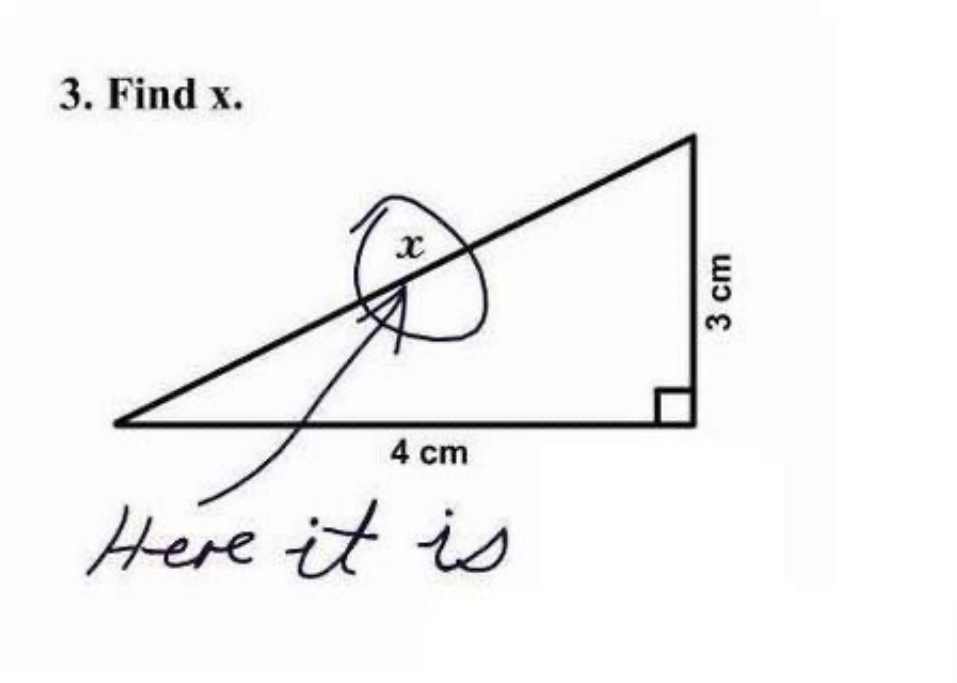
FROM DATA TO FUNCTIONALITY - TOWARDS AN INFORMATION THEORETIC ONTOLOGY

Muriel Médard

RLE

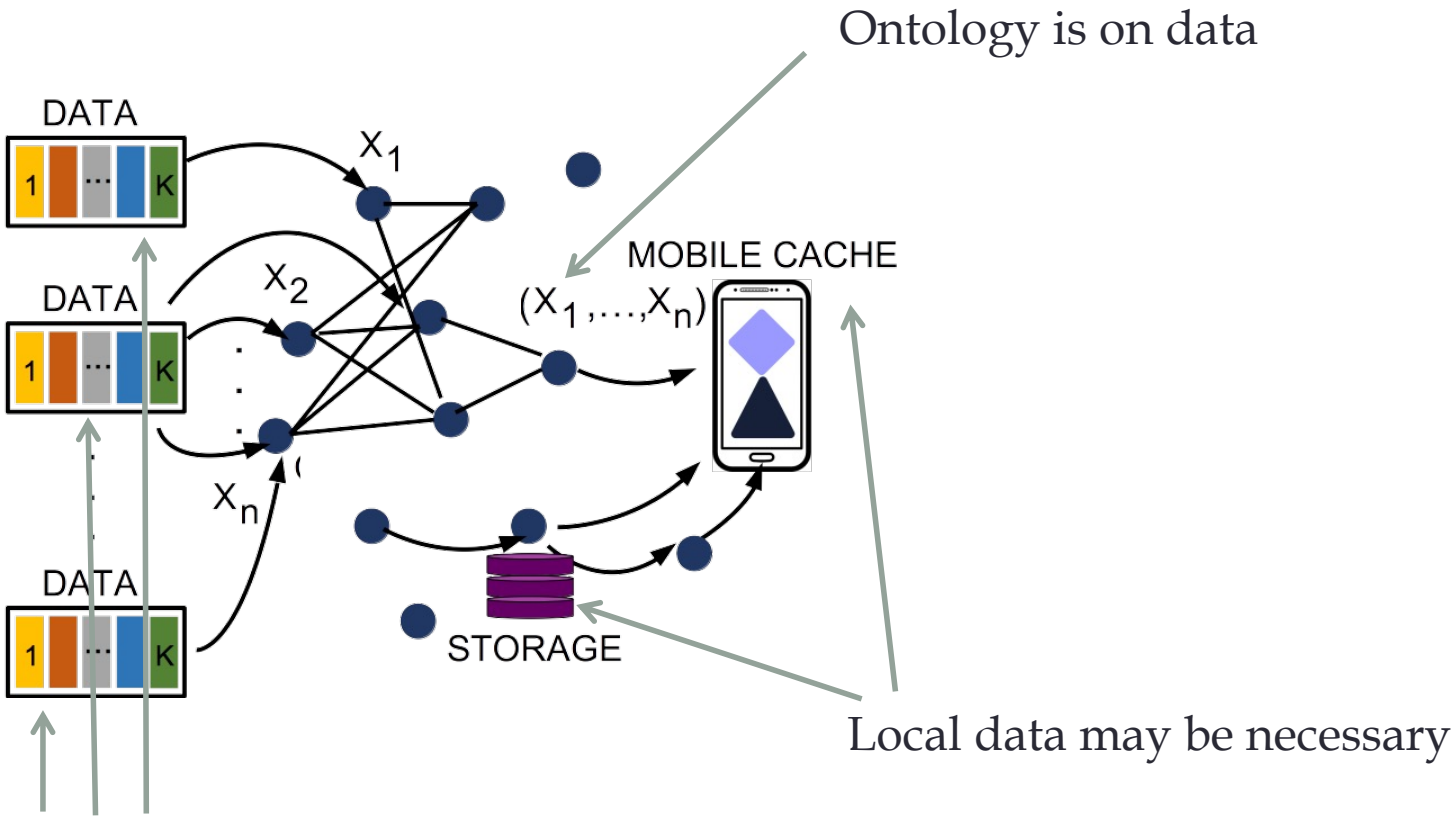
MIT

What is x?



memecenter.com

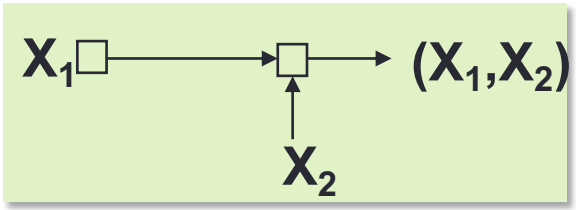
What does data mean - ontology



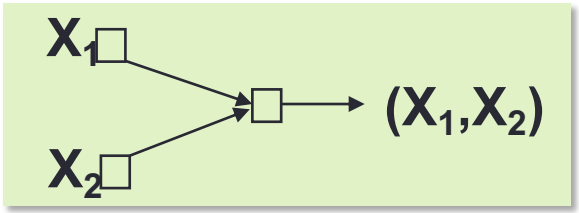
The data resides across different domains and locations

Data

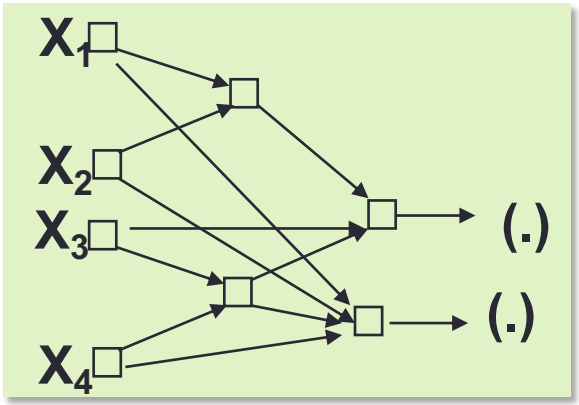
- Side information problem (Wyner-Ziv)



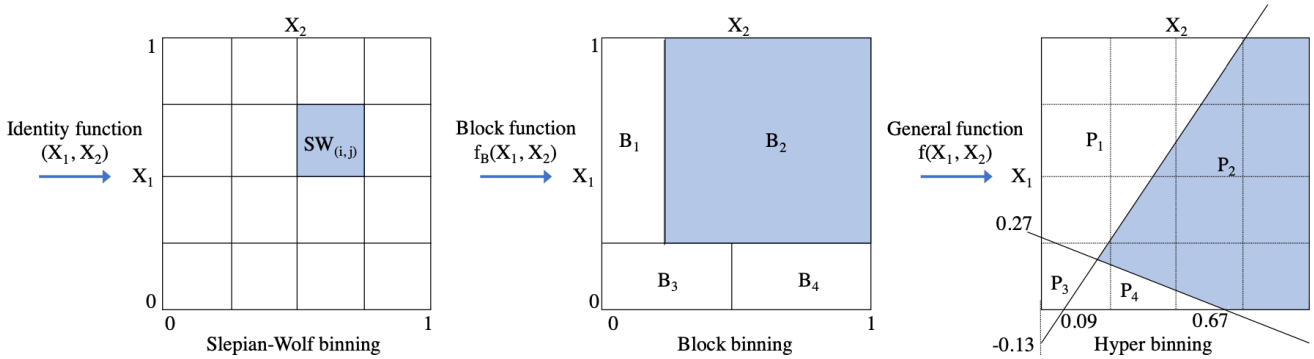
- Depth one trees (Slepian-Wolf)
  - Can be generalized to trees



