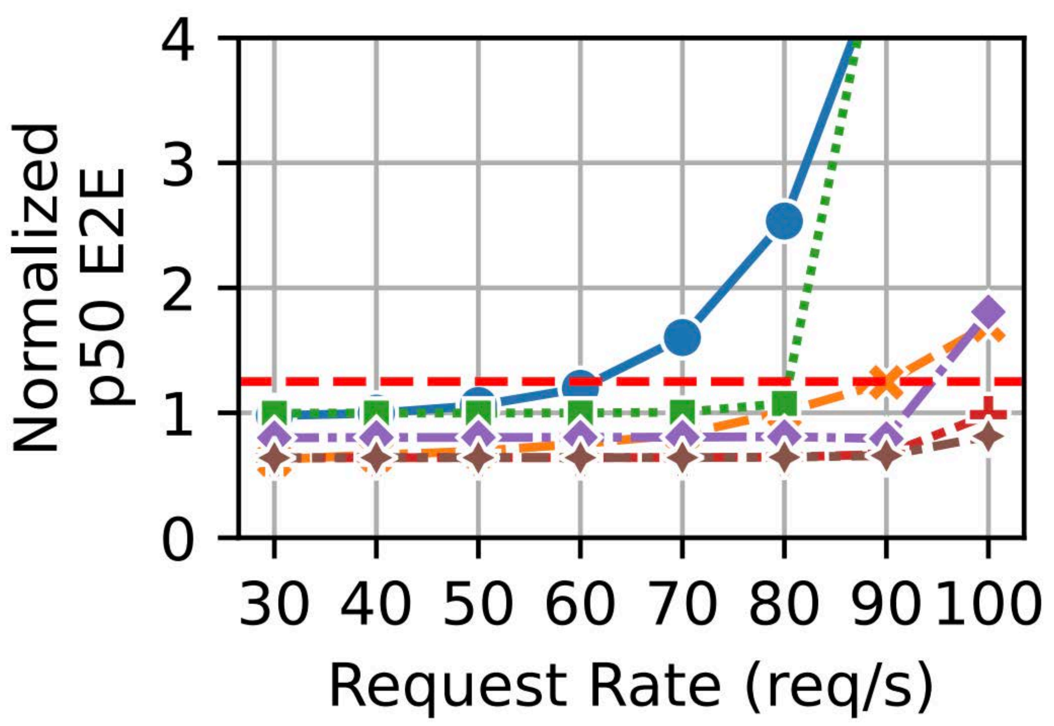


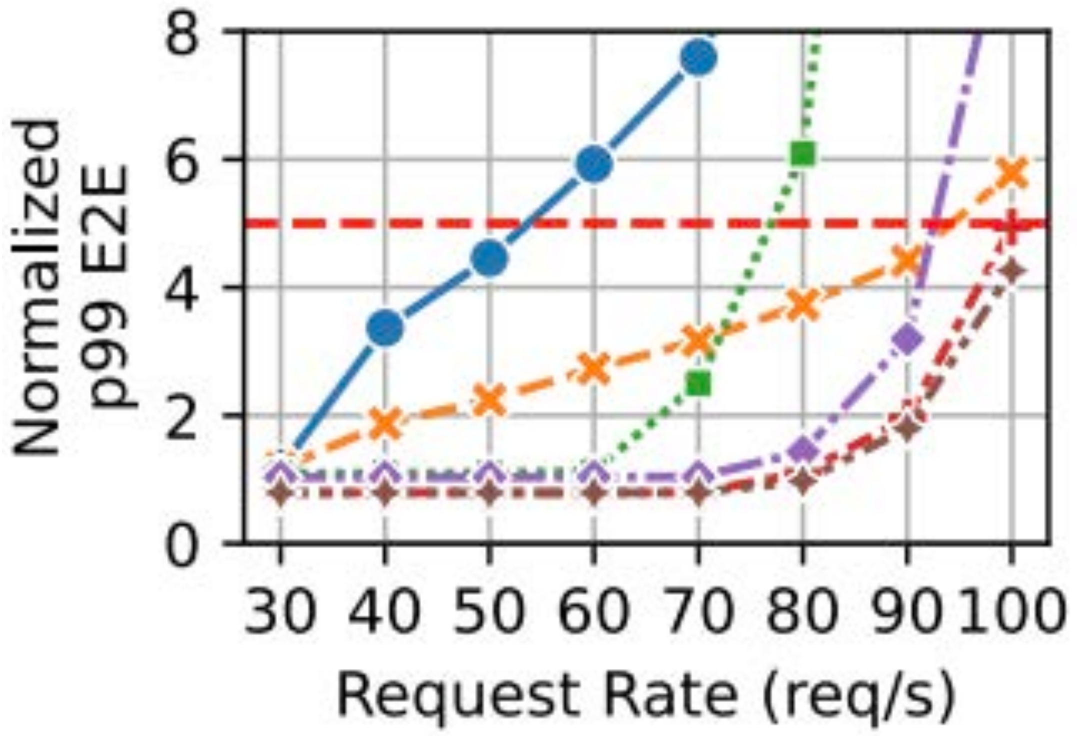


Energy tolls of large language models

Inference: Configurations matters

Tensor Parallelism		TP2				TP4				TP8			
GPU Frequency (GHz)		0.8	1.2	1.6	2.0	0.8	1.2	1.6	2.0	0.8	1.2	1.6	2.0
Input	Output												
Short	Short		**0.77**	0.97	1.03	0.94	0.79	0.91	1.01	1.35	1.19	1.29	1.49
Short	Medium		**2.78**	3.45	3.68	3.39	2.82	3.37	3.81	4.55	4.15	4.43	4.74
Short	Long					4.84	**4.17**	4.97	5.52	6.37	5.62	5.59	6.95
Medium	Short			**1.02**	1.09		1.08	1.07	1.20	1.51	1.29	1.34	1.73
Medium	Medium						4.23	**3.91**	4.08	5.34	4.39	4.56	5.44
Medium	Long						4.99	4.66	**4.53**	6.86	5.79	6.52	7.12
Long	Short						**1.51**	1.64	1.76	2.55	2.53	2.83	2.94
Long	Medium										**7.71**	8.81	9.17
Long	Long										12.99	**11.89**	13.21





Energy tolls of large language models

Inference

	GPT-4o	Llama 3 8B
Write an email (170 tokens)	14.9 Wh	0.641 Wh
Small Conversation (400 tokens)	35.1 Wh	1.510 Wh

[1] Online Calculator https://huggingface.co/spaces/genai-impact/ecologits-calculator

Energy tolls of large language models

Inference

	🔥 GPT-4	🔥 Llama 3 70B
Write an email (170 tokens)	190 Wh	2.13 Wh
Small Conversation (400 tokens)	447 Wh	5.02 Wh

[1] Online Calculator https://huggingface.co/spaces/genai-impact/ecologits-calculator

Energy tolls of large language models

Inference

	🔥 GPT-4	🔥 Llama 3 70B
Write an email (170 tokens)	190 Wh	2.13 Wh
Small Conversation (400 tokens)	447 Wh	5.02 Wh

[1] Online Calculator https://huggingface.co/spaces/genai-impact/ecologits-calculator

Energy tolls of large language models

Inference

	GPT-4	Llama 3 70B
Write an email (170 tokens)	190 Wh	2.13 Wh
Small Conversation (400 tokens)	447 Wh	5.02 Wh

21.1 Hours

11 Cycle

2.89 Cycle

0.76 Miles

[1] Online Calculator https://huggingface.co/spaces/genai-impact/ecologits-calculator

Energy tolls of large language models

Inference

	🔥 GPT-4	🔥 Llama 3 70B
Write an email (170 tokens)	190 Wh	2.13 Wh
Small Conversation (400 tokens)	447 Wh	5.02 Wh

21.1 Hours

11 Cycle

2.89 Cycle

60 emails ≈ NCAR 🚗 DEN

[1] Online Calculator https://huggingface.co/spaces/genai-impact/ecologits-calculator

What's under the hood

What’s under the hood

Carbon

What’s under the hood

Carbon

Water

What’s under the hood

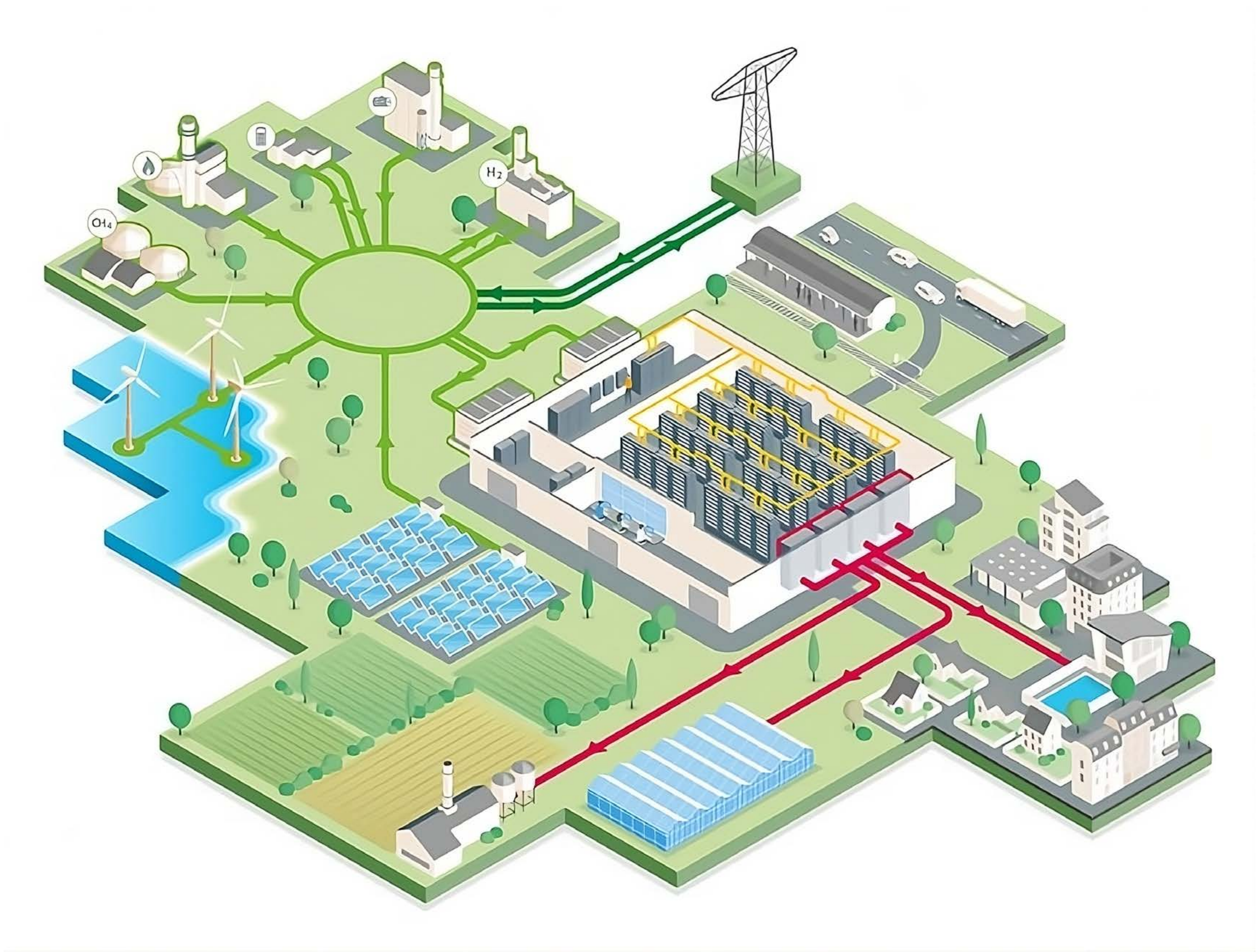
Carbon

Water

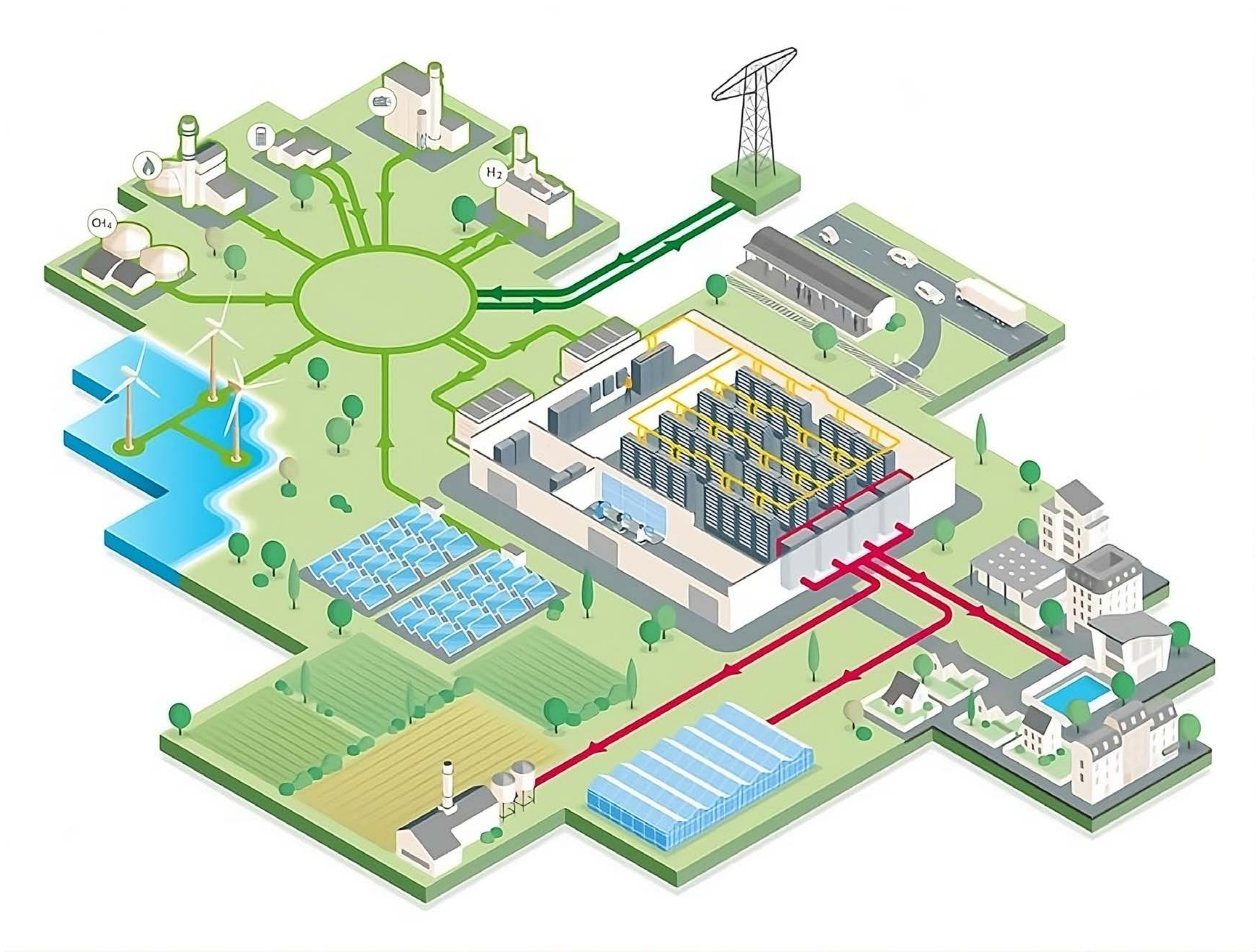
Air Pollution

Greenhouse gas emission

Greenhouse gas emission



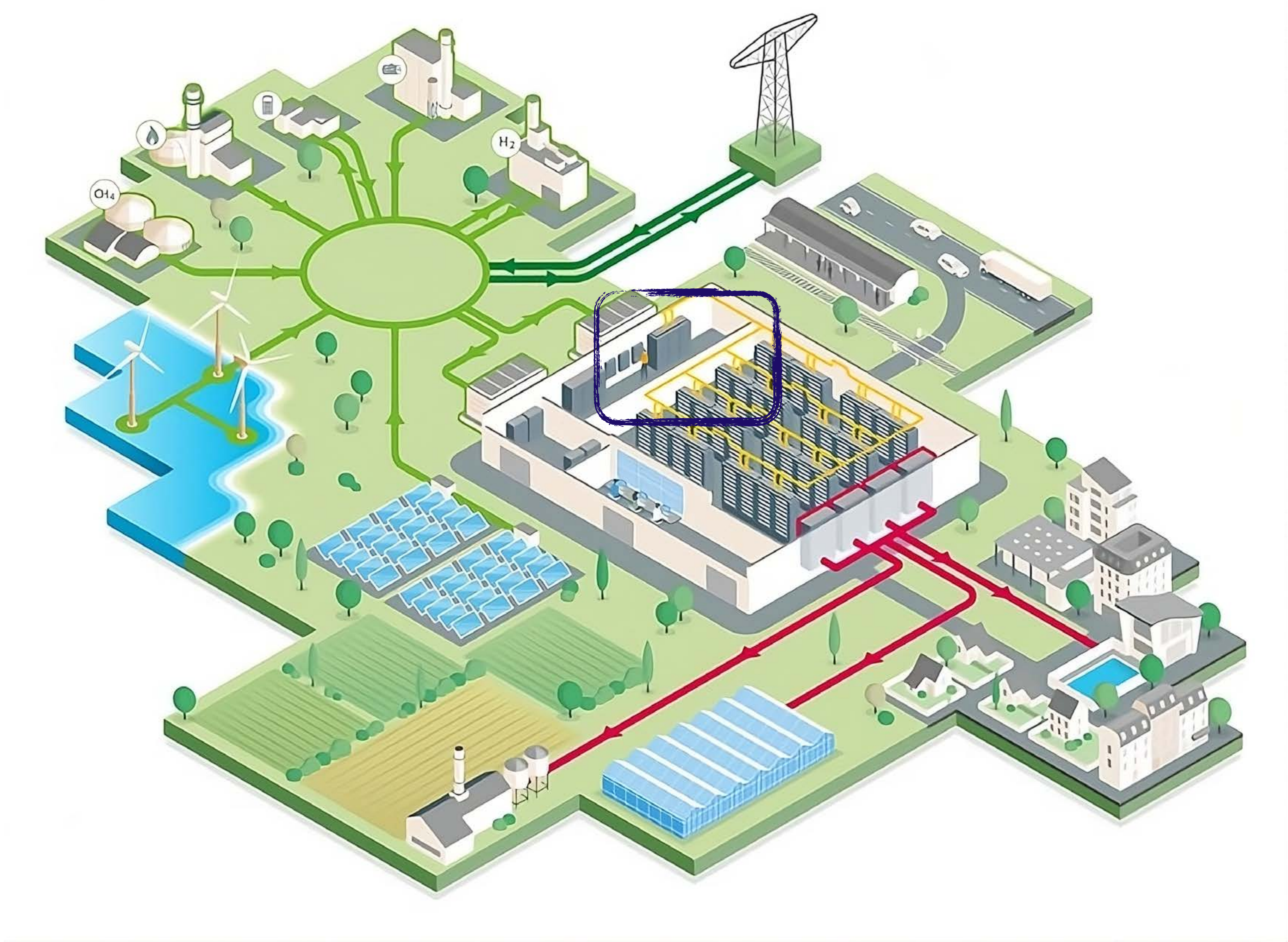
Greenhouse gas emission



CH4 N20 HFCs CO2

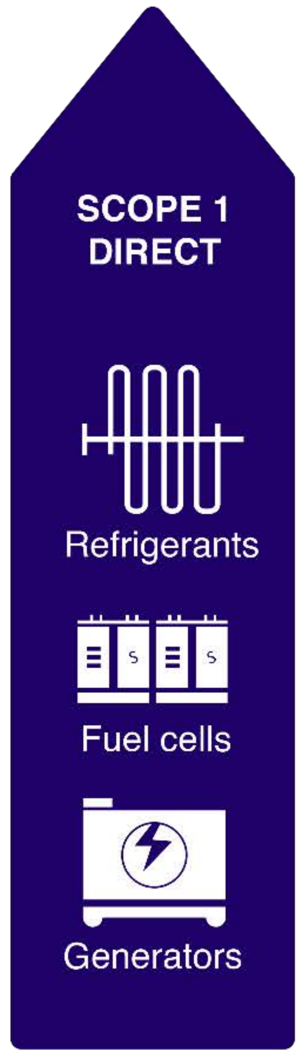
PFCs SF6 NF3

Greenhouse gas emission

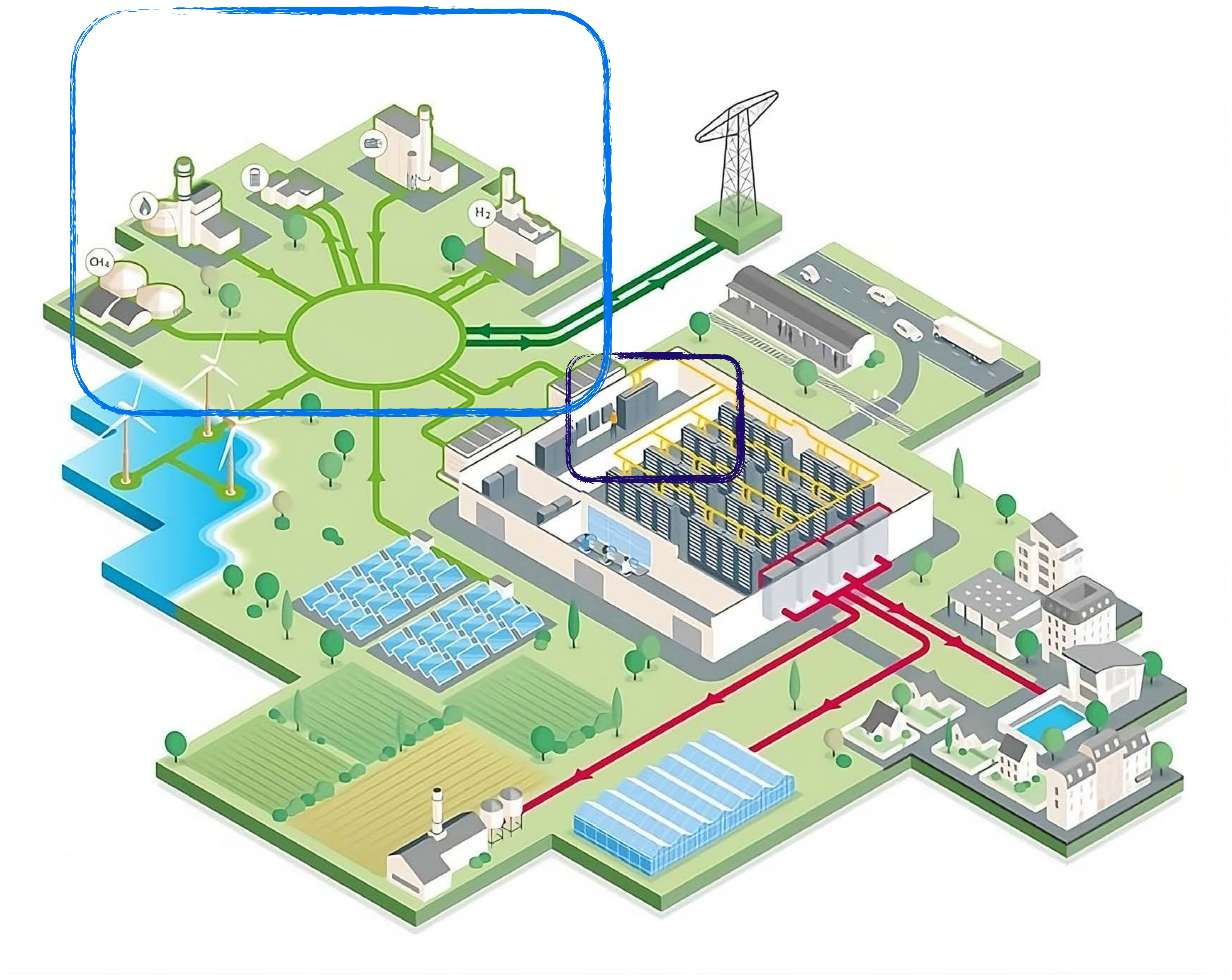


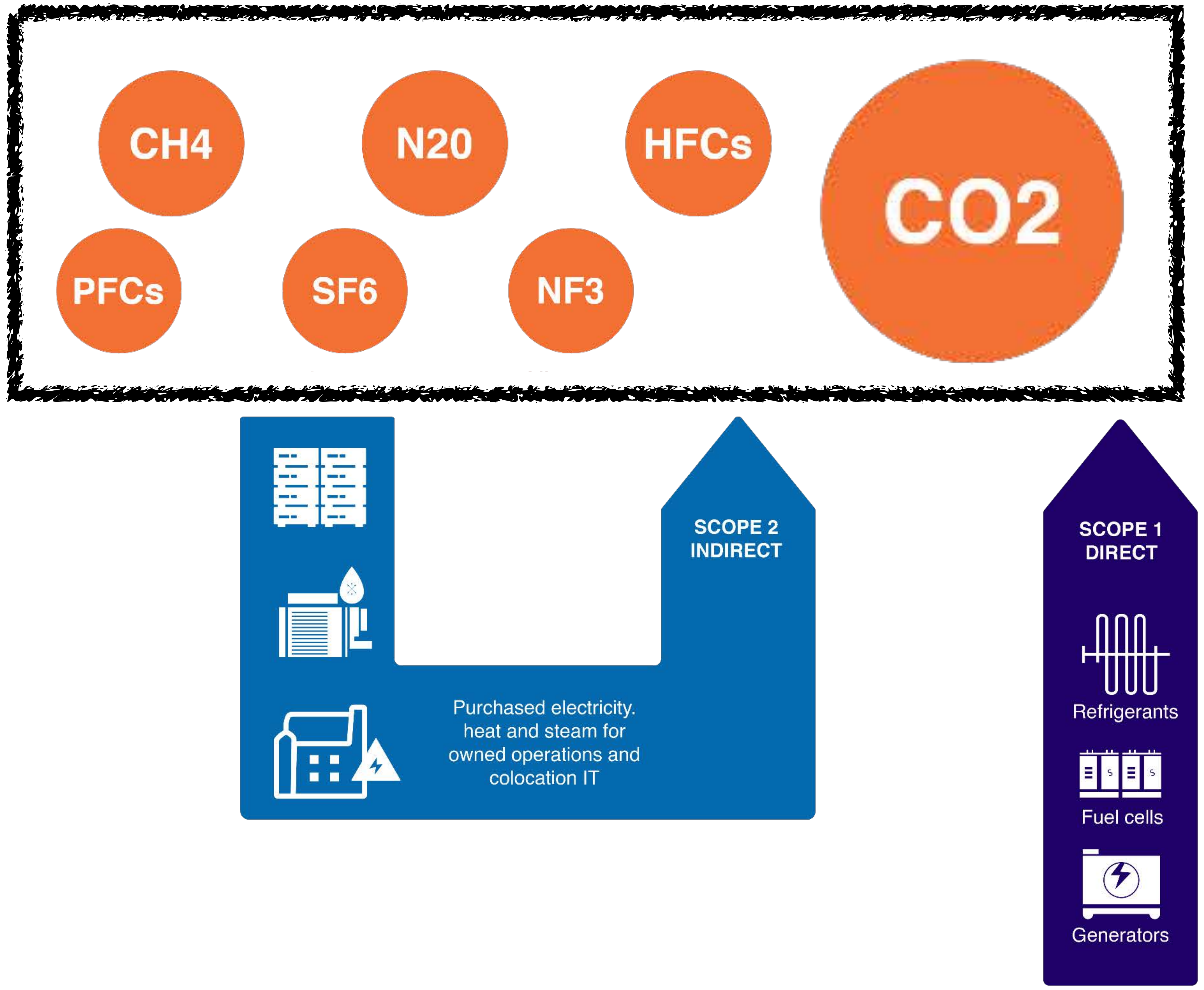
CH4 N20 HFCs CO2

PFCs SF6 NF3

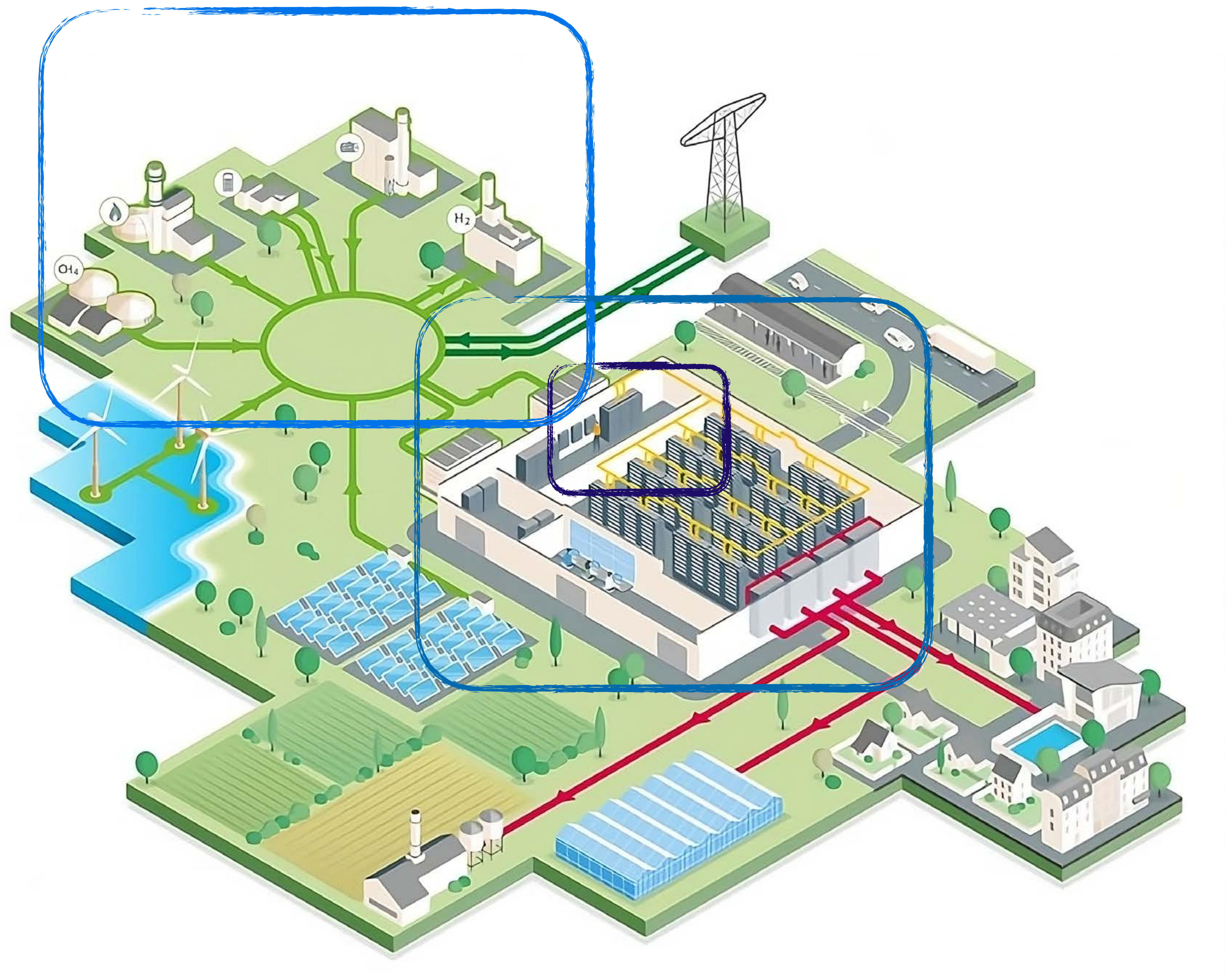


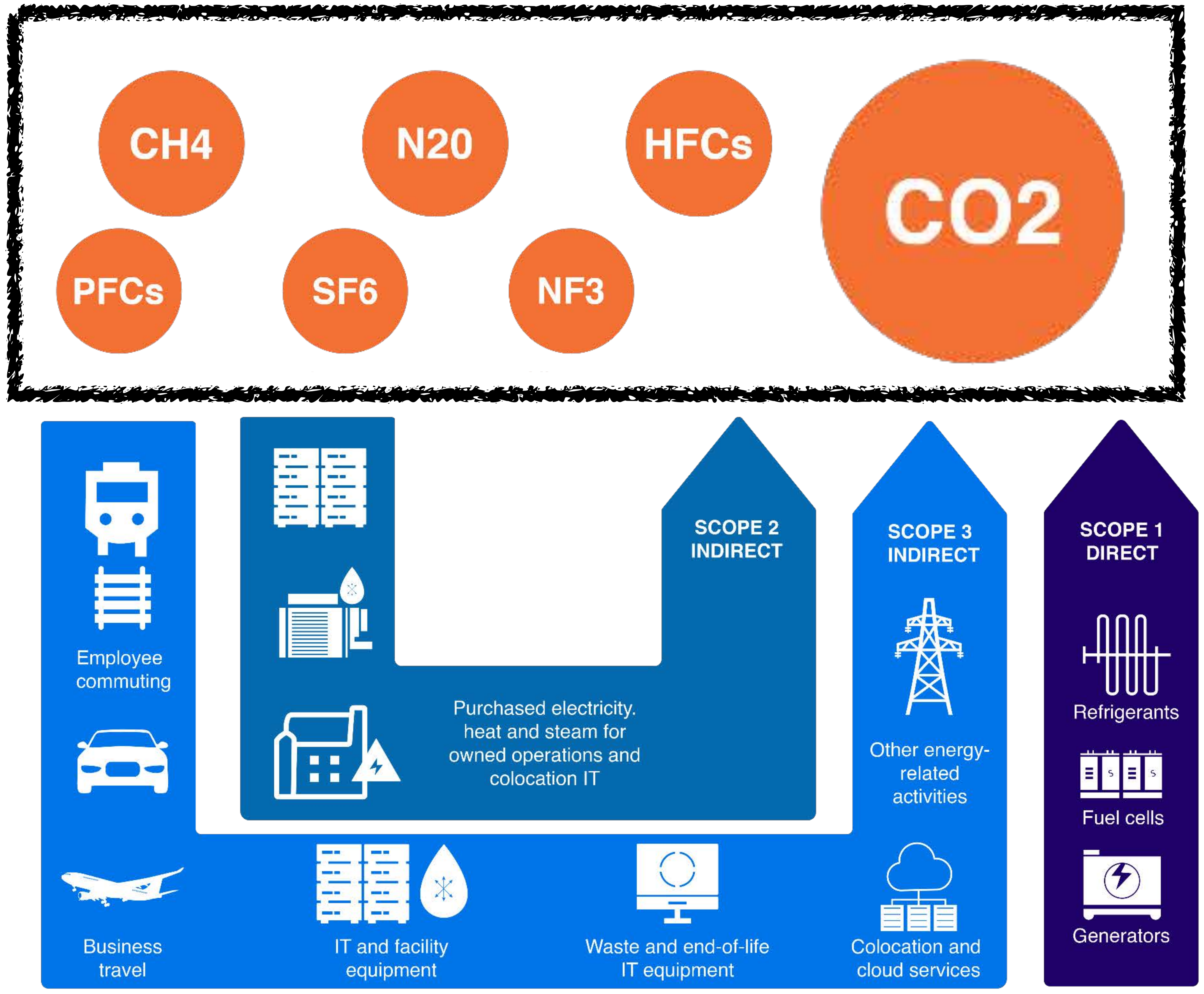
Greenhouse gas emission





Greenhouse gas emission





Carbon footprint of LLM

Training

Llama 3.1 405B

8930 tons CO_2eq

≈

308,000 Apple Watches

160,000 iPhones

89,000 iPads

56,000 Surface Laptop

Carbon footprint of LLM

Location-based vs market-based

	Training Time (GPU hours)	Training Power Consumption (W)	Training Location-Based Greenhouse Gas Emissions (tons CO2eq)	Training Market-Based Greenhouse Gas Emissions (tons CO2eq)
Llama 3.1 8B	1.46M	700	420	0
Llama 3.1 70B	7.0M	700	2,040	0
Llama 3.1 405B	30.84M	700	8,930	0
Total	39.3M		11,390	0

Carbon footprint of LLM

Location-based vs market-based

Location Based

> 🔥 ***Example**: If your company consumes 100,000 kWh of **electricity** in the UK, with a **grid emission factor** of 0.21233 $kgCO_2e/kWh$, the calculation would be:*
>
> *100,000 kWh x 0.21233 $kgCO_2e/kWh$ = 21,233 $kgCO_2e$ or 21.23 tCO_2e.*

Carbon footprint of LLM

Location-based vs market-based

Market Based

> 🔥 ***Example:***
> *If your company consumes 100,000 kWh of electricity but buys 100% renewable energy through a REC, the emissions factor is 0 $kgCO_2e/kWh$. Your Scope 2 emissions would be:*
> *100,000 kWh x 0 $kgCO_2e/kWh$ = 0 $kgCO_2e$.*

