A New Perspective on *Effective Goal-Oriented Communication* in the Era of Networked Intelligence

Marios Kountouris

Communication Systems Department

EURECOM

Sophia-Antipolis, France

ITN Windmill Training Week
ETHZ, Zürich

September 7, 2022

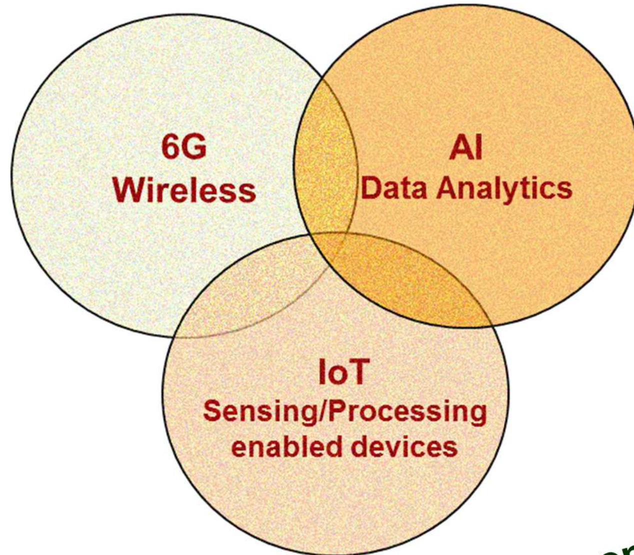
European Research Council
Established by the European Commission

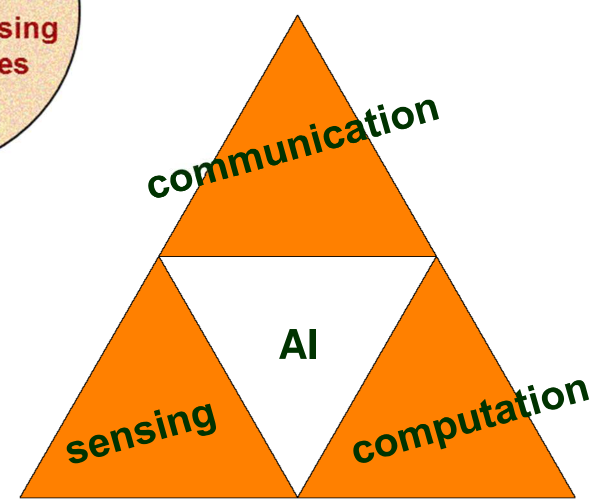
The Future of Communication Systems

A Consensus View on Future Wireless

New Services & Use Cases

- Multisensory AR/VR/XR
- Hi-Res imaging, sensing, 3D mapping
- Metaverse
- 3D teleportation, Digital twins
- Tactile Internet
- Precise Positioning
- Consumer Robotics
- Telehealth



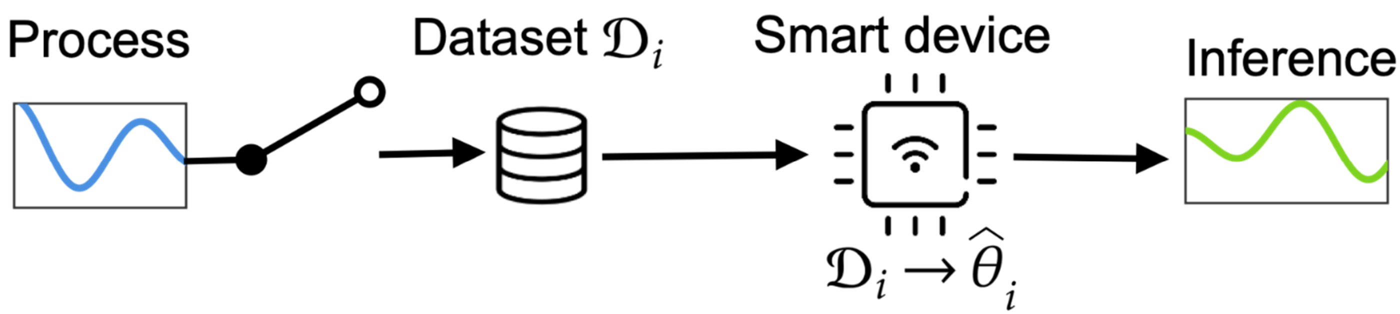


2G	3G	4G	5G	6G
Voice	Visio-phony	Mobile Internet	Wireless for Things	Wireless for ???