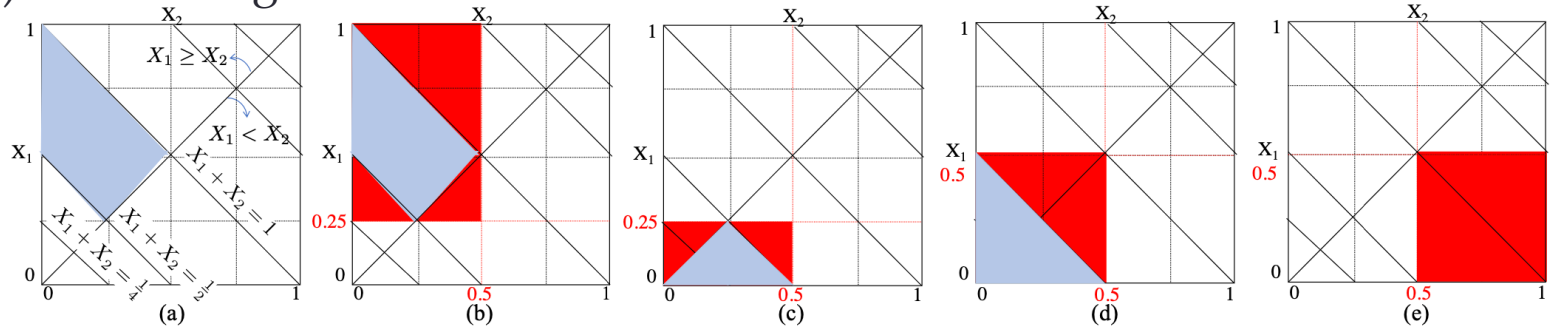
- General Networks (Ho et al)
  - Multicast
  - Random Linear Network Coding
  - Error exponents via method of types generalize Csiszar error exponents



Quantizing



(a) Computing a convex region.
(b)-(d) The source hyperplanes are emphasized (red) to represent the regions corresponding to several different function outcomes.
Each outcome is the intersection of the tessellation with the red shaded region.
(e) A decoding error occurs.



D. Malak, M. Médard, "A Distributed Computationally Aware Quantizer Design via Hyper Binning," *IEEE Transactions on Signal Processing*, vol. 71, pp. 76-9, January 2023

Low influence functions

- Low influence functions: No coordinate has too much control on the function.
- This is effectively what allows the relatively high level of tolerance to noise or data in inference.

Central Limit Theorem: For $X_1, X_2, \ldots, X_n$ i.i.d. with mean 0 and finite variance σ^2,

$$\sqrt{n} \cdot \overline{X}_n \xrightarrow{d} \mathcal{N}(0, \sigma^2).$$

Berry-Esseen Theorem: For $X_1, X_2, \ldots, X_n$ i.i.d. with mean 0, finite variance σ^2, and third moment bounded by ρ, let F_n be the CDF of $\sqrt{n} \cdot \overline{X}_n/\sigma$ and Φ be the standard Gaussian CDF. Then, for all x and n and some constant $C > 0$,

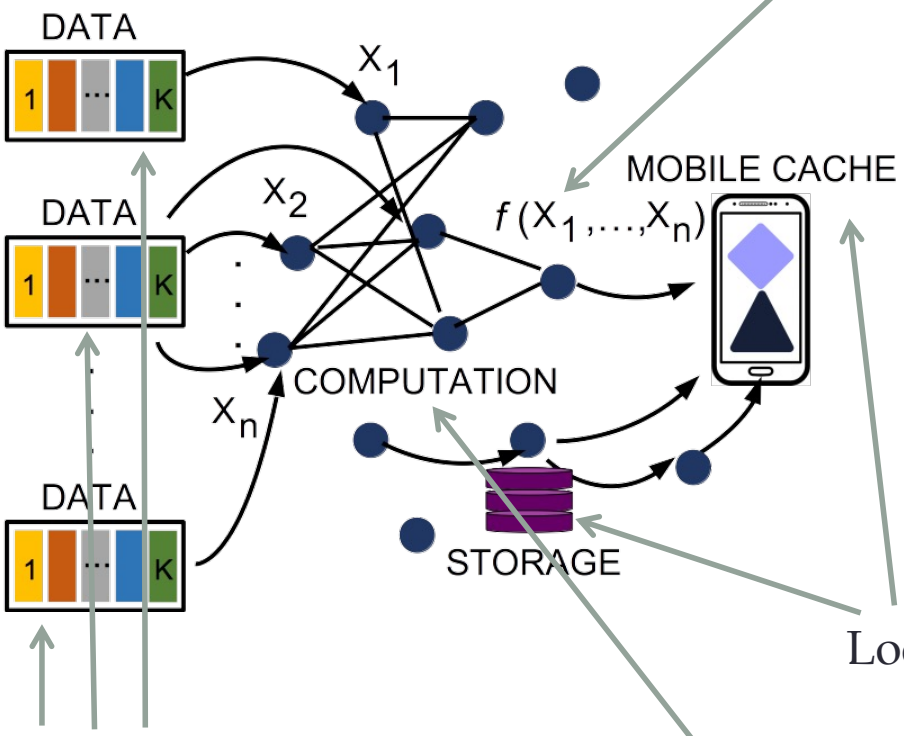
$$|F_n(x) - \Phi(x)| \leq \frac{C\rho}{\sigma^3\sqrt{n}}.$$

Rafael G. L. D'Oliveira, Salim El Rouayheb, Muriel Médard:, "The Computational Wiretap Channel". *Allerton Conference*, 2018

A. Mariona, H. Esfahanizadeh, R. G. L. D'Oliveira and M. Médard, "A Bivariate Invariance Principle," *IEEE Information Theory Workshop (ITW)*, pp. 756-761, Oct. 2022.

What does data mean - ontology

Your work requires you to exploit data



Computation storage, communications are inseparable

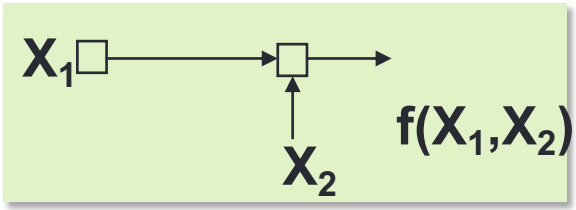
Local data may be necessary

The data resides across different domains and locations

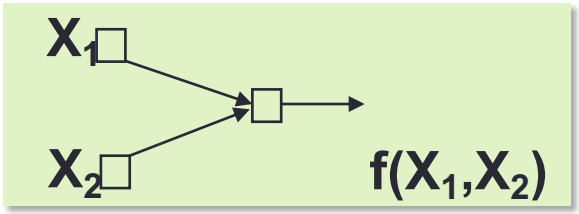
For reasons of dimensioning, scaling, security Computation needs to occur in network

Ontology

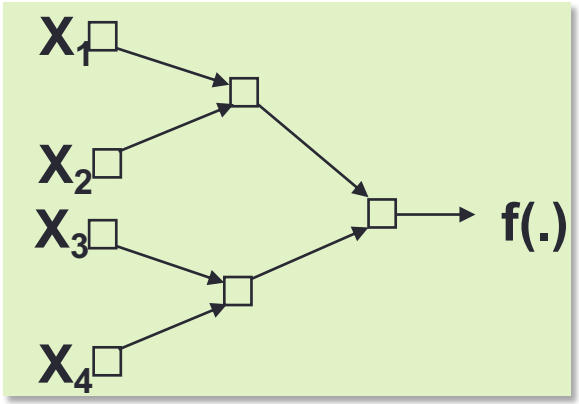
- Side Information Problem
  (Orlitsky and Roche, Ortlitsky and Alon)



- Depth One Trees (Doshi et al.)
  - Characterizing the rate region
  - A necessary and sufficient condition for achievability



- General Trees (Feizi and Médard)
  - Rate lower bounds for a general case
  - For independent sources: optimal coding schemes
  - Polynomial time (practical) coding schemes for some conditions
  - Feedback can improve rate bounds



V. Doshi, Shah, D., Médard, M., and Effros, M., "Functional Compression Through Graph Coloring", *IEEE Transactions on Information Theory*, Vol. 56, No. 8, Aug. 2010
S. Feizi, and Médard, M., "On Network Functional Compression", *Transactions on Information Theory,* vol. 60, no. 9, Sept.2014.