Carbon footprint of LLM

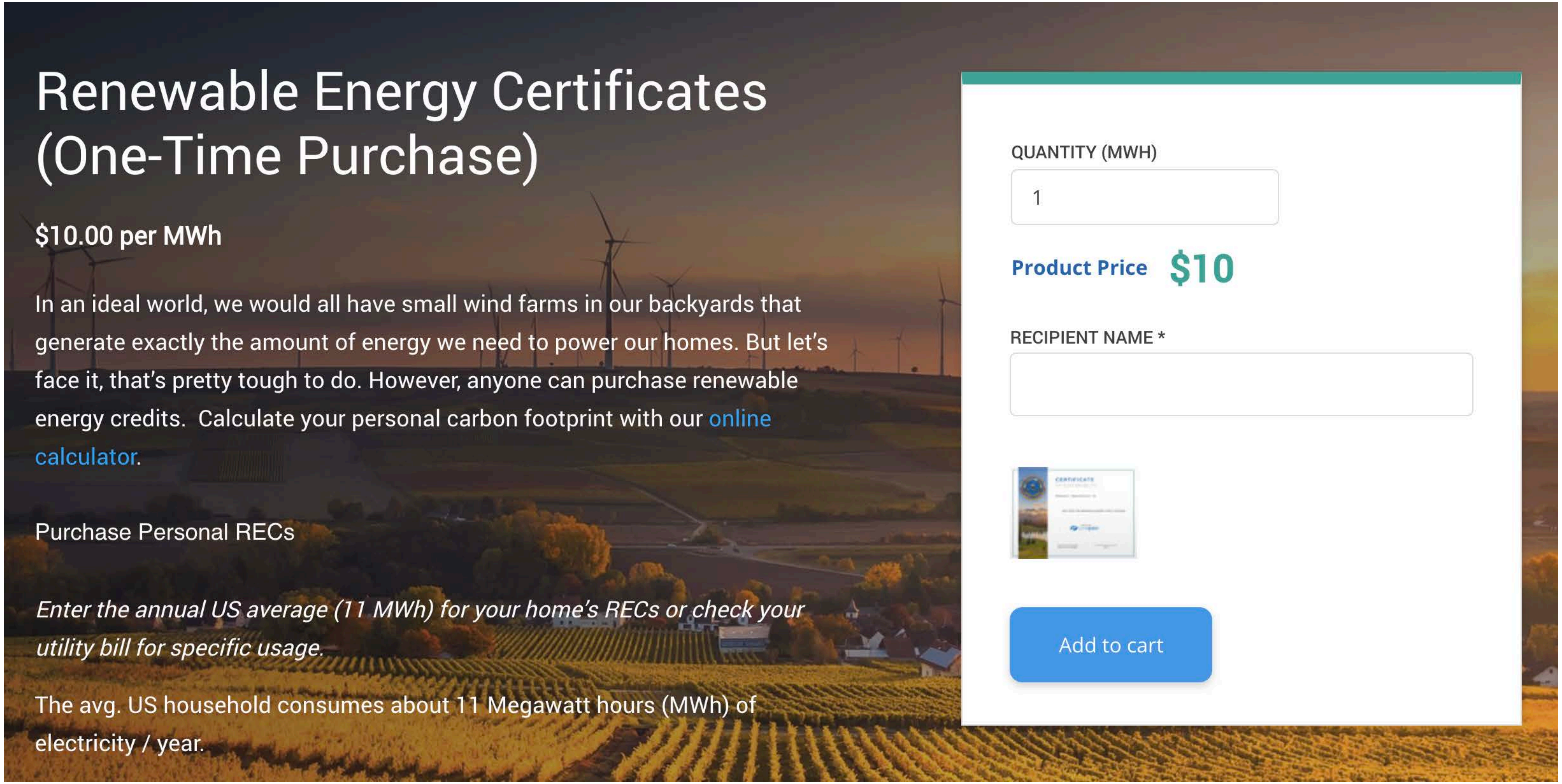
Location-based vs market-based

Market Based

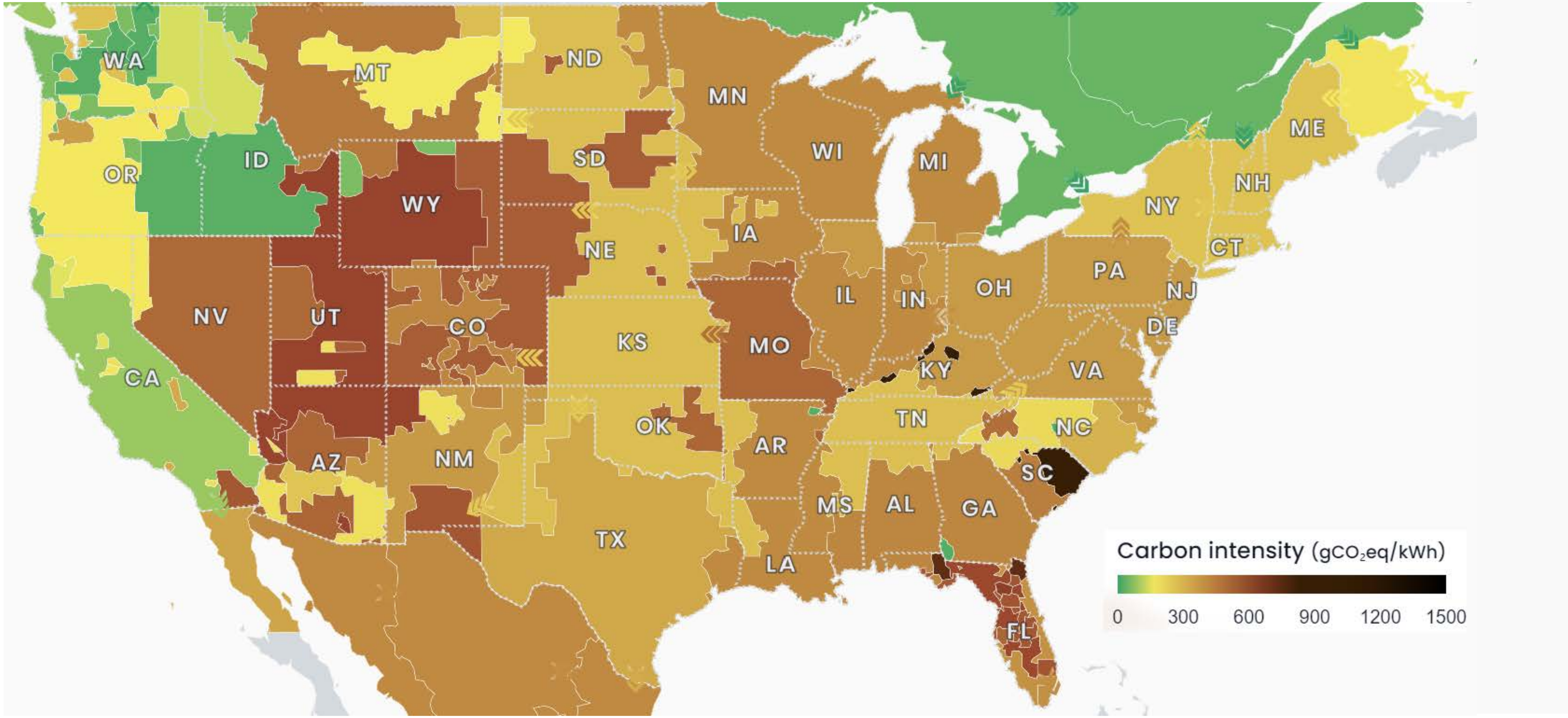
> 🔥 ***Example:***
>
> *100,000 kWh x 0.316 kgCO₂e/kWh = 31,600 kgCO₂e or 31.60 tCO₂e.*

Carbon footprint of LLM

Location-based vs market-based

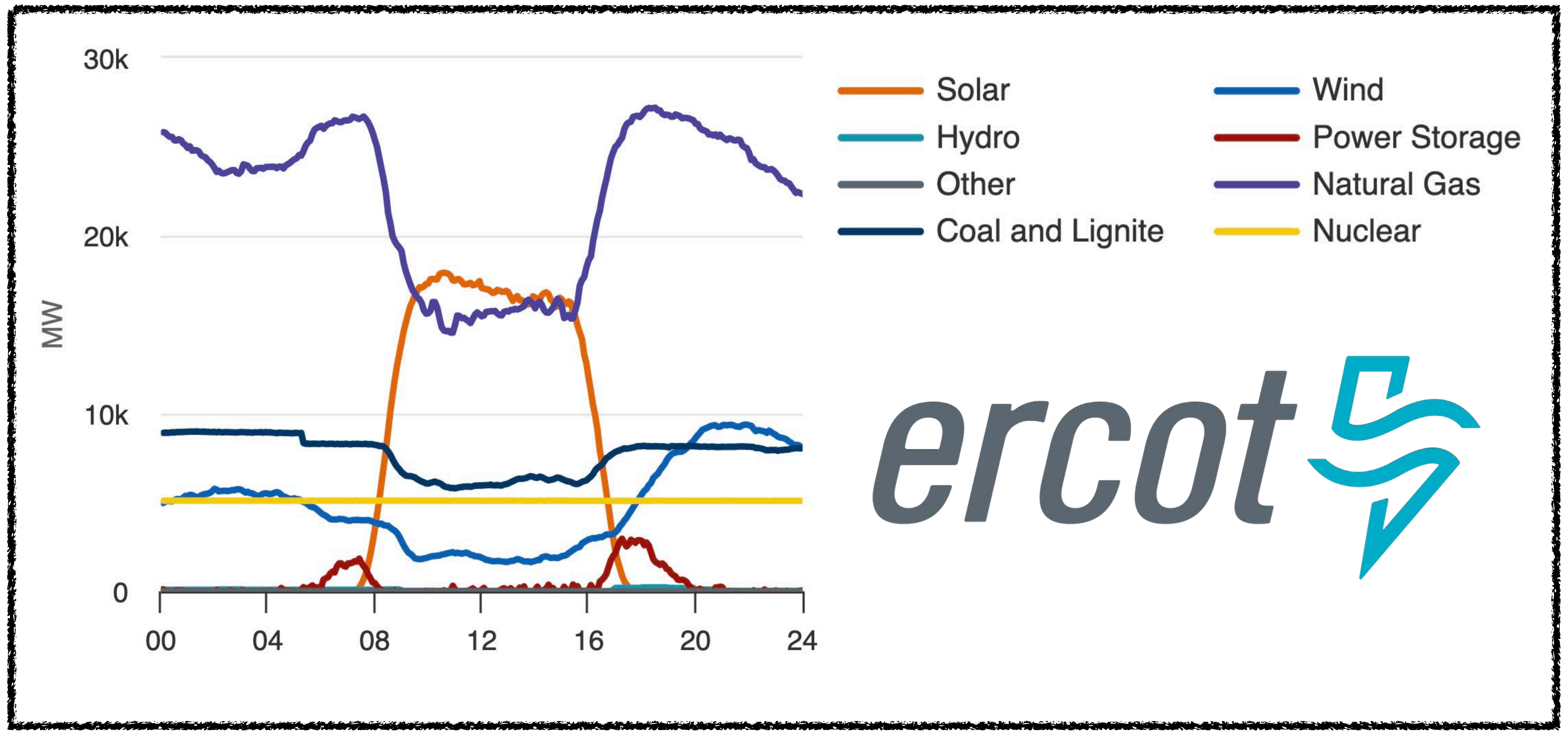


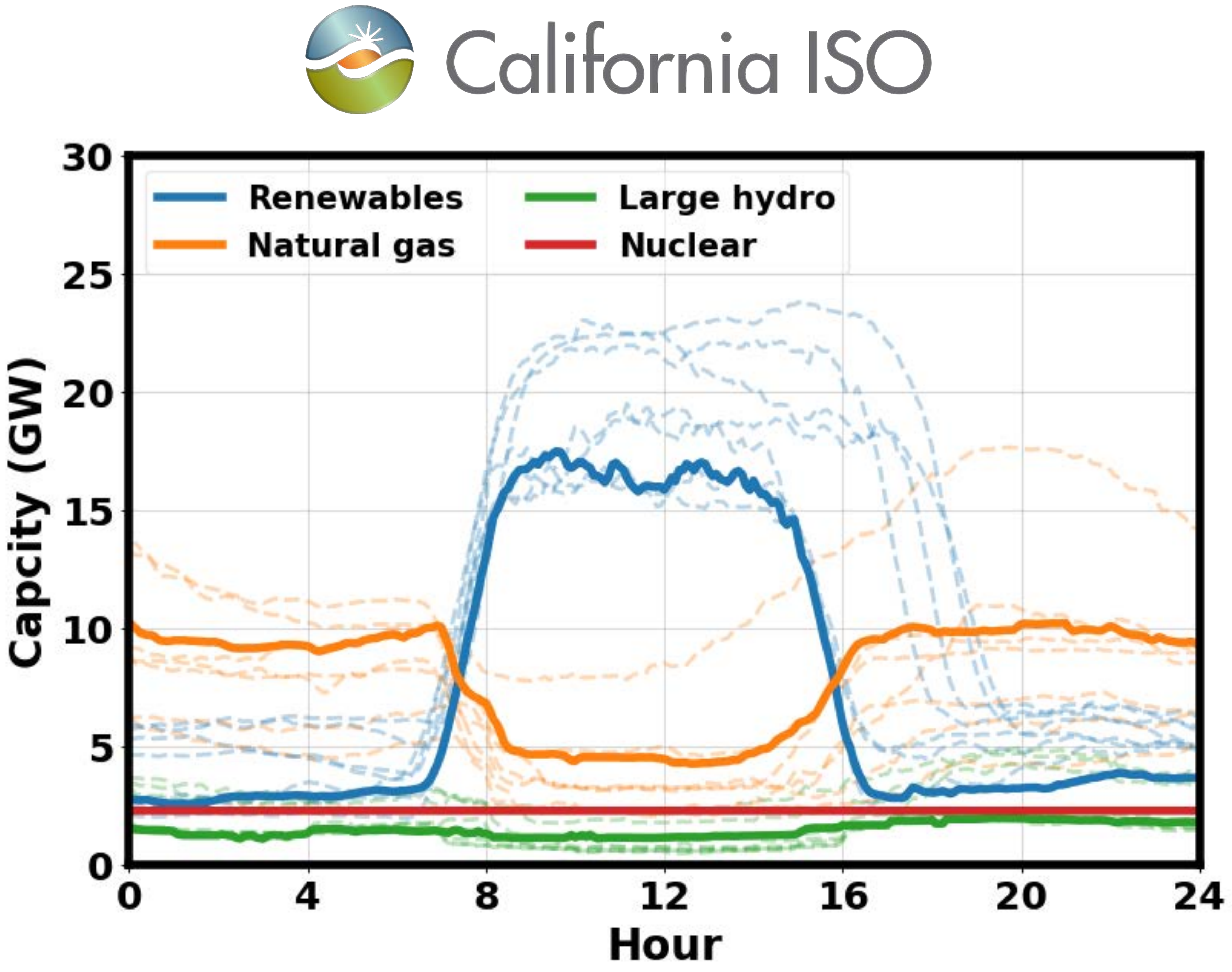
All regions are not equal



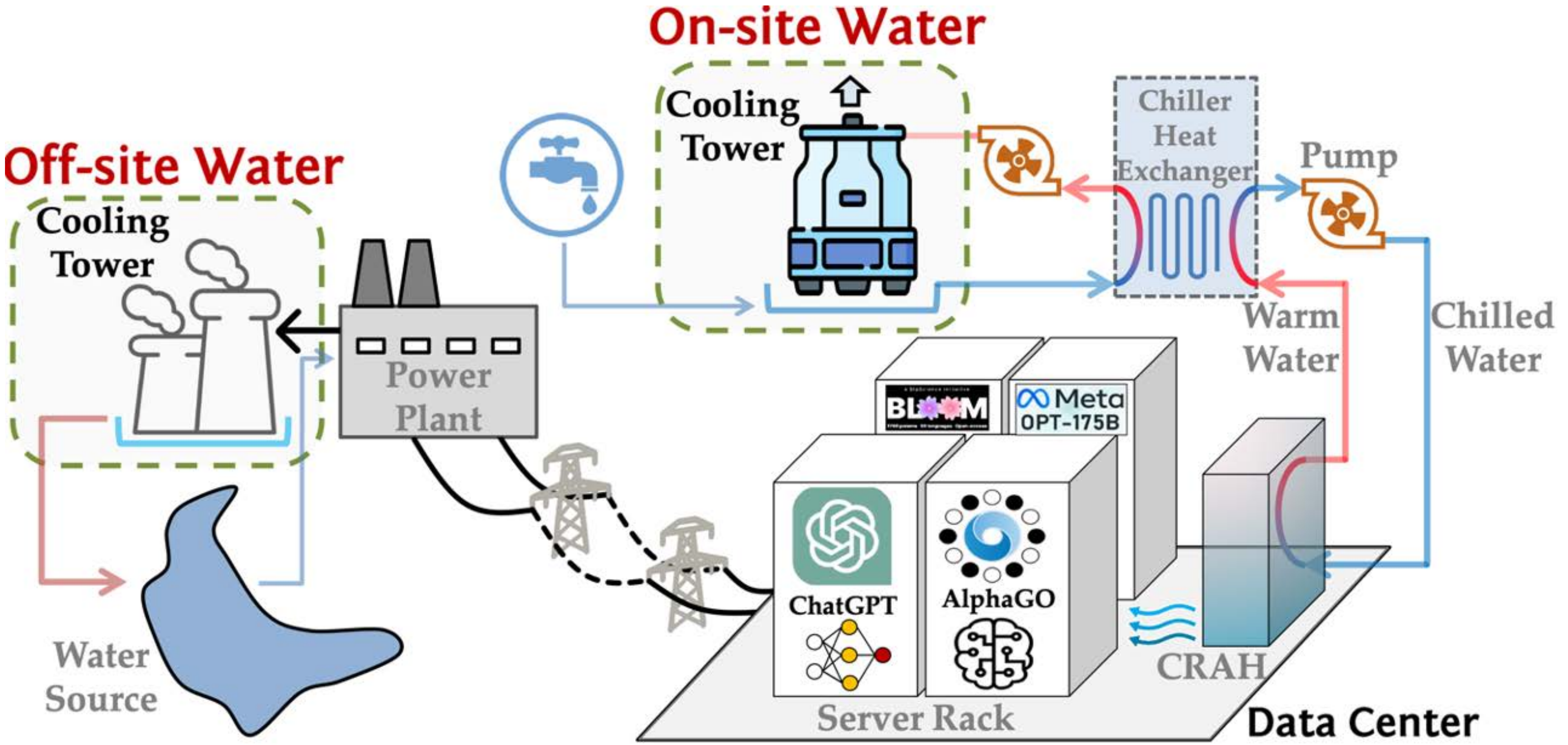
Source: ElectricityMaps (April 27, 2024)

All regions are not equal

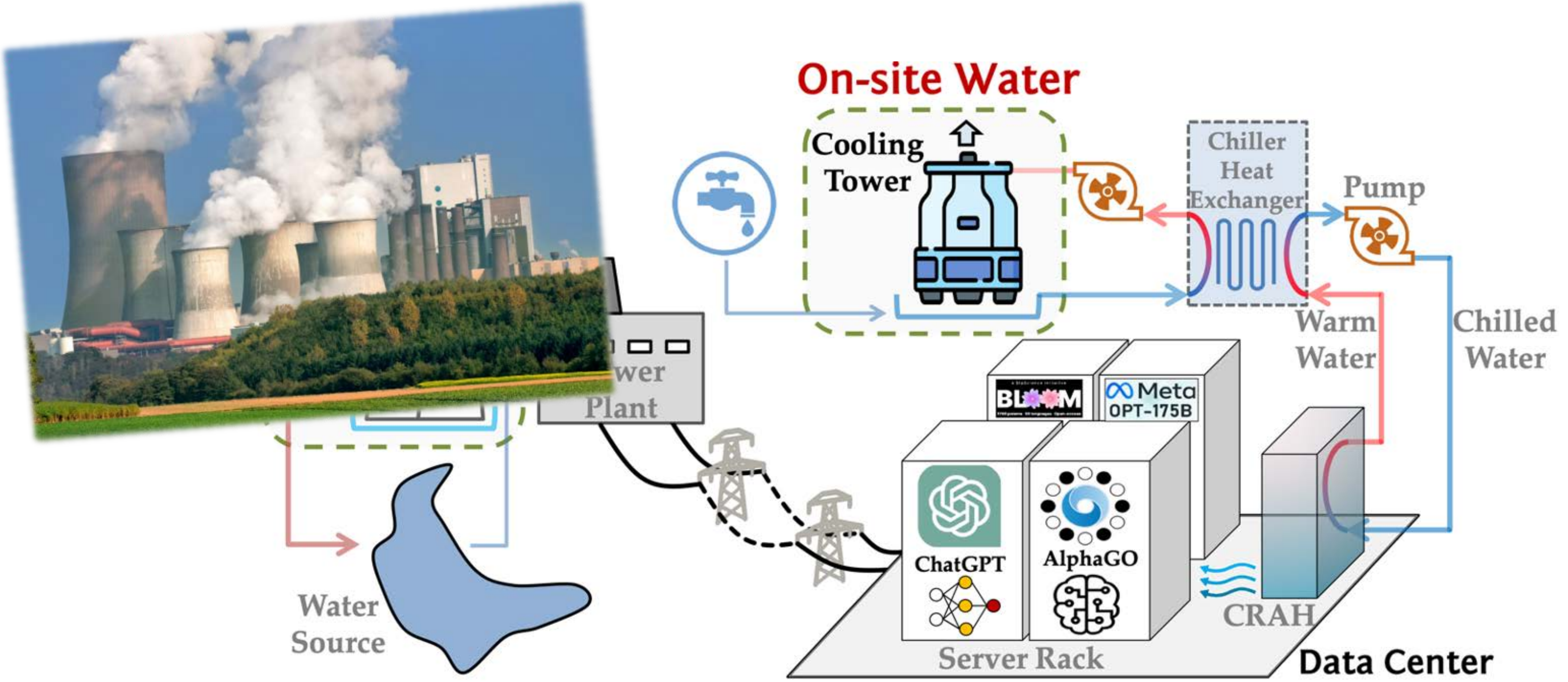




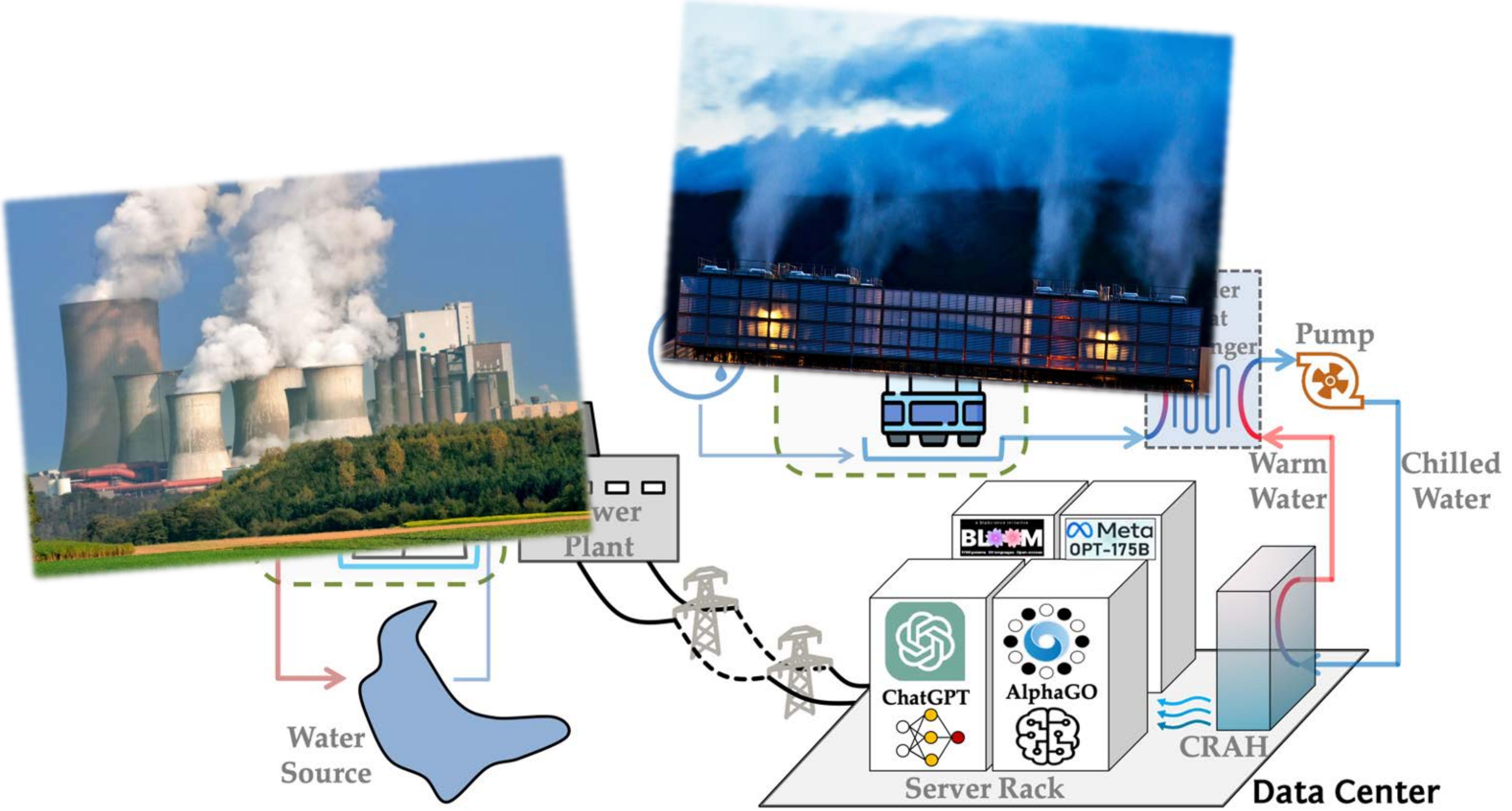
Data center water footprint



Data center water footprint



Data center water footprint



Water footprint estimation

Onsite Water WUE (based on an example cooling tower)

$$WUE_{on} = \frac{S}{S-1}\left(6 \times 10^{-5} \cdot T_{\mathrm{w}}^{3} - 0.01 \cdot T_{\mathrm{w}}^{2} + 0.61 \cdot T_{\mathrm{w}} - 10.40\right)$$

Offsite Water WUE

$$WUE_{off} = \frac{\sum_k b_k \times EWIF_k}{\sum_k b_k}$$

Water footprint estimation

Onsite Water WUE (based on an example cooling tower)

$$WUE_{on} = \frac{S}{\boxed{S} - 1}\left(6 \times 10^{-5} \cdot T_{\mathrm{w}}^{3} - 0.01 \cdot T_{\mathrm{w}}^{2} + 0.61 \cdot T_{\mathrm{w}} - 10.40\right)$$

Number of Cycles

Offsite Water WUE

$$WUE_{off} = \frac{\sum_k b_k \times EWIF_k}{\sum_k b_k}$$

Water footprint estimation

Onsite Water WUE (based on an example cooling tower)

$$WUE_{on} = \frac{S}{\boxed{S} - 1}\left(6 \times 10^{-5} \cdot \boxed{T_{\mathrm{w}}^3} - 0.01 \cdot T_{\mathrm{w}}^2 + 0.61 \cdot T_{\mathrm{w}} - 10.40\right)$$

Number of Cycles

Outside Wetbulb Temperature

Offsite Water WUE

$$WUE_{off} = \frac{\sum_k b_k \times EWIF_k}{\sum_k b_k}$$

Water footprint estimation

Onsite Water WUE (based on an example cooling tower)

$$WUE_{on} = \frac{S}{\boxed{S} - 1}\left(6 \times 10^{-5} \cdot \boxed{T_{\mathrm{w}}^3} - 0.01 \cdot T_{\mathrm{w}}^2 + 0.61 \cdot T_{\mathrm{w}} - 10.40\right)$$

Number of Cycles

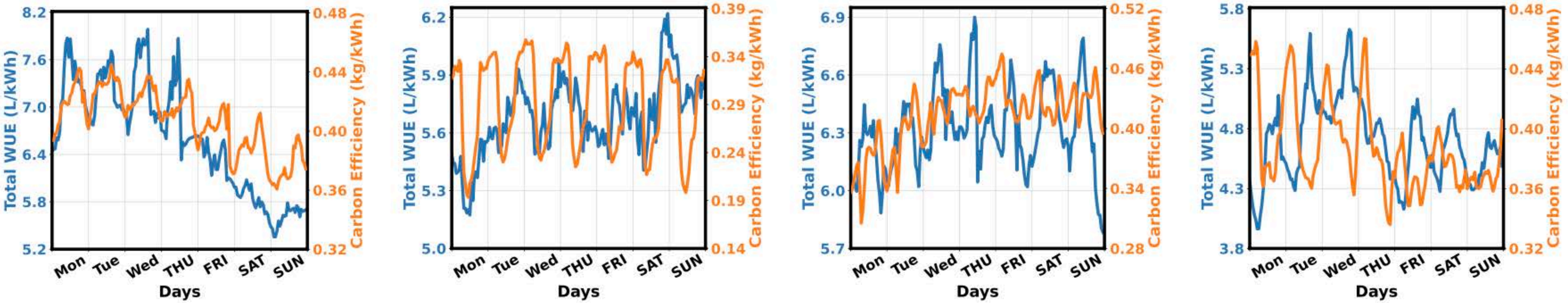
Outside Wetbulb Temperature

Offsite Water WUE

$$WUE_{off} = \frac{\sum_k b_k \times \boxed{EWIF_k}}{\sum_k b_k}$$

Estimated energy water intensity factor (EWIF) of each energy source

Hourly carbon efficiency and WUE



(a) Virginia (b) Nevada (c) Texas (d) Oregon

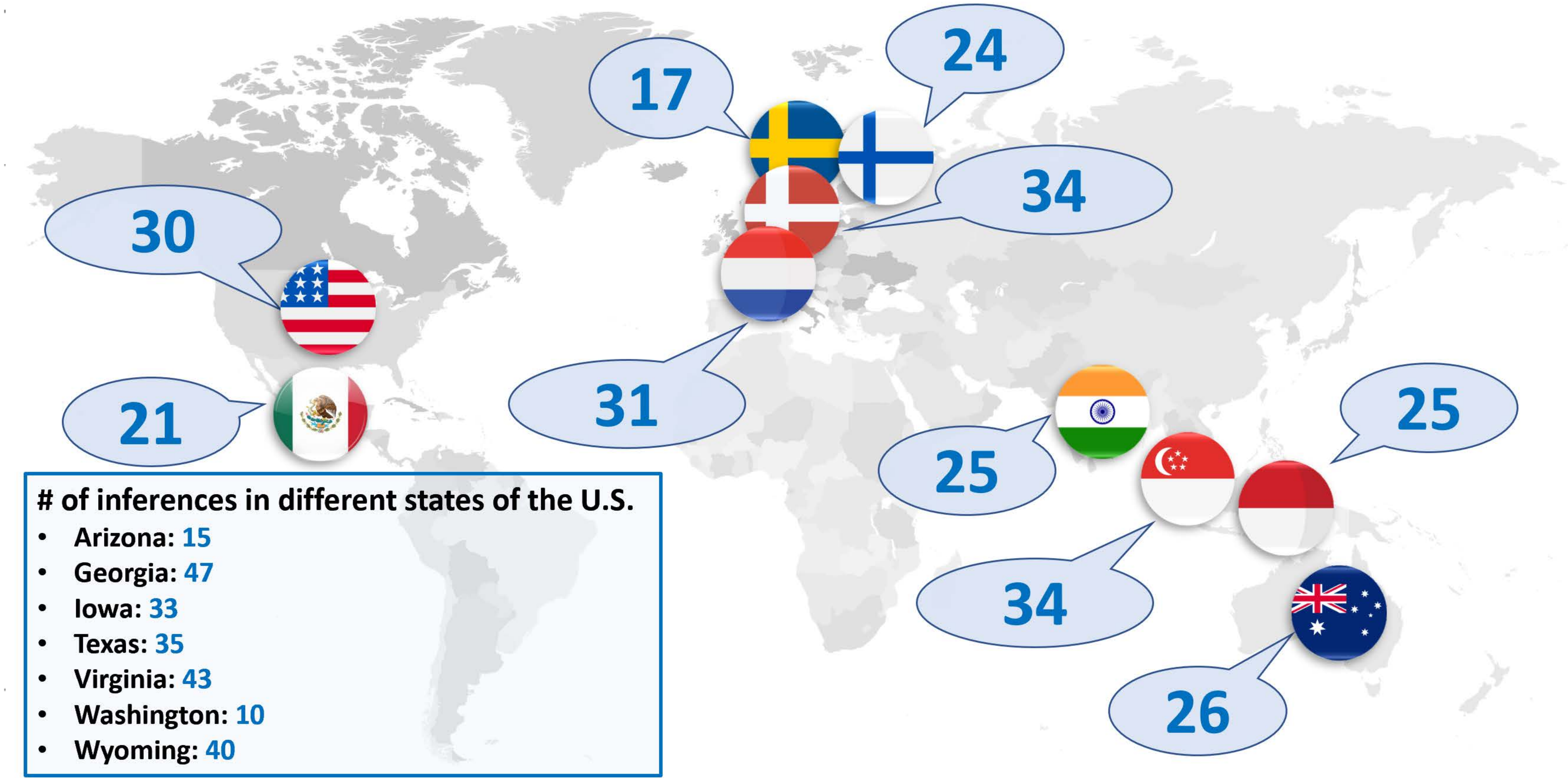
Hourly carbon efficiency and total WUE for the first week of August 2022.

Estimated # of GPT-3 response for 500mL water

Location	PUE	WUE (L/kWh)	EWIF (L/kWh)	Water for Training (Million L)	Water for Each Request (mL)	# of Requests for 500ml Water
U.S. Average	1.17	0.55	3.142	5.439	16.904	29.6
Wyoming	1.125	0.23	2.574	4.023	12.503	40
Iowa	1.16	0.19	3.104	4.879	15.163	33
Arizona	1.223	2.24	4.959	10.688	33.219	15.1
Washington	1.156	1.09	9.501	15.539	48.294	10.4
Virginia	1.144	0.17	2.385	3.73	11.593	43.1
Texas	1.307	1.82	1.287	4.507	14.009	35.7
Singapore	1.358	2.06	1.199	4.747	14.753	33.9
Ireland	1.197	0.03	1.476	2.313	7.189	69.6
Netherlands	1.158	0.08	3.445	5.237	16.276	30.7
Sweden	1.172	0.16	6.019	9.284	28.856	17.3

Pengfei Li, Jianyi Yang, Mohammad A. Islam, Shaolei Ren, "Making AI Less 'Thirsty': Uncovering and Addressing the Secret Water Footprint of AI Models," 2023.

Estimated # of GPT-3 response for 500mL water



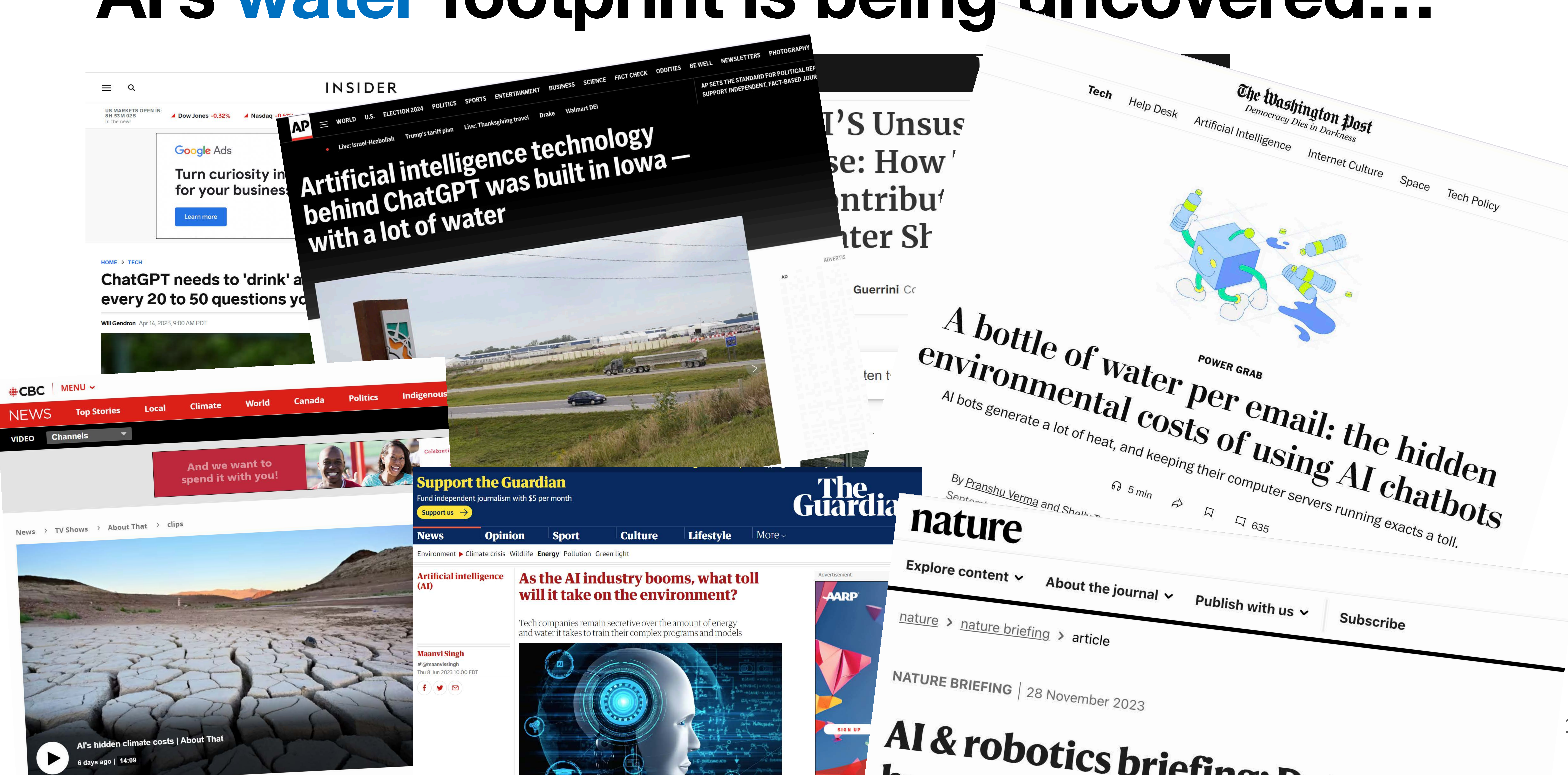
# of inferences in different states of the U.S.

- **Arizona:** 15
- **Georgia:** 47
- **Iowa:** 33
- **Texas:** 35
- **Virginia:** 43
- **Washington:** 10
- **Wyoming:** 40

Pengfei Li, Jianyi Yang, Mohammad A. Islam, Shaolei Ren, "Making AI Less 'Thirsty': Uncovering and Addressing the Secret Water Footprint of AI Models," 2023.

