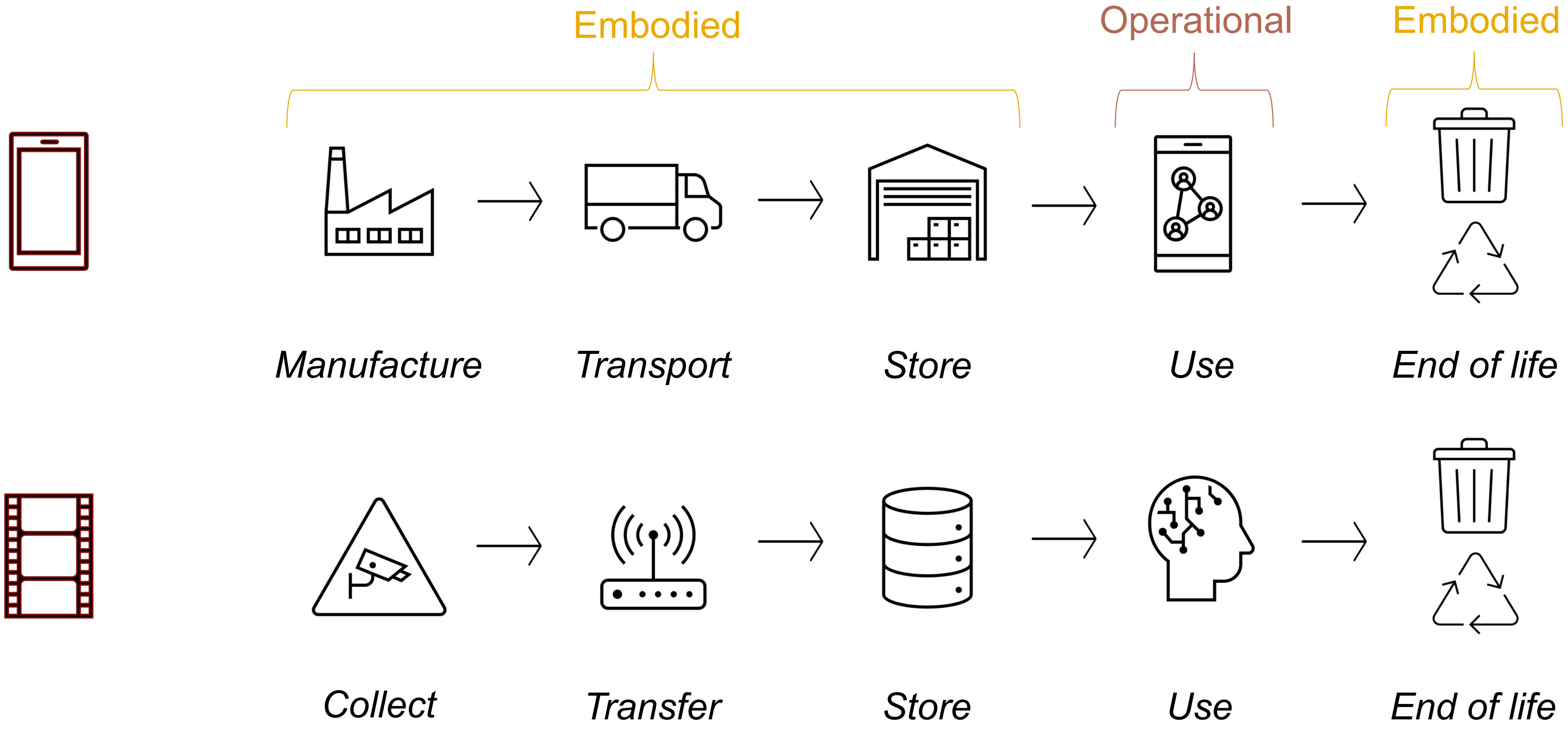
Toward a Life Cycle Assessment for the Carbon Footprint of Data