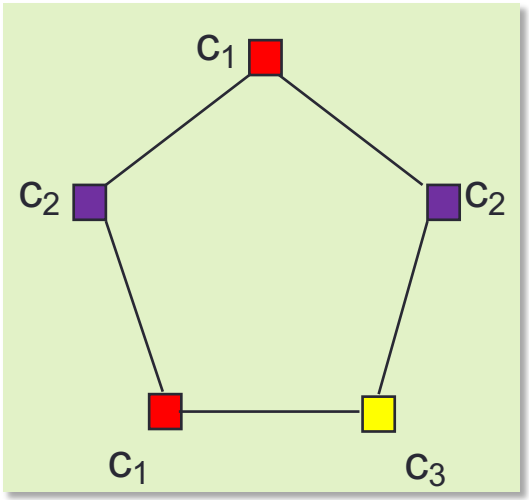
Characteristic Graph

- **Example:** X_1 is a RV with uniform distribution over {0,1,2,3,4}. X_2 is a RV such that we have the following graph G_{X_1} for

$$c_{G_{X_1}} = \{c_1, c_2, c_3\}.$$

$$P(c_1) = P(c_2) = 2/5 \qquad P(c_3) = 1/5$$

$$H(c_{G_{X_1}}) \approx 1.52$$

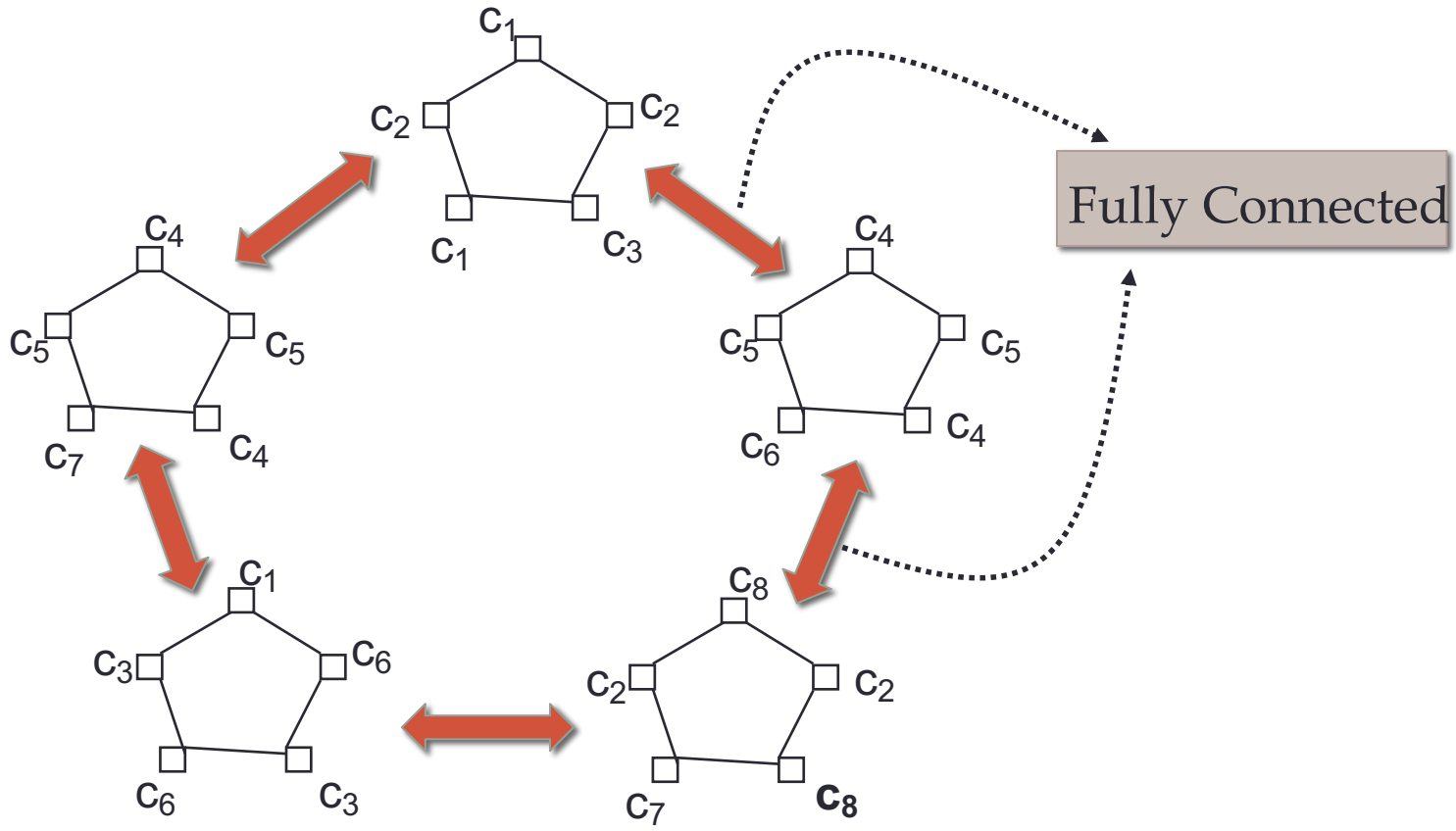


- X_1^2 can take 25 values {00,01,…,44}.
- To make characteristic graph of X_1^2, we connect two vertices if at least one of coordinates are connected in G_{X_1}

Characteristic graph:

- Vertices are different sample values
- Two vertices are connected if they should be distinguished

Power Graph Example

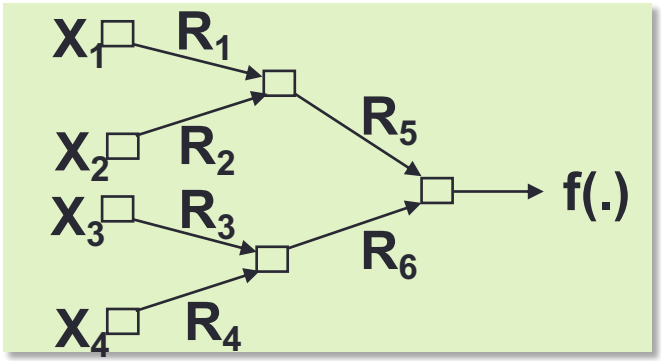


- One can color this graph by using **8 colors**.

$$\frac{1}{2}H(c_{G^2_{X_1}}) \approx 1.48 < H(c_{G_{X_1}}) \approx 1.52$$

Trees

- Intermediate nodes are allowed to compute some functions.
- Rate lower bounds by using cut-set bounds on graph entropies:



$$R_1 + R_2 + R_3 + R_4 \geq H_{G_{X_1}, G_{X_2}, G_{X_3}, G_{X_4}}(X_1, X_2, X_3, X_4)$$

$$R_1 + R_2 \geq H_{G_{X_1}, G_{X_2}}(X_1, X_2 | X_3, X_4)$$

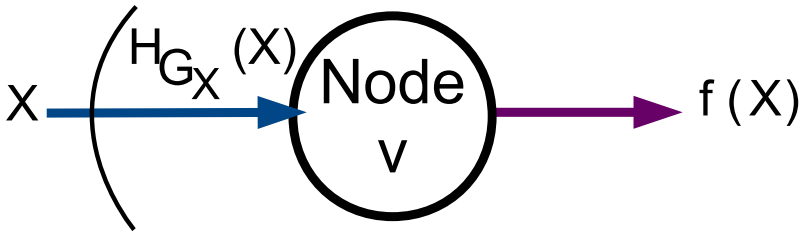
$$R_5 + R_6 \geq H_{G_{X_1, X_2}, G_{X_3, X_4}}(X_1, X_2, X_3, X_4), \ldots$$

- **Theorem**
  When sources are independent, these bounds are tight.

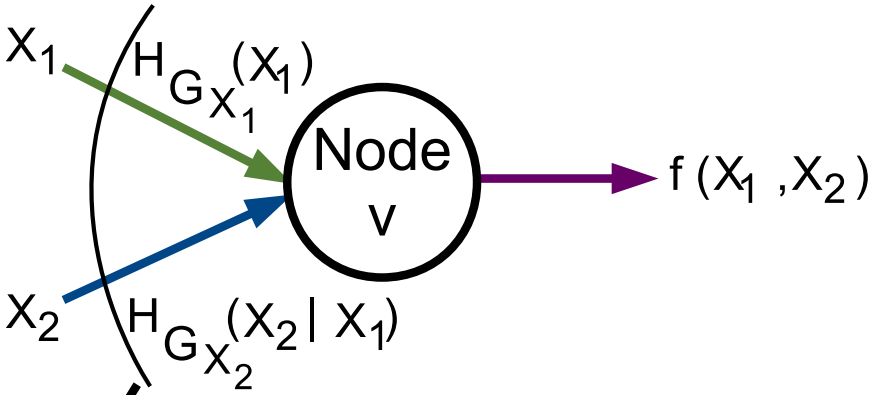
- For independent sources, functions to be computed at intermediate nodes are **coloring functions**.
- Unlike regular entropy, chain rule **does not hold** for graph entropies in general:

$$H_{G_{X_1}, G_{X_2}}(X_1, X_2) \neq H_{G_{X_1, X_2}}(X_1, X_2)$$

Rate region – ontology meets compression



$$R_X > H_{G_X}(X)$$

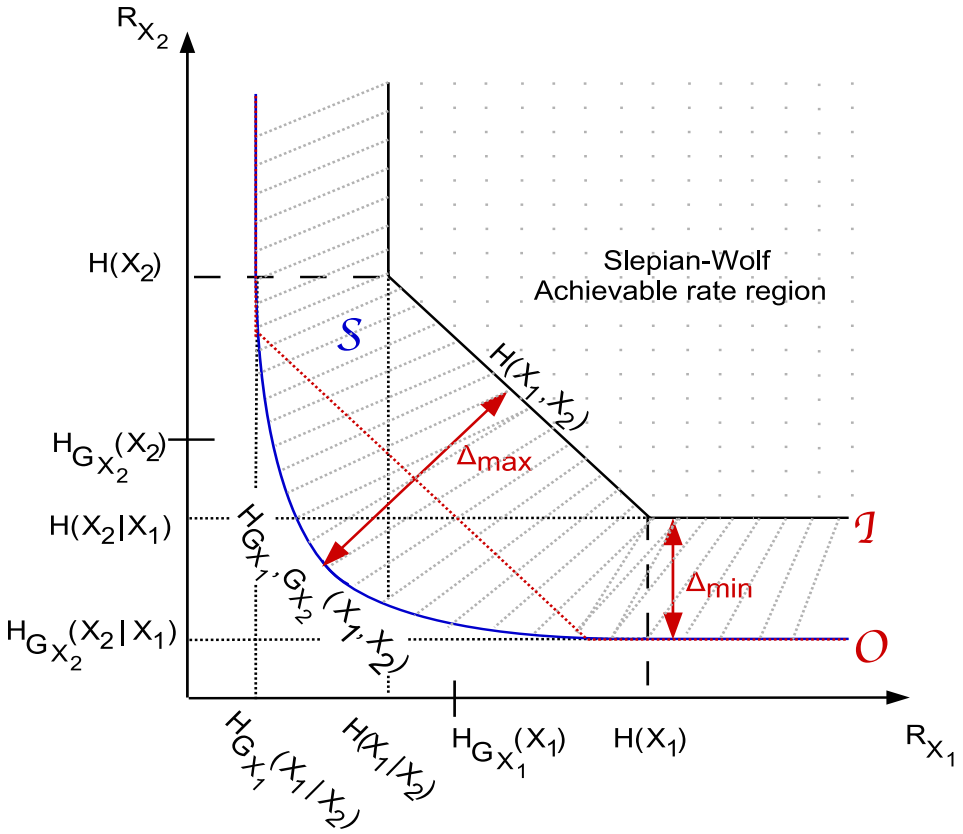


$$R_{X_1} + R_{X_2} > H_{G_{X_1}}(X_1) + H_{G_{X_2}}(X_2 | X_1)$$

$$> H_{G_{X_1} G_{X_2}}(X_1, X_2)$$

Exploit Körner's graph entropy to compute the true rate region for distributed functional compression.

Ontology versus data (Slepian-Wolf)



Entropic surjectivity

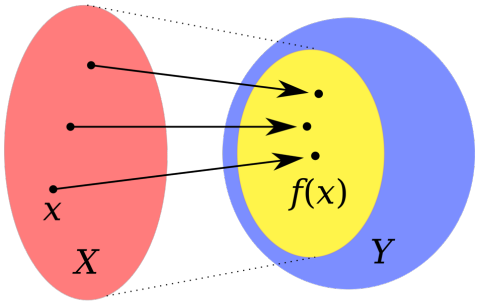
Defn. Entropic surjectivity of a function is how well can be compressed wrt the compression rate of its domain :

$$\Gamma_c(f) = \Gamma_c = \frac{\mathbf{H}(f_c(X))}{H(X)}$$

A minimal representation of function *f*, e.g. if coloring is used

$$\mathbf{H}(f(X)) = H_{G_X}(X)$$

A non-surjective function has less redundancy vs surjective function: $\frac{\mathbf{H}(f_c(X))}{H(X)} \approx 0$



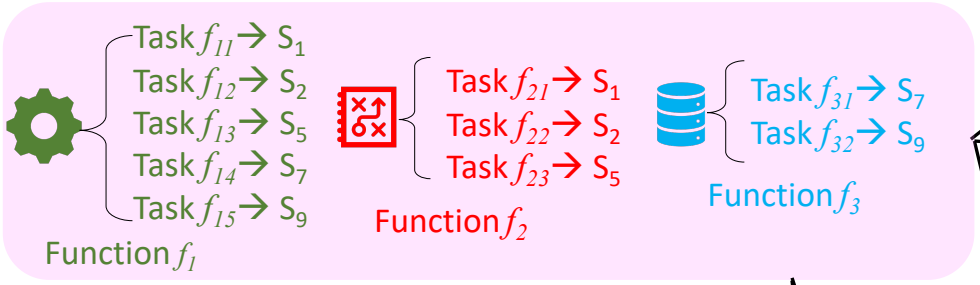
$f: X \to Y$

D. Malak, A. Cohen and M. Médard, "How to Distribute Computation in Networks," *IEEE INFOCOM 2020 - IEEE Conference on Computer Communications*

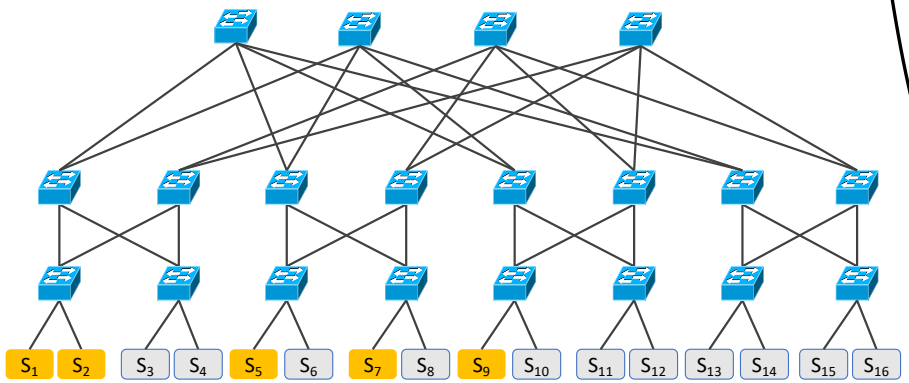
D. Malak, Médard, M., "Function Load Balancing Over Networks," in *IEEE Journal on Selected Areas in Information Theory*, vol. 2, no. 3, pp. 1041-1056, Sept. 2021

D. Malak, Y. Li, S. Ioannidis, E. Yeh, M. Médard, "Joint Optimization of Storage and via Coding Traffic Flows for Content Distribution" *IEEE WiOpt 2023 – Best Paper Award*