AI's water footprint is being uncovered…

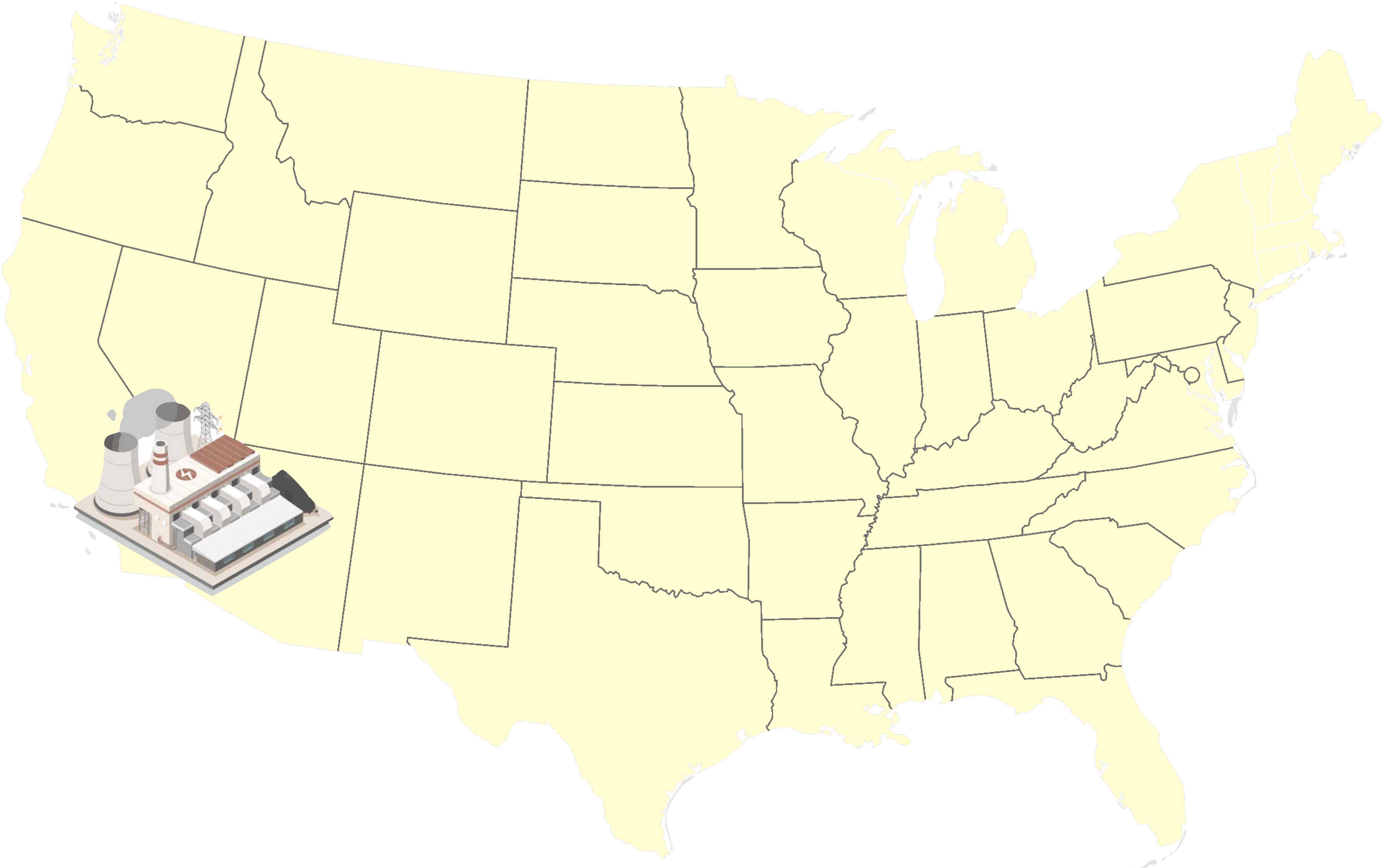
AI’s water footprint is being uncovered…



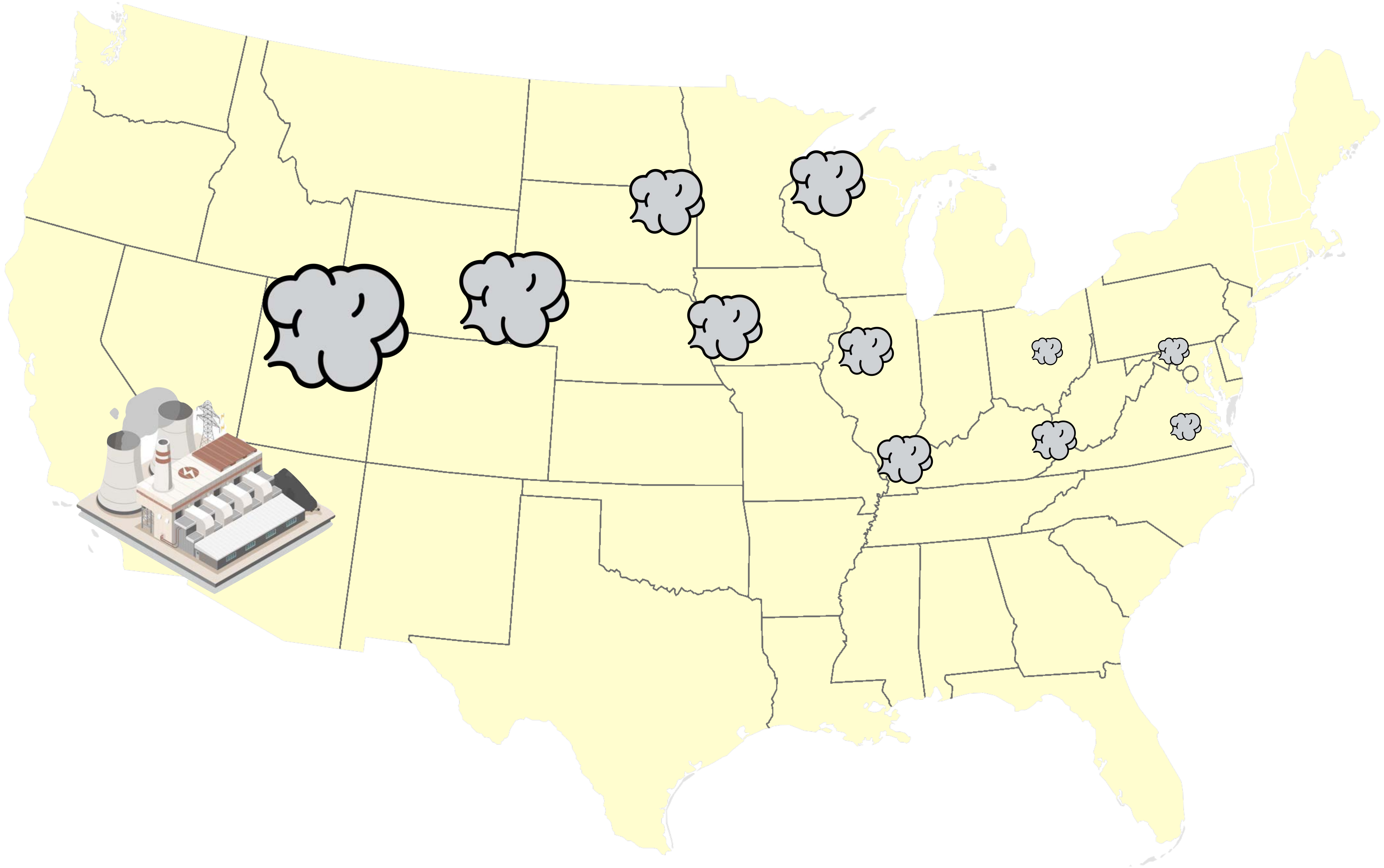
Health risk of AI computing

Health risk of AI computing

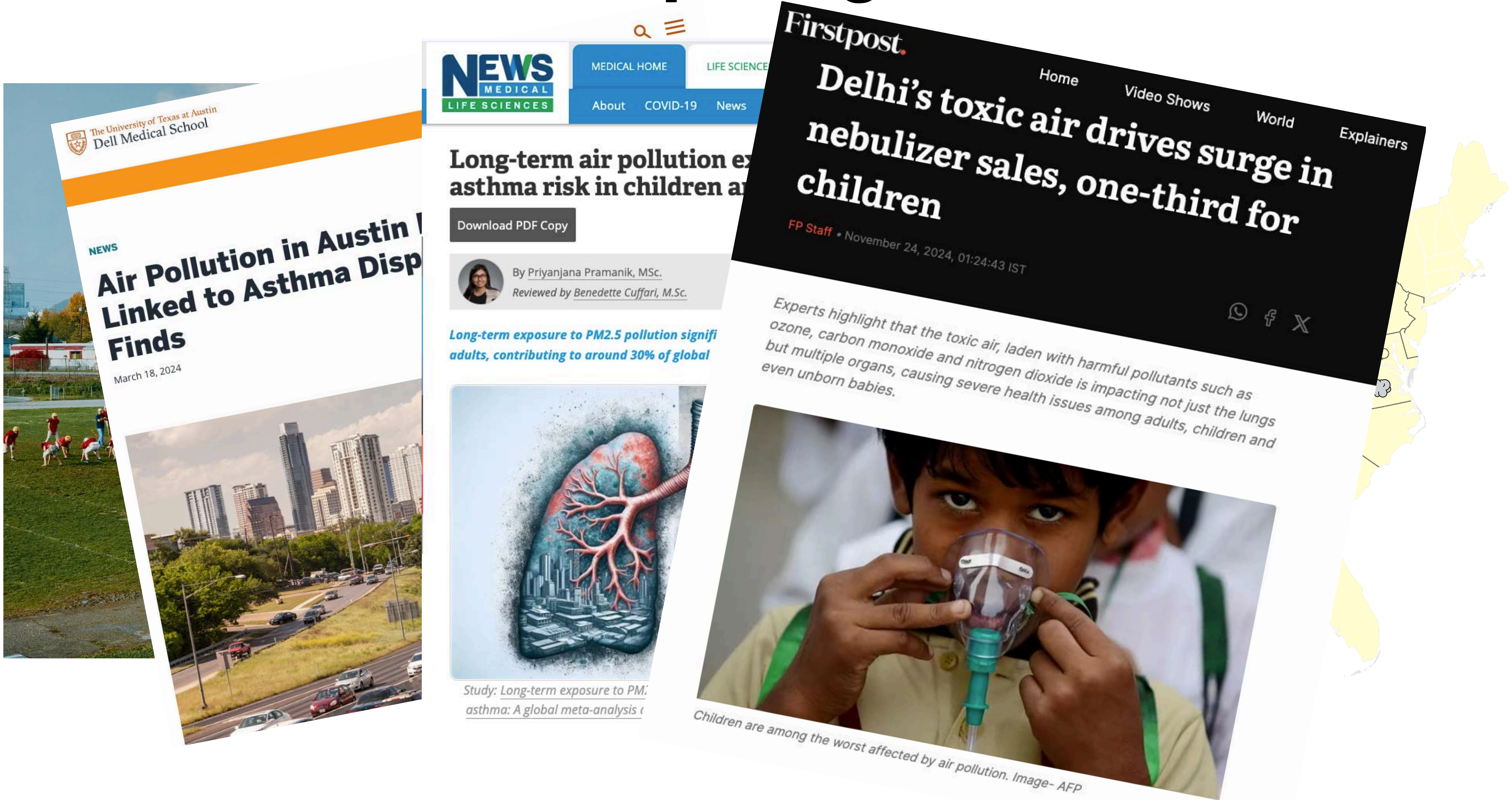
Health risk of AI computing



Health risk of AI computing



Health risk of AI computing



What's under the hood

Carbon

Water

Air Pollution

Intelligent AI workload management

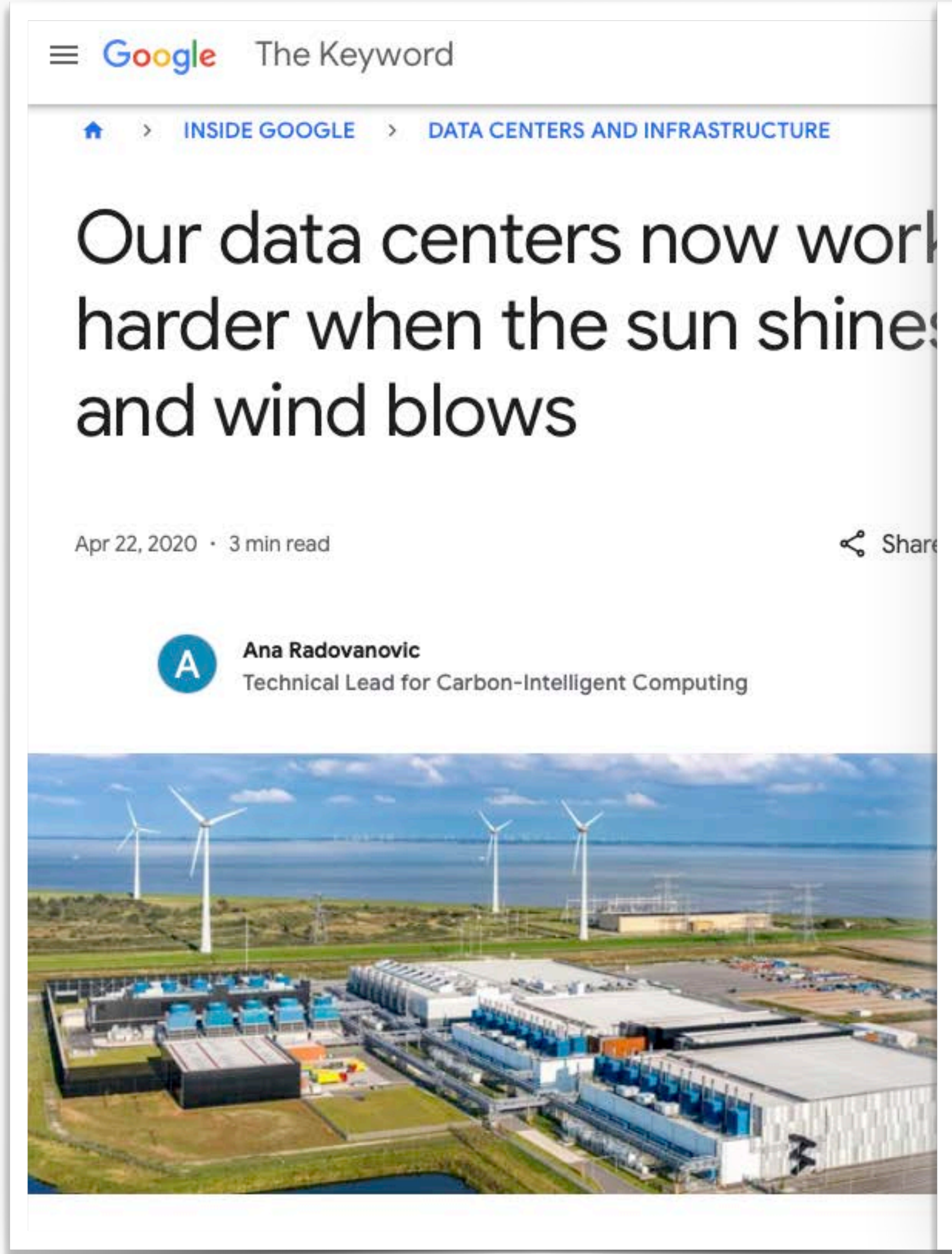
Example: carbon-ware computing

Intelligent AI workload management

Example: carbon-ware computing

Intelligent AI workload management

Example: carbon-ware computing



Google The Keyword

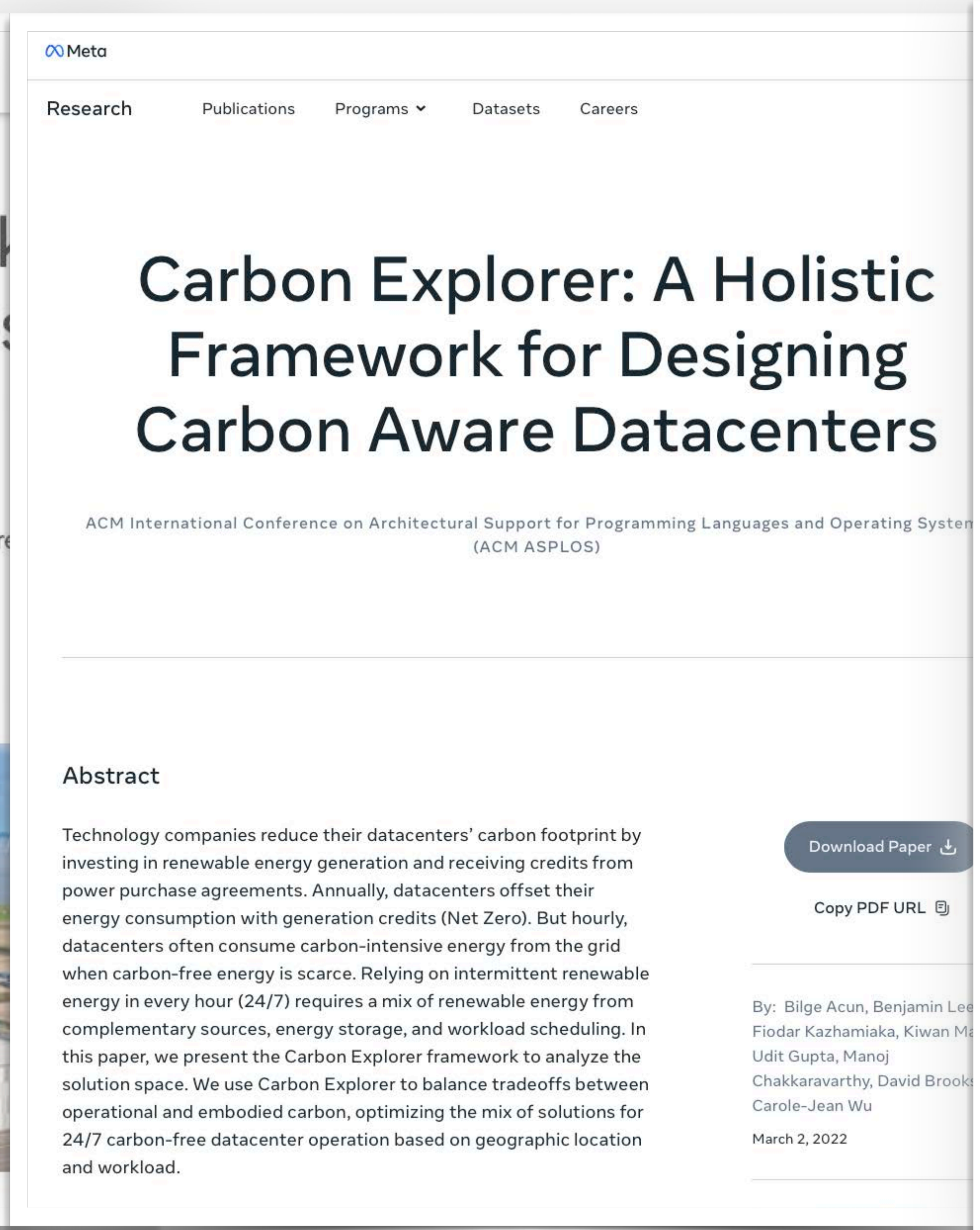
INSIDE GOOGLE > DATA CENTERS AND INFRASTRUCTURE

Our data centers now work harder when the sun shines and wind blows

Apr 22, 2020 · 3 min read

Share

Ana Radovanovic
Technical Lead for Carbon-Intelligent Computing



Meta

Research Publications Programs Datasets Careers

Carbon Explorer: A Holistic Framework for Designing Carbon Aware Datacenters

ACM International Conference on Architectural Support for Programming Languages and Operating System (ACM ASPLOS)

Abstract

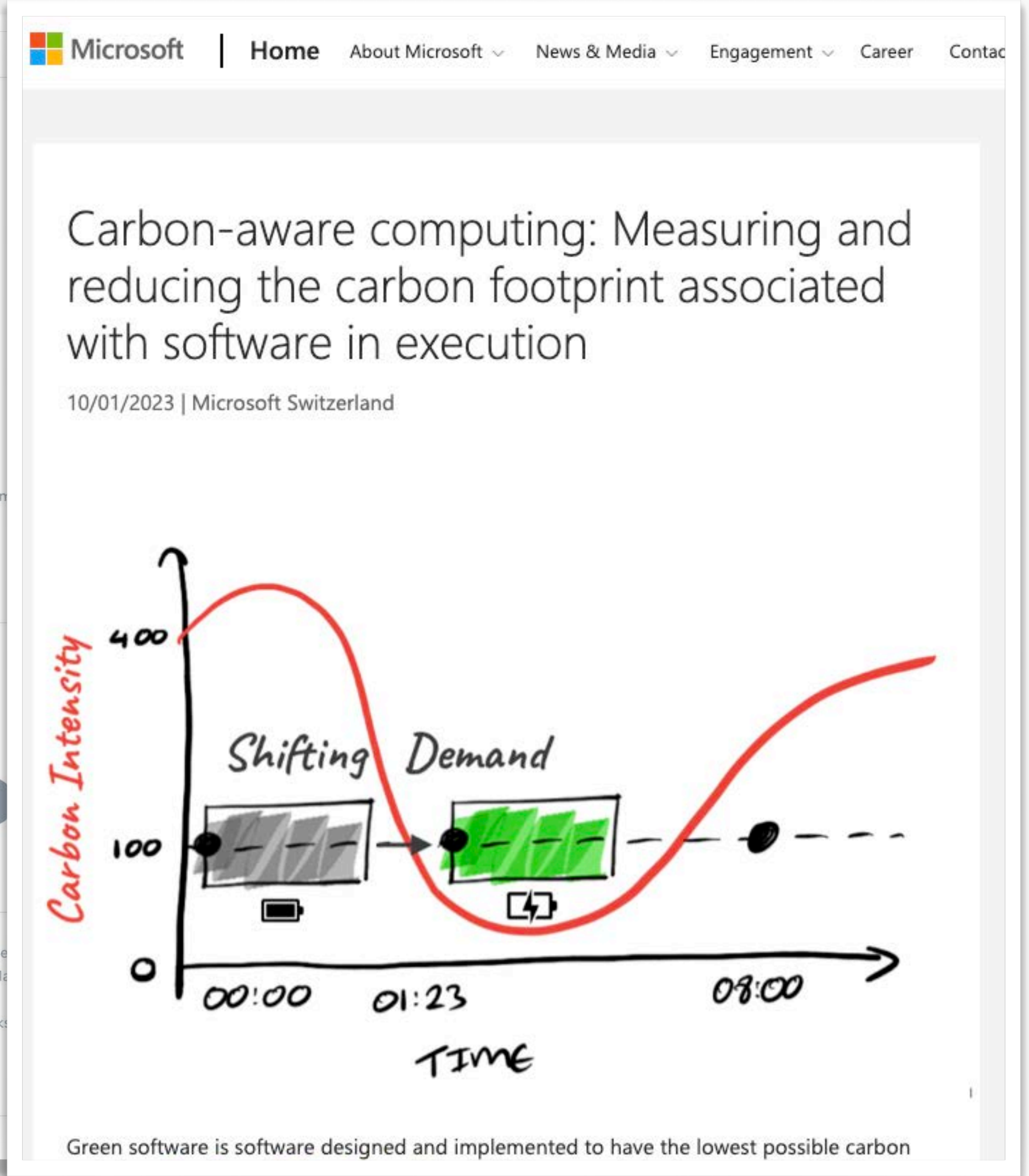
Technology companies reduce their datacenters' carbon footprint by investing in renewable energy generation and receiving credits from power purchase agreements. Annually, datacenters offset their energy consumption with generation credits (Net Zero). But hourly, datacenters often consume carbon-intensive energy from the grid when carbon-free energy is scarce. Relying on intermittent renewable energy in every hour (24/7) requires a mix of renewable energy from complementary sources, energy storage, and workload scheduling. In this paper, we present the Carbon Explorer framework to analyze the solution space. We use Carbon Explorer to balance tradeoffs between operational and embodied carbon, optimizing the mix of solutions for 24/7 carbon-free datacenter operation based on geographic location and workload.

Download Paper

Copy PDF URL

By: Bilge Acun, Benjamin Lee, Fiodar Kazhamiaka, Kiwan Ma, Udit Gupta, Manoj Chakkaravarthy, David Brooks, Carole-Jean Wu

March 2, 2022



Microsoft | Home About Microsoft News & Media Engagement Career Contac

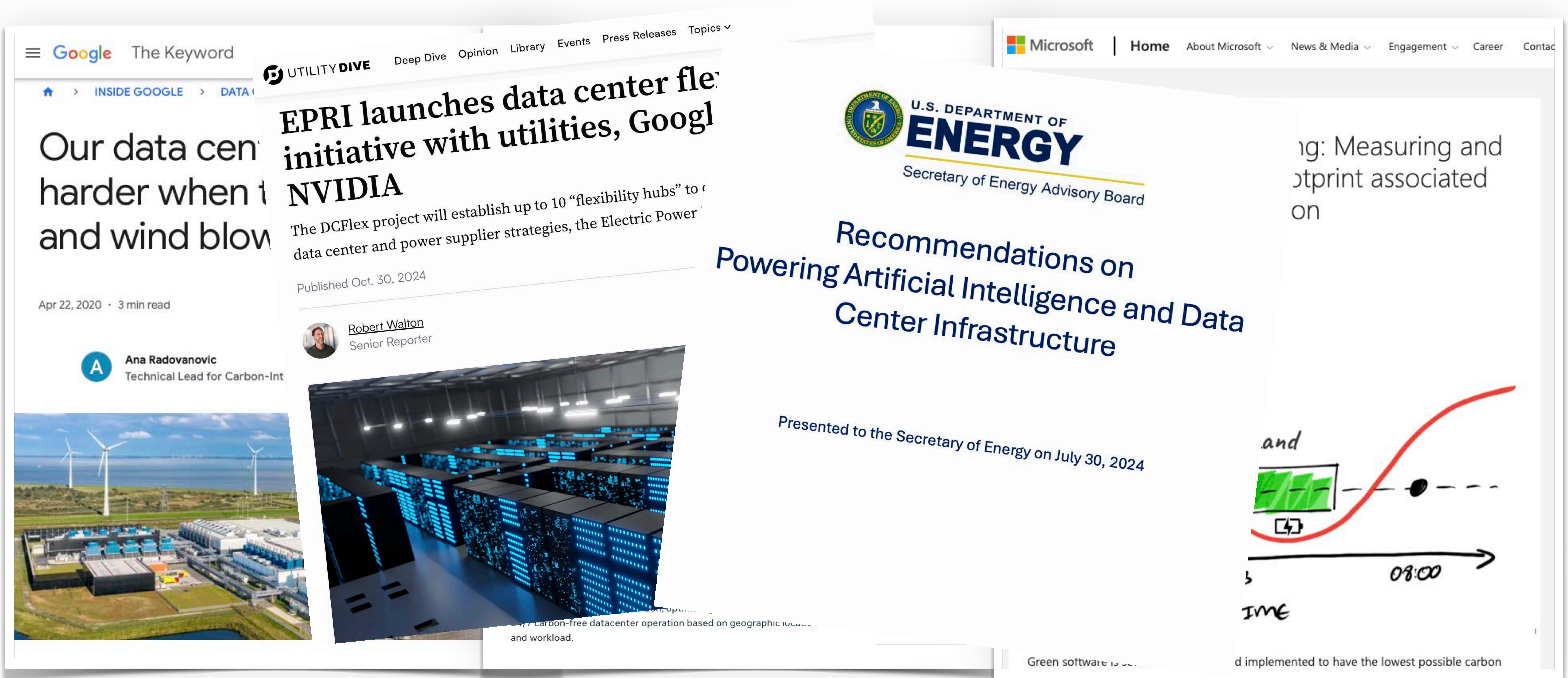
Carbon-aware computing: Measuring and reducing the carbon footprint associated with software in execution

10/01/2023 | Microsoft Switzerland

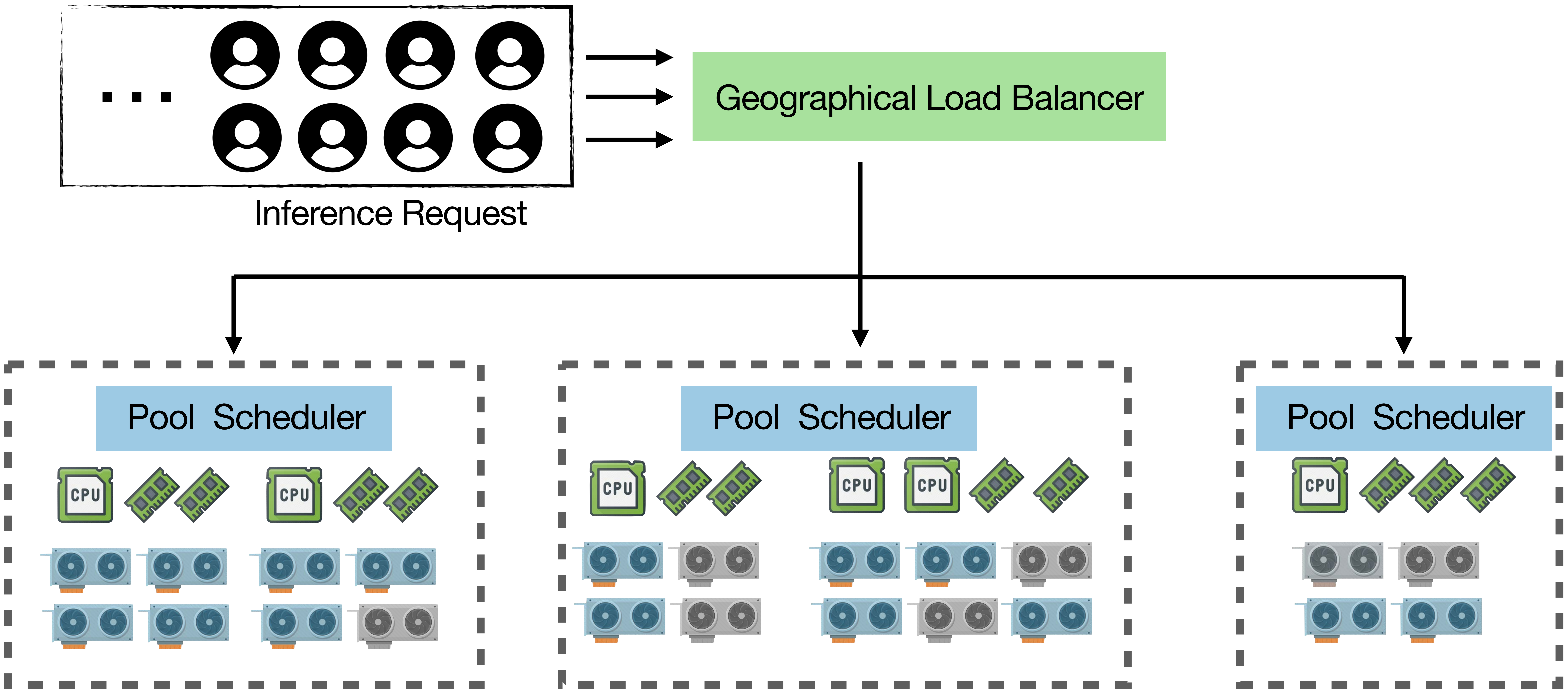
Green software is software designed and implemented to have the lowest possible carbon