Gabriel Mersy and Sanjay Krishnan

HotCarbon ’23

Data as a good with a life cycle



Embodied

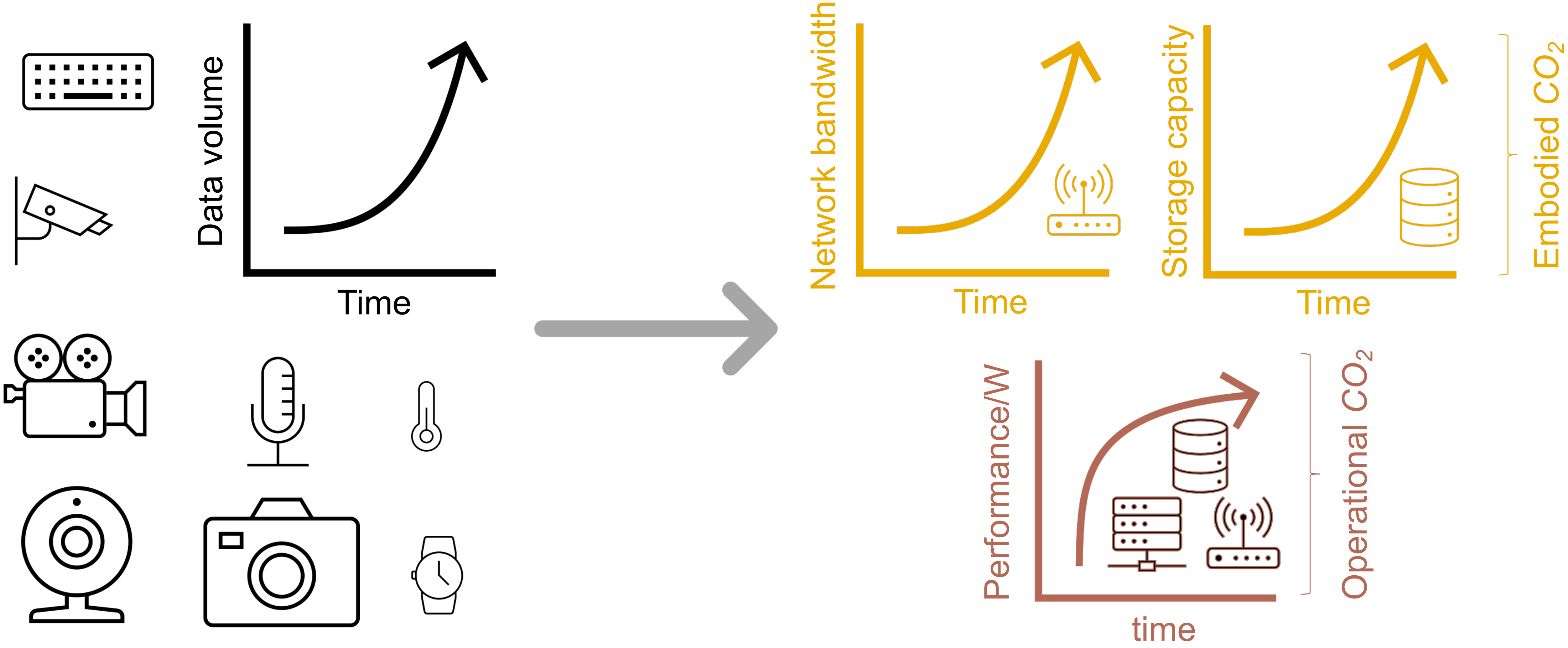