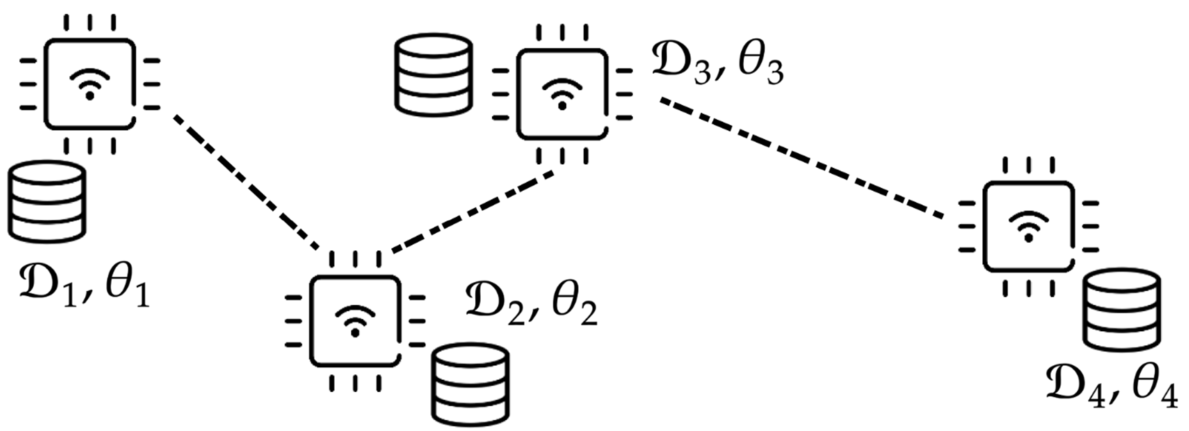
Wireless in the Era of Networked Intelligence

Two key features of Emerging Wireless Networks

- Smart edge devices, collecting data and extrapolating concepts.



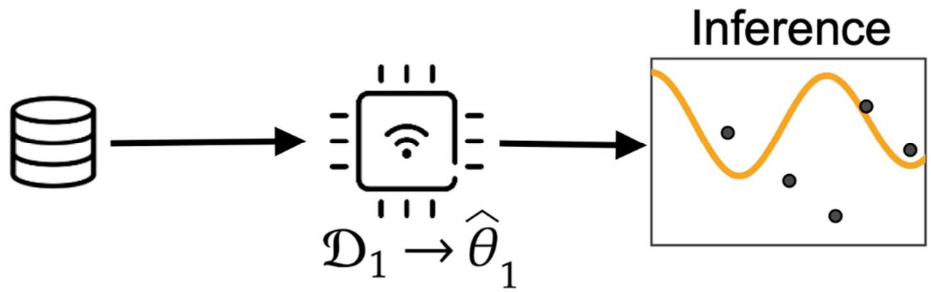
- Device-to-Device (D2D) connectivity and Multi-Agent Systems.