Intelligent AI workload management

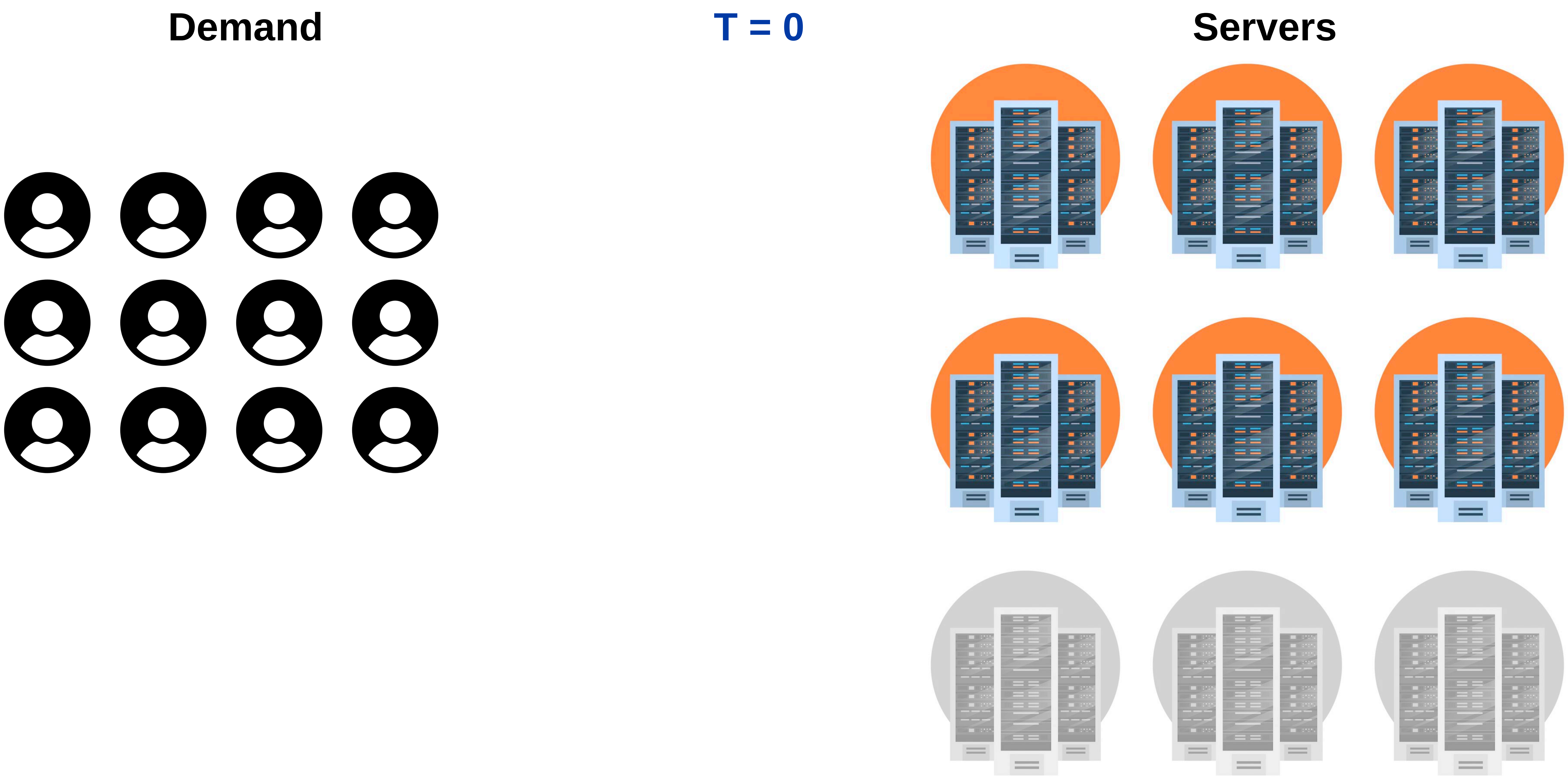
Example: carbon-ware computing



Dynamic server provisioning for LLM

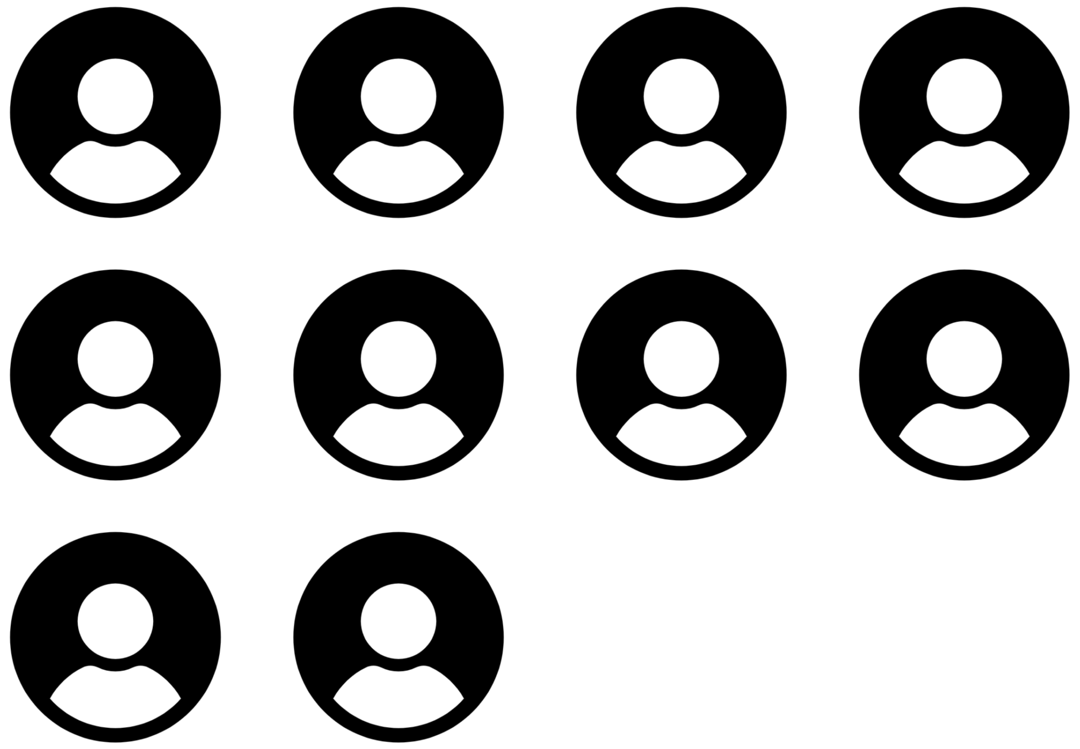


Dynamic server provisioning



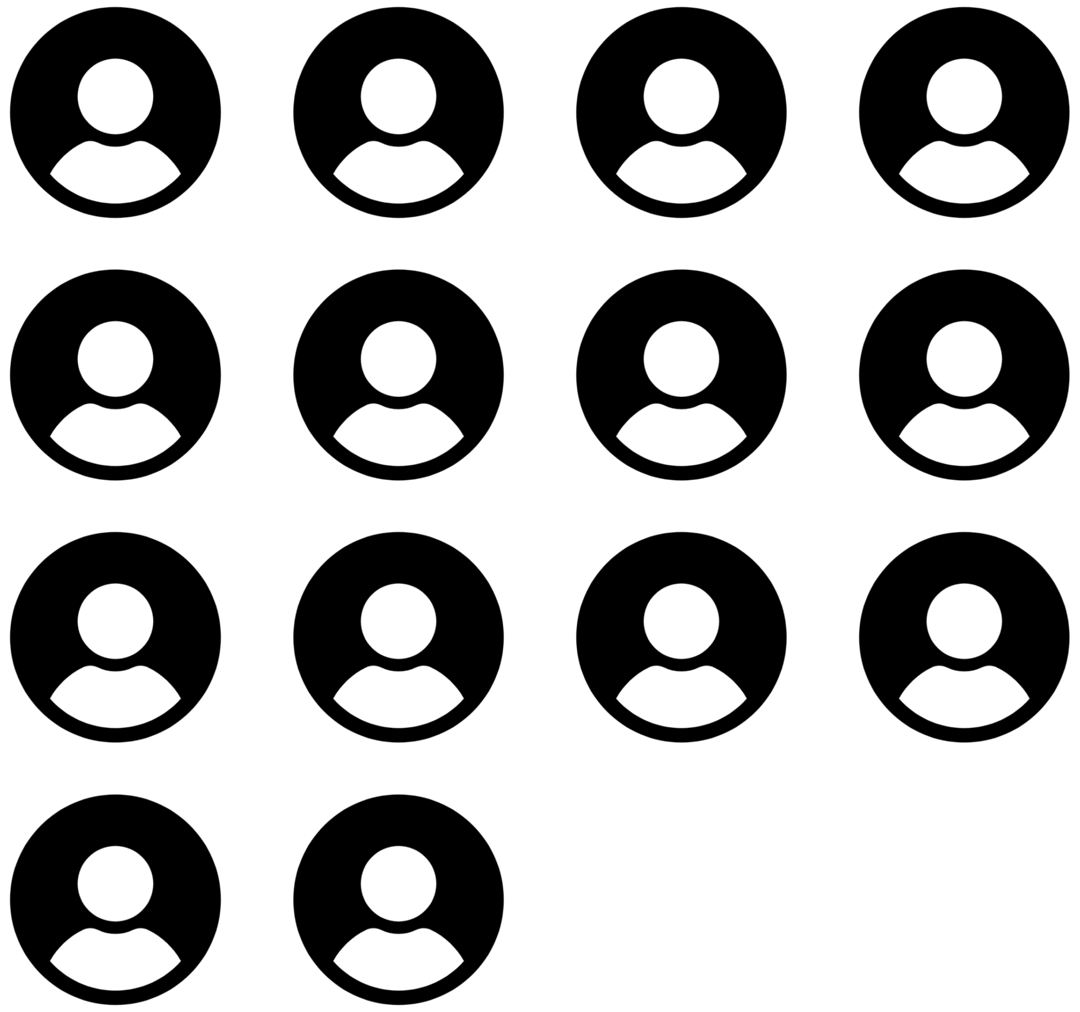
Dynamic server provisioning

Demand **T = 1** **Servers**



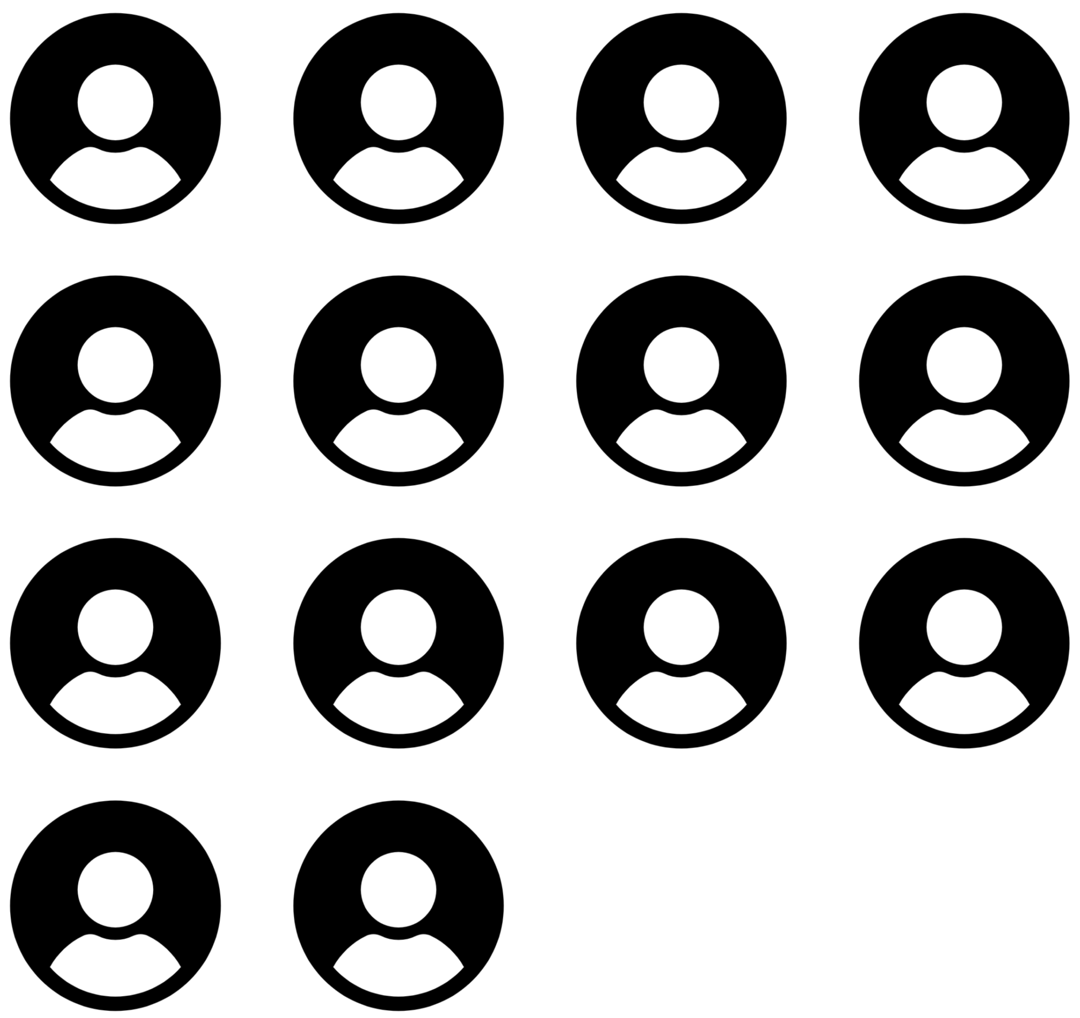
Dynamic server provisioning

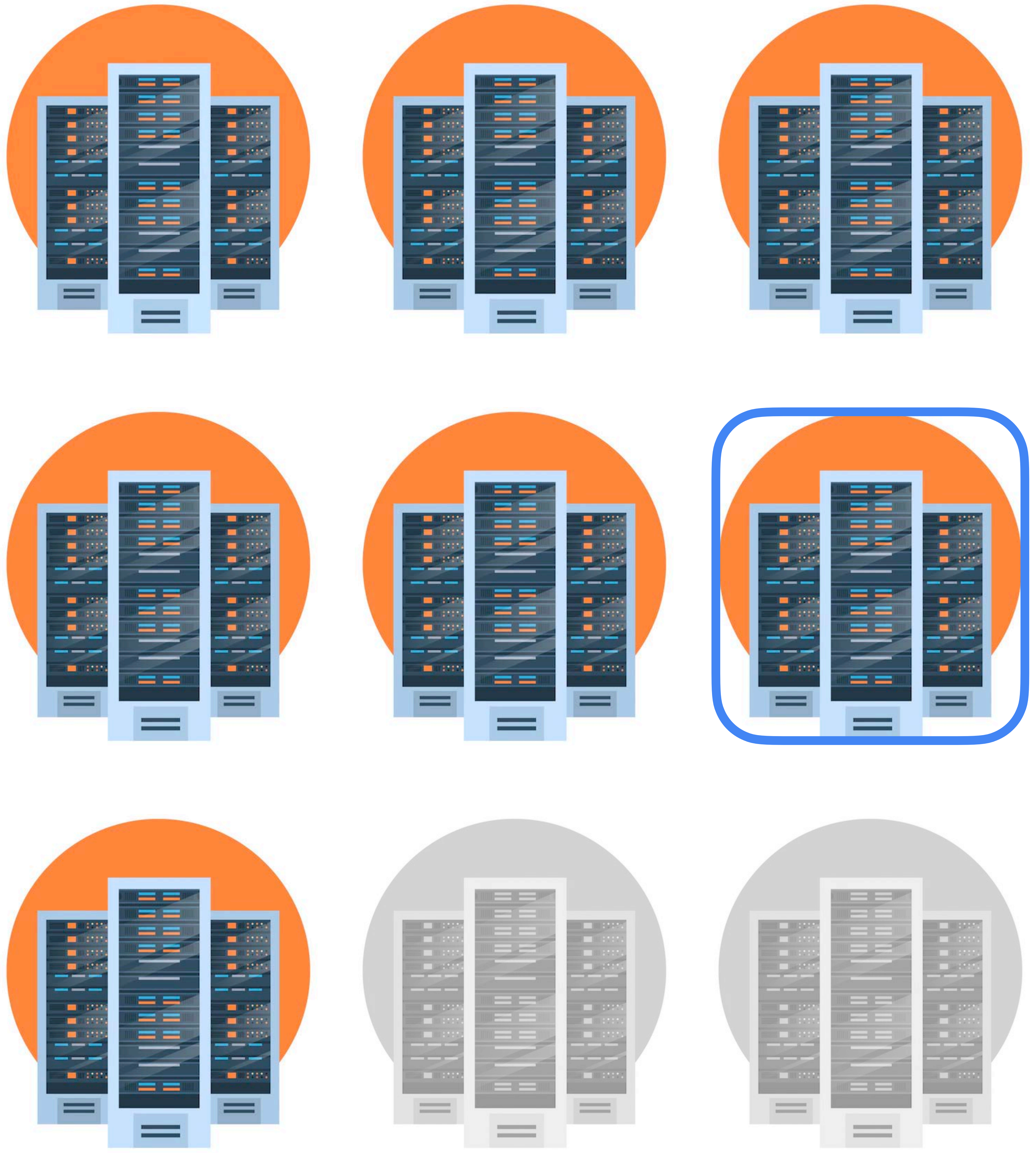
Demand **T = 2** **Servers**



Dynamic server provisioning

Demand **T = 2** **Servers**





Responsbile AI computing

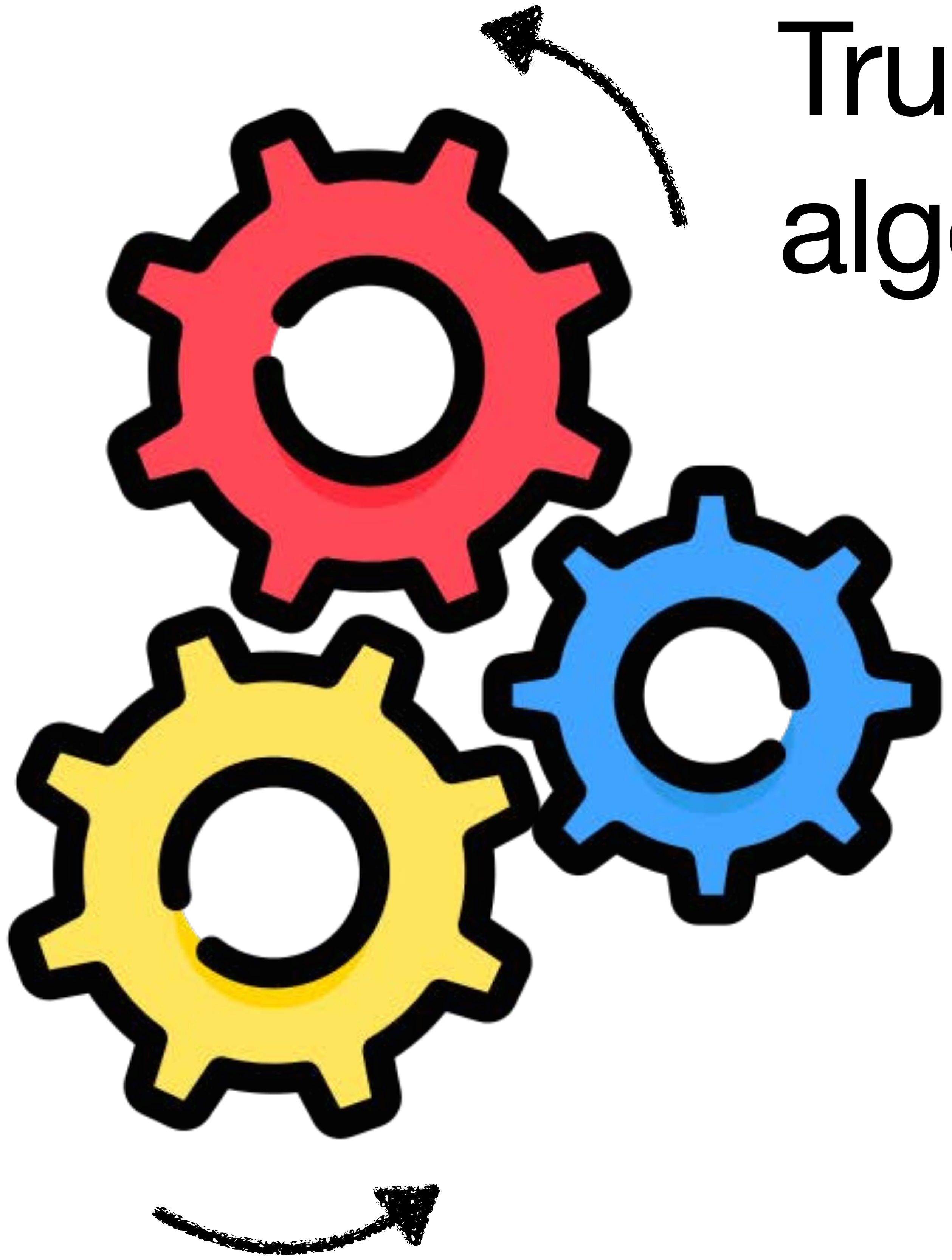
Three main component

Sustainable AI

[**_L_**YIR, CACM’ 23]
[GH**_L_**IR, eEnergy’ 24]
[**_L_**LWR, HotCarbon’ 24]
[L**_L_**WR, HotEthics’ 24]

Equitable AI system

[LY**_L_**LR, ICML’ 24]
[**_L_**YWR, eEnergy’ 24]



Trustworthy ML-augmented algorithms

[**_L_**YR, SIGMETRICS’ 22]
[Y**_L_**R, NeurIPS’ 23]
[**_L_**YR, NeurIPS’ 23]
[**_L_**YR, ICML’ 23]
[**_L_**YR, INFOCOM’ 23]
[Y**_L_**IR, SIGMETRICS’ 24]
[**_L_**YWR, SIGMETRICS’ 25]

Smooth Online Convex Optimization (SOCO)

Problem formulation

Smoothed Online Convex Optimization (SOCO)

Goal

$$\min_{x_t \in \mathbb{X}} \sum_{t=1}^{T} f(x_t, y_t) + c(x_t, x_{t-1})$$

[1] The switching cost can also be written as $c(x_t, x_{t-p:t-1})$ to encode a multi-step structured memory cost.

Problem formulation

Smoothed Online Convex Optimization (SOCO)

Goal

Hitting cost

$$\min_{x_t \in \mathbb{X}} \sum_{t=1}^{T} f(x_t, y_t) + c(x_t, x_{t-1})$$

[1] The switching cost can also be written as $c(x_t, x_{t-p:t-1})$ to encode a multi-step structured memory cost.

Problem formulation

Smoothed Online Convex Optimization (SOCO)

Goal

$$\min_{x_t \in \mathbb{X}} \sum_{t=1}^{T} \overbrace{f(x_t, y_t)}^{\text{Hitting cost}} + \overbrace{c(x_t, x_{t-1})}^{\text{Switching cost}^{[1]}}$$

[1] The switching cost can also be written as $c(x_t, x_{t-p:t-1})$ to encode a multi-step structured memory cost.

Problem formulation

Smoothed Online Convex Optimization (SOCO)

Goal

Hitting cost Switching cost[1]

$$\min_{x_t \in \mathbb{X}} \sum_{t=1}^{T} f(x_t, y_t) + c(x_t, x_{t-1})$$

Online Decision Making

$y_1, x_1, y_2, x_2, y_3, x_3 \ldots$

[1] The switching cost can also be written as $c(x_t, x_{t-p:t-1})$ to encode a multi-step structured memory cost.

Problem formulation

Smoothed Online Convex Optimization (SOCO)

Goal

$$\min_{x_t \in \mathbb{X}} \sum_{t=1}^{T} \underbrace{f(x_t, y_t)}_{\text{Hitting cost}} + \underbrace{c(x_t, x_{t-1})}_{\text{Switching cost[1]}}$$

Online Decision Making

$y_1, x_1, y_2, x_2, y_3, x_3 \ldots$

Metrics

$$\boxed{\mathrm{AVG}(\pi) = \mathbb{E}\left[\mathrm{cost}(\pi, s)\right]}$$

Average Cost

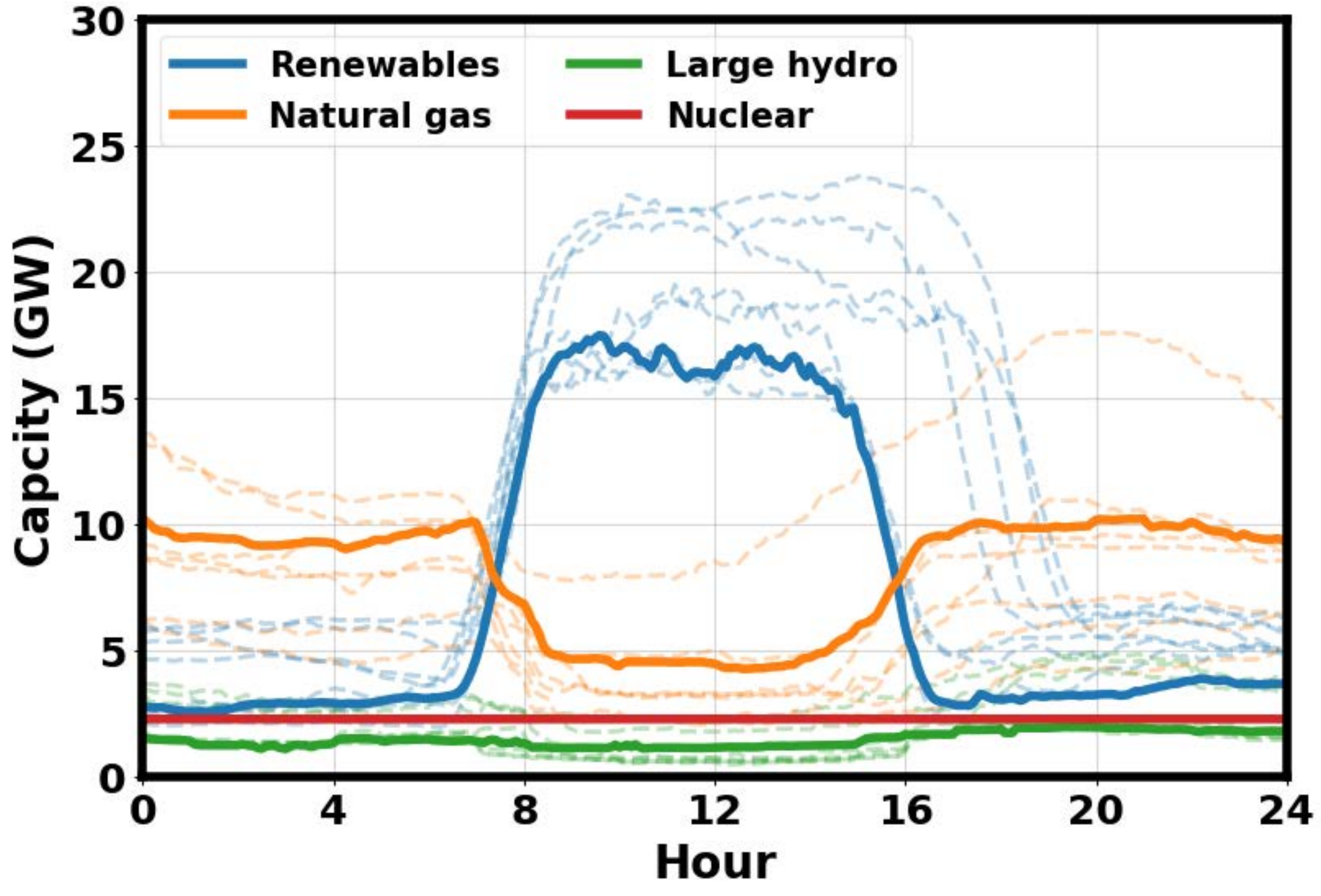
$$\boxed{\mathrm{CR}(\pi) = \sup_{s \in \mathcal{S}} \frac{\mathrm{cost}(\pi, s)}{\mathrm{cost}(\pi^*, s)}}$$

Competitive Ratio

[1] The switching cost can also be written as $c(x_t, x_{t-p:t-1})$ to encode a multi-step structured memory cost.

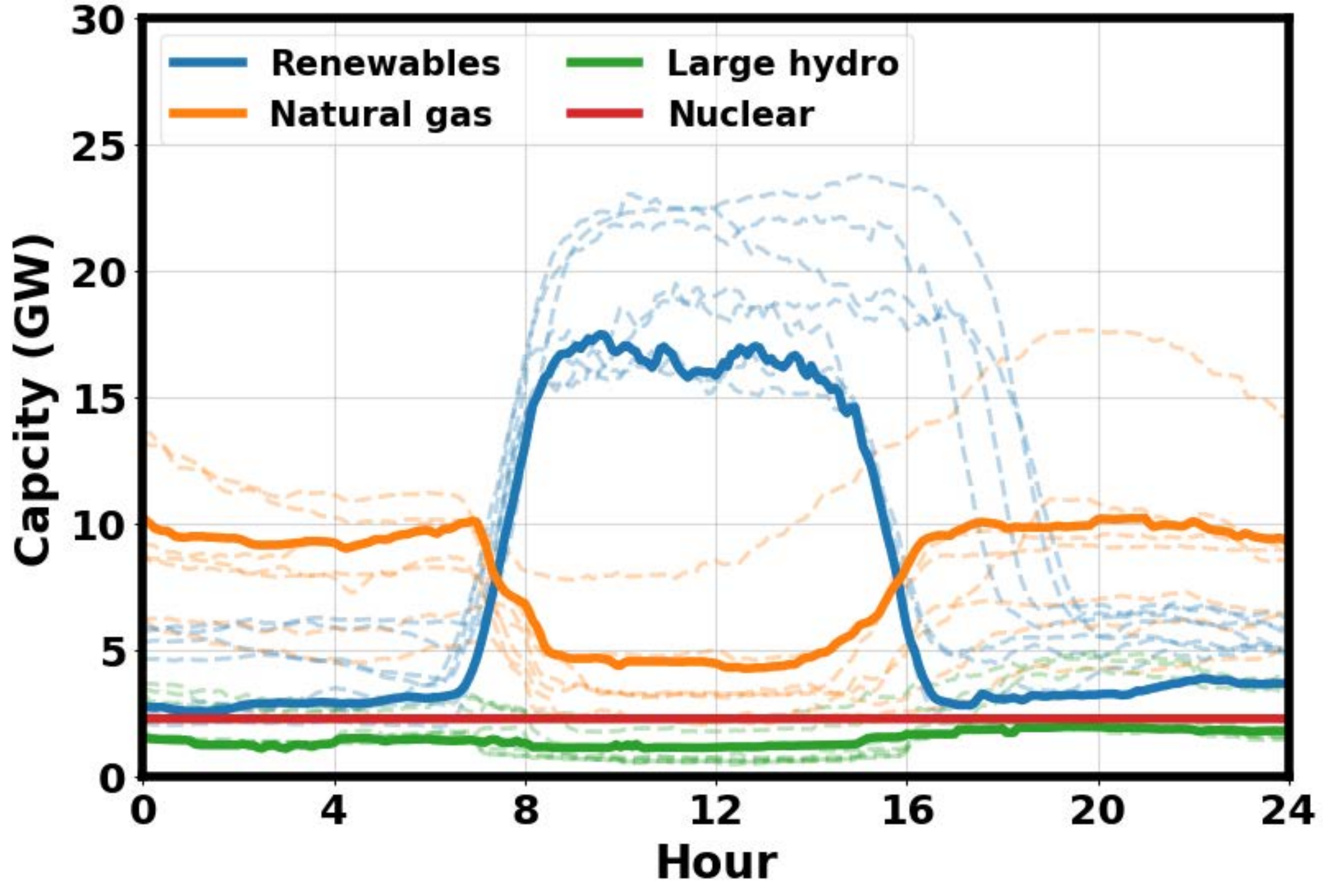
A quick example in responsible AI computing