Operational

Embodied

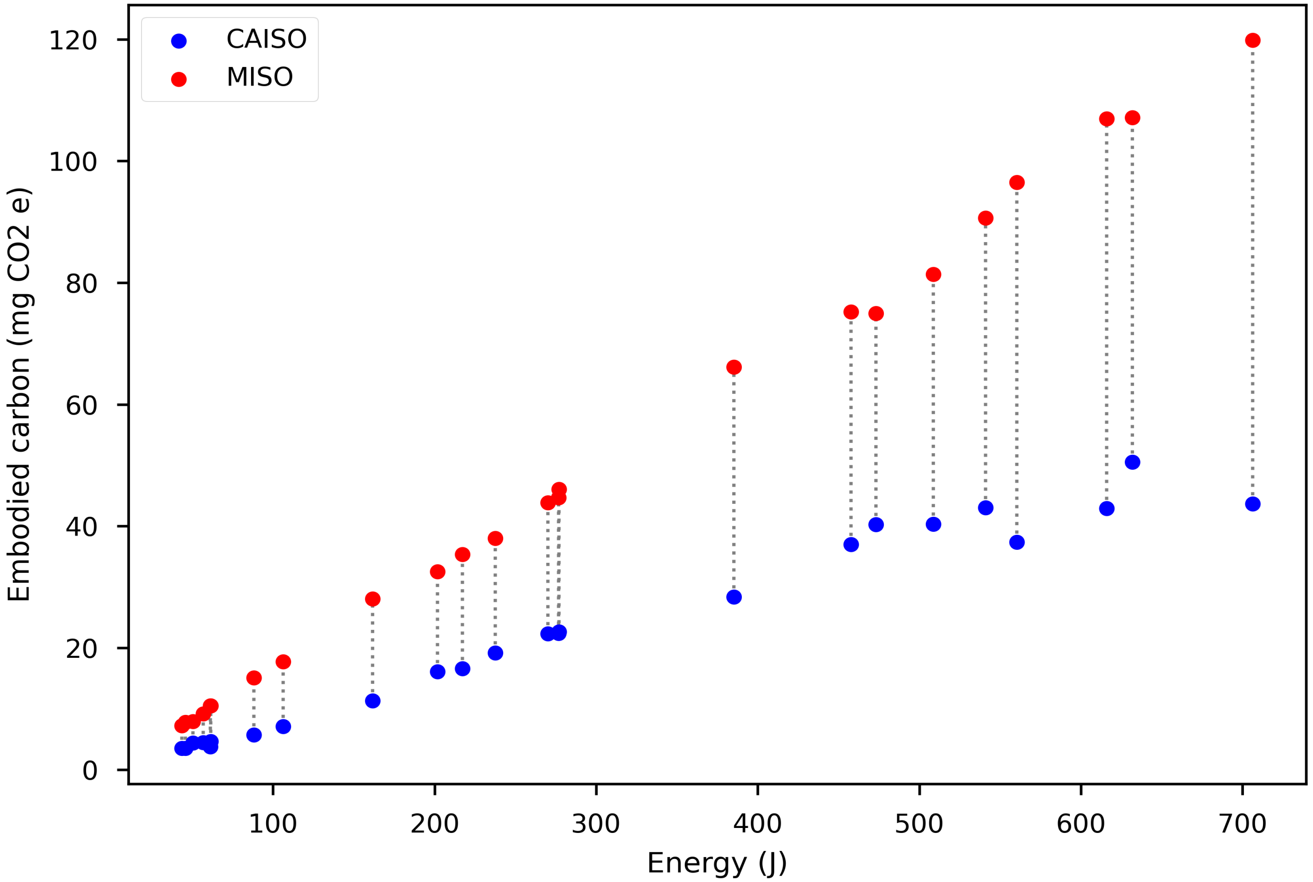
Manufacture → Transport → Store → Use → End of life