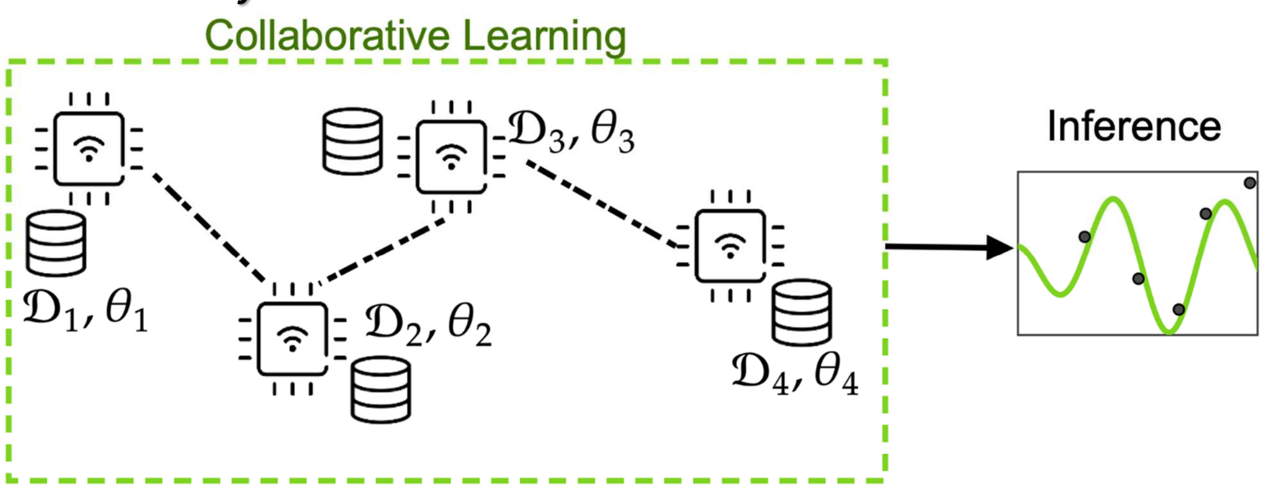
Decentralized Learning

One  *does not make Spring*

- A single device may not have a large enough **dataset $\mathfrak{D}_i$** or **computational resource** to **timely** optimize the model θ_i with satisfactory inference performance.



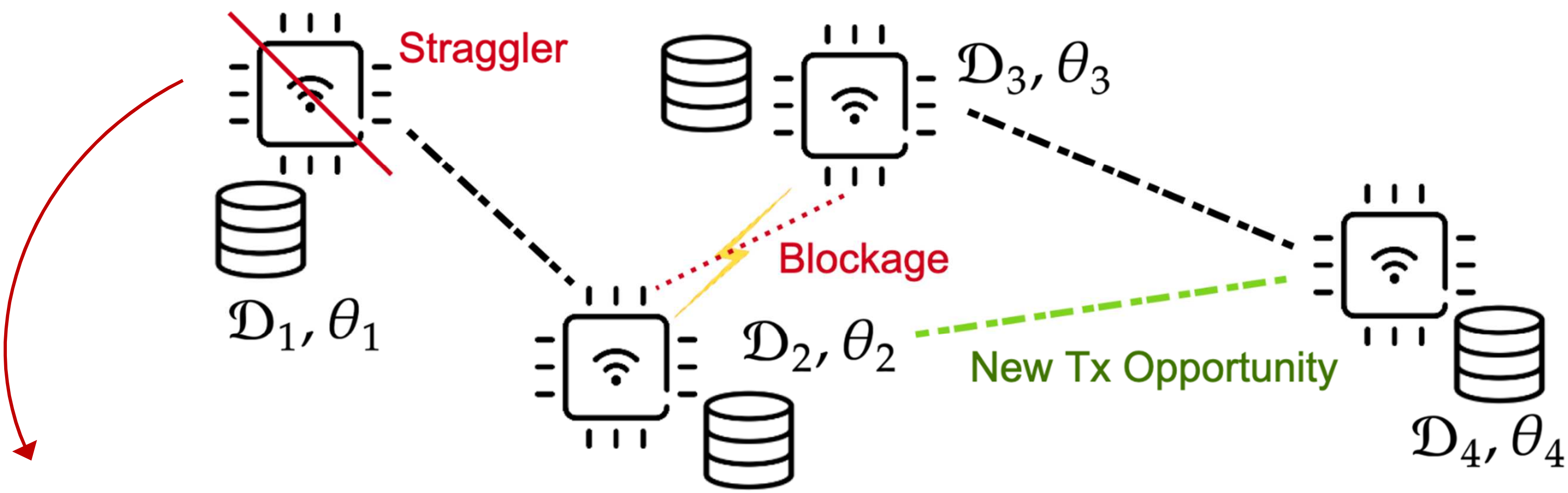
- **Decentralized learning** empowers smart edge devices to share data and computational resources to *collaboratively* train an ML model.