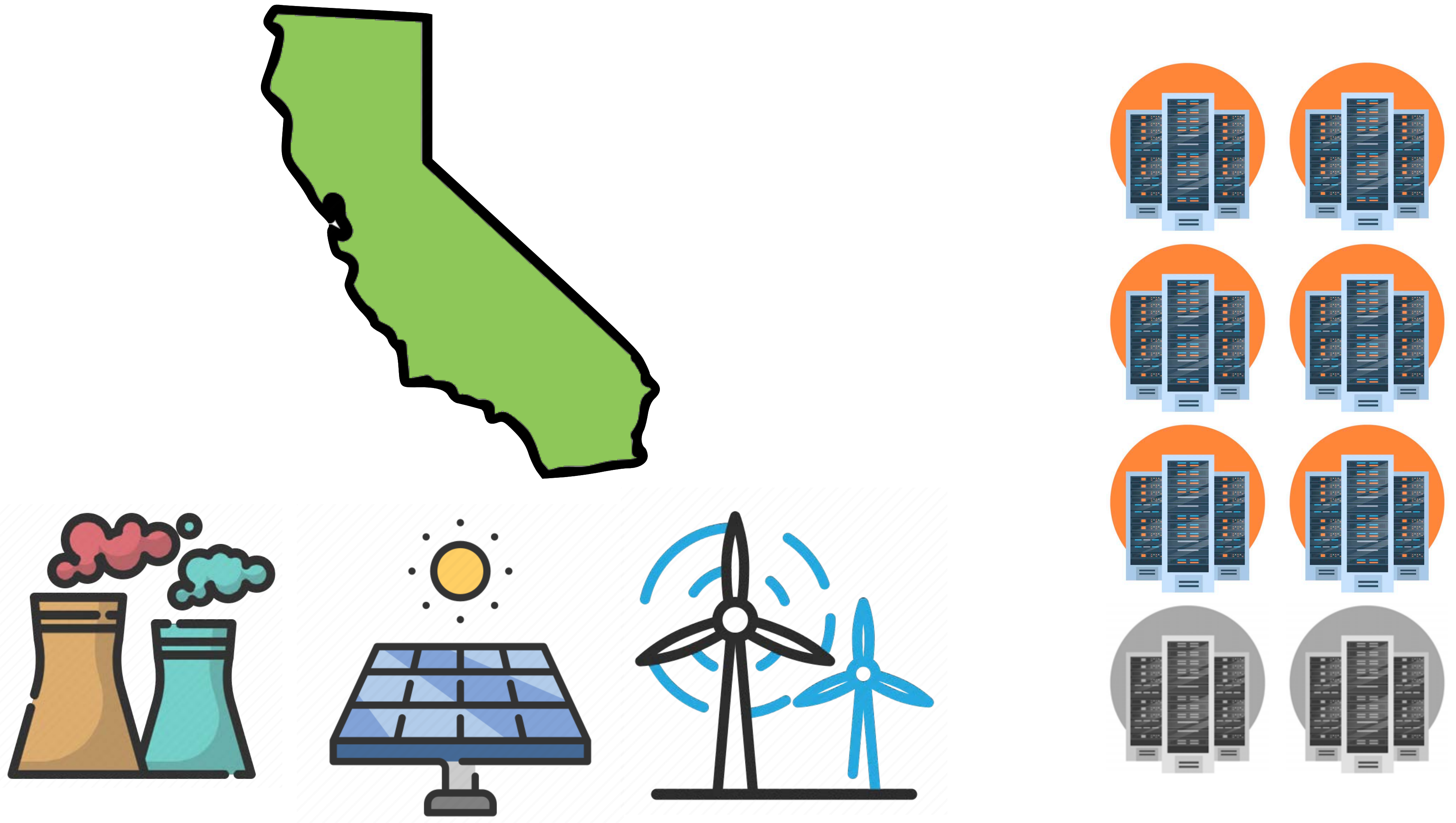
California ISO



A quick example in responsible AI computing

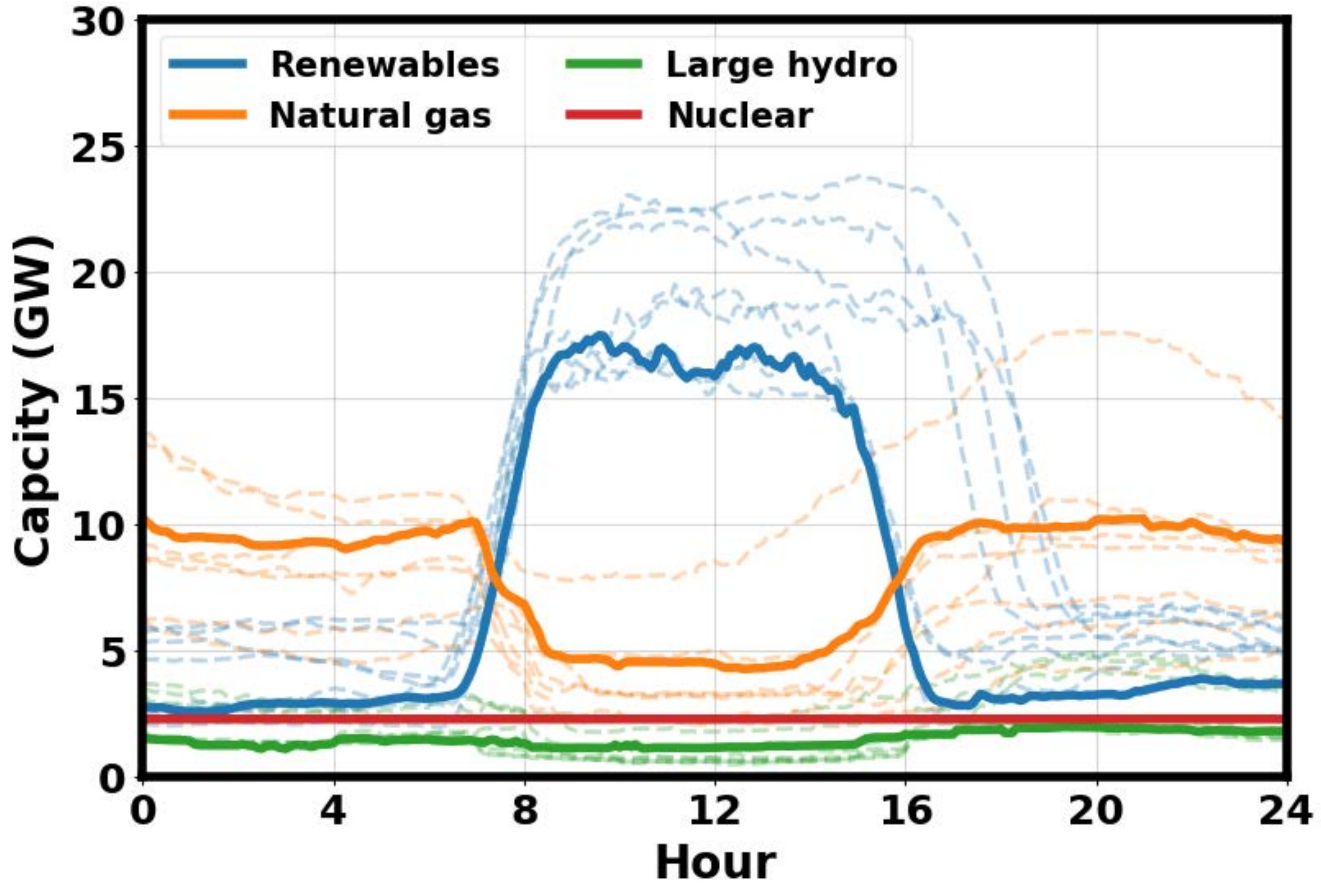
California ISO

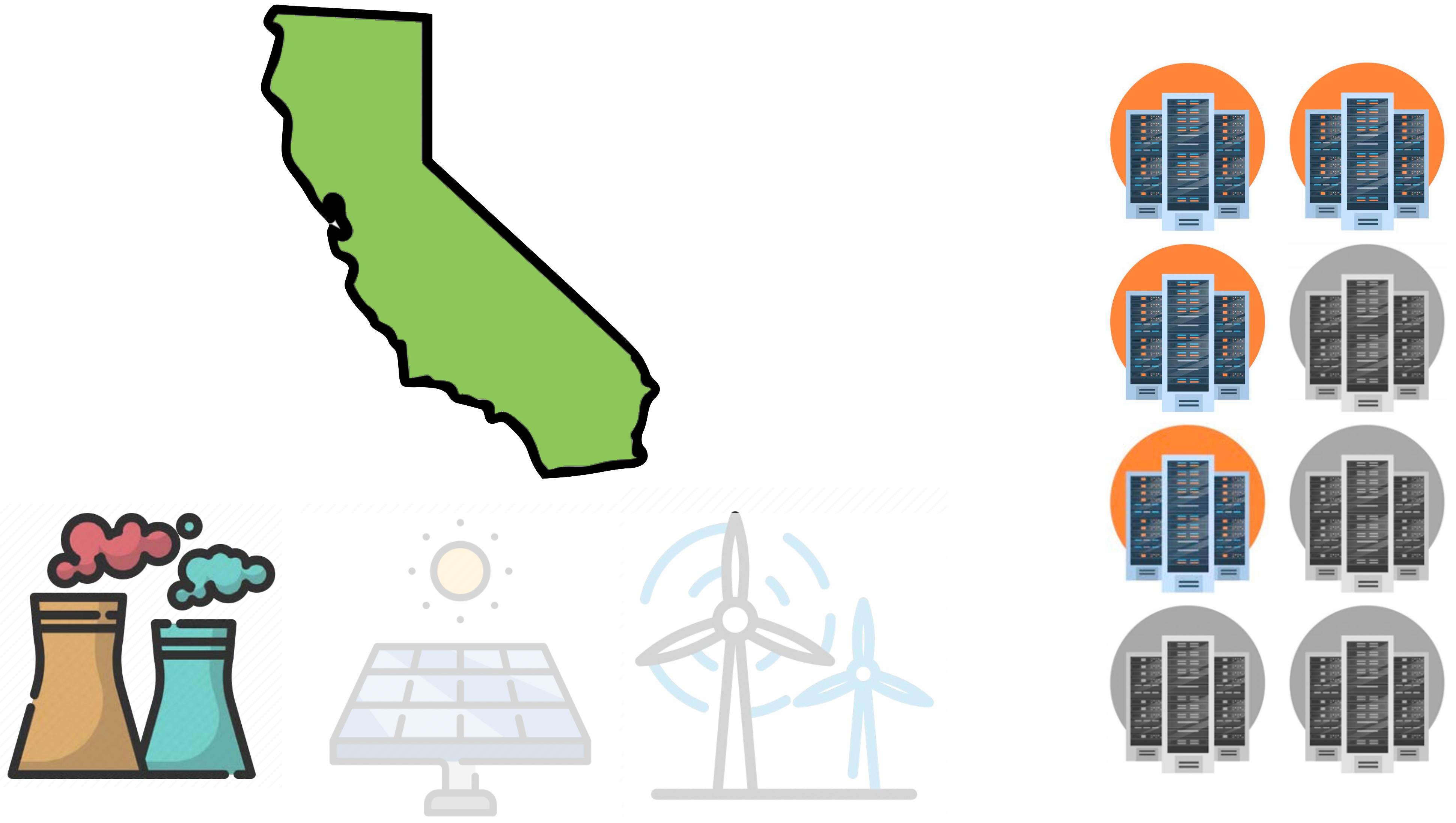




A quick example in responsible AI computing

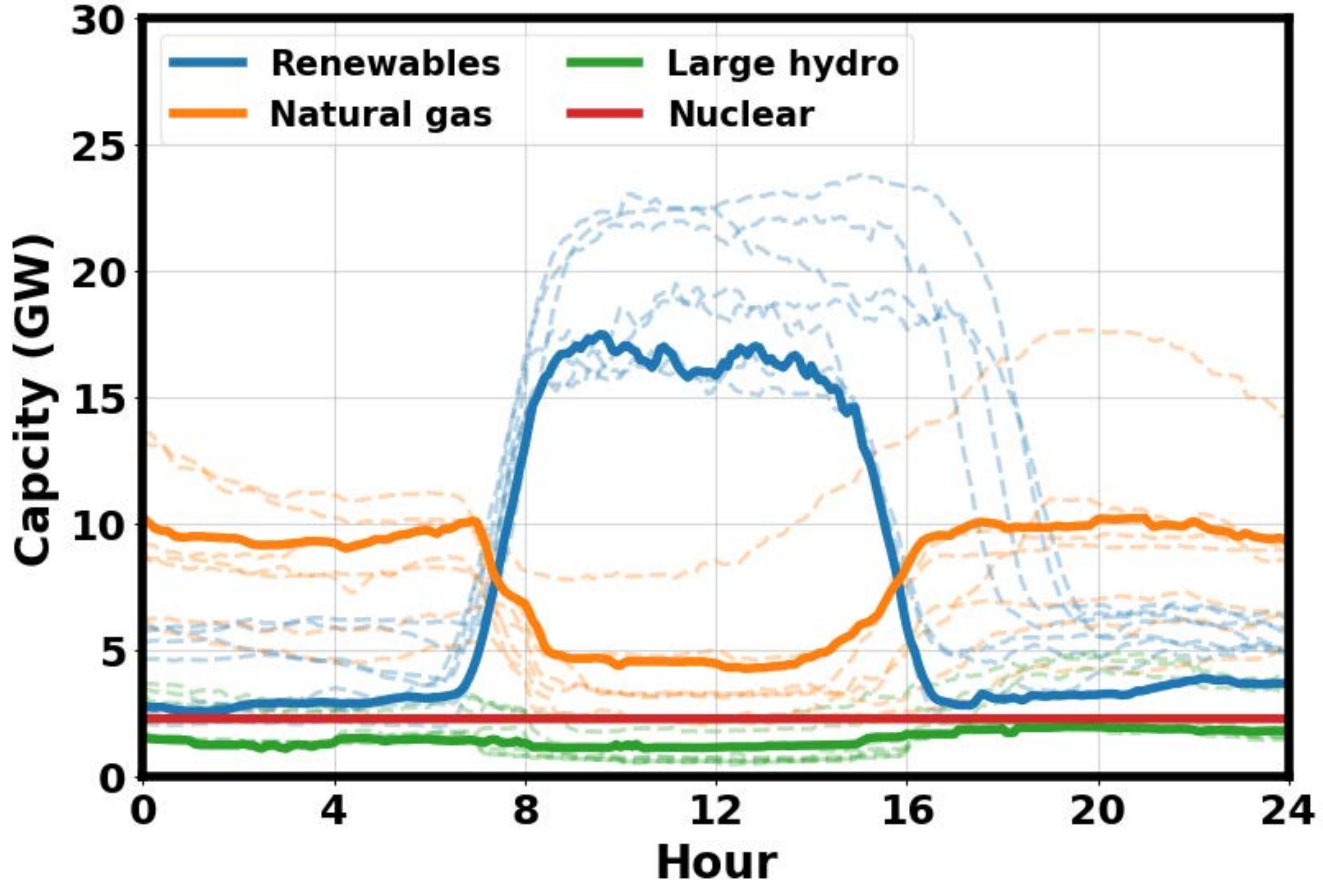
California ISO

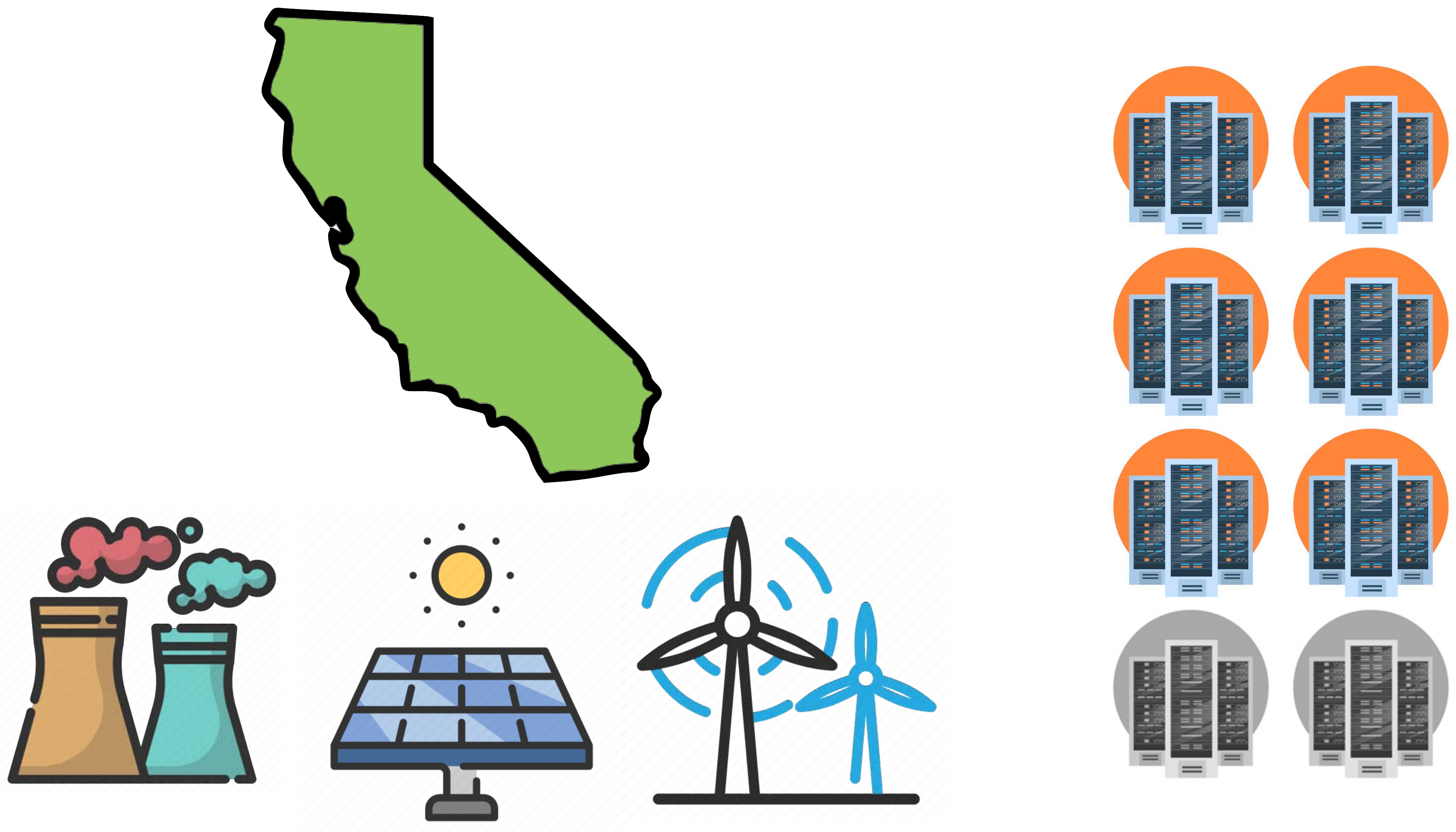




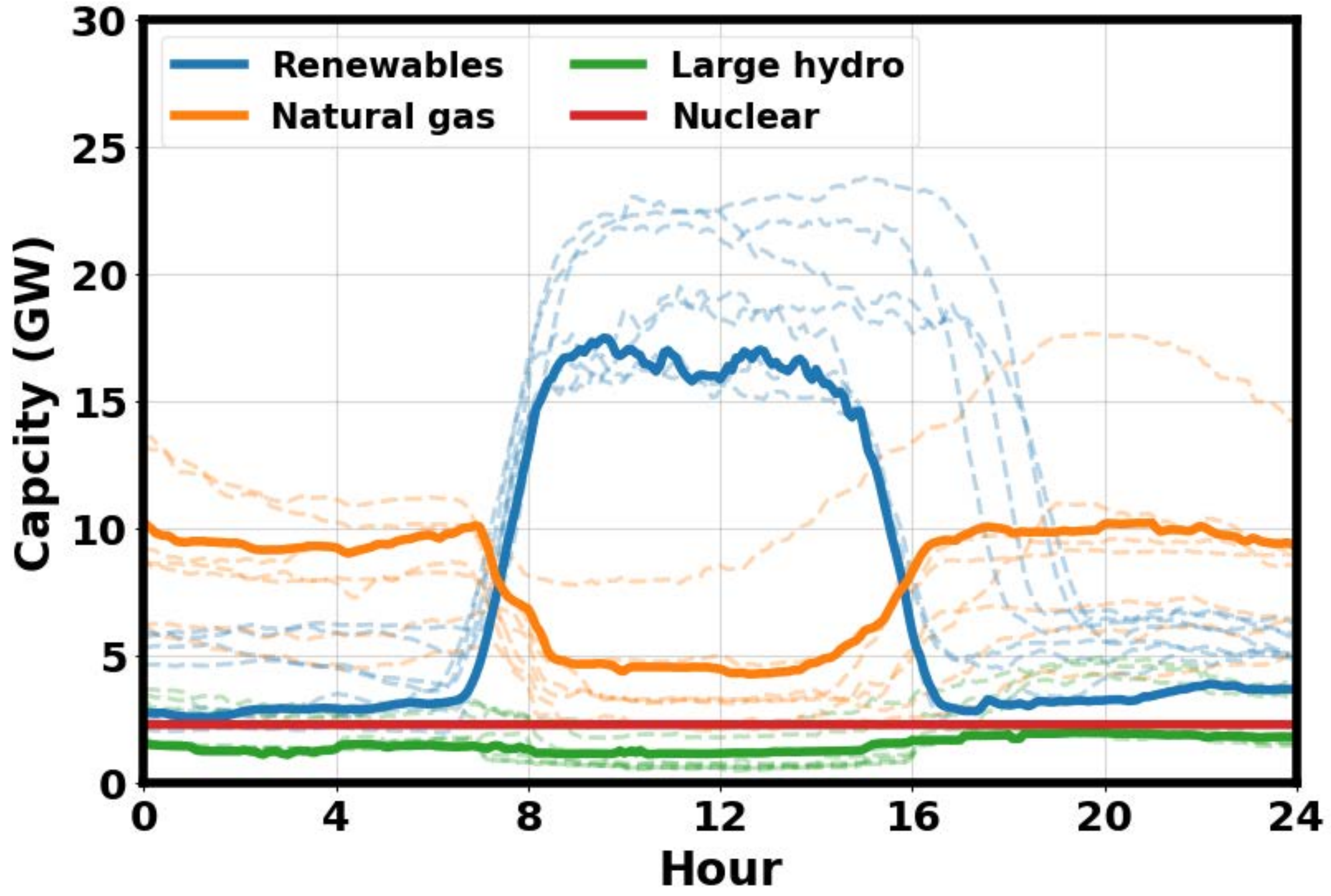
A quick example in responsible AI computing

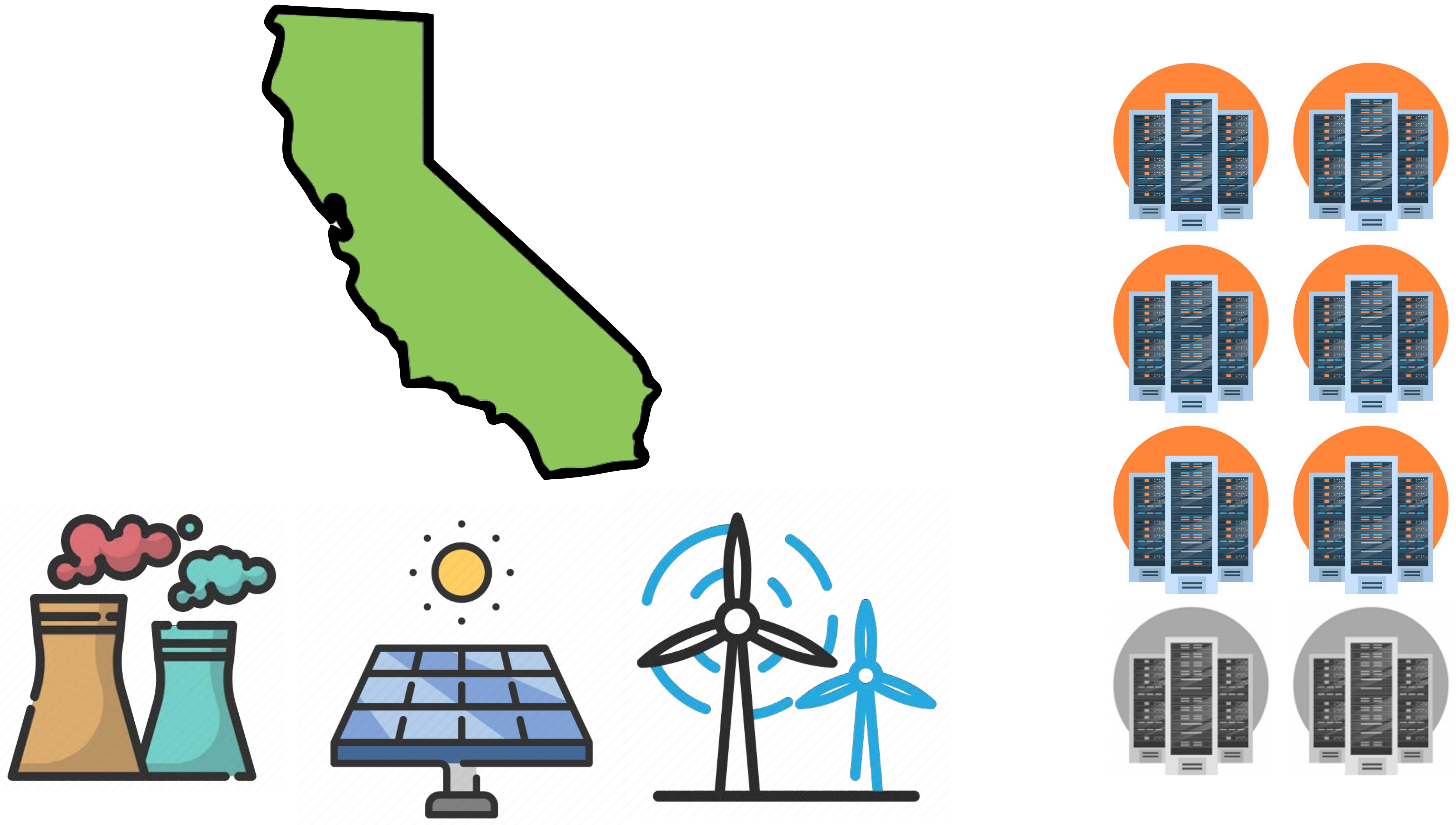
California ISO





A quick example in responsible AI computing

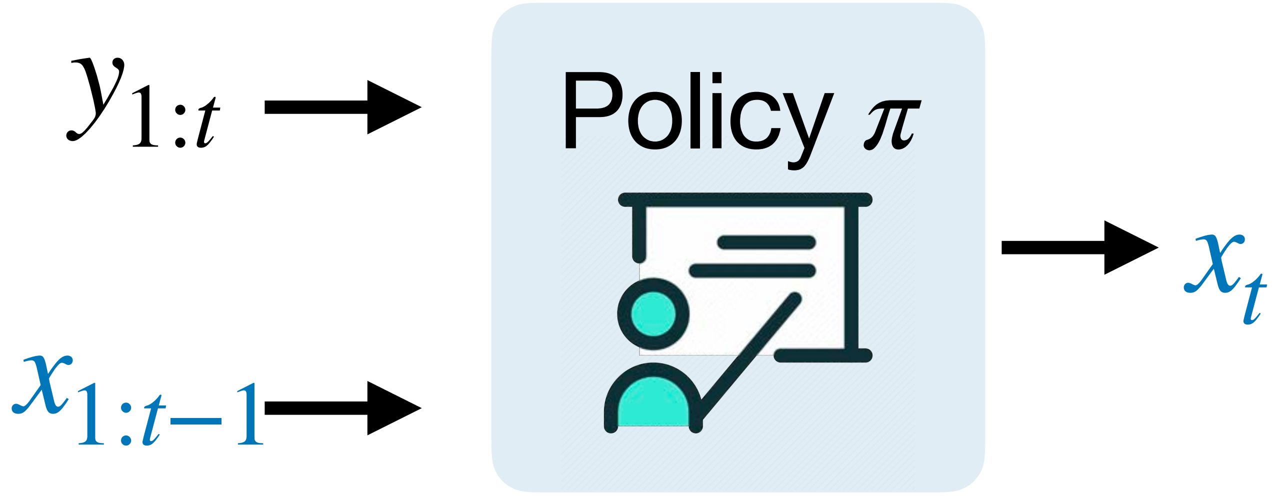




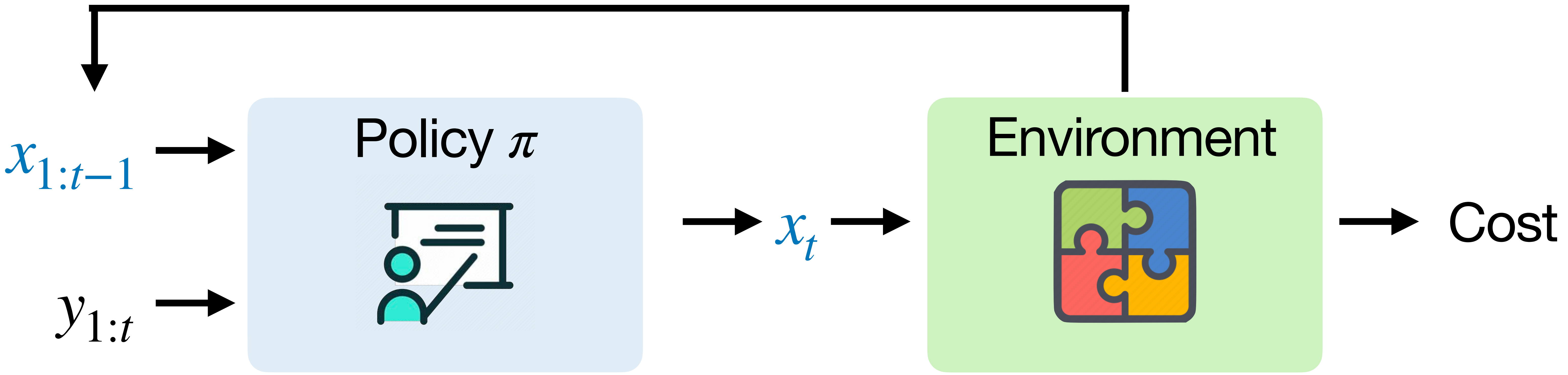
A Greedy policy: Minimize hitting cost

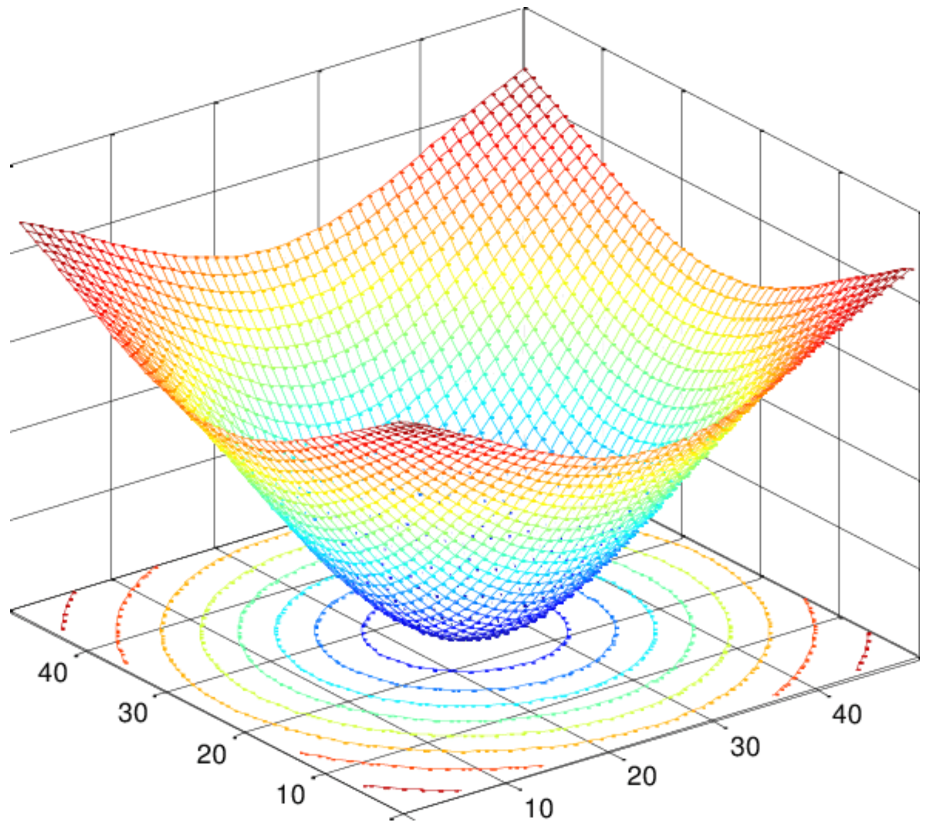
$$x_t = \arg\min_{x_t \in \mathbb{X}} f(x_t, y_t)$$

The hitting cost is minimized, but we may pay too much switching cost.

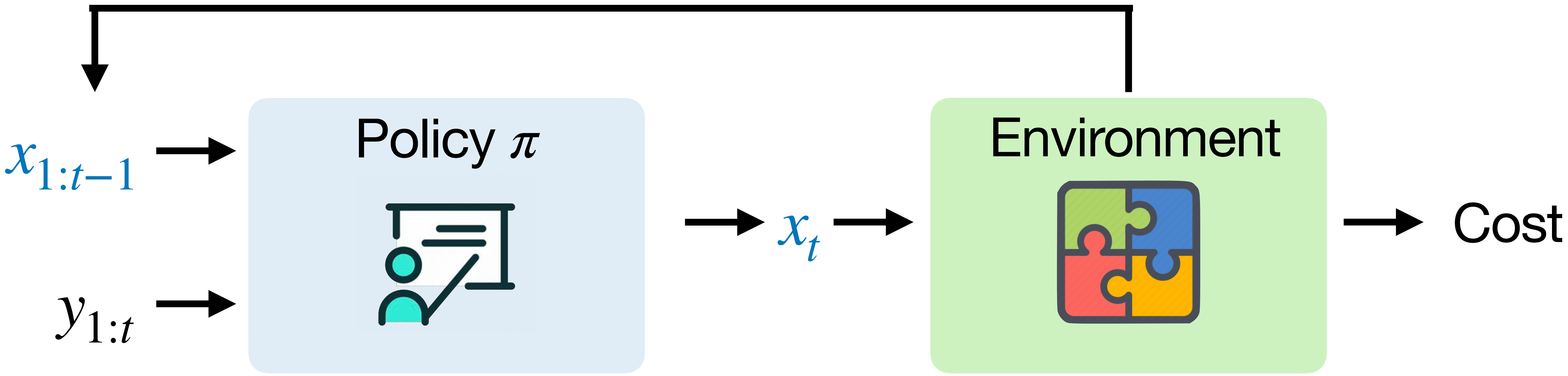


Worst-case vs average-case





Worst-case vs average-case

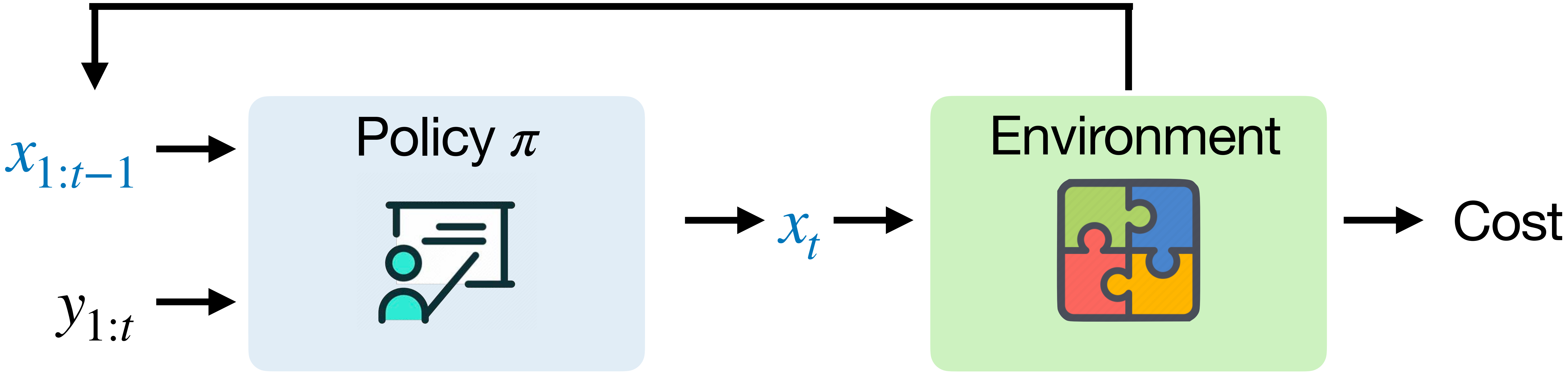


Strong **worst-case** guarantee

Sub-optimal average performance

Expert algorithm

Worst-case vs average-case

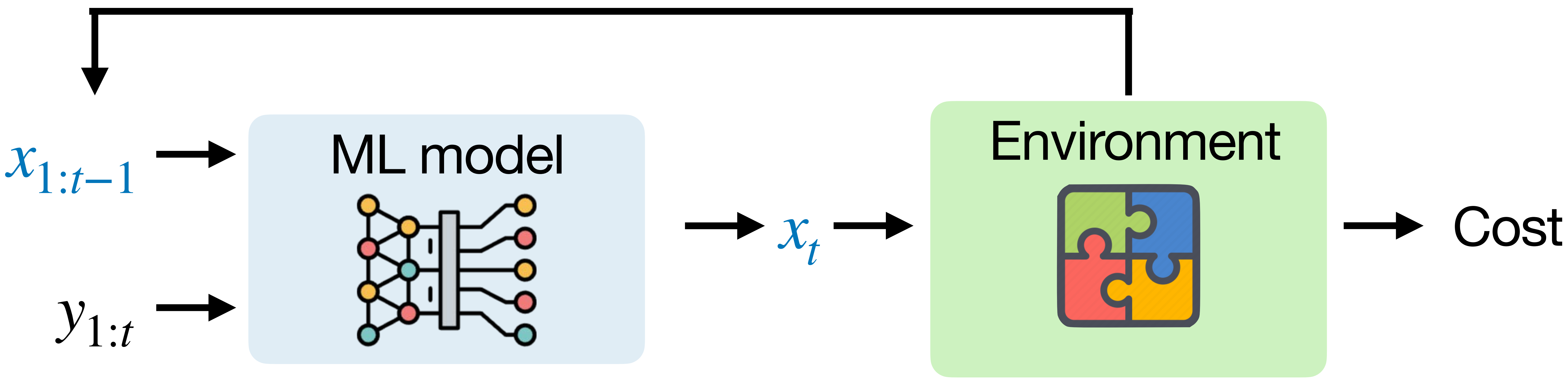


Strong **worst-case** guarantee

Sub-optimal average performance

Expert algorithm

Worst-case vs average-case

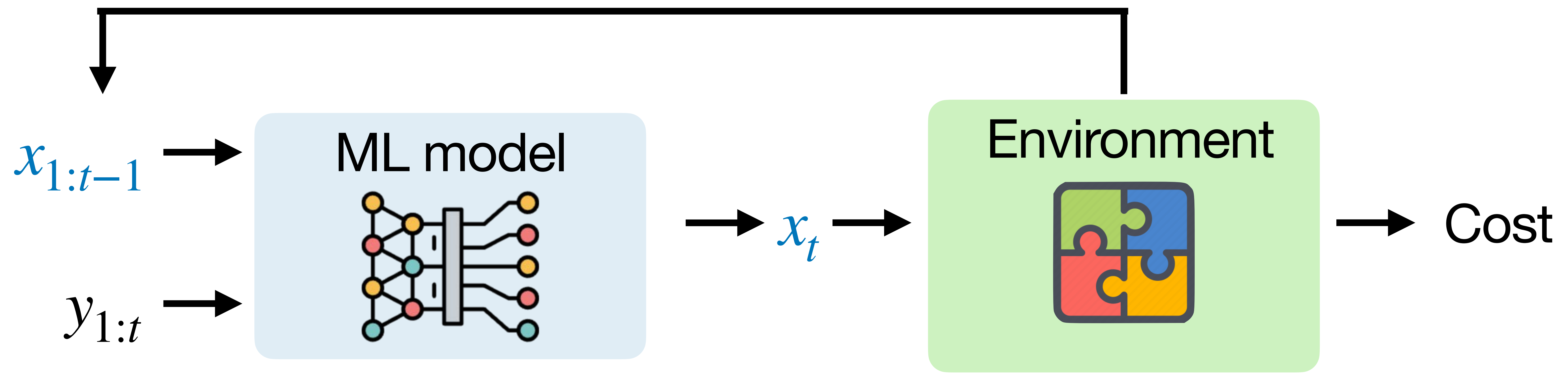


Strong **worst-case** guarantee

Sub-optimal average performance

Expert algorithm

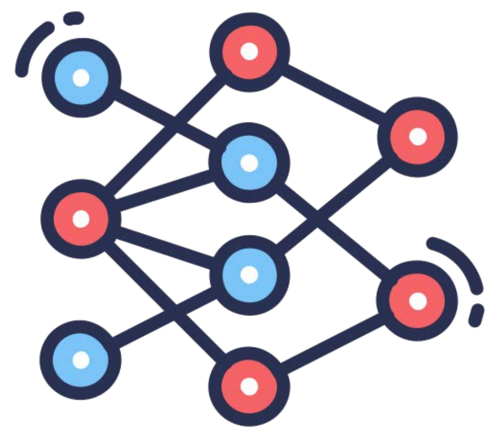
Worst-case vs average-case



Strong **worst-case** guarantee

Sub-optimal average performance

Expert algorithm



Good **average** performance

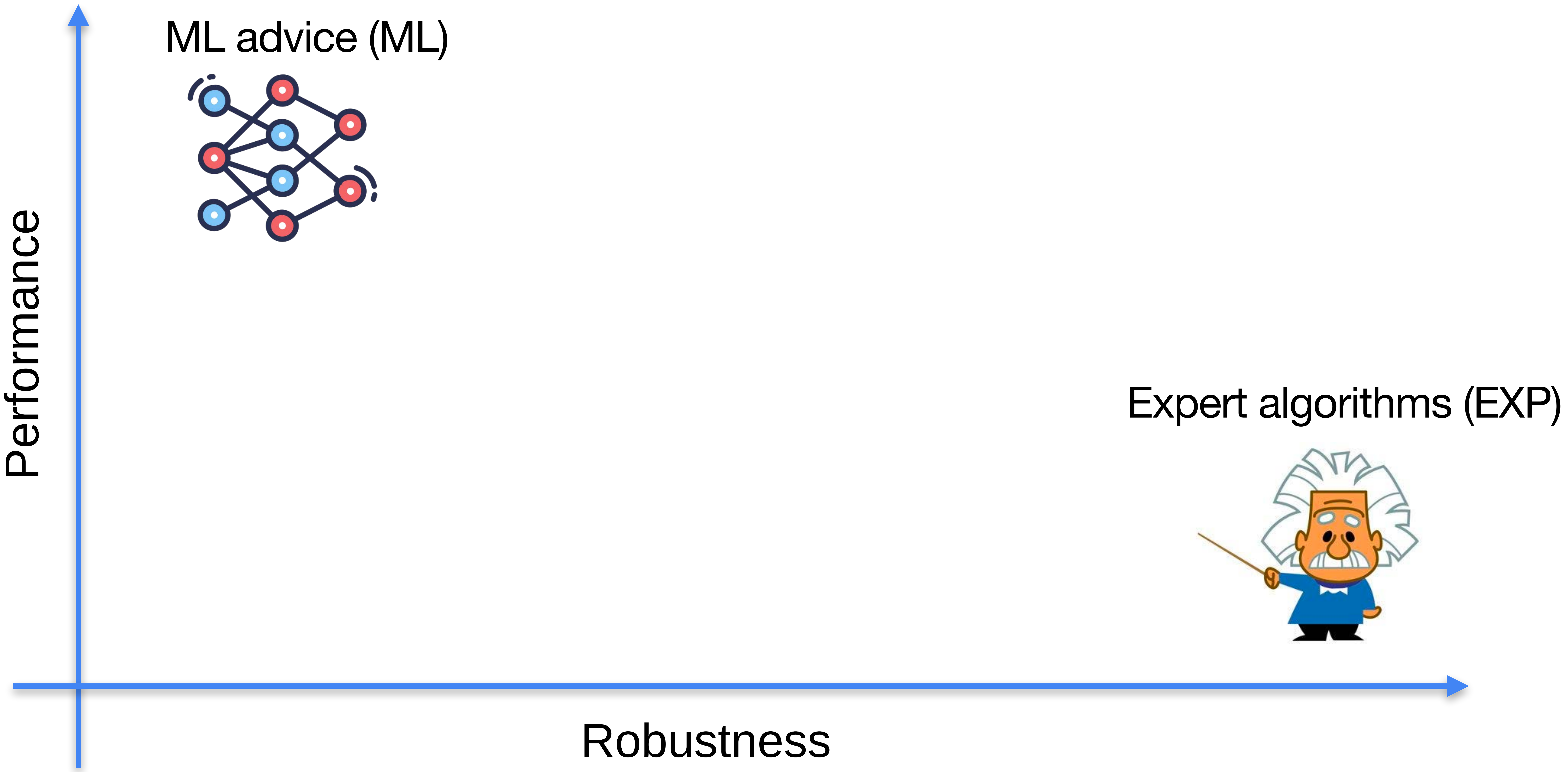
Vulnerable to worst case context

Learn to optimize (L2O)

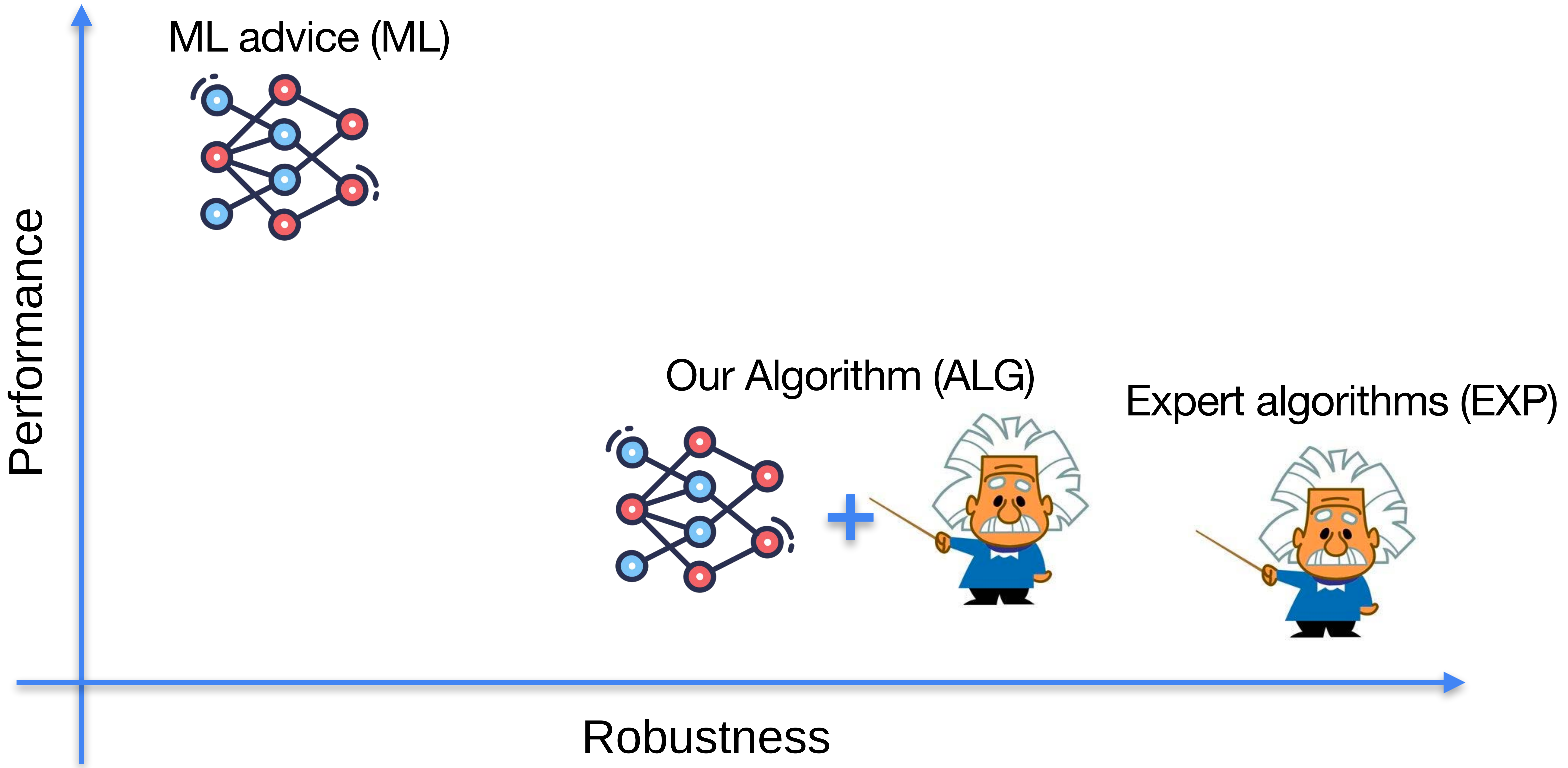
Best of both worlds?

Best of both worlds?

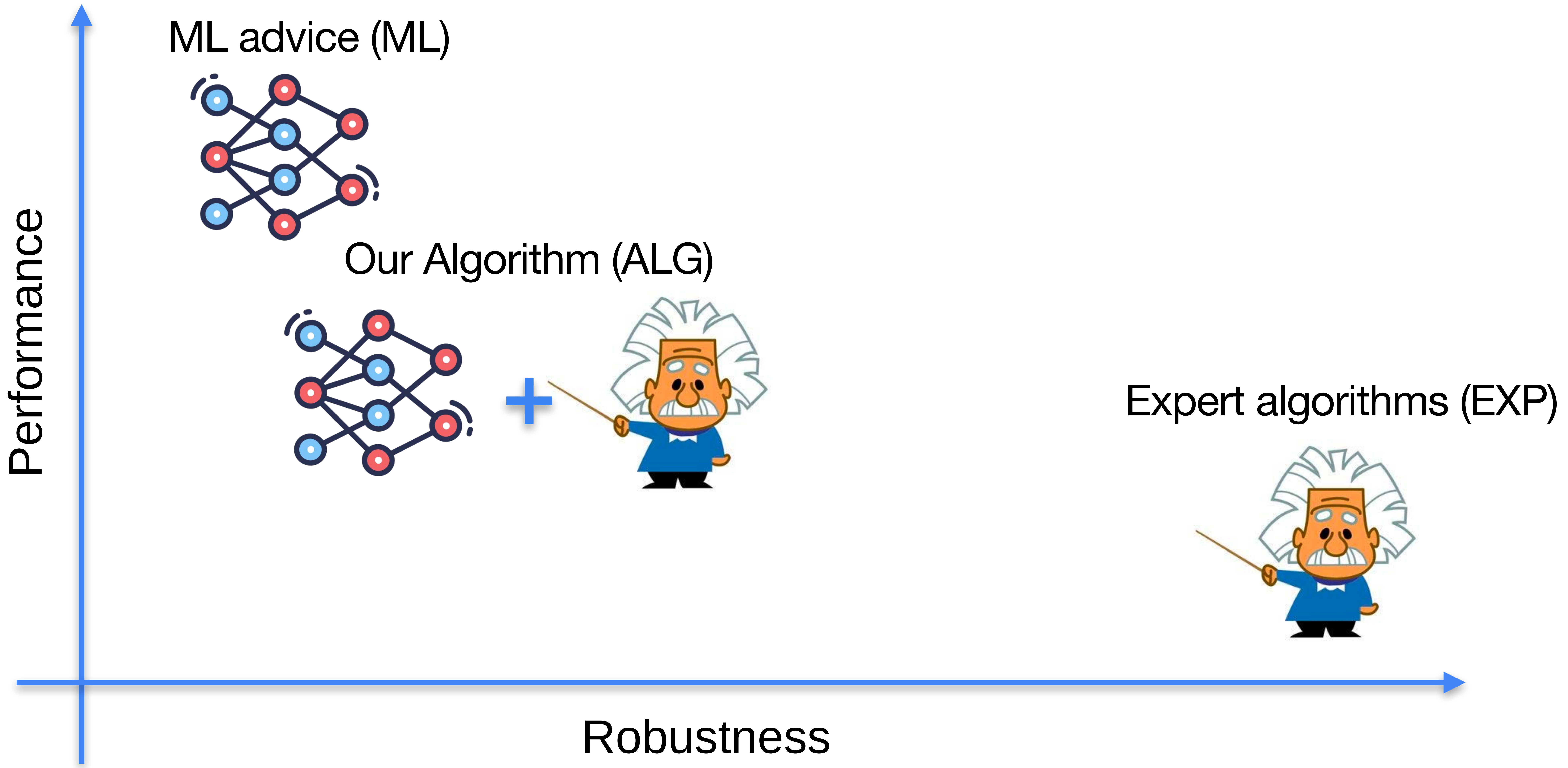
Best of both worlds?