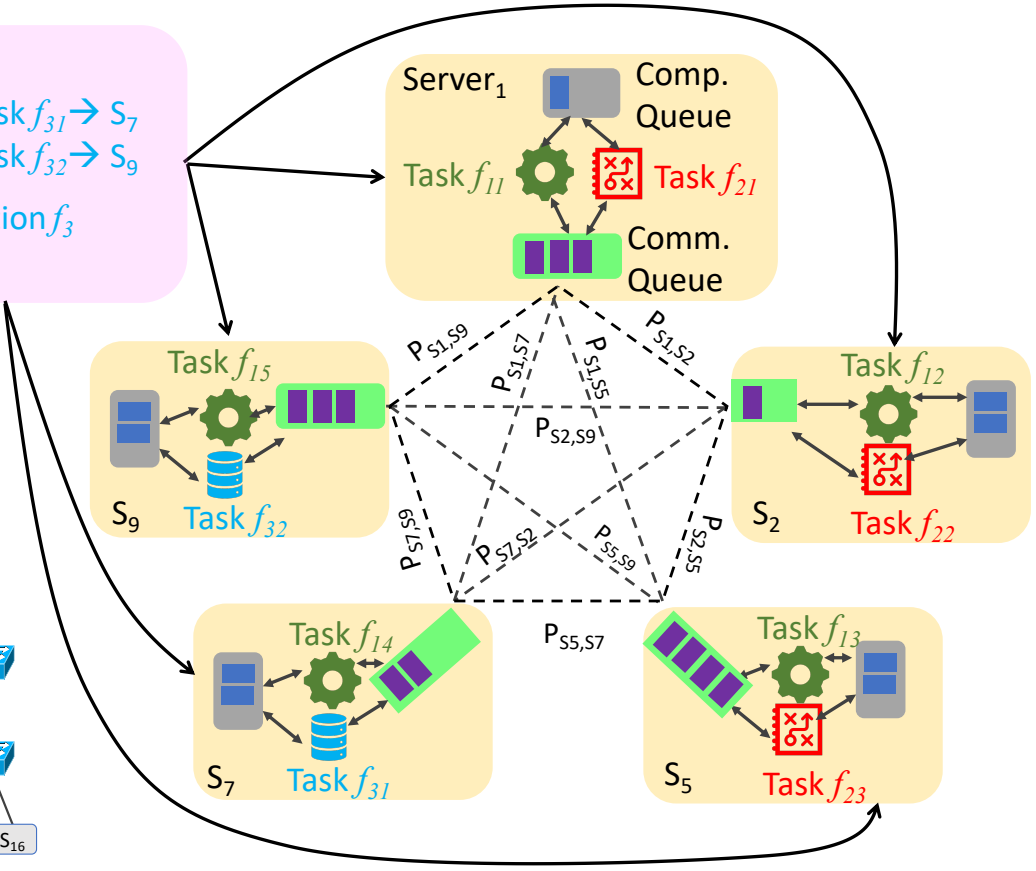
Network resource allocation



(b) Function Manager

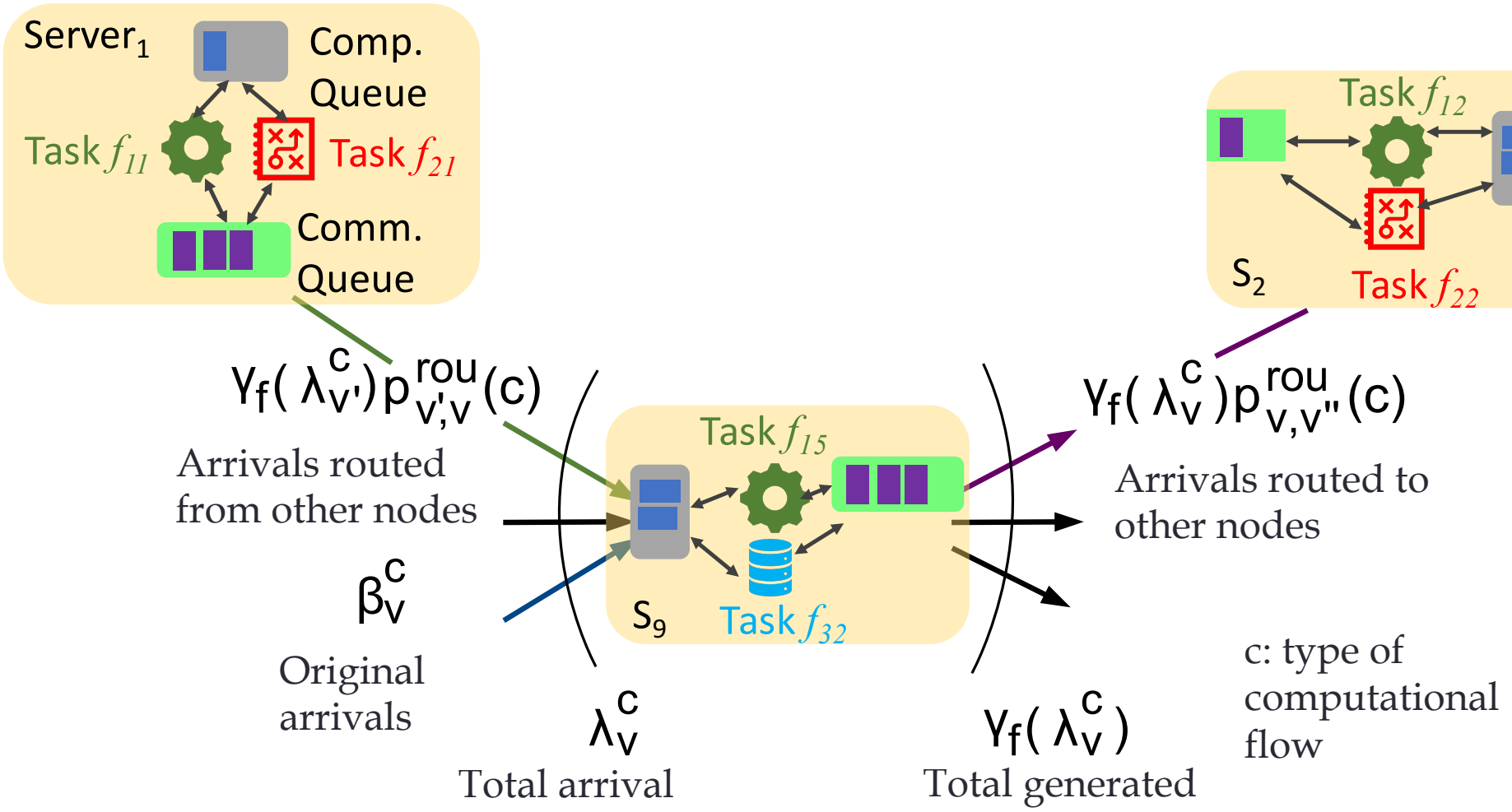


(a) Physical topology of the network



(c) Logical topology of workload distribution

Routing for computing



Network is in product form. Hence, nodes (servers) can be considered in isolation.

Average delay (per node)

The total delay of computation and communications for processing functions of type $c \in \mathcal{C}$ at node v

$$W_v^c = C_{v,comp}^c + C_{v,comm}^c$$

time complexity of computing

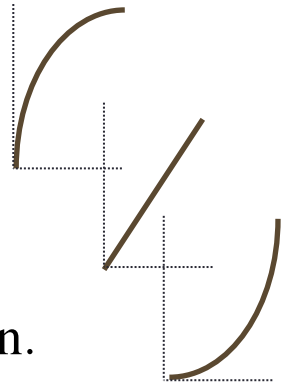
time complexity of communicating

$$C_{v,comp}^c = \frac{1}{\lambda_v^c} d_f(M_v^c)$$

$$C_{v,comm}^c = \frac{1}{\mu_v^c - \gamma_f(\lambda_v^c)}$$



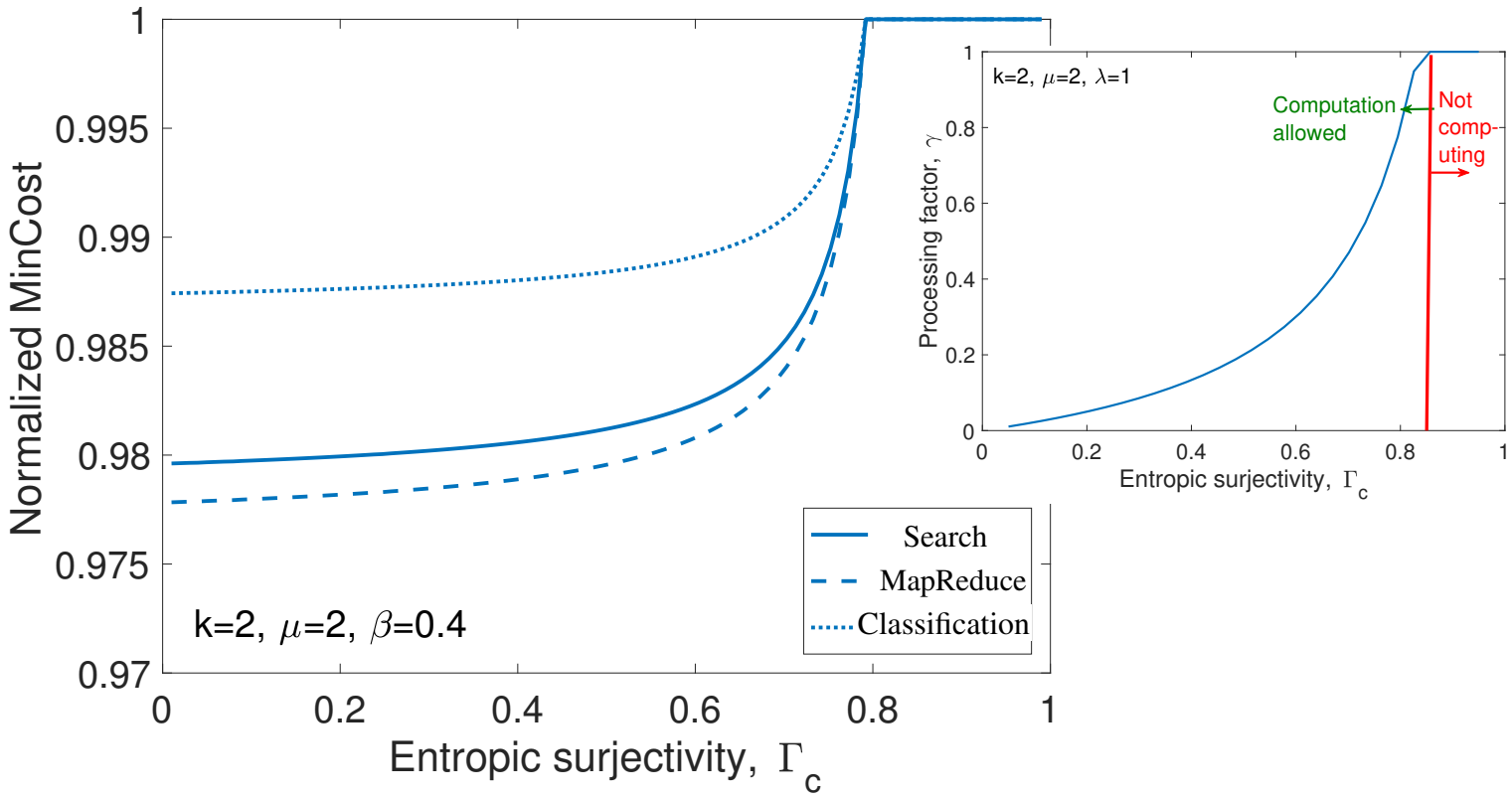
$$d_f(M_v^c) = \begin{cases} O(\log(M_v^c)), & \text{Search,} \\ O(M_v^c), & \text{MapReduce,} \\ O(\exp(M_v^c)), & \text{Classification.} \end{cases}$$



$$\rho_v^c = \lambda_v^c / \mu_v^c \in [0, 1)$$

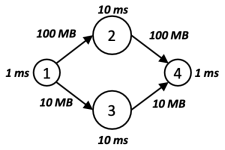
M_v^c is the long-term average number of packets waiting for service.