Collect → Transfer → Store → Use → End of life

Data sustainability: the hidden carbon costs

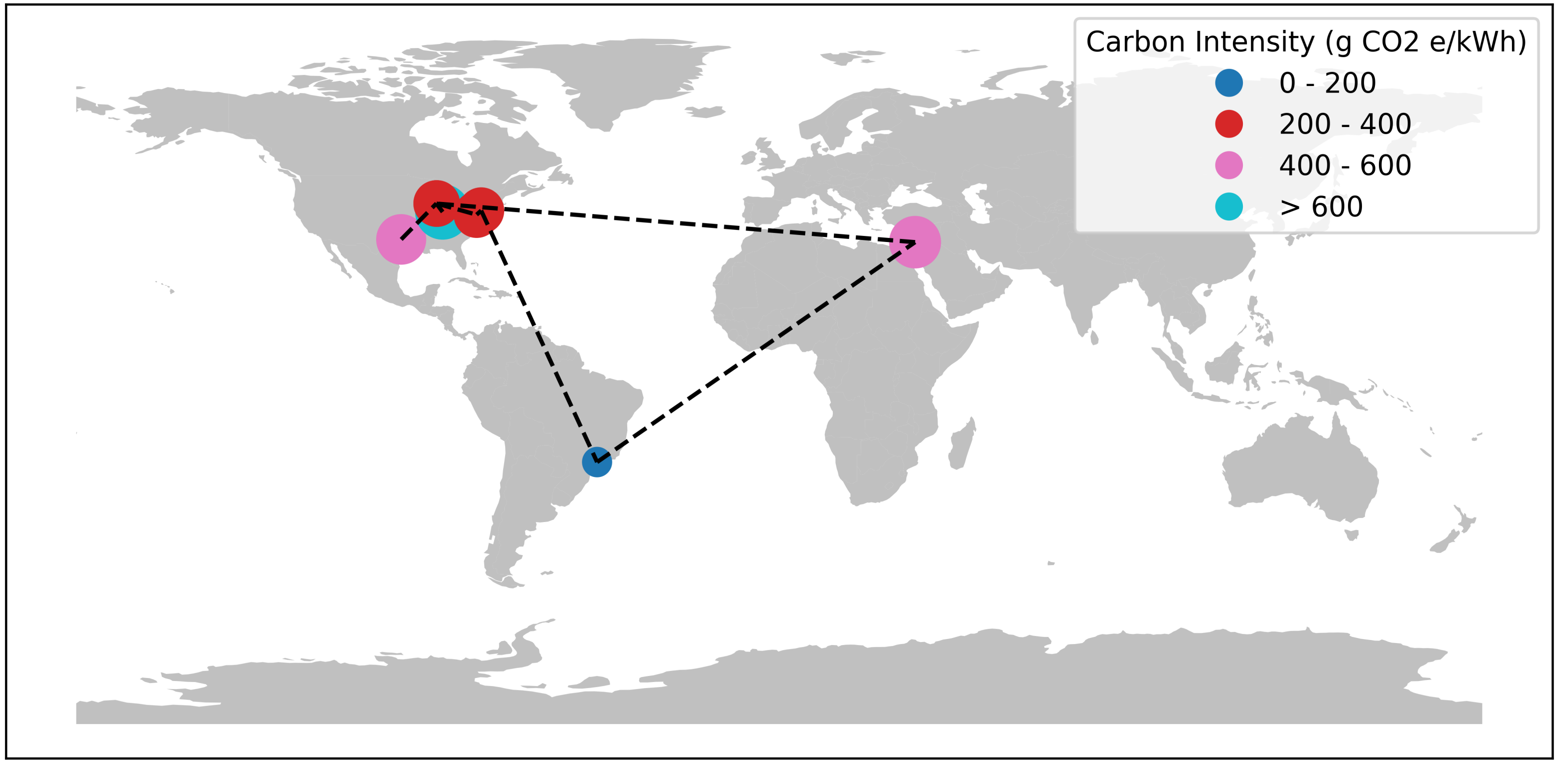


Data collection carbon costs are often overlooked



e.g., 26 second webcam video: 37 mg – 119 mg CO_2 e

Communication (also) matters



Core path to data center: 1.51 g CO_2 e/GB

Talk outline

Carbon provenance: tracking carbon costs across the data life cycle