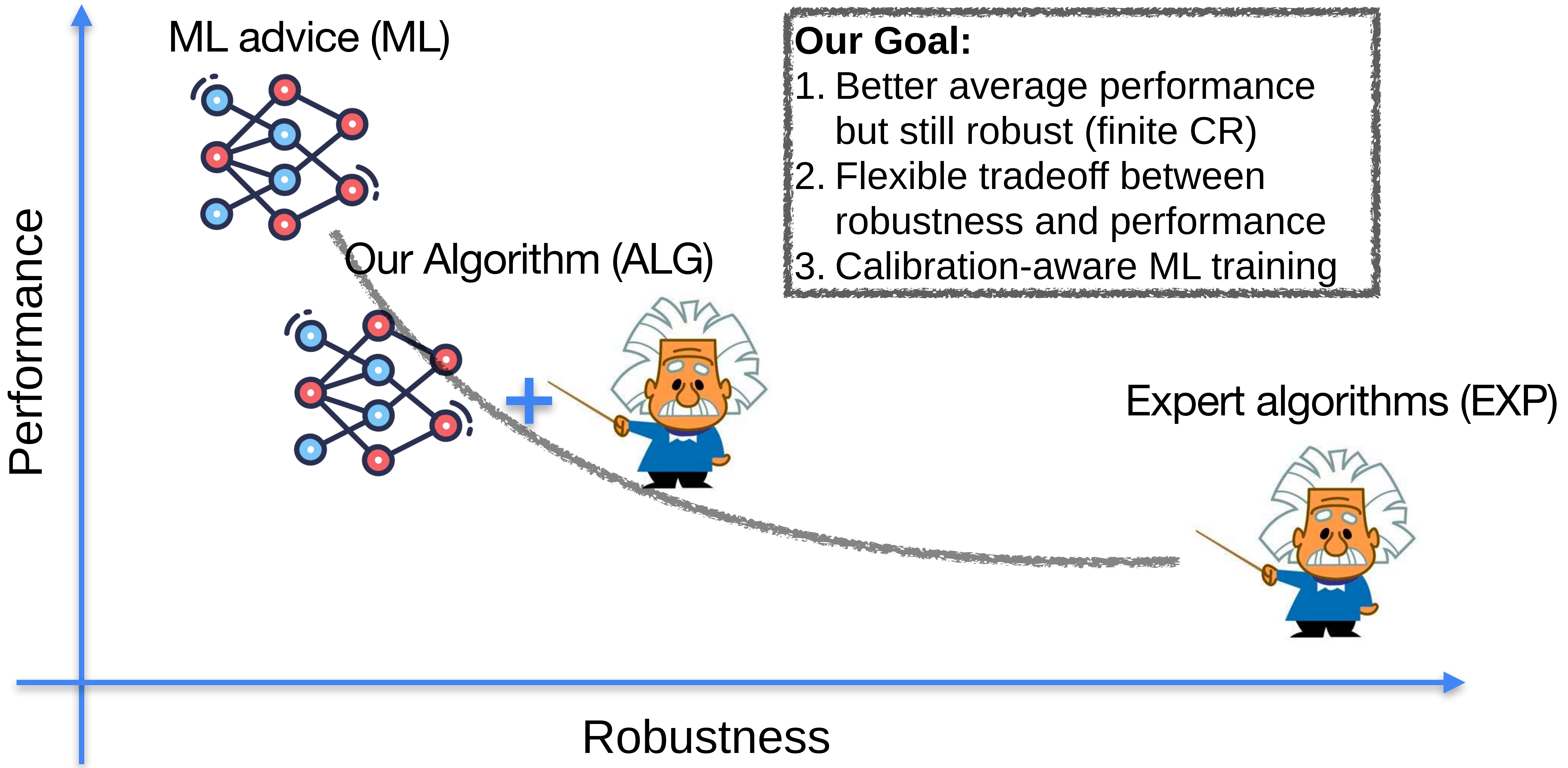
Best of both worlds?



Best of both worlds?



Best of both worlds?



Our Goal:
1. Better average performance but still robust (finite CR)
2. Flexible tradeoff between robustness and performance
3. Calibration-aware ML training

Mathematical formulation of the goal

Robustness $cost(ALG) \leq (1 + \lambda) \cdot cost(EXP)$ $\forall y_{1:T} \in \mathscr{Y}$

Consistency $cost(ALG) \leq C(\lambda) \cdot cost(ML)$ $\forall y_{1:T} \in \mathscr{Y}$

Mathematical formulation of the goal

Robustness $cost(ALG) \leq (1 + \lambda) \cdot cost(EXP)$ $\forall y_{1:T} \in \mathscr{Y}$

Tradeoff parameter λ

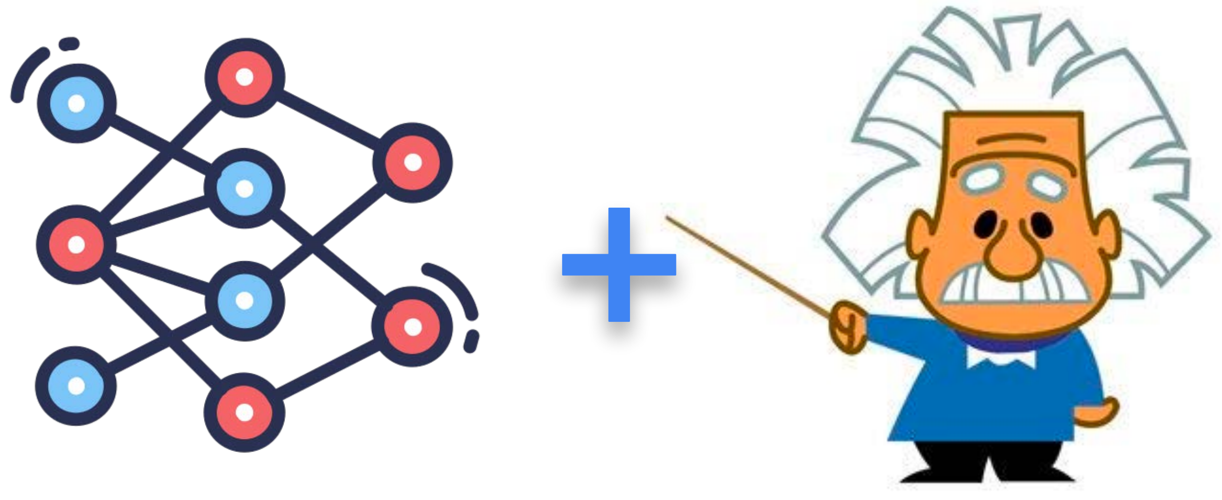
Consistency $cost(ALG) \leq C(\lambda) \cdot cost(ML)$ $\forall y_{1:T} \in \mathscr{Y}$

Mathematical formulation of the goal

Robustness $\quad cost(ALG) \leq (1+\lambda) \cdot cost(EXP) \qquad \forall y_{1:T} \in \mathcal{Y}$

Tradeoff parameter λ

Our Algorithm (ALG)



Consistency $\quad cost(ALG) \leq C(\lambda) \cdot cost(ML) \qquad \forall y_{1:T} \in \mathcal{Y}$

A quick idea about robustness

Constraint. **(Given $\lambda \geq 1$)**

$$\sum_{\tau=1}^{t} f(x_\tau, y_\tau) + c(x_\tau, x_{\tau-1}) \leq \lambda \left(\sum_{\tau=1}^{t} f(x_\tau^\pi, y_\tau) + c(x_\tau^\pi, x_{\tau-1}^\pi) \right) + B$$

A quick idea about robustness

Constraint. **(Given $\lambda \geq 1$)**

Our algorithm cost

Expert cost

$$\sum_{\tau=1}^{t} f(x_\tau, y_\tau) + c(x_\tau, x_{\tau-1}) \leq \lambda \left(\sum_{\tau=1}^{t} f(x_\tau^\pi, y_\tau) + c(x_\tau^\pi, x_{\tau-1}^\pi) \right) + B$$

A quick idea about robustness

Constraint. **(Given** $\lambda \geq 1$**)**

$$\underbrace{\sum_{\tau=1}^{t} f(x_\tau, y_\tau) + c(x_\tau, x_{\tau-1})}_{\text{Our algorithm cost}} \leq \lambda \Big(\underbrace{\sum_{\tau=1}^{t} f(x_\tau^\pi, y_\tau) + c(x_\tau^\pi, x_{\tau-1}^\pi)}_{\text{Expert cost}} \Big) + B$$

✗

Why?

What if

$$\sum_{\tau=1}^{t} f(x_\tau, y_\tau) + c(x_\tau, x_{\tau-1}) = \lambda \Big(\sum_{\tau=1}^{t} f(x_\tau^\pi, y_\tau) + c(x_\tau^\pi, x_{\tau-1}^\pi) \Big) + B$$

$$f(x_{t+1}, y_{t+1}) + c(x_{t+1}, x_t) > \lambda \left(f(x_{t+1}^\pi, y_{t+1}) + c(x_{t+1}^\pi, x_t^\pi) \right) \quad \forall x_{t+1} \in \mathcal{X}$$

A quick idea about robustness

Constraint. **(Given $\lambda \geq 1$)**

Our algorithm cost

Expert cost

$$\underbrace{\textstyle\sum_{\tau=1}^{t} f(x_\tau, y_\tau) + c(x_\tau, x_{\tau-1})} \leq \lambda \left(\underbrace{\textstyle\sum_{\tau=1}^{t} f(x_\tau^\pi, y_\tau) + c(x_\tau^\pi, x_{\tau-1}^\pi)} \right) + B$$

✗

Why?

What if

$$\textstyle\sum_{\tau=1}^{t} f(x_\tau, y_\tau) + c(x_\tau, x_{\tau-1}) = \lambda \left(\textstyle\sum_{\tau=1}^{t} f(x_\tau^\pi, y_\tau) + c(x_\tau^\pi, x_{\tau-1}^\pi) \right) + B$$

$$f(x_{t+1}, y_{t+1}) + c(x_{t+1}, \boxed{x_t}) > \lambda \left(f(x_{t+1}^\pi, y_{t+1}) + c(x_{t+1}^\pi, \boxed{x_t^\pi}) \right) \quad \forall x_{t+1} \in \mathcal{X}$$

A quick idea about robustness

Constraint. **(Given $\lambda \geq 1$)**

Our algorithm cost

Expert cost

$$\sum_{\tau=1}^{t} f(x_\tau, y_\tau) + c(x_\tau, x_{\tau-1}) \leq \lambda \left(\sum_{\tau=1}^{t} f(x_\tau^\pi, y_\tau) + c(x_\tau^\pi, x_{\tau-1}^\pi) \right) + B$$

Why?

What if

$$\sum_{\tau=1}^{t} f(x_\tau, y_\tau) + c(x_\tau, x_{\tau-1}) = \lambda \left(\sum_{\tau=1}^{t} f(x_\tau^\pi, y_\tau) + c(x_\tau^\pi, x_{\tau-1}^\pi) \right) + B$$

$$f(x_{t+1}, y_{t+1}) + c(x_{t+1}, \boxed{x_t}) > \lambda \left(f(x_{t+1}^\pi, y_{t+1}) + c(x_{t+1}^\pi, \boxed{x_t^\pi}) \right) \quad \forall x_{t+1} \in \mathcal{X}$$

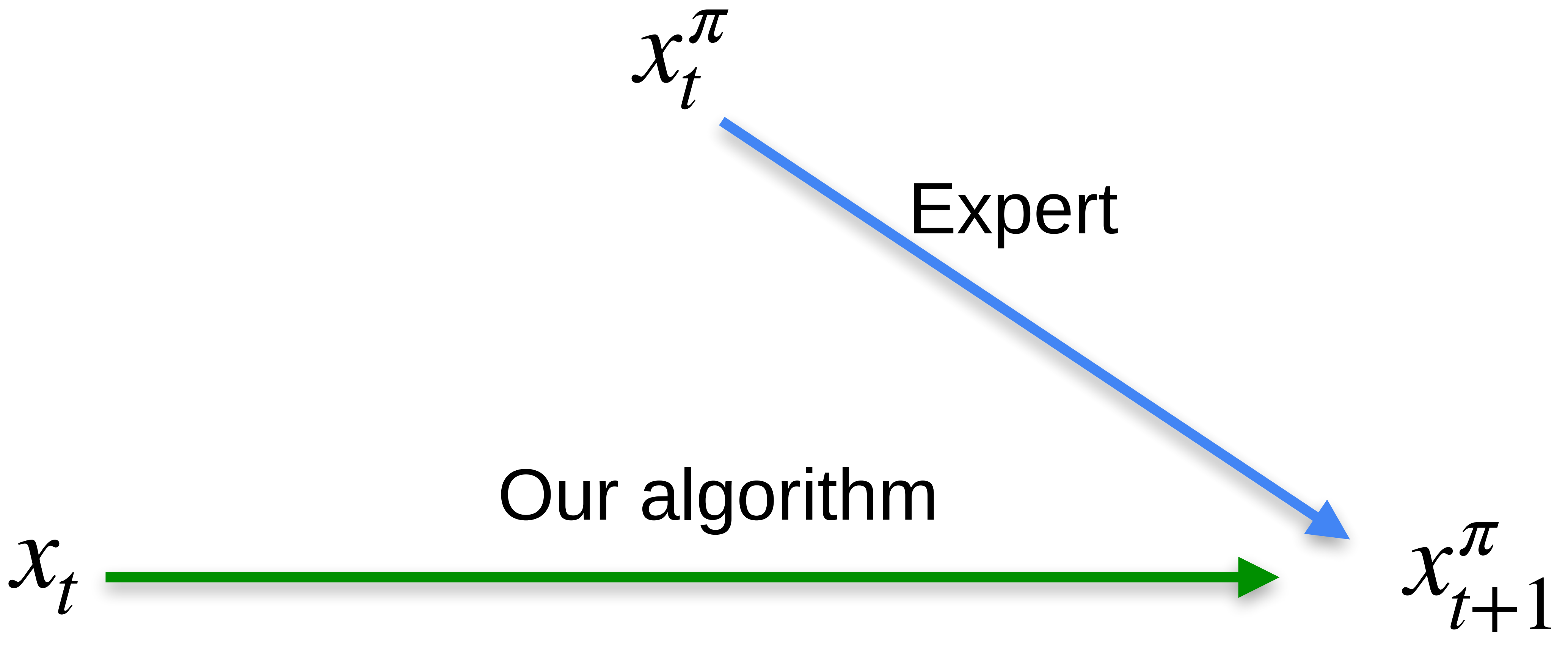
Take home message

Your current action will have some unknown future impacts. Only considering the current cost is not enough

Robustified learning for SOCO

Consider the L_p norm $d(\cdot,\cdot)$ as switching cost. For each step $t = 1,2,\cdots$

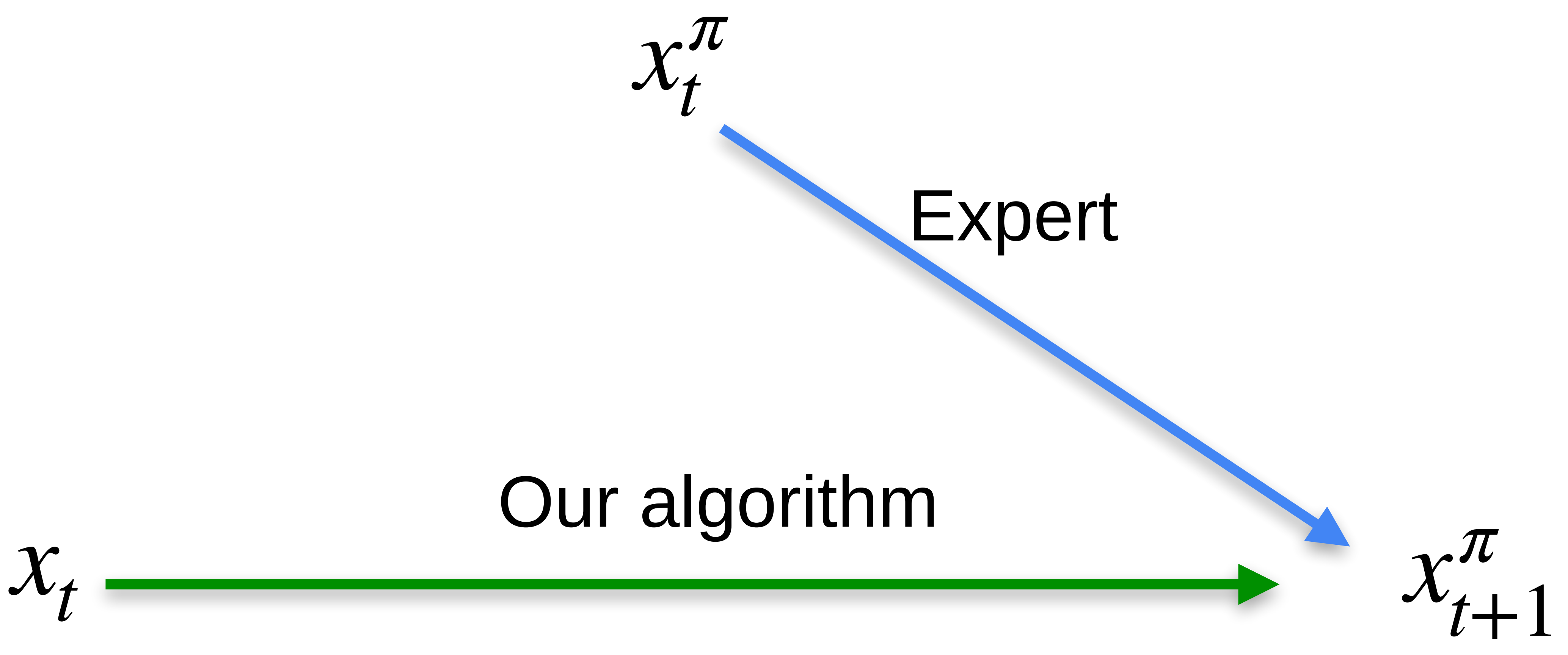
$$x_t = \arg\min_{x \in \mathcal{X}} \|x - \tilde{x}_t\|^2$$
$$s.t., \text{cost}(x_{1:t-1}) + f(x, y_t) + c(x, x_{t-1}) + G(x, x_t^{\pi}) \leq (1+\lambda)\text{cost}(x_{1:t}^{\pi})$$



Robustified learning for SOCO

Consider the L_p norm $d(\cdot,\cdot)$ as switching cost. For each step $t = 1,2,\cdots$

$$x_t = \arg\min_{x\in\mathcal{X}} \|x - \tilde{x}_t\|^2$$
$$s.t., \mathrm{cost}(x_{1:t-1}) + f(x, y_t) + c(x, x_{t-1}) + G(x, x_t^\pi) \leq (1+\lambda)\mathrm{cost}(x_{1:t}^\pi)$$



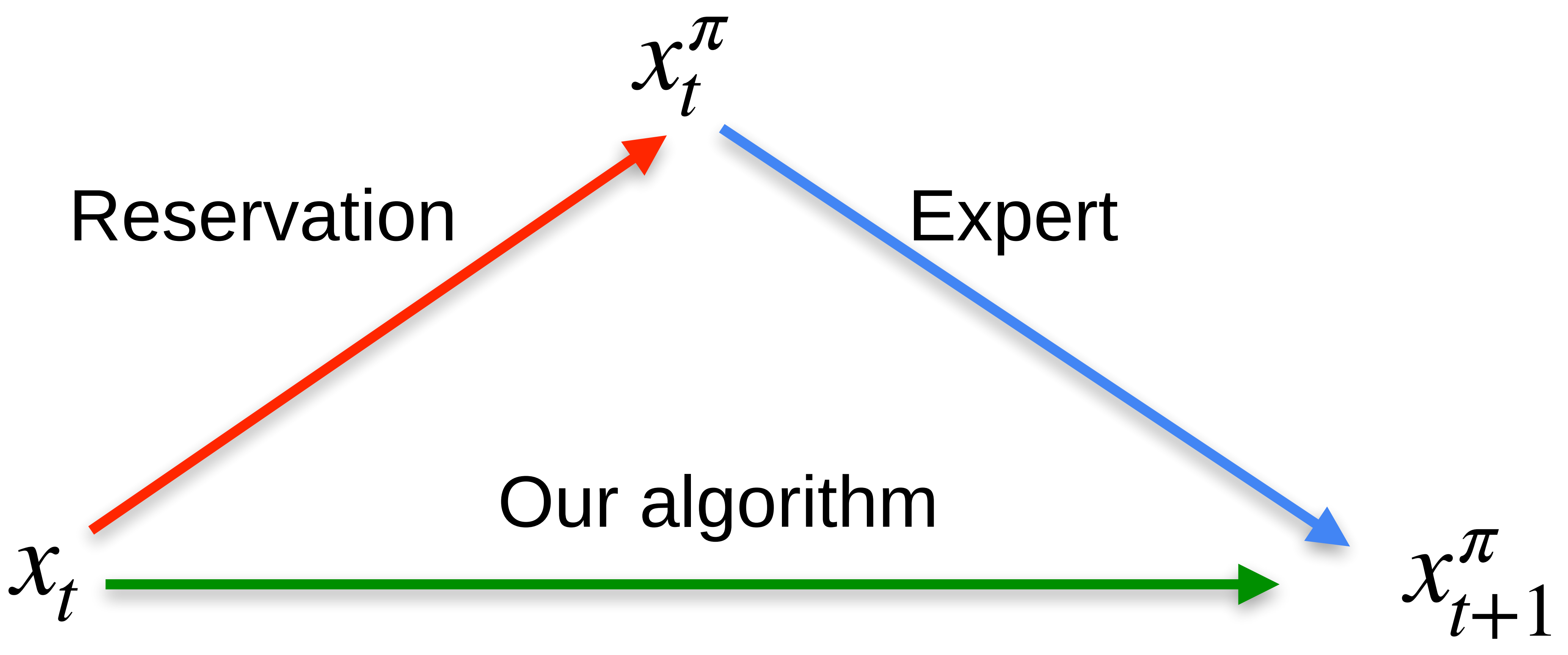
Robustified learning for SOCO

Consider the L_p norm $d(\cdot,\cdot)$ as switching cost. For each step $t = 1,2,\cdots$

$$x_t = \arg\min_{x\in\mathcal{X}} \|x - \tilde{x}_t\|^2$$

$$s.t., \text{cost}(x_{1:t-1}) + f(x, y_t) + c(x, x_{t-1}) + G(x, x_t^\pi) \le (1+\lambda)\text{cost}(x_{1:t}^\pi)$$

$$\boxed{G(x, x_t^\pi) = \|x - x_t^\pi\|}$$



Robustified learning for SOCO

Consider the L_p norm $d(\cdot,\cdot)$ as switching cost. For each step $t = 1,2,\cdots$

$$x_t = \arg\min_{x \in \mathcal{X}} \|x - \tilde{x}_t\|^2$$

$$s.t.,\ \text{cost}(x_{1:t-1}) + f(x, y_t) + c(x, x_{t-1}) + G(x, x_t^{\pi}) \leq (1+\lambda)\text{cost}(x_{1:t}^{\pi})$$

Robustified learning for SOCO

Consider the L_p norm $d(\cdot,\cdot)$ as switching cost. For each step $t = 1,2,\cdots$

$$x_t = \arg\min_{x\in\mathcal{X}} \|x - \tilde{x}_t\|^2$$

$$s.t.,\ \text{cost}(x_{1:t-1}) + f(x, y_t) + c(x, x_{t-1}) + G(x, x_t^\pi) \le (1+\lambda)\text{cost}(x_{1:t}^\pi)$$

$$\boxed{G(x, x_t^\pi) = \|x - x_t^\pi\|}$$

Robustified learning for SOCO

Consider the L_p norm $d(\cdot\,,\cdot\,)$ as switching cost. For each step $t = 1,2,\cdots$

$$x_t = \arg\min_{x \in \mathcal{X}} \|x - \tilde{x}_t\|^2$$

$$s.t., \text{cost}(x_{1:t-1}) + f(x, y_t) + c(x, x_{t-1}) + G(x, x_t^\pi) \le (1+\lambda)\text{cost}(x_{1:t}^\pi)$$

$$\boxed{G(x, x_t^\pi) = \|x - x_t^\pi\|}$$

Robustified learning for SOCO

Consider the L_p norm $d(\cdot,\cdot)$ as switching cost. For each step $t = 1,2,\cdots$

$$x_t = \arg\min_{x\in\mathcal{X}} \|x - \tilde{x}_t\|^2$$

$$s.t.,\ \mathrm{cost}(x_{1:t-1}) + f(x, y_t) + c(x, x_{t-1}) + G(x, x_t^\pi) \le (1+\lambda)\mathrm{cost}(x_{1:t}^\pi)$$

$$\boxed{G(x, x_t^\pi) = \|x - x_t^\pi\|}$$

Expert Robustified Learning: ERL

Algorithm 1 Expert-Robustified Learning for Online Optimization with Memory Cost (ERL)

Input: $\lambda \geq 1$, $B \geq 0$, initial x_0, trained ML model (Section 4.3), and expert online algorithm π

for $t = 1, \cdots, T$
    Receive the context y_t
    Expert chooses x_t^π
    $\tilde{x}_t \leftarrow h(x_{t-1}, y_t)$ `//Action output from ML`
    $x_t \leftarrow \text{proj}(\tilde{x}_t, x_t^\pi, \text{cost}(x_{1:t-1}), \text{cost}(x_{1:t}^\pi))$ based on Eqn. (2) `//Robustification`

$x_{t-1} \longrightarrow$

$y_t \longrightarrow$

Expert Robustified Learning: ERL

Algorithm 1 Expert-Robustified Learning for Online Optimization with Memory Cost (ERL)

Input: $\lambda \geq 1$, $B \geq 0$, initial x_0, trained ML model (Section 4.3), and expert online algorithm π

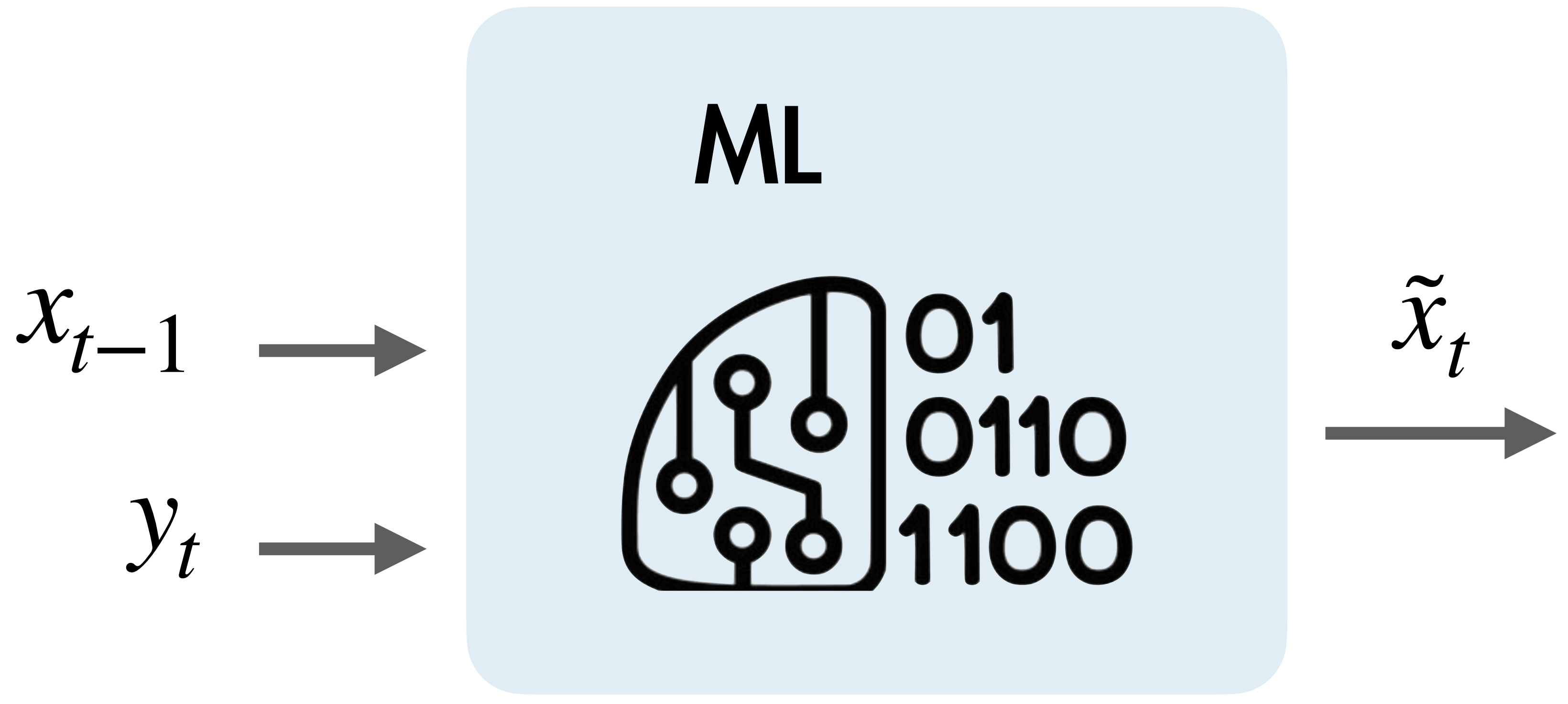
for $t = 1, \cdots, T$

  Receive the context y_t

  Expert chooses x_t^π

  $\boxed{\tilde{x}_t \leftarrow h(x_{t-1}, y_t)}$ `//Action output from ML`

  $x_t \leftarrow \text{proj}(\tilde{x}_t, x_t^\pi, \text{cost}(x_{1:t-1}), \text{cost}(x_{1:t}^\pi))$ based on Eqn. (2) `//Robustification`



Expert Robustified Learning: ERL

Algorithm 1 Expert-Robustified Learning for Online Optimization with Memory Cost (ERL)

Input: $\lambda \geq 1$, $B \geq 0$, initial x_0, trained ML model (Section 4.3), and expert online algorithm π

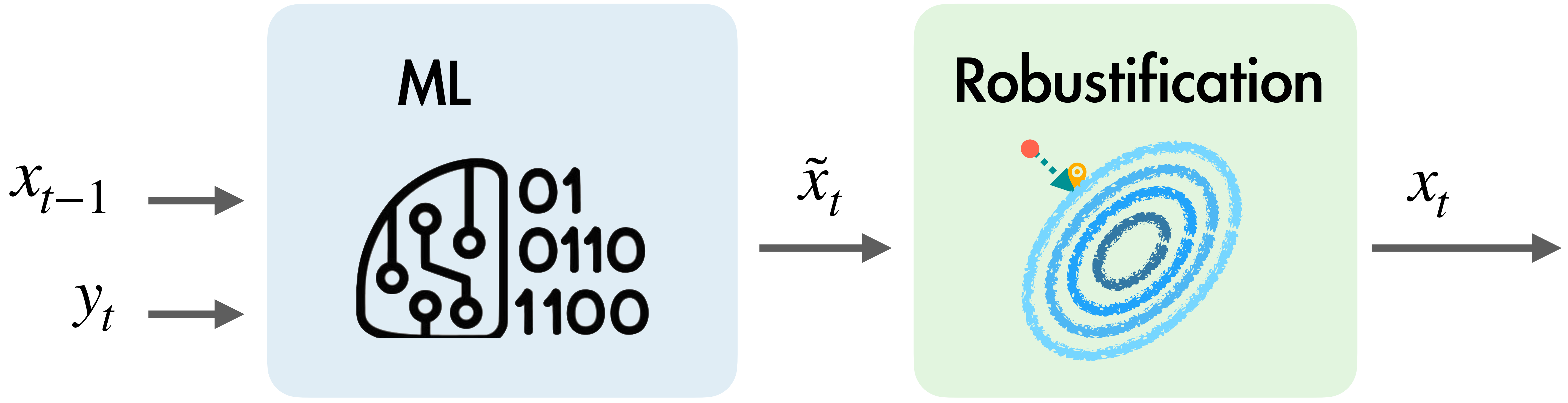
for $t = 1, \cdots, T$

  Receive the context y_t

  Expert chooses x_t^π

  $\boxed{\tilde{x}_t \leftarrow h(x_{t-1}, y_t)}$ `//Action output from ML`

  $\boxed{x_t \leftarrow \text{proj}(\tilde{x}_t, x_t^\pi, \text{cost}(x_{1:t-1}), \text{cost}(x_{1:t}^\pi)) \text{ based on Eqn. (2)}}$ `//Robustification`



Expert Robustified Learning: ERL

Algorithm 1 Expert-Robustified Learning for Online Optimization with Memory Cost (ERL)

Input: $\lambda \geq 1$, $B \geq 0$, initial x_0, trained ML model (Section 4.3), and expert online algorithm π

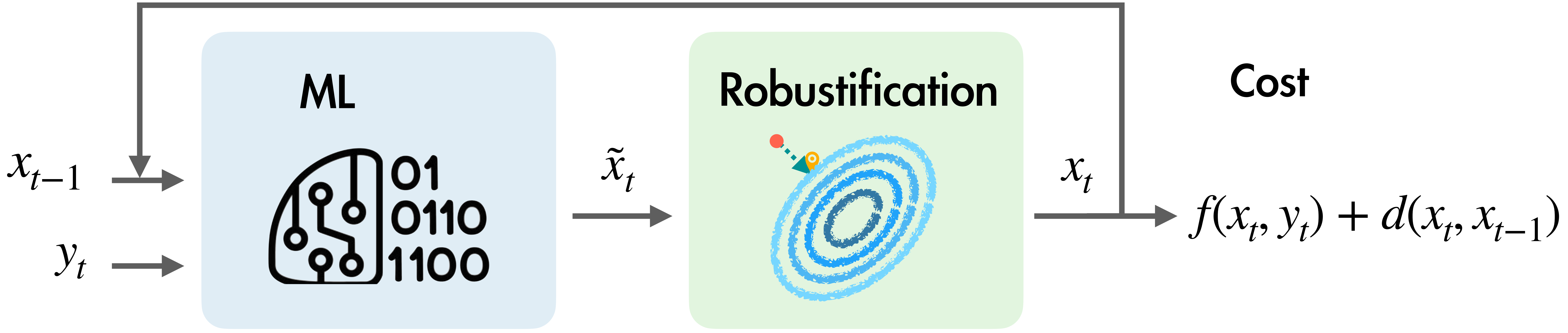
1: for $t = 1, \cdots, T$

2: Receive the context y_t

3: Expert chooses x_t^π

4: $\boxed{\tilde{x}_t \leftarrow h(x_{t-1}, y_t)}$ `//Action output from ML`

5: $\boxed{x_t \leftarrow \text{proj}(\tilde{x}_t, x_t^\pi, \text{cost}(x_{1:t-1}), \text{cost}(x_{1:t}^\pi)) \text{ based on Eqn. (2)}}$ `//Robustification`