Cost change with surjectivity

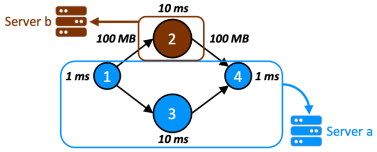


The normalized MinCost versus Γ_c .

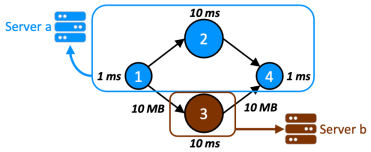
Allocation over networks



(a) A Sample DAG

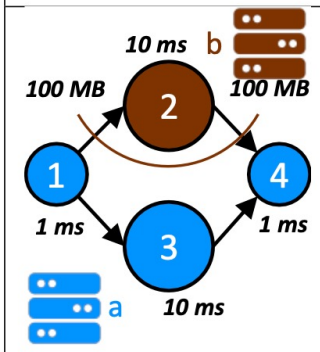
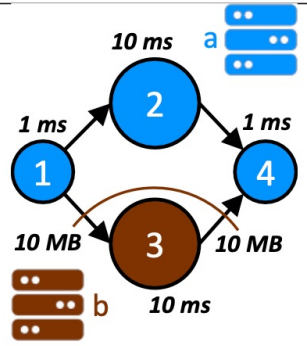
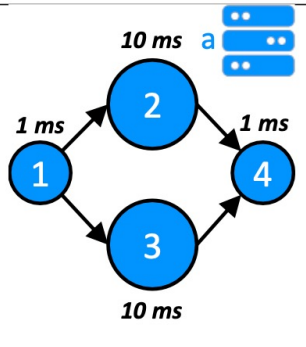
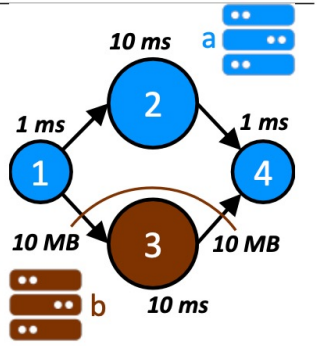


(b) Allocation 1



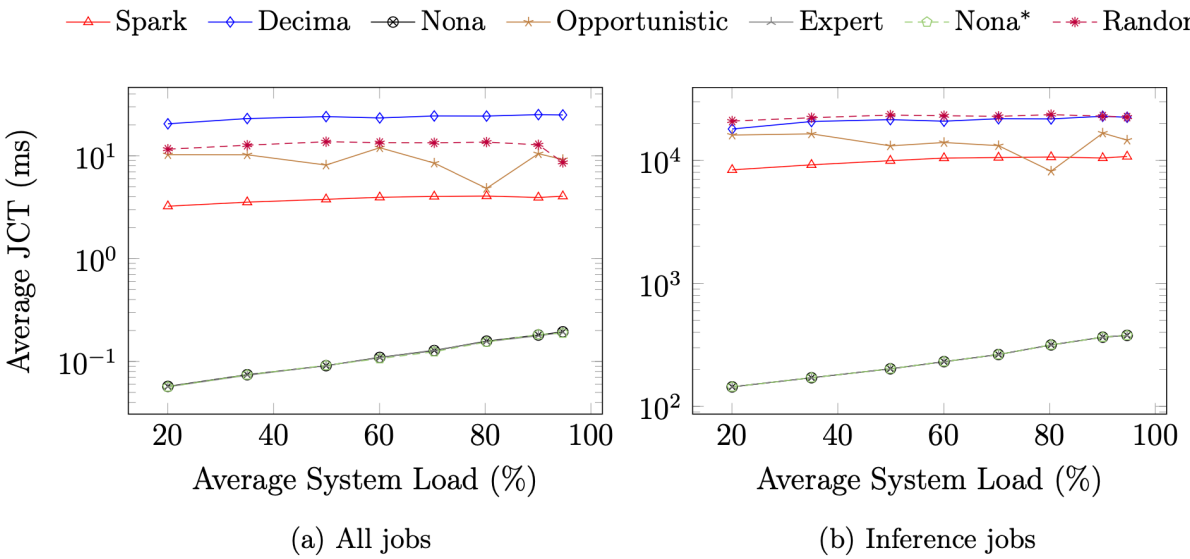
(c) Allocation 2

- Stochastic schedulers have the benefit of effectively displacing the complexity of the problem to an offline phase.
- When a new job arrives in the system, the scheduling decision consists in a series of samples from precomputed probability distributions.

Cluster Load	Congestion-Oblivious Schedulers		Nona	
				
Low	50%	50%	20%	80%
Medium	50%	50%	50%	50%
High	50%	50%	80%	20%

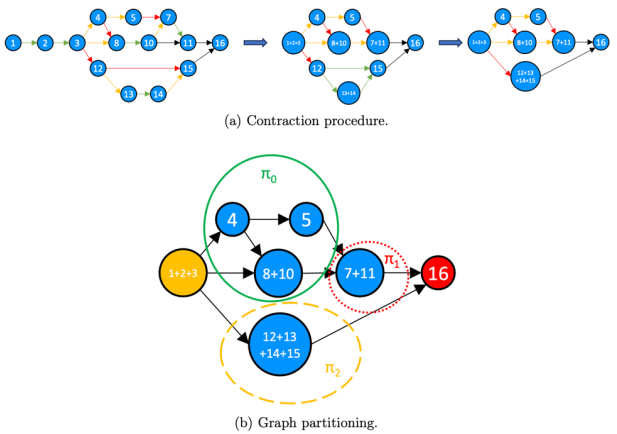
B. Pit-Claudel, D. Malak, A. Cohen, M. Médard and M. Ghobadi,
"Nona: A Stochastic Congestion-Aware Job Scheduler for Real-Time Inference Queries,"
2024 IEEE 13th International Conference on Cloud Networking (CloudNet), Rio de Janeiro, Brazil, 2024, pp. 1-10