Carbon-responsive data: reducing carbon costs by approximating data

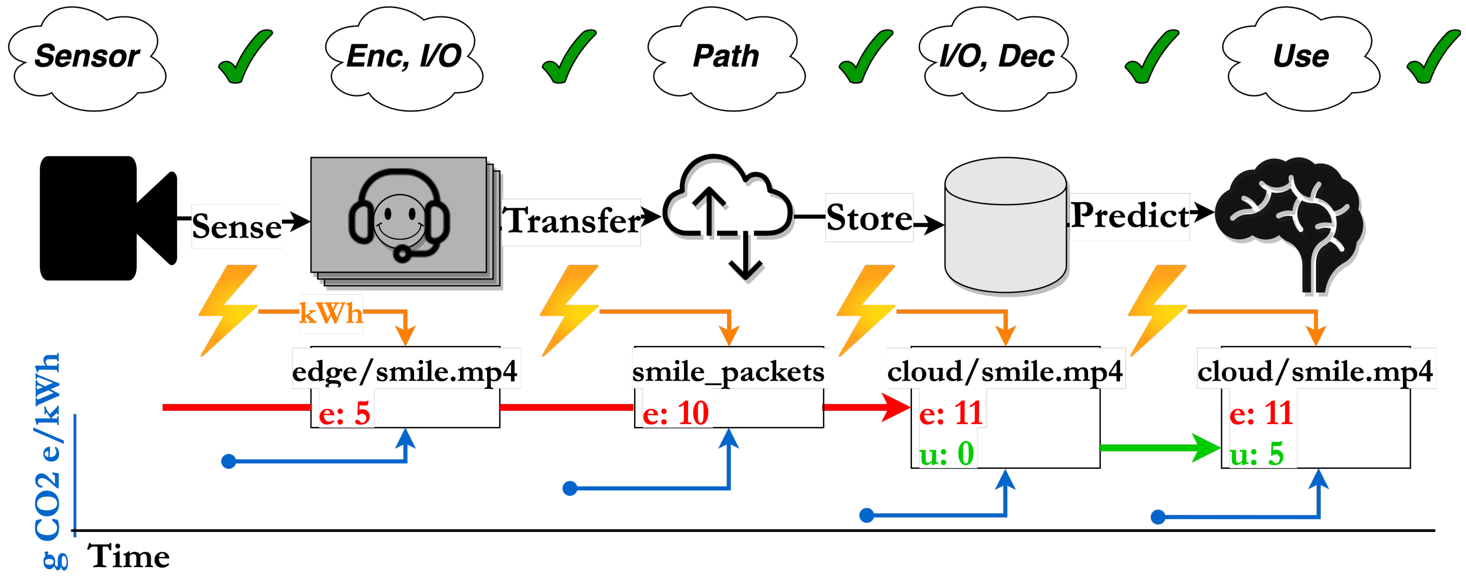
Talk outline

Carbon provenance: tracking carbon costs across the data life cycle

Carbon-responsive data: reducing carbon costs by approximating data

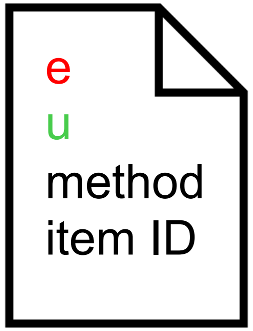
Carbon provenance: a carbon LCA for data

- Associate two annotations with each data item
  - Embodied (collection, transfer, storage)
  - Operational (use)