Theoretical Analysis

ML-Expert Discrepancy

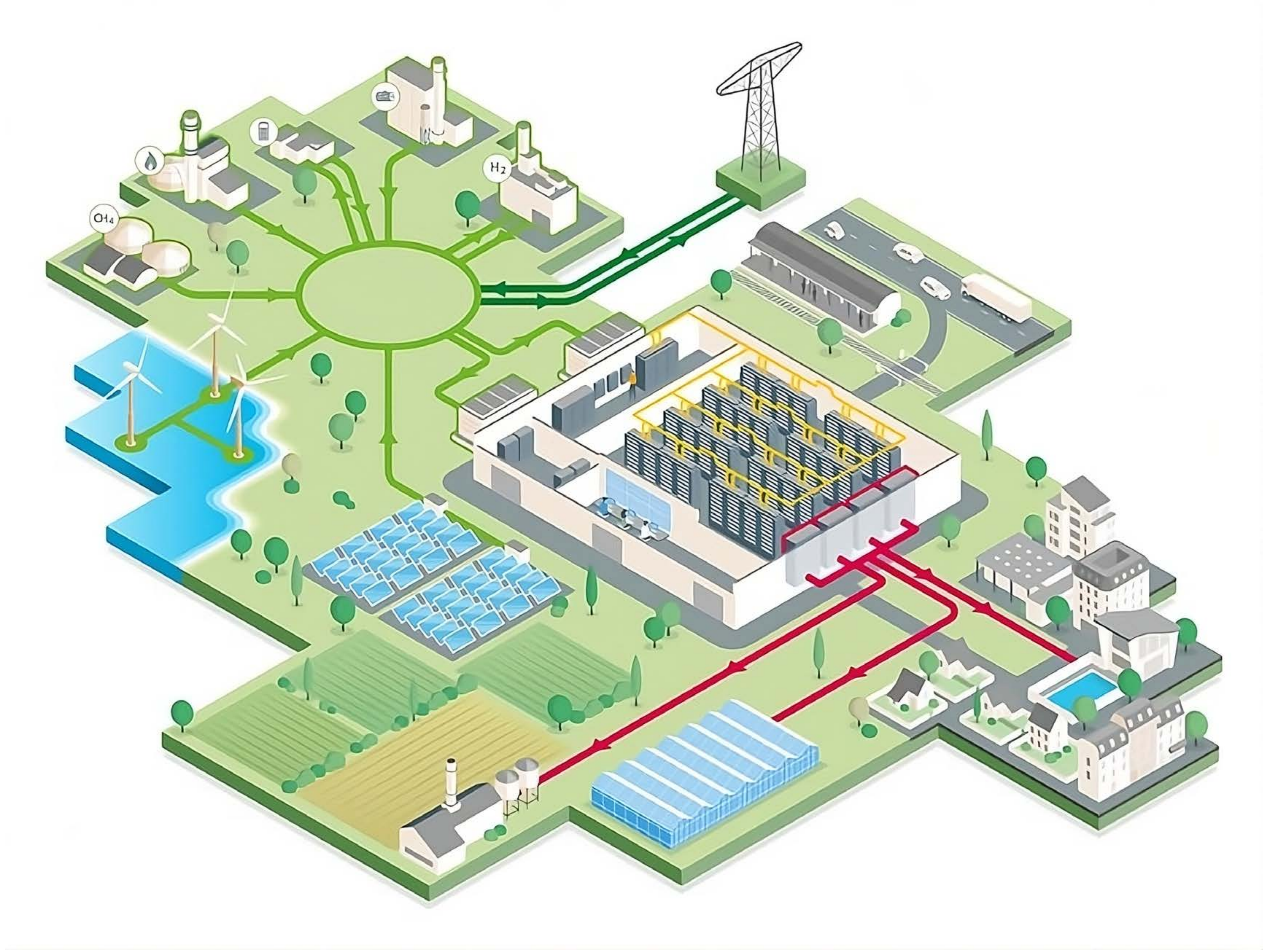
$$\rho(y) = \max_{t=1,\cdots,T} \frac{\|\tilde{x}_t - x_t^{\pi}\|^2}{f(x_t^{\pi}, y_t) + d(x_t^{\pi}, x_{t-1}^{\pi})}.$$

Bi-Competitive Ratio

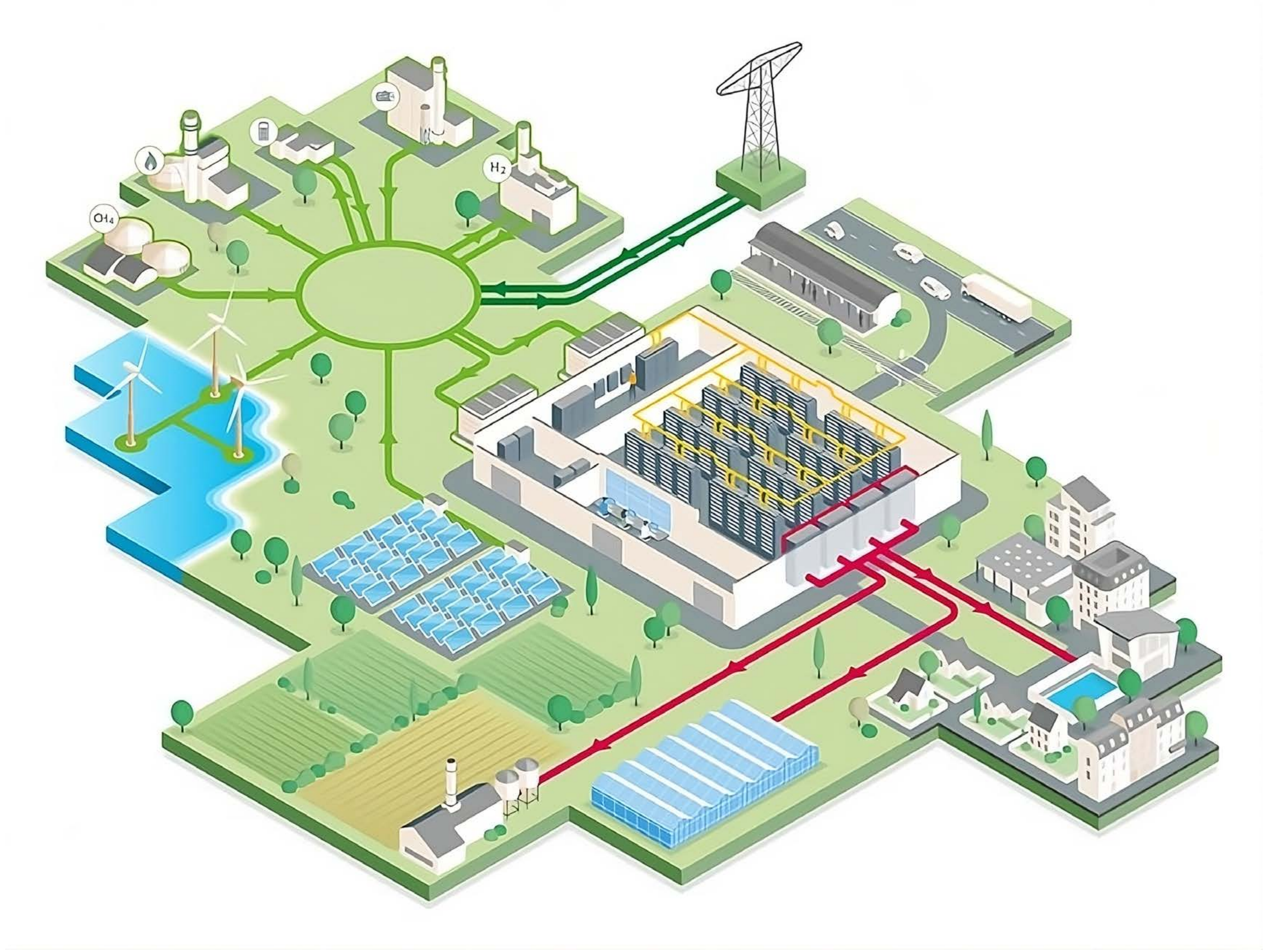
$$CR \leq \min\left\{\underbrace{(1+\lambda)CR^{\pi}}_{\text{Robustness}}, \underbrace{\left(\sqrt{\tilde{CR}} + \sqrt{CR^{\pi}}\left[1 + \sqrt{\frac{\beta_h + (1+L_1)^2}{2}\rho} - \sqrt{1+\lambda}\right]^{+}\right)^2}_{\text{Consistency}}\right\},$$

Case study: datacenter demand response

Case study: datacenter demand response



Case study: datacenter demand response

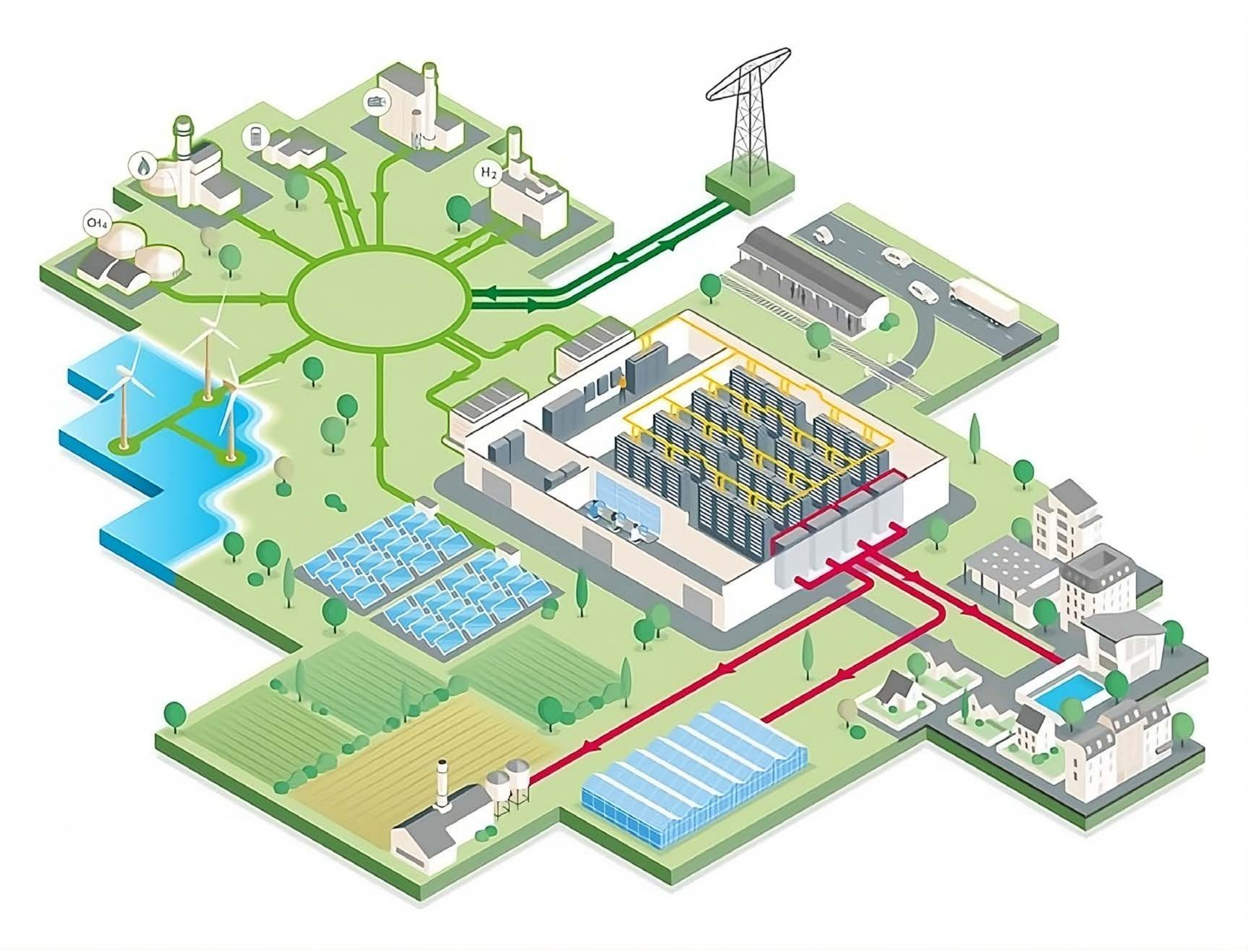


Hitting Cost: $f(x_t, y_t)$

$y_t = [$  ,  ,  , 

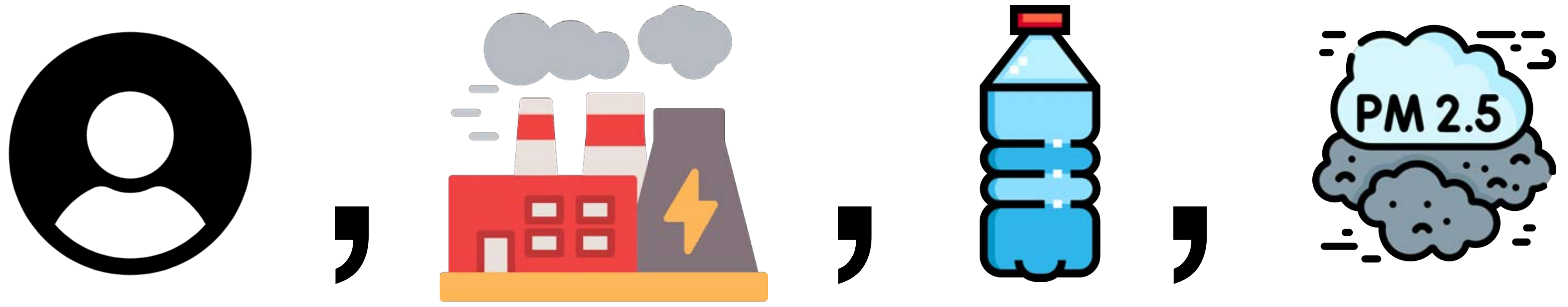
$]$

Case study: datacenter demand response



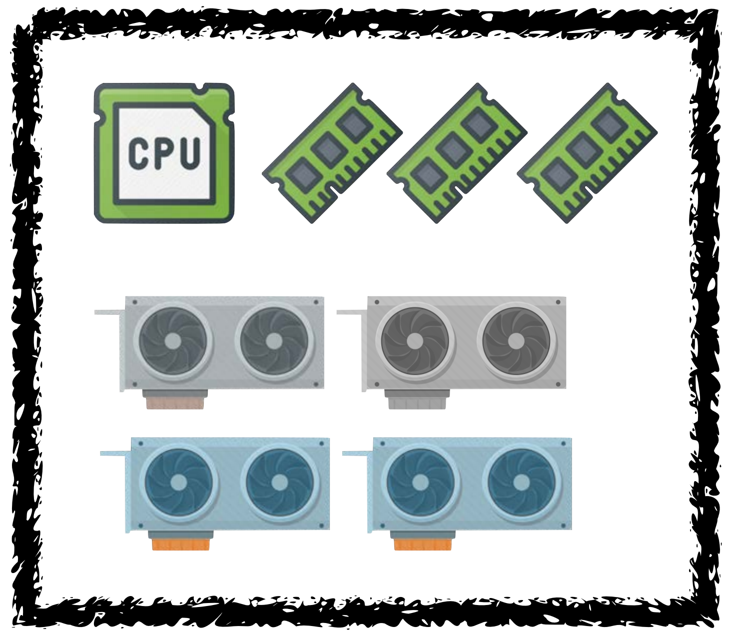
Hitting Cost: $f(x_t, y_t)$

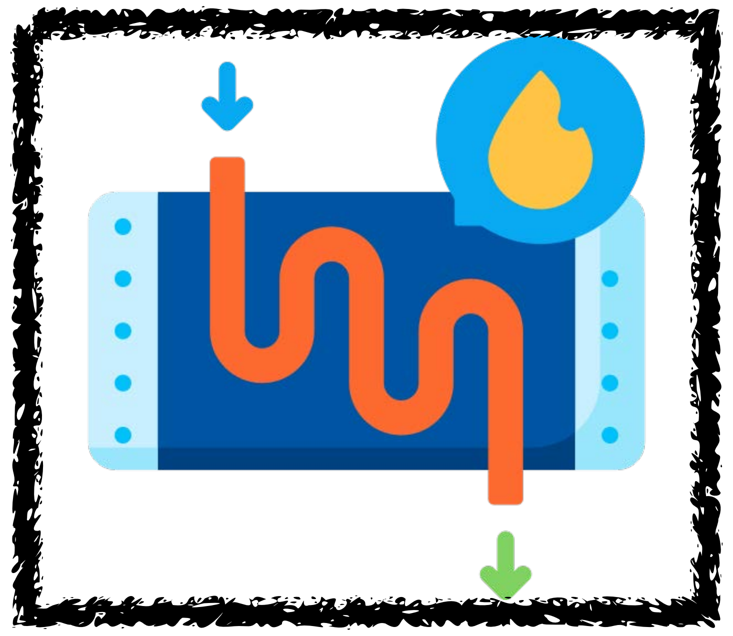
y_t = [



]

Switching Cost: $c(x_t, x_{t-1})$

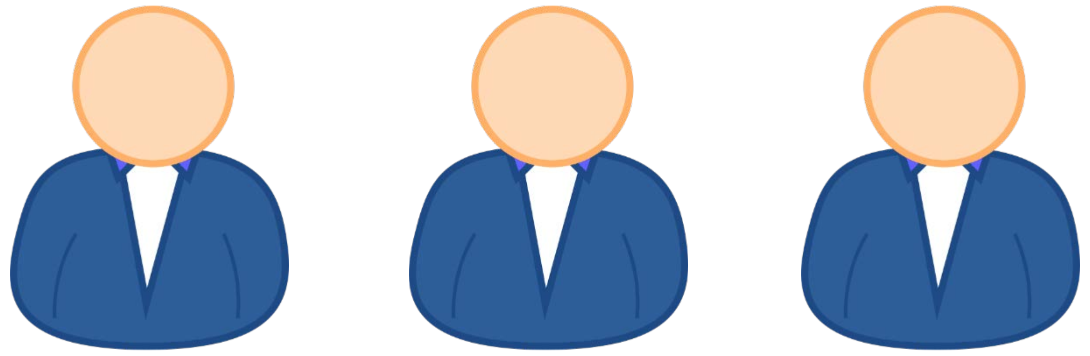




A toy example for equitable AI

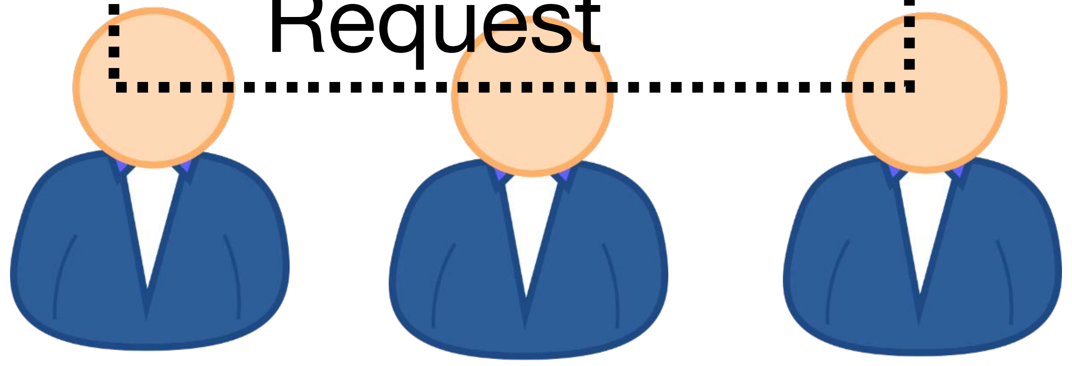
$$\frac{\text{Health Cost}}{\text{Request}} = 1$$

$$\frac{\text{Health Cost}}{\text{Request}} = 2$$



A toy example for equitable AI

$$\frac{\text{Health Cost}}{\text{Request}} = 1$$

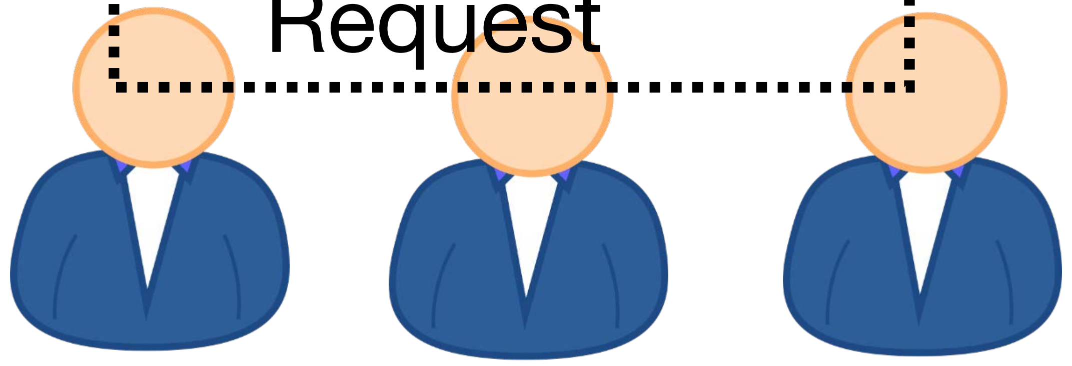


$$\frac{\text{Health Cost}}{\text{Request}} = 2$$

A toy example for equitable AI

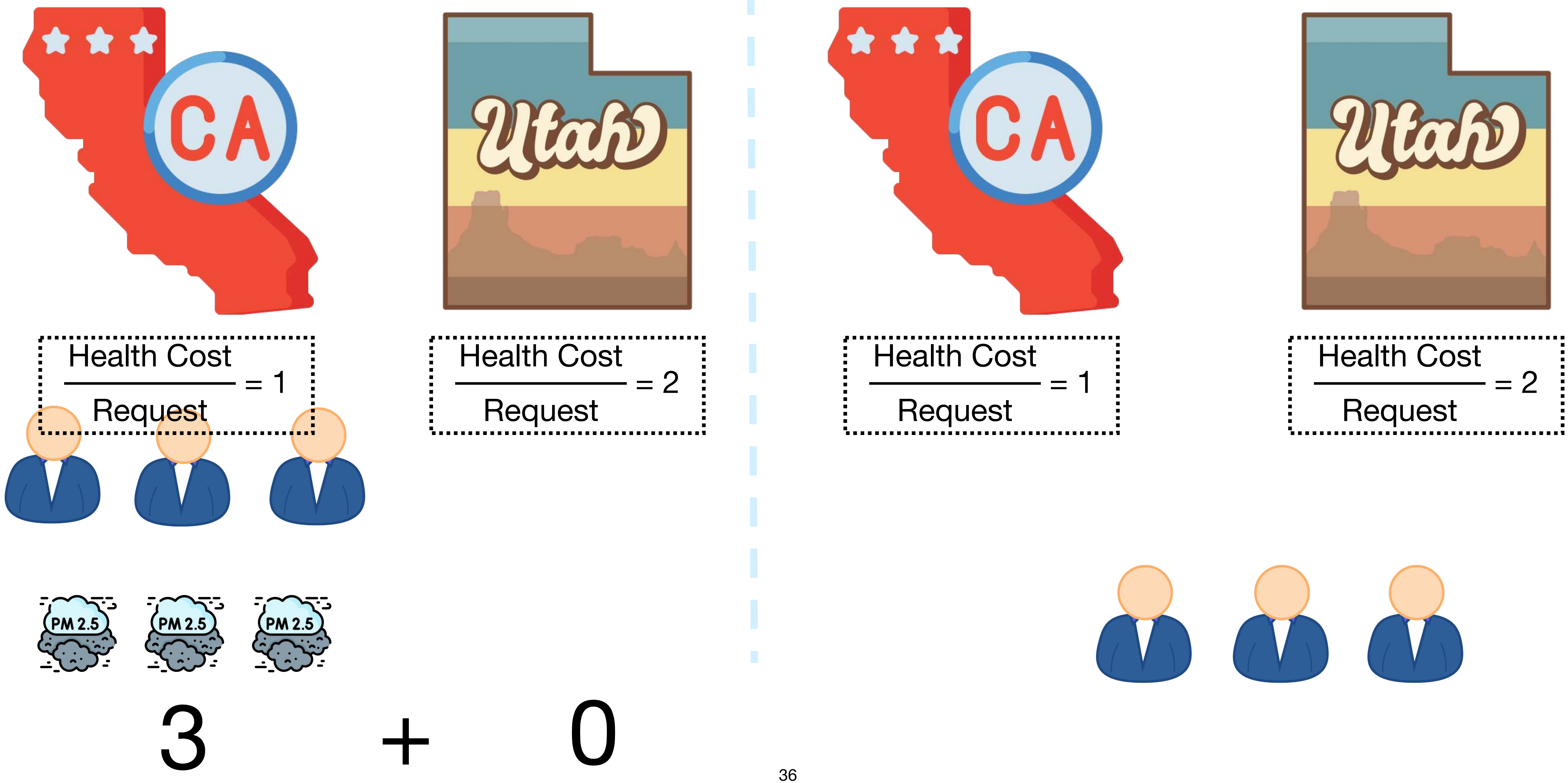
$$\frac{\text{Health Cost}}{\text{Request}} = 1$$

$$\frac{\text{Health Cost}}{\text{Request}} = 2$$



3 + 0

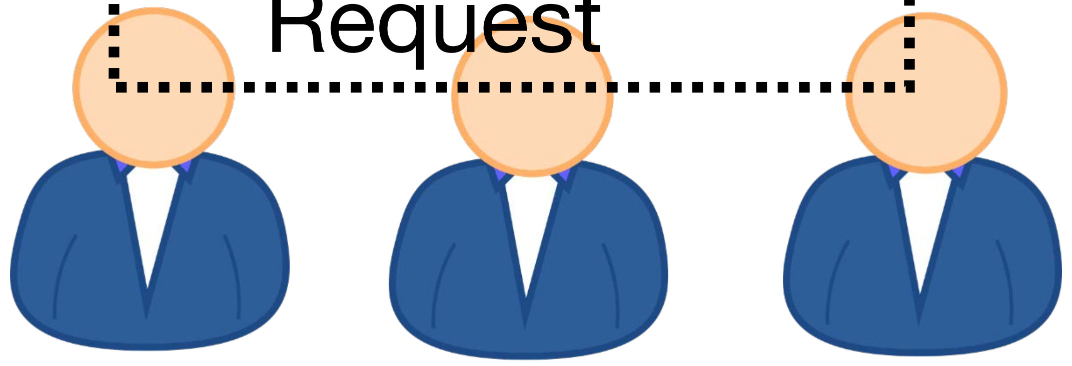
A toy example for equitable AI



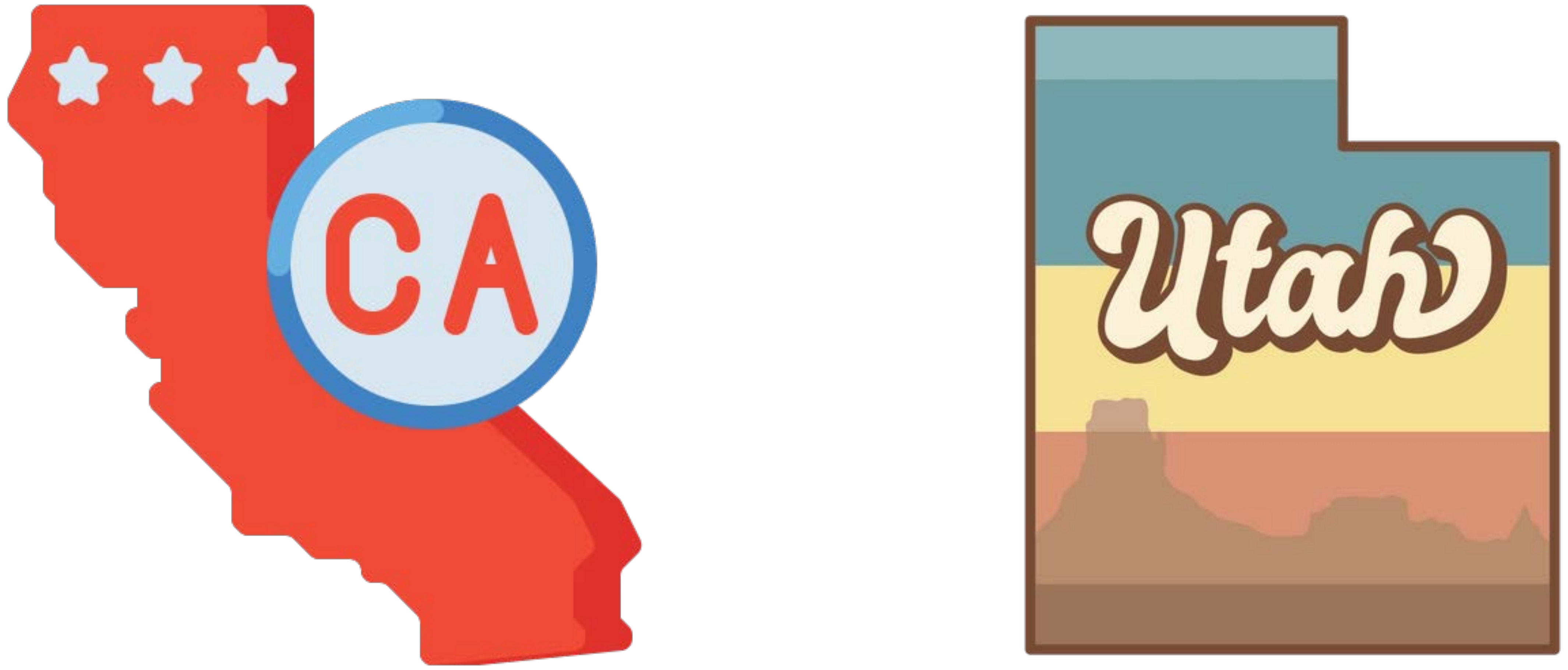
A toy example for equitable AI

$\frac{\text{Health Cost}}{\text{Request}} = 1$

$\frac{\text{Health Cost}}{\text{Request}} = 2$

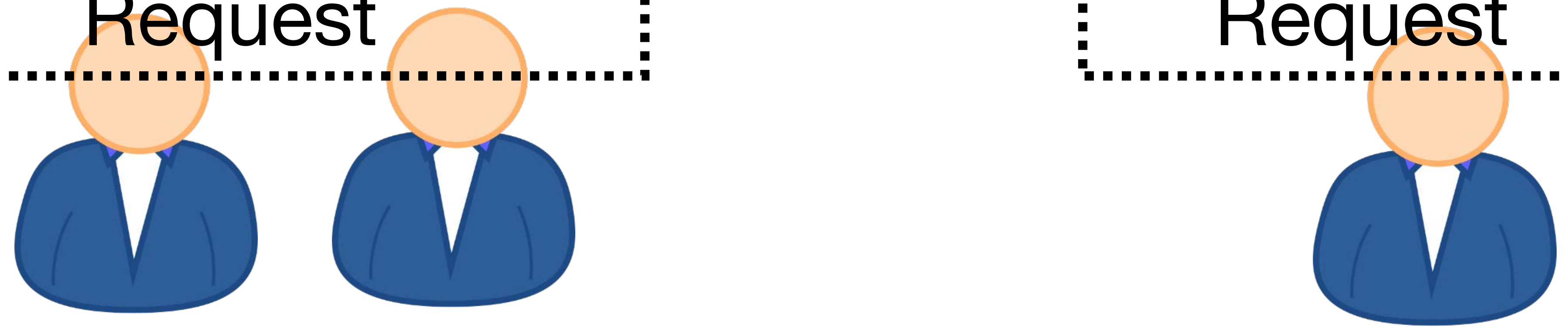


3 + 0

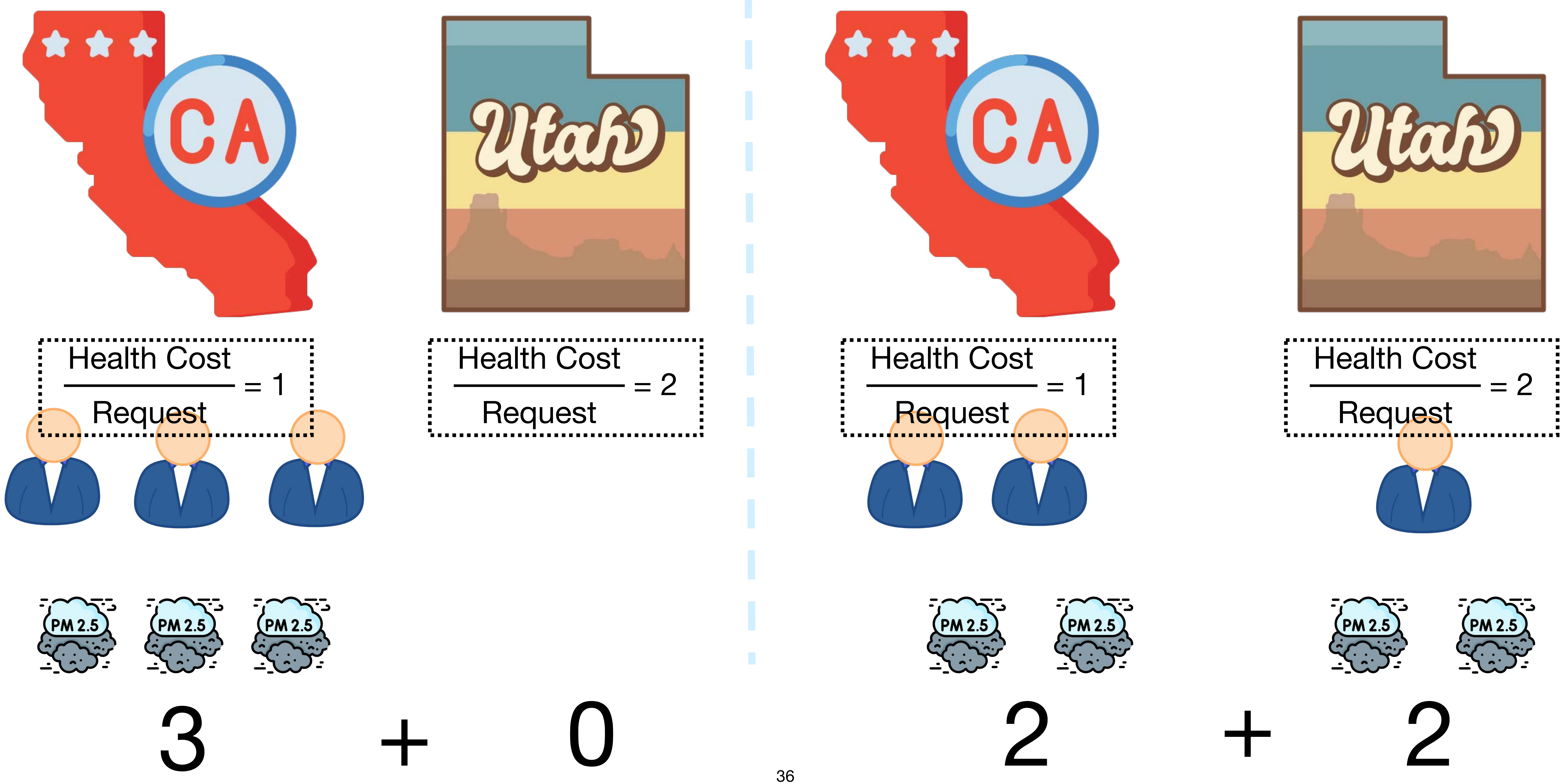


$\frac{\text{Health Cost}}{\text{Request}} = 1$

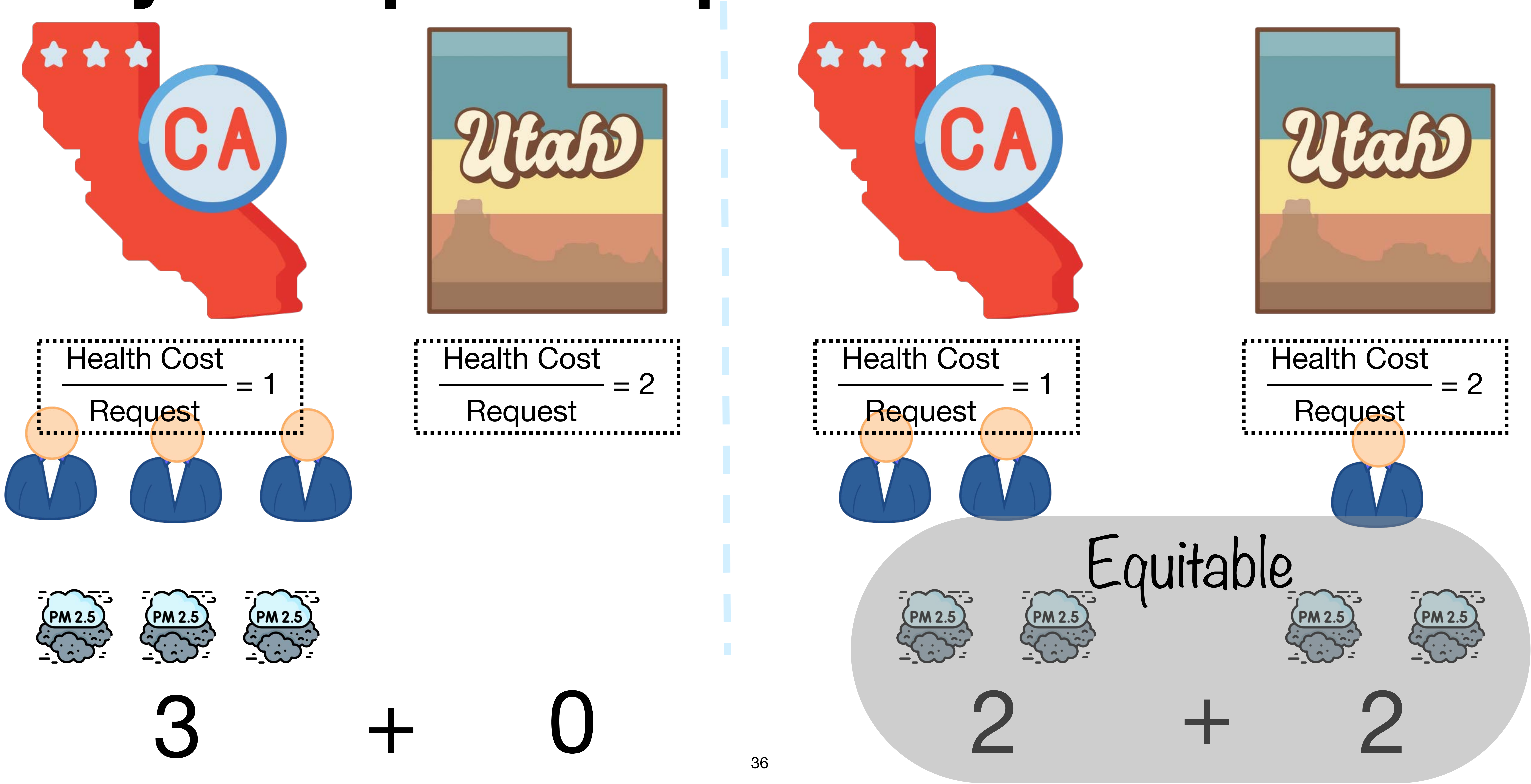
$\frac{\text{Health Cost}}{\text{Request}} = 2$