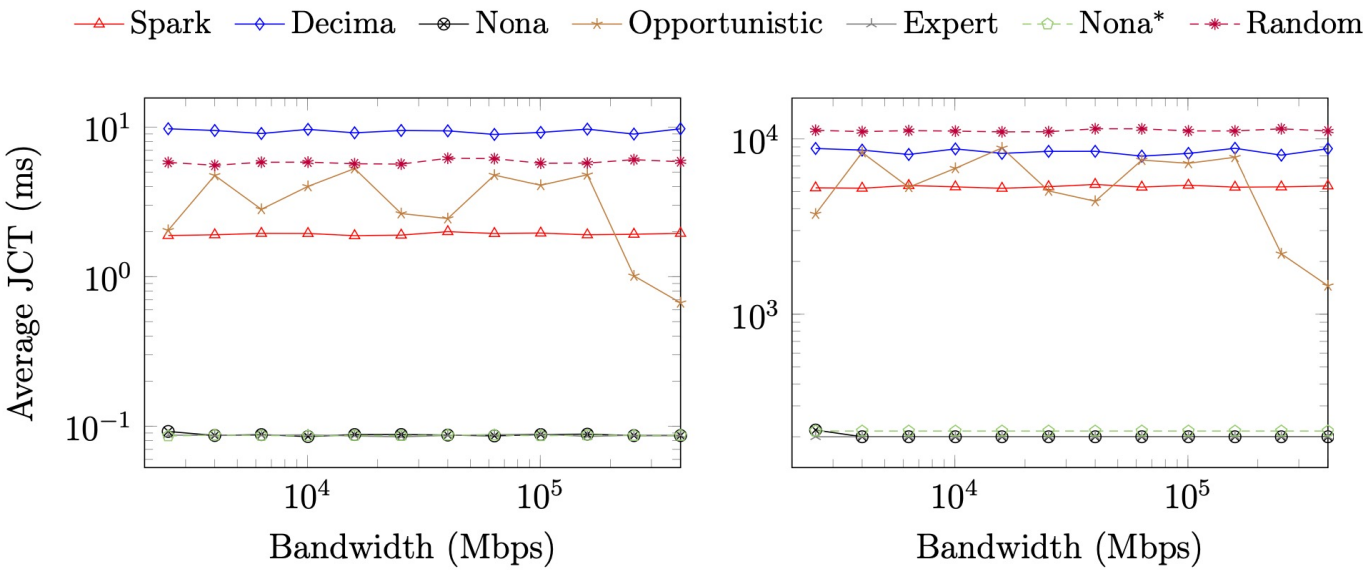
Allocation over networks



(a) All jobs

(b) Inference jobs

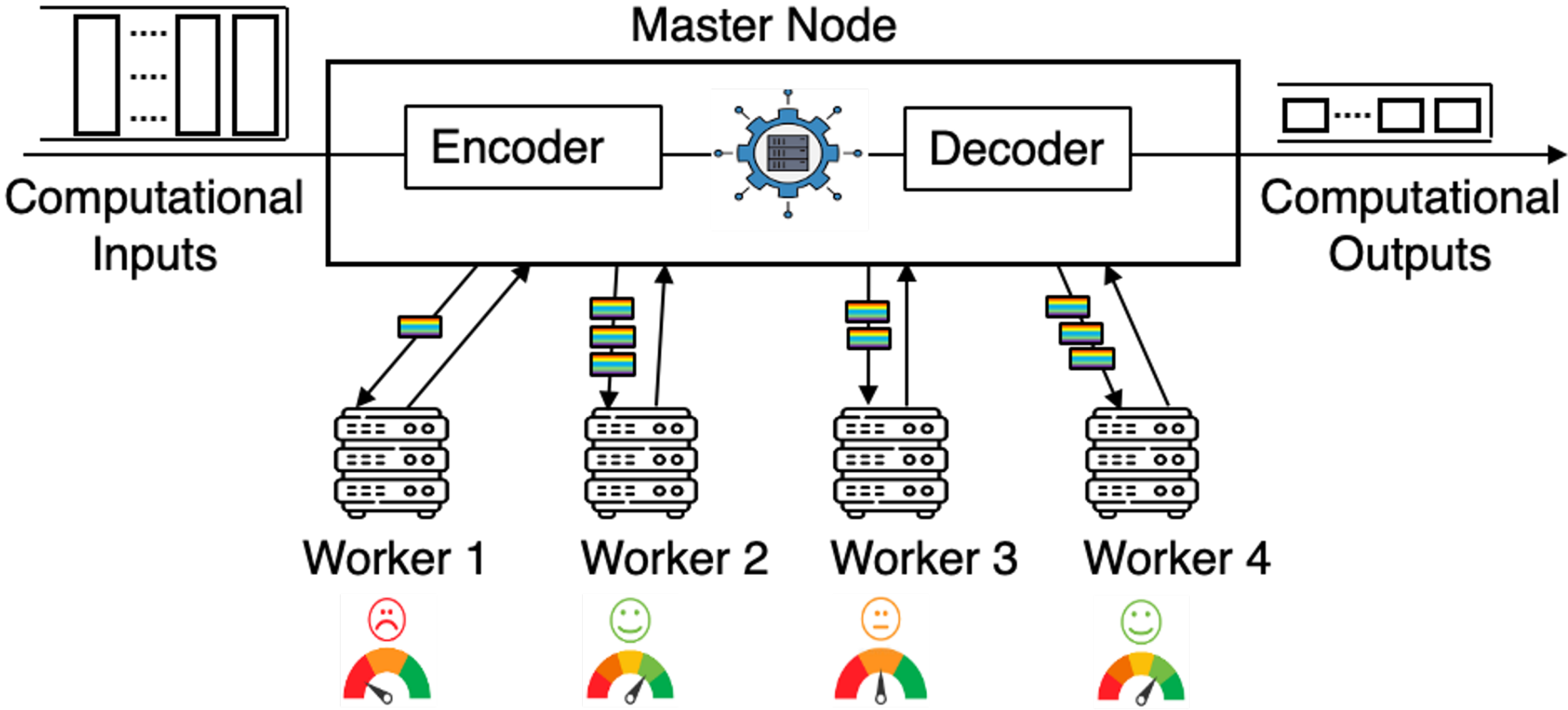




(a) All jobs

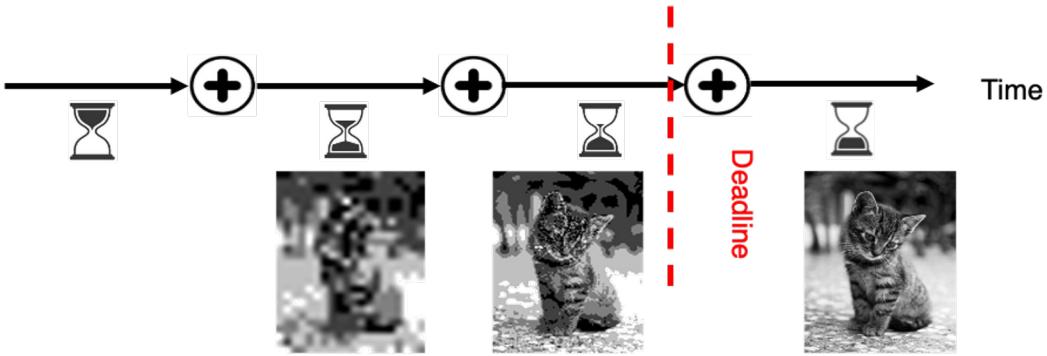
(b) Inference jobs

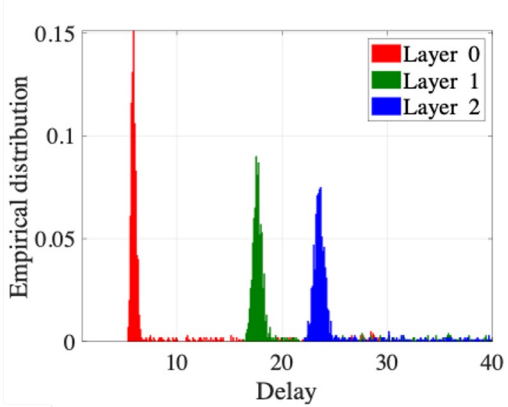
Temporal aspects

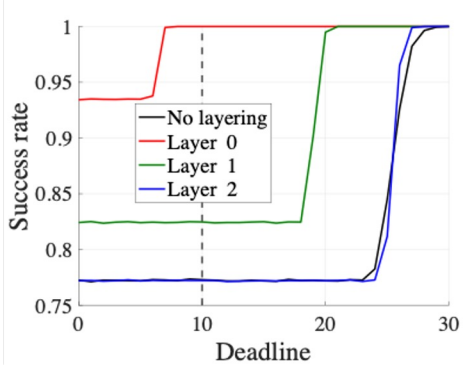


H. Esfahanizadeh, A. Cohen, and M. Médard, "Stream iterative distributed coded computing for learning applications in heterogeneous systems," *IEEE Infocom* 2022.
A. Cohen, G. Thiran, H. Esfahanizadeh and M. Médard, "Stream distributed coded computing," *IEEE JSAIT* 2021

Temporal multi-resolution







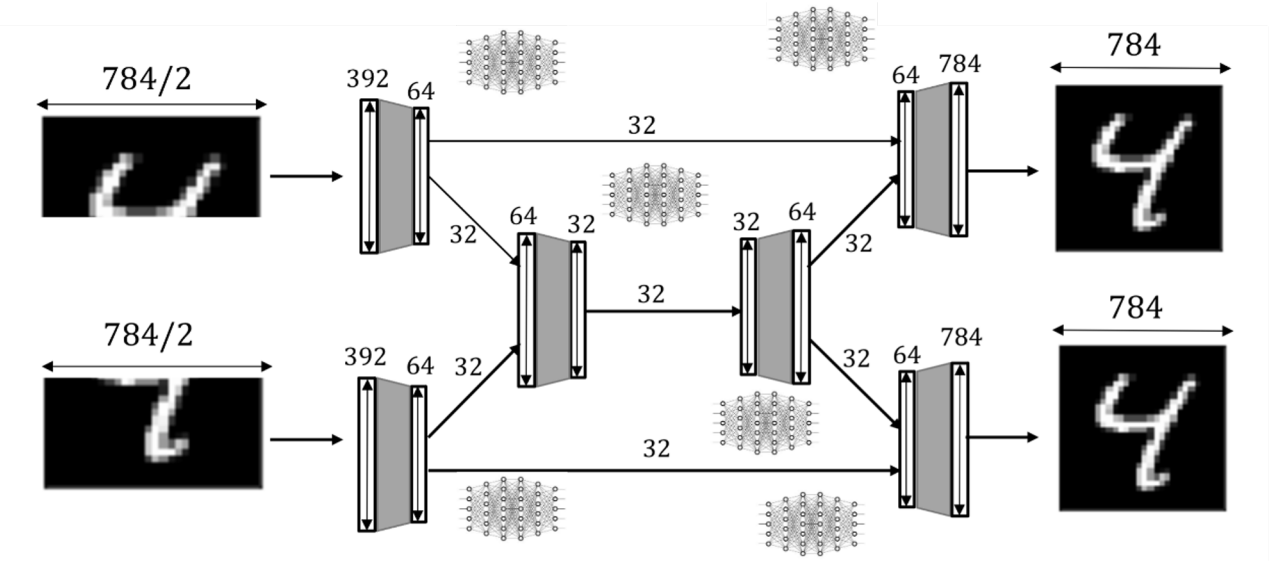
Coarser approximations are released at earlier stages than the full resolution

H. Esfahanizadeh, A. Cohen, M. Médard, and S. Shamai, "Distributed computations with layered resolution," *IEEE Cloudnet* 2022.