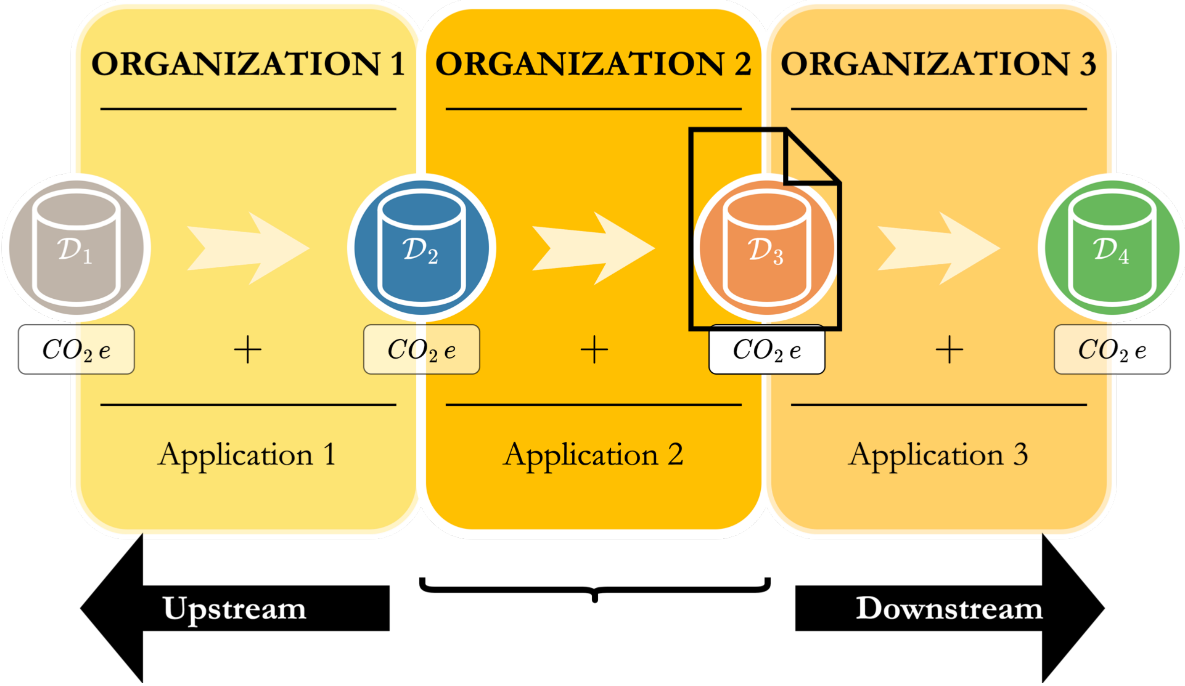


What about purchased data?

- Value chain accounting
- Goal: link carbon costs across entities when a data item is sold
- Idea: carbon header



e
u
method
item ID



Talk outline

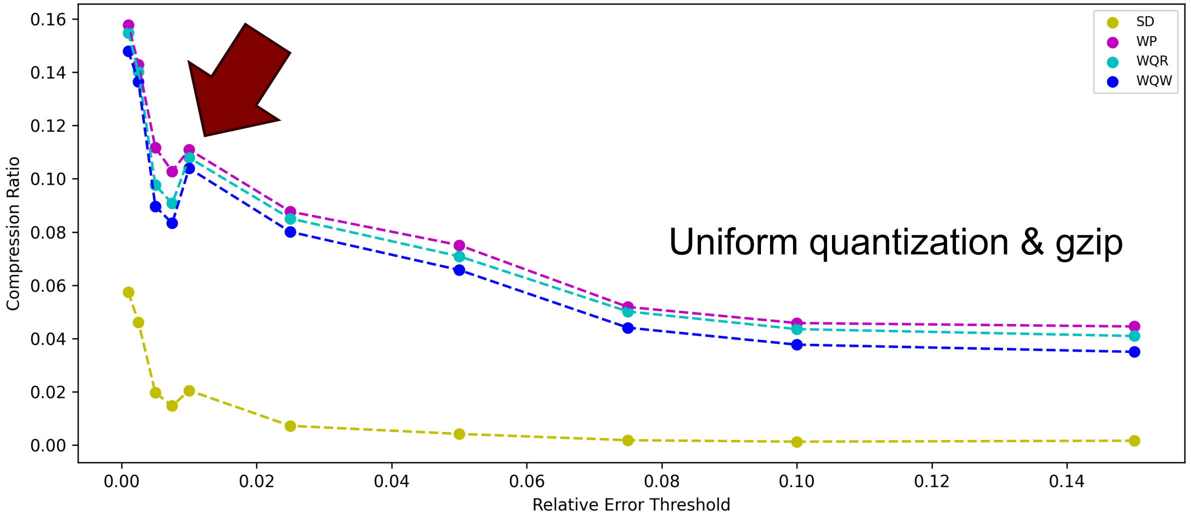
Carbon provenance: tracking carbon costs across the data life cycle