A toy example for equitable AI



A toy example for equitable AI

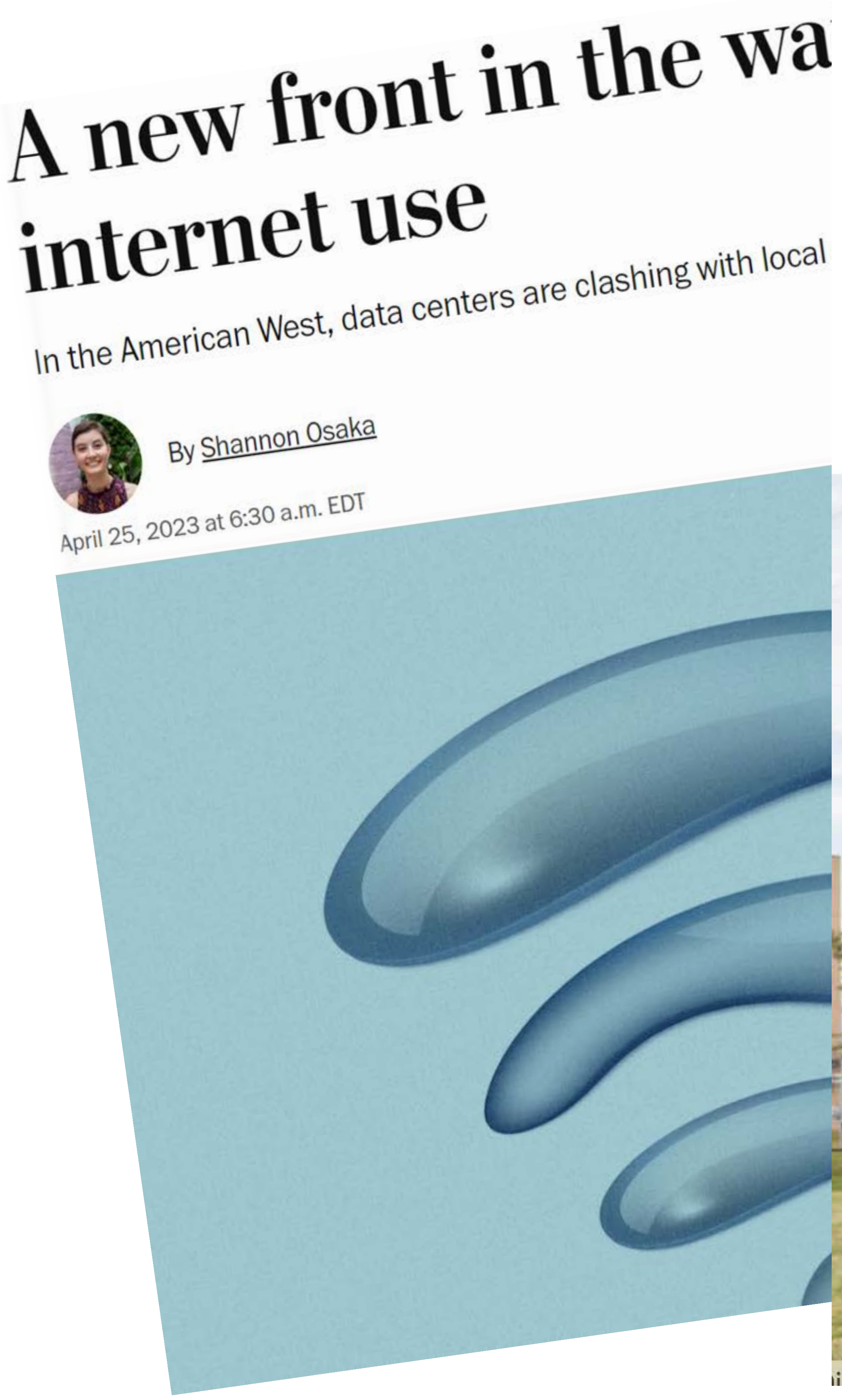


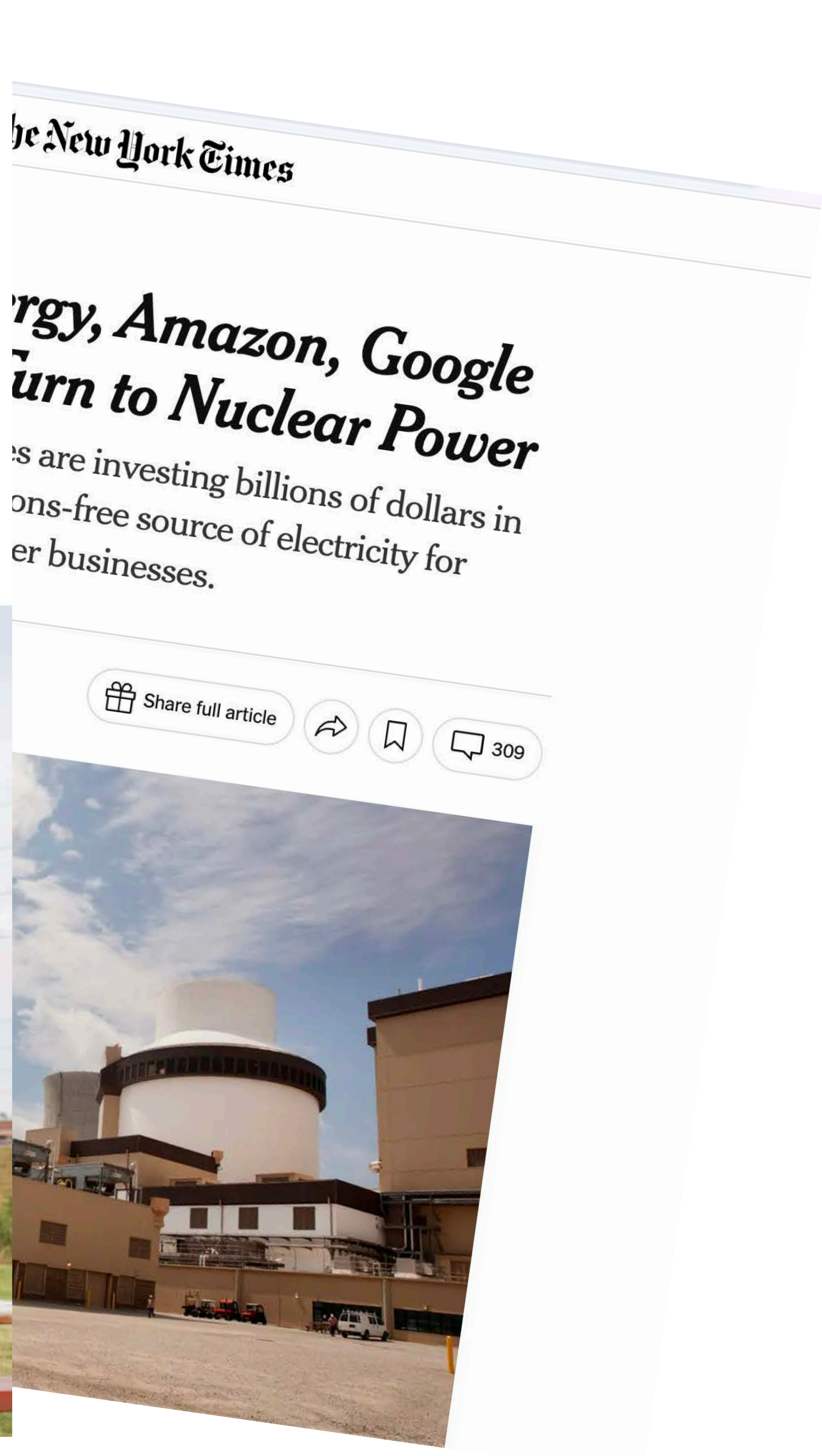
Listen to the local voice

Localized environmental burdens

Listen to the local voice

Localized environmental burdens





Long-term Regularized Online Optimization

Problem formulation

Goal

Hitting cost

$$\min_{x_t \in \mathbb{X}} \sum_{t=1}^{T} f(x_t, y_t) + c(x_t, x_{t-1}) + h\left(\sum_{t=1}^{T} p(x_t, a_t)\right)$$

Switching cost

Metrics

$$\boxed{\mathrm{AVG}(\pi) = \mathbb{E}\left[\mathrm{cost}(\pi, s)\right]}$$

Average Cost

$$\boxed{\mathrm{CR}(\pi) = \sup_{s \in \mathcal{S}} \frac{\mathrm{cost}(\pi, s)}{\mathrm{cost}(\pi^*, s)}}$$

Competitive Ratio

Problem formulation

Goal

Hitting cost

Long-term cost

$$\min_{x_t \in \mathbb{X}} \sum_{t=1}^{T} f(x_t, y_t) + c(x_t, x_{t-1}) + h\left(\sum_{t=1}^{T} p(x_t, a_t)\right)$$

Switching cost

Online Decision Making

$(y_1, c_1), x_1, (y_2, c_2), x_2 \ldots$

Metrics

$$\boxed{\mathrm{AVG}(\pi) = \mathbb{E}\left[\mathrm{cost}(\pi, s)\right]}$$

Average Cost

$$\boxed{\mathrm{CR}(\pi) = \sup_{s \in \mathcal{S}} \frac{\mathrm{cost}(\pi, s)}{\mathrm{cost}(\pi^*, s)}}$$

Competitive Ratio

A premier algorithm (R - OBD)

Goal

Hitting cost

Switching cost

$$\min_{x_t \in \mathbb{X}} \sum_{t=1}^{T} f(x_t, y_t) + c(x_t, x_{t-1})$$

Online Decision Making

$y_1, x_1, y_2, x_2 \ldots$

At each time t, receive context y_t

- $v_t \leftarrow \arg\min_x f(x, y_t)$
- $x_t \leftarrow \arg\min_x f(x, y_t) + \lambda_1 c(x, x_{t-1}) + \lambda_2 c(x, v_t)$

A premier algorithm (R - OBD)

Goal

Hitting cost

Switching cost

$$\min_{x_t \in \mathbb{X}} \sum_{t=1}^{T} f(x_t, y_t) + c(x_t, x_{t-1}) + h(\sum_{t=1}^{T} p(x_t, a_t))$$

Online Decision Making

$(y_1, a_1), x_1, (y_2, a_2), x_2 \ldots$

At each time t, receive context y_t

- $v_t \leftarrow \arg\min_x f(x, y_t)$
- $x_t \leftarrow \arg\min_x f(x, y_t) + \lambda_1 c(x, x_{t-1}) + \lambda_2 c(x, v_t)$

A premier algorithm (R - OBD)

Goal

Hitting cost

Switching cost

$$\min_{x_t \in \mathbb{X}} \sum_{t=1}^{T} f(x_t, y_t) + c(x_t, x_{t-1}) + h(\sum_{t=1}^{T} p(x_t, a_t))$$

Online Decision Making

$(y_1, a_1), x_1, (y_2, a_2), x_2 \ldots$

At each time t, receive context y_t

- $v_t \leftarrow \arg\min_x f(x, y_t)$
- $x_t \leftarrow \arg\min_x f(x, y_t) + \lambda_1 c(x, x_{t-1}) + \lambda_2 c(x, v_t)$

Problem reformulation

Objective function

$$\min_{x_t \in \mathbb{X}} \sum_{t=1}^{T} f(x_t, y_t) + c(x_t, x_{t-1}) + \sum_{t=1}^{T} h(z_t)$$

$$s.t. \quad \sum_{t=1}^{T} p(x_t, c_t) \le \boxed{\sum_{t=1}^{T} z_t}$$

z_t **is the location-wise "budget"**

Problem reformulation

Objective function

Hitting cost

$$\min_{x_t \in \mathbb{X}} \sum_{t=1}^{T} f(x_t, y_t) + c(x_t, x_{t-1}) + \sum_{t=1}^{T} h(z_t)$$

$$s.t. \quad \sum_{t=1}^{T} p(x_t, c_t) \leq \sum_{t=1}^{T} z_t$$

z_t is the location-wise "budget"

Problem reformulation

Objective function

Hitting cost

Decoupled cost

$$\min_{x_t \in \mathbb{X}} \sum_{t=1}^{T} f(x_t, y_t) + c(x_t, x_{t-1}) + \sum_{t=1}^{T} h(z_t)$$

$$s.t. \quad \sum_{t=1}^{T} p(x_t, c_t) \leq \boxed{\sum_{t=1}^{T} z_t}$$

z_t is the location-wise "budget"

Equity-aware online optimization

$$\min_{x_t \in \mathbb{X}} \frac{1}{T} \sum_{t=1}^{T} \left[f(x_t, y_t) + c(x_t, x_{t-1}) + h(z_t) + \mu \cdot \Big(p(x_t, a_t) - z_t \Big) \right]$$

Equity-aware online optimization

Dual variable

$$\min_{x_t \in \mathbb{X}} \frac{1}{T} \sum_{t=1}^{T} \left[f(x_t, y_t) + c(x_t, x_{t-1}) + h(z_t) + \boxed{\mu \cdot} \left(p(x_t, a_t) - z_t \right) \right]$$

Equity-aware online optimization

Dual variable

$$\min_{x_t \in \mathbb{X}} \frac{1}{T} \sum_{t=1}^{T} \left[f(x_t, y_t) + c(x_t, x_{t-1}) + h(z_t) + \mu \cdot \Big(p(x_t, a_t) - z_t \Big) \right]$$

At each time t, receive context (y_t, a_t)

- $x_t \leftarrow \arg\min_{x \in \mathcal{X}} f(x, y_t) + \lambda_1 c(x, x_{t-1}) + \lambda_2 c(x, v_t) + \mu_t \cdot p(x, a_t)$
- $z_t = \min_{z \in \mathcal{Z}} h(z) - \mu_t z_t$
- $\mu = \arg\min_{\mu} \langle z_t - p(x, a_t) \rangle + \frac{1}{\eta} V_h(\mu, \mu_t)$

Equity-aware online optimization

Dual variable

$$\min_{x_t \in \mathbb{X}} \frac{1}{T} \sum_{t=1}^{T} \left[f(x_t, y_t) + c(x_t, x_{t-1}) + h(z_t) + \boxed{\mu \cdot} \Big(p(x_t, a_t) - z_t \Big) \right]$$

At each time t, receive context (y_t, a_t)

- $x_t \leftarrow \arg\min_{x \in \mathcal{X}} f(x, y_t) + \lambda_1 c(x, x_{t-1}) + \lambda_2 c(x, v_t) + \mu_t \cdot p(x, a_t)$
- $z_t = \min_{z \in \mathcal{Z}} h(z) - \mu_t z_t$
- $\mu = \arg\min_{\mu} \langle z_t - p(x, a_t) \rangle + \frac{1}{\eta} V_h(\mu, \mu_t)$

Equity-aware online optimization

Dual variable

$$\min_{x_t \in \mathbb{X}} \frac{1}{T} \sum_{t=1}^{T} \left[f(x_t, y_t) + c(x_t, x_{t-1}) + h(z_t) + \mu \cdot \Big(p(x_t, a_t) - z_t \Big) \right]$$

At each time t, receive context (y_t, a_t)

- $x_t \leftarrow \arg\min_{x \in \mathcal{X}} f(x, y_t) + \lambda_1 c(x, x_{t-1}) + \lambda_2 c(x, v_t) + \mu_t \cdot p(x, a_t)$
- $z_t = \min_{z \in \mathcal{Z}} h(z) - \mu_t z_t$
- $\mu = \arg\min_{\mu} \langle z_t - p(x, a_t) \rangle + \frac{1}{\eta} V_h(\mu, \mu_t)$

Equity-aware online optimization

Dual variable

$$\min_{x_t \in \mathbb{X}} \frac{1}{T} \sum_{t=1}^{T} \left[f(x_t, y_t) + c(x_t, x_{t-1}) + h(z_t) + \mu \cdot \Big(p(x_t, a_t) - z_t \Big) \right]$$

At each time t, receive context (y_t, a_t)

- $x_t \leftarrow \arg\min_{x \in \mathcal{X}} f(x, y_t) + \lambda_1 c(x, x_{t-1}) + \lambda_2 c(x, v_t) + \mu_t \cdot p(x, a_t)$
- $z_t = \min_{z \in \mathcal{Z}} h(z) - \mu_t z_t$
- $\mu = \arg\min_{\mu} \langle z_t - p(x, a_t) \rangle + \frac{1}{\eta} V_h(\mu, \mu_t)$

Performance analysis

Theorem (informal)

When $T \to \infty$, for any finite R, the cost of eGLB satisfies

$$cost(eGLB) \leq C \cdot cost(\text{OPT}^*) + O\left(\sqrt{\frac{1}{T}}\right) + \frac{L \cdot \delta}{T}$$

Performance analysis

Theorem (informal)

When $T \to \infty$, for any finite R, the cost of eGLB satisfies

$$cost(eGLB) \le C \cdot cost(\text{OPT}^*) + O\left(\sqrt{\frac{1}{T}}\right) + \frac{L \cdot \delta}{T}$$

(R, δ)-constrained offline algorithm:

$$\min_{x_{1:T}} \frac{1}{T} \sum_{t=1}^{T} cost_t(x_t) + g\left(\frac{1}{T} \sum c_t x_t\right)$$

$$\text{s.t.,} \quad \sum_k \left\| \sum_{(k-1)R+1}^{kR} c_t x_t - \frac{R}{T} \sum_1^T c_t x_t \right\| \le \delta$$

Performance analysis

Theorem (informal)

When $T \rightarrow \infty$, for any finite R, the cost of eGLB satisfies

$$cost(eGLB) \leq C \cdot cost(\mathrm{OPT}^{*}) + O\left(\sqrt{\frac{1}{T}}\right) + \frac{L \cdot \delta}{T}$$

(R, δ)-constrained offline algorithm:

$$\min_{x_{1:T}} \frac{1}{T} \sum_{t=1}^{T} cost_t(x_t) + g\left(\frac{1}{T} \sum c_t x_t\right)$$

$$\text{s.t.,} \quad \sum_{k} \left|\left| \sum_{(k-1)R+1}^{kR} c_t x_t - \frac{R}{T} \sum_{1}^{T} c_t x_t \right|\right| \leq \delta$$

Performance analysis

Theorem (informal)

When $T \to \infty$, for any finite R, the cost of eGLB satisfies

$$cost(eGLB) \leq C \cdot cost(\mathrm{OPT}^*) + O\left(\sqrt{\frac{1}{T}}\right) + \frac{L \cdot \delta}{T}$$

(R, δ)-constrained offline algorithm:

$$\min_{x_{1:T}} \frac{1}{T} \sum_{t=1}^{T} cost_t(x_t) + g\left(\frac{1}{T} \sum c_t x_t\right)$$

$$\text{s.t.,} \quad \sum_k || \sum_{(k-1)R+1}^{kR} c_t x_t - \frac{R}{T} \sum_1^T c_t x_t || \leq \delta$$

~ **"Total variation"**

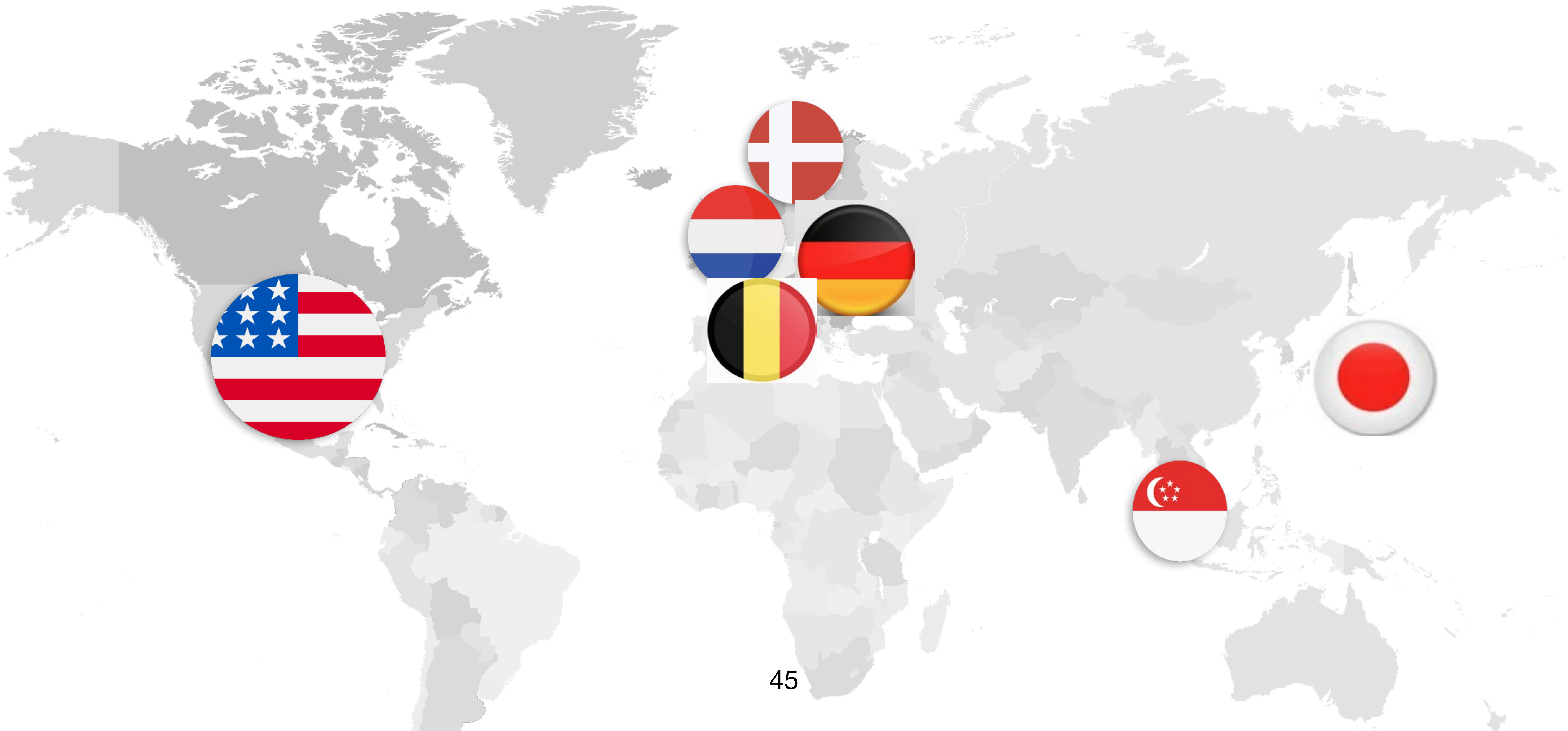
A case study

Country	State/Province	City	Total WUE (m^3/MWh)	Carbon Intensity (ton/MWh)	Energy Price ($/MWh)
U.S.	Texas	Midlothian	5.7397	0.4011	64.931
U.S.	Virginia	Loudoun	5.9755	0.3741	77.793
U.S.	Georgia	Douglas	5.9001	0.4188	80.566
U.S.	Nevada	Storey	4.9306	0.2980	84.738
Germany	Hessen	Frankfurt	4.5889	0.3295	315.233
Belgium	Hainaut	Saint-Ghislain	4.9316	0.4802	247.083
Netherlands	Groningen	Eemshaven	3.0928	0.4454	248.258
Denmark	Fredericia	Fredericia	3.8900	0.1391	213.773
Japan	Chiba Prefecture	Inzai	2.4989	0.3280	129.269
Singapore	Singapore	Jurong West	5.8652	0.5260	155.462

- 10 different data center locations (4 in the US, 4 in Europe, and 2 in Asia)
- BLOOM inference trace (scaled up)
- Environmental costs: Water and carbon footprints

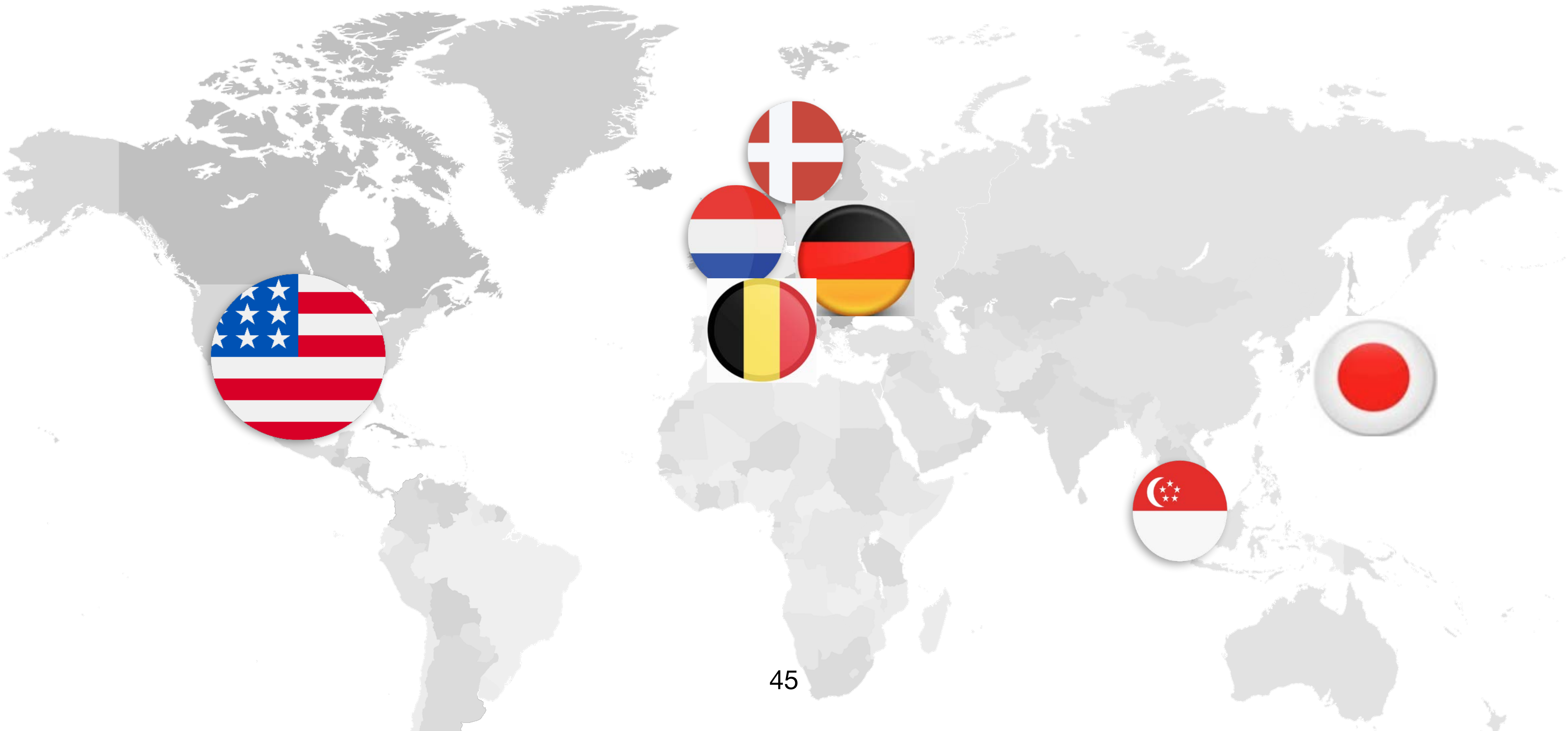
A case study

Metric		Algorithm							
		GLB-Energy	GLB-Carbon	GLB-Water	GLB-C2	GLB-All	GLB-Nearest	eGLB-Off	**eGLB**
Energy (US$)	avg	279620	454608	539847	326104	312372	450992	341998	**359433**
Water (m^3)	avg	14329.6	12992.8	11694.2	13822.4	13338.4	13584.9	13439.3	**13591.5**
	max	23753.4	24779.5	19478.0	25154.2	21307.6	19662.3	16339.6	**18199.0**
	max/avg	1.66	1.91	1.67	1.82	1.60	1.45	1.22	**1.34**
Carbon (ton)	avg	1098.29	830.66	947.89	925.28	975.76	1035.97	951.91	**977.92**
	max	1868.37	1544.89	2110.61	1566.99	1656.06	1342.44	1202.91	**1294.23**
	max/avg	1.70	1.86	2.23	1.69	1.70	1.30	1.26	**1.32**



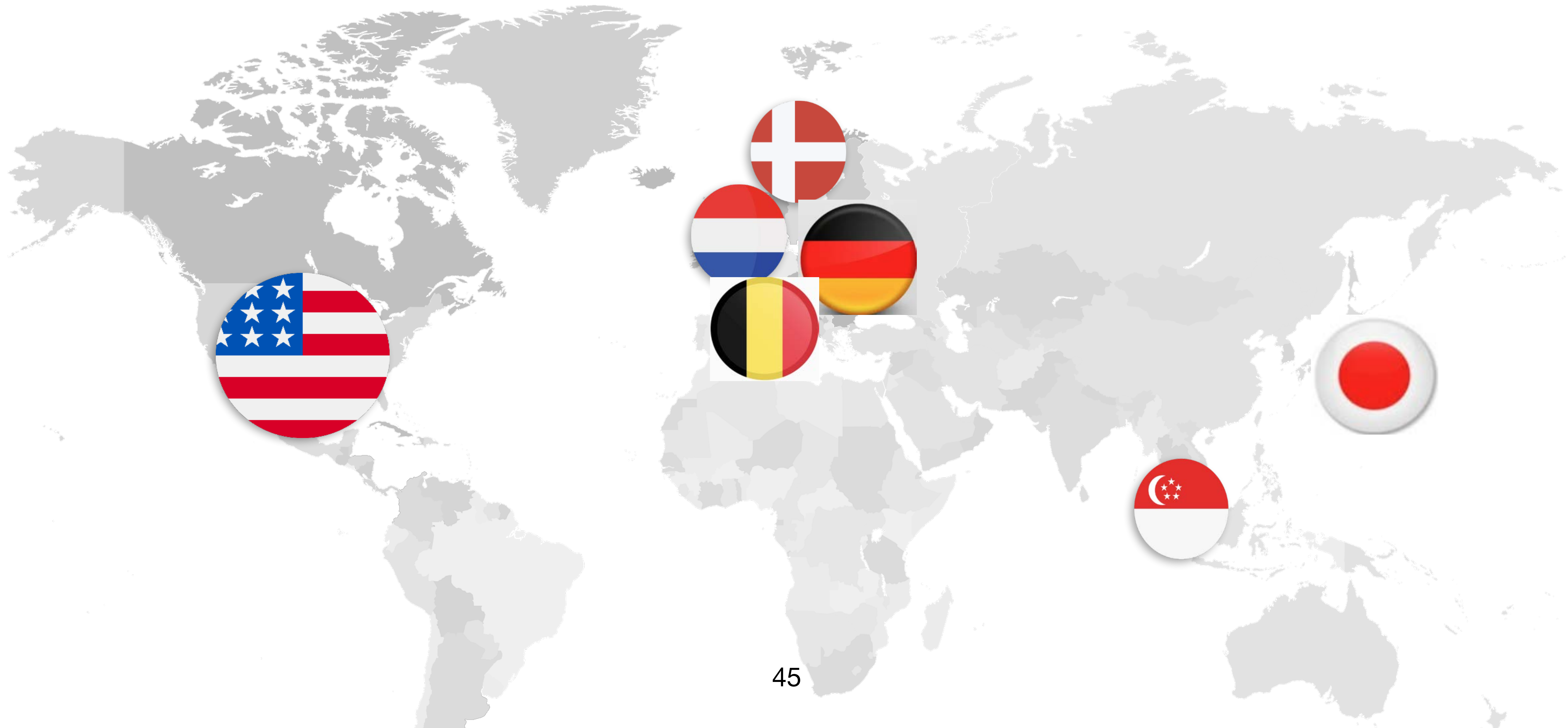
A case study

Metric		Algorithm							
		GLB-Energy	GLB-Carbon	GLB-Water	GLB-C2	GLB-All	GLB-Nearest	eGLB-Off	eGLB
Energy (US$)	avg	279620	454608	539847	326104	312372	450992	341998	**359433**
Water (m^3)	avg	14329.6	12992.8	11694.2	13822.4	13338.4	13584.9	13439.3	**13591.5**
	max	23753.4	24779.5	19478.0	25154.2	21307.6	19662.3	16339.6	**18199.0**
	max/avg	1.66	1.91	1.67	1.82	1.60	1.45	1.22	**1.34**
Carbon (ton)	avg	1098.29	830.66	947.89	925.28	975.76	1035.97	951.91	**977.92**
	max	1868.37	1544.89	2110.61	1566.99	1656.06	1342.44	1202.91	**1294.23**
	max/avg	1.70	1.86	2.23	1.69	1.70	1.30	1.26	**1.32**



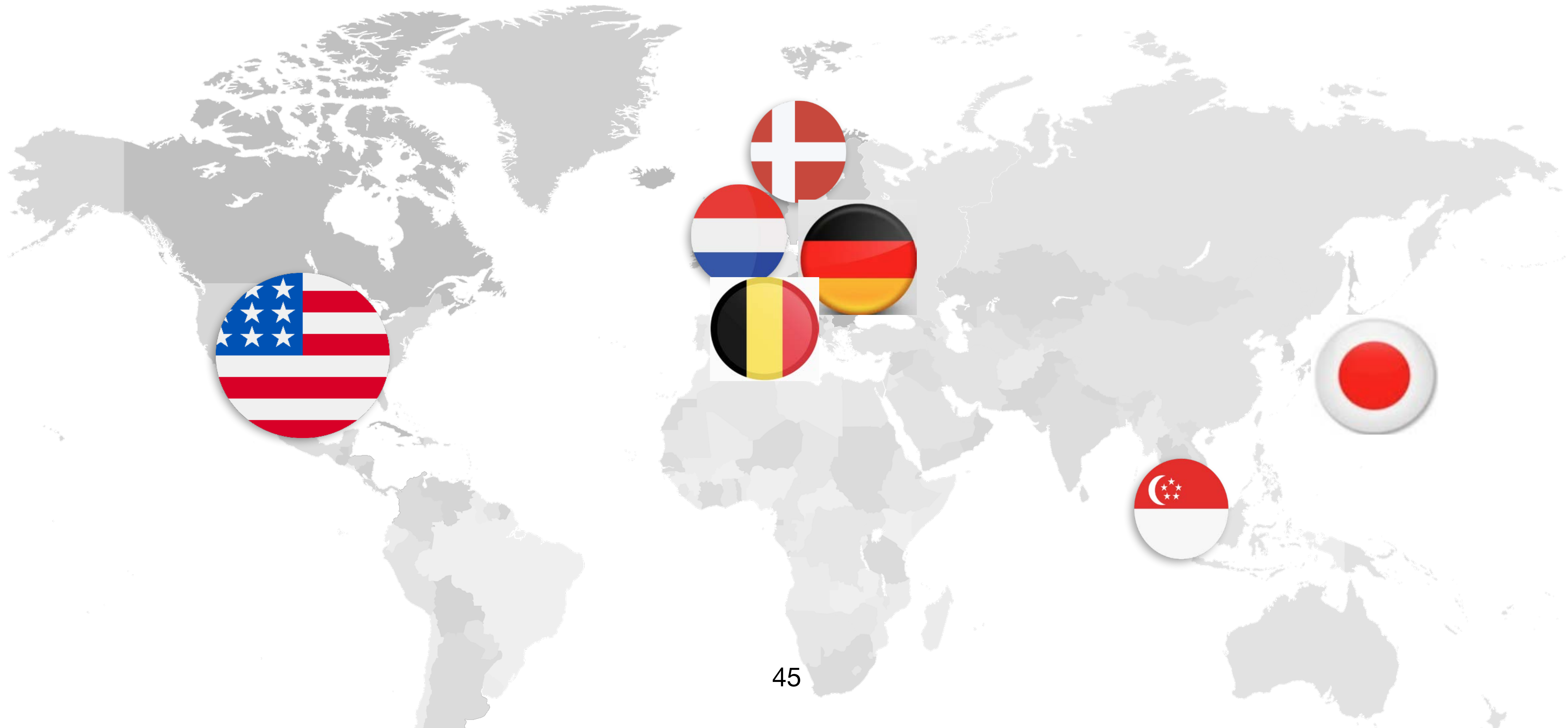
A case study

Metric		Algorithm							
		GLB-Energy	GLB-Carbon	GLB-Water	GLB-C2	GLB-All	GLB-Nearest	eGLB-Off	eGLB
Energy (US$)	avg	279620	454608	539847	326104	312372	450992	341998	359433
Water (m^3)	avg	14329.6	12992.8	11694.2	13822.4	13338.4	13584.9	13439.3	13591.5
	max	23753.4	24779.5	19478.0	25154.2	21307.6	19662.3	16339.6	18199.0
	max/avg	1.66	1.91	1.67	1.82	1.60	1.45	1.22	1.34
Carbon (ton)	avg	1098.29	830.66	947.89	925.28	975.76	1035.97	951.91	977.92
	max	1868.37	1544.89	2110.61	1566.99	1656.06	1342.44	1202.91	1294.23
	max/avg	1.70	1.86	2.23	1.69	1.70	1.30	1.26	1.32



A case study

Metric		Algorithm							
		GLB-Energy	GLB-Carbon	GLB-Water	GLB-C2	GLB-All	GLB-Nearest	eGLB-Off	eGLB
Energy (US$)	avg	279620	454608	539847	326104	312372	450992	341998	**359433**
Water (m^3)	avg	14329.6	12992.8	11694.2	13822.4	13338.4	13584.9	13439.3	**13591.5**
	max	23753.4	24779.5	19478.0	25154.2	21307.6	19662.3	16339.6	**18199.0**
	max/avg	1.66	1.91	1.67	1.82	1.60	1.45	1.22	**1.34**
Carbon (ton)	avg	1098.29	830.66	947.89	925.28	975.76	1035.97	951.91	**977.92**
	max	1868.37	1544.89	2110.61	1566.99	1656.06	1342.44	1202.91	**1294.23**
	max/avg	1.70	1.86	2.23	1.69	1.70	1.30	1.26	**1.32**



A case study

Metric		Algorithm							
		GLB-Energy	GLB-Carbon	GLB-Water	GLB-C2	GLB-All	GLB-Nearest	eGLB-Off	eGLB
Energy (US$)	avg	279620	454608	539847	326104	312372	450992	341998	359433
Water (m^3)	avg	14329.6	12992.8	11694.2	13822.4	13338.4	13584.9	13439.3	13591.5
	max	23753.4	24779.5	19478.0	25154.2	21307.6	19662.3	16339.6	18199.0
	max/avg	1.66	1.91	1.67	1.82	1.60	1.45	1.22	1.34
Carbon (ton)	avg	1098.29	830.66	947.89	925.28	975.76	1035.97	951.91	977.92
	max	1868.37	1544.89	2110.61	1566.99	1656.06	1342.44	1202.91	1294.23
	max/avg	1.70	1.86	2.23	1.69	1.70	1.30	1.26	1.32

eGLB mitigates AI's environmental inequity (at a small cost)

Thanks

Q & A