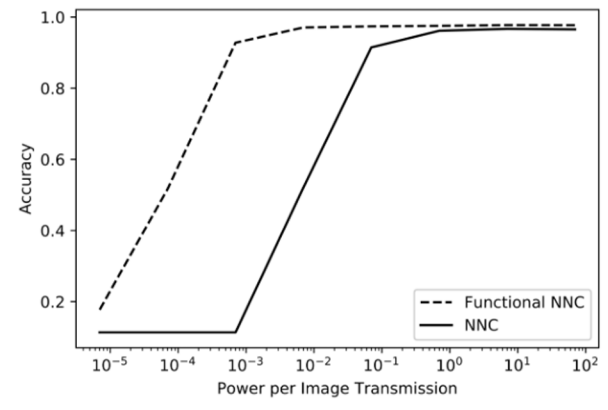
Neural network coding

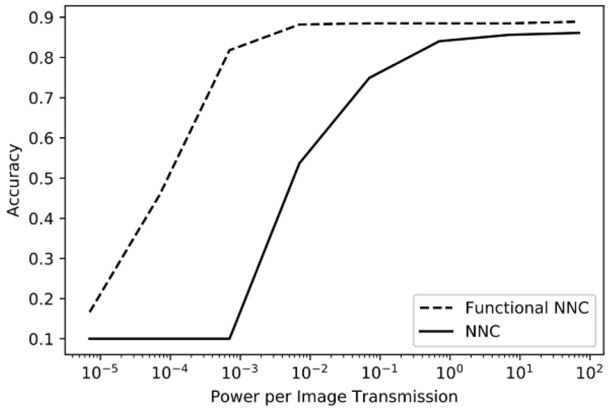
- Reconstruction metric

$$pSNR = 10\log_{10}\left(\frac{\max\{X_i\}^2}{\mathbb{E}[(\vec{X}-\vec{X})^2]}\right)$$

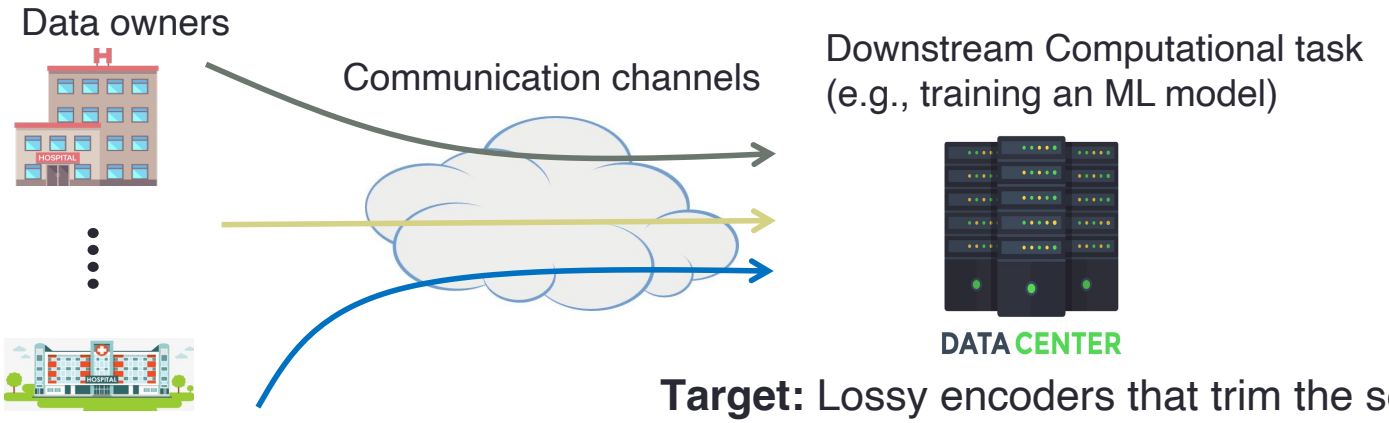


L. Liu, A. Solomon, S. Salamatian and M. Médard, "Neural Network Coding," *in* Machine Learning and Wireless Communications 2022, also ICC 2022



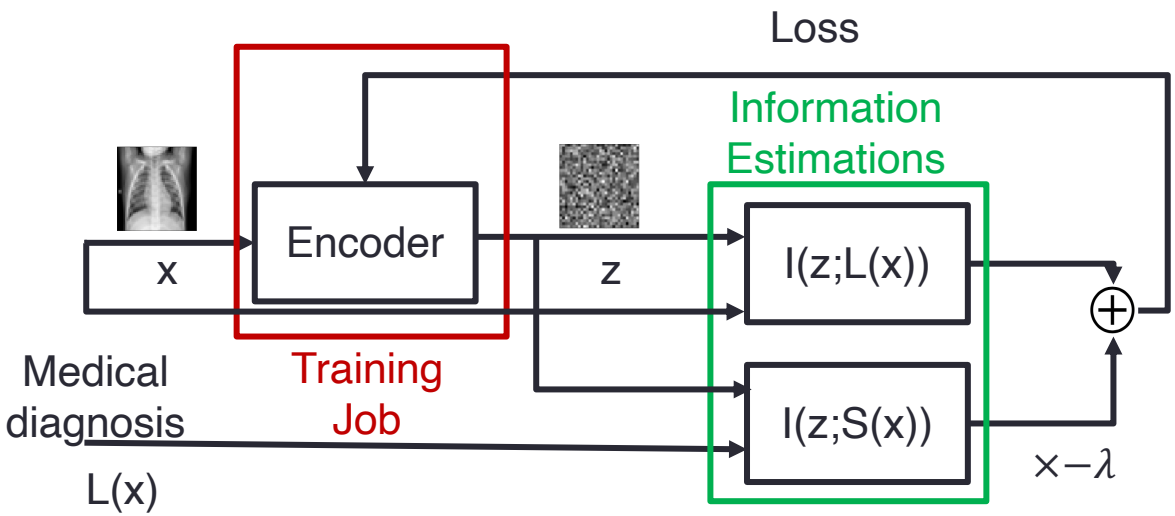


Neural network coding



Target: Lossy encoders that trim the sensitive unnecessary data from each sample (task-based compressors).

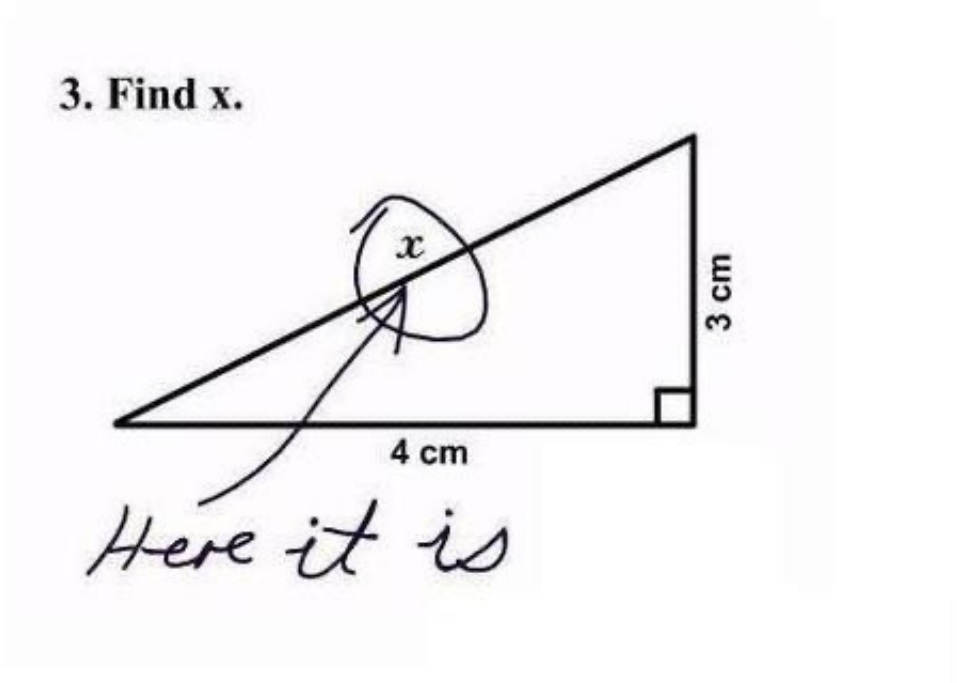
H. Esfahanizadeh, W. Wu, M. Ghobadi, R. Barzilay, M. Medard,
"InfoShape: Task-based neural data shaping via mutual information," *IEEE ICASSP* 2023.



Data-owners would like to outsource their data to the cloud for a specific purpose, e.g., training a machine learning model.

- More data and advanced computing units.

What is x? Ask us in a few years



Where to?

- What is the data, how we treat the data and where we treat the data depend on how we are using the data – ontology
- Computation, networking and communication are connected through the ontology that they support and affect ontology through what is possible
- Information theory has tools to study an exploit this unification

Compression and Communication

Distributed source compression [Slepian and Wolf, 73], [Pradhan and Ramchandran, 13], [Coleman et al, 06], [Wyner and Ziv, 76]

Rate region and graph entropy [Körner, 73], [Alon and Orlitsky, 96], [Orlitsky and Roche, 01], [Doshi et al, 10], [Feizi and Médard, 14], [Feng et al, 04], [Gallager, 88], [Kamath and D. Manjunath, 08], [Shah et al, 13]

Network coding and linear functions [Ho et al, 06], [Kowshik and Kumar, 10, 12], [Appuswamy and M. Franceschetti, 14], [Koetter et al, 04], [Koetter and Médard, 03], [Huang et al, 18], [Li et al, 03]

Coding for computation/communications [Li et al, 18], [Kamran et al, 19], [Yu, Maddah-Ali, and Avestimehr, 18]

Functions with special structures [Shen et al, 18], [Giridhar and Kumar, 05], [Gorodilova, 19]

Principle: Use underlying redundancy both in data and functions, and recover a sparse representation, or labeling, at the destination.