Carbon-responsive data: reducing carbon costs by approximating data

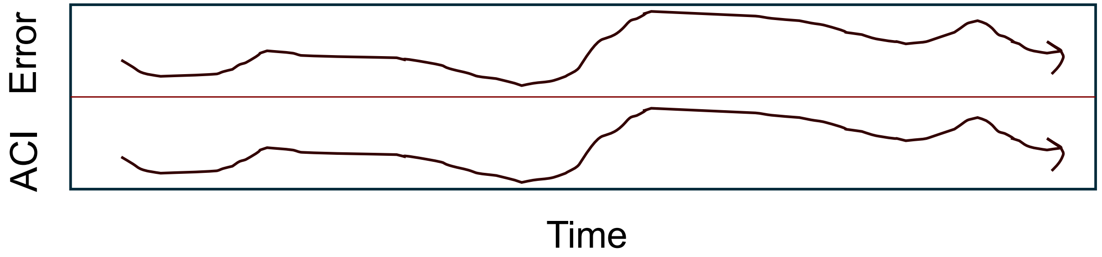
Approximation: a little error can go a long way

- Trade **error** for cost savings (e.g., **energy,** storage size, latency) in certain non-mission-critical use cases
  - Lossy compression: JPEG/H.264/MP3
  - AQP: sketching/sampling
  - ML: neural network pruning
- **Dynamically use error to reduce carbon costs**
  - No workload shifting necessary
- Required: an **error policy**

1% worst-case error, ≈10x space savings!

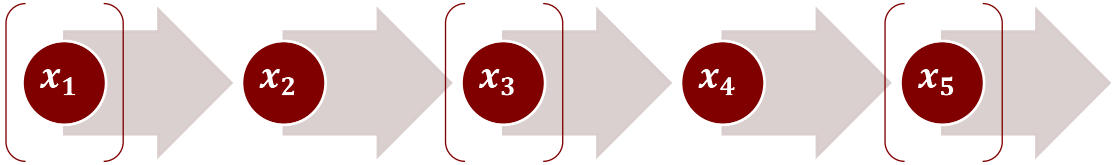


Uniform quantization & gzip



Carbon-adaptive data science

- Adapt error in certain DS workloads according to carbon intensity
  - Queries
  - ML inference
- Example: mean of a stream
- Error policy
  - High ACI ⟶ sample 3/5 values
  - Low ACI ⟶ use 5/5 values

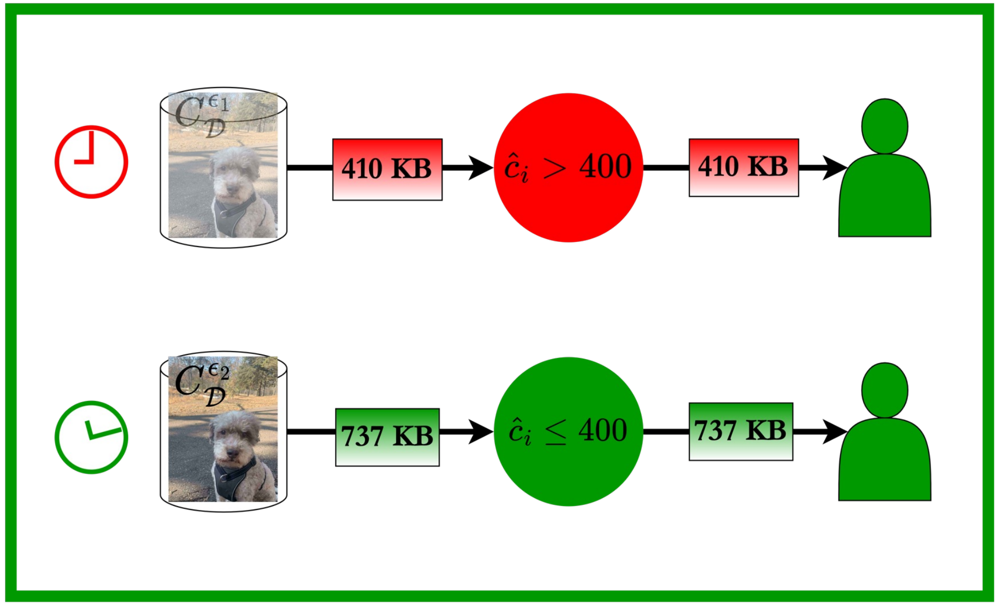


Carbon-adaptive compression

- Multiresolution compression [SIGMOD ‘23]: encoding that combines sub-encodings with different errors (& sizes)

$$C_{\mathcal{D}} = \boxed{C_{\mathcal{D}}^{\epsilon_1}} \bigoplus \boxed{C_{\mathcal{D}}^{\epsilon_2}} \bigoplus \dots \bigoplus C_{\mathcal{D}}^{\epsilon_l}$$

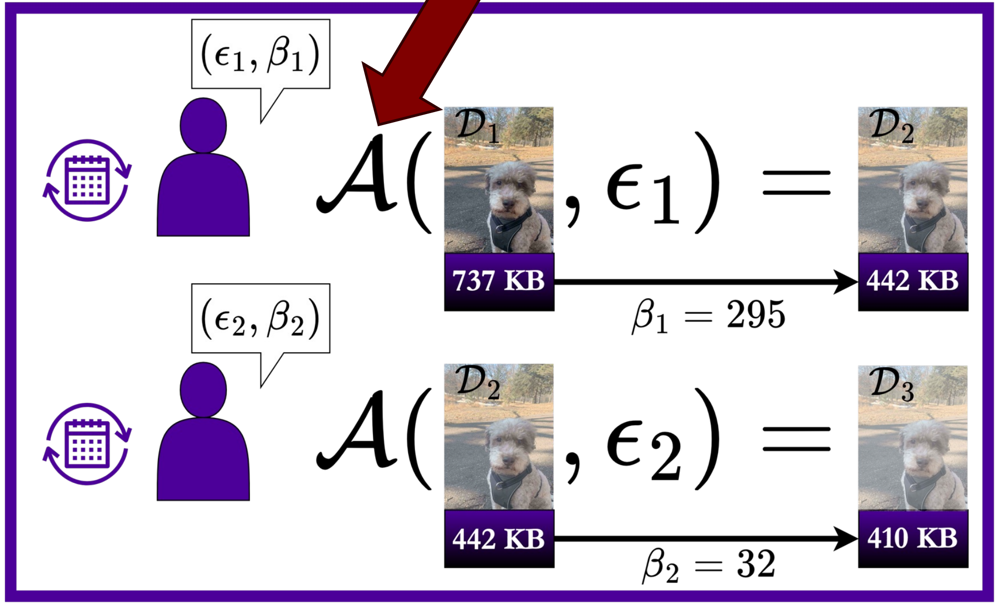
- Error policy: choose sub-encoding according to path carbon intensity



Data wrinkles: lossy data aging

- More data ⟶ more storage ⟶ more manufacturing carbon
- **Data disposal/fungi** [Milo 2019, Kersten 2015]: policies to discard or reduce quality
- Q: What is the grey area between retention and deletion?
- A: **recursively apply approximation operations over time**
  - (ϵ, β)-data wrinkle: ϵ error, $\beta > 0$ space

Approximation, e.g., JPEG



(ϵ_1, β_1)
$\mathcal{A}(\mathcal{D}_1, \epsilon_1) = \mathcal{D}_2$
737 KB
442 KB
$\beta_1 = 295$
(ϵ_2, β_2)
$\mathcal{A}(\mathcal{D}_2, \epsilon_2) = \mathcal{D}_3$
442 KB
410 KB
$\beta_2 = 32$

Summary

- **Data sustainability**: volume comes at a cost to the environment
- **Carbon provenance**: an **LCA for data**
  - **Embodied** and **operational** categories, just like hardware
- **Carbon-responsive data**: **error** can reduce carbon costs
  - Carbon-adaptive data science
  - Carbon-adaptive compression
  - Data wrinkles

Thanks!

Code:

https://github.com/gmersy/data-carbon

Email:

gmersy@uchicago.edu

Twitter:

@gabemersy