Decentralized Learning over Unreliable Edge Networks

- Collaborative learning over wireless networks limited to scenarios with
  - **fixed** network connectivity
  - **reliable** network devices (always available for computations, no channel impairments).

- But, wireless network connectivity is time-varying and edge devices are unreliable workers.



Stragglers introduce asynchronicity/inconsistency by disseminating outdated model estimates

Decentralized Learning Problem

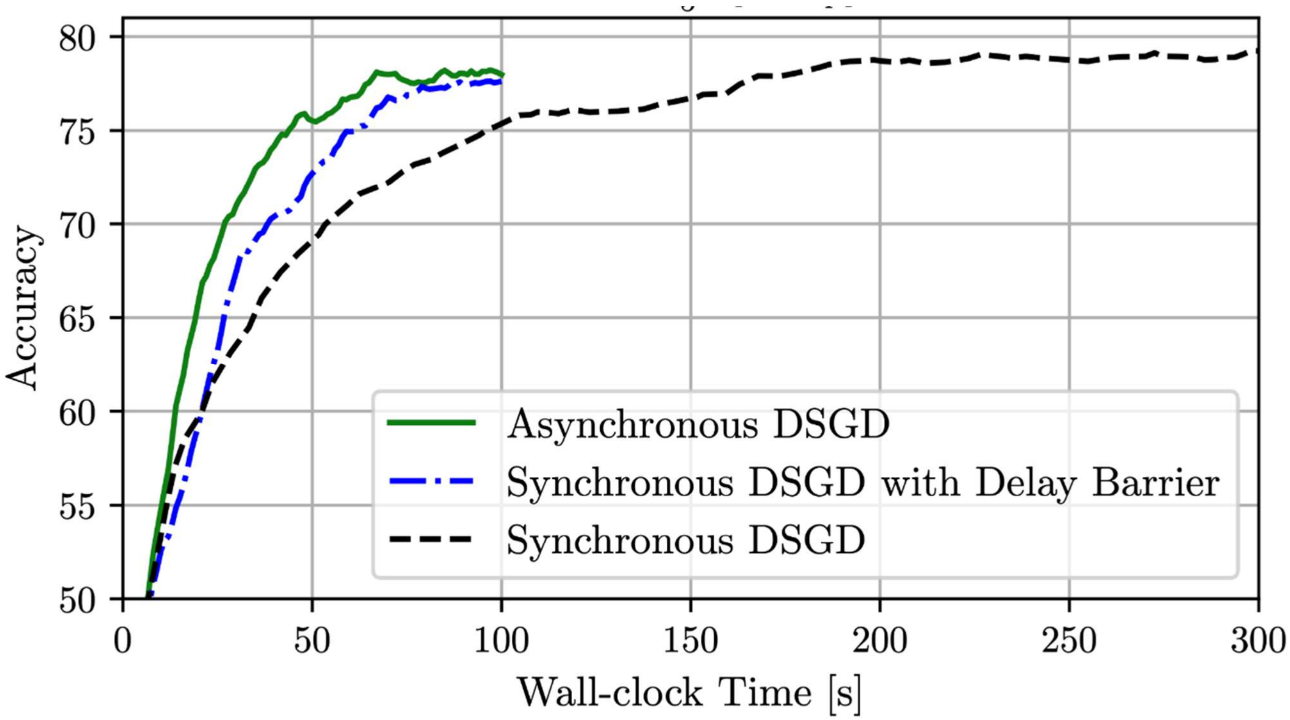
- Network of m smart edge devices, each with a local dataset $\mathcal{D}_i$,
  a local model estimate $\theta_i \in \mathbb{R}^d$,
  and loss function $f_i: \theta_i \to \mathbb{R}^+$
- Network goal: minimize the network aggregate empirical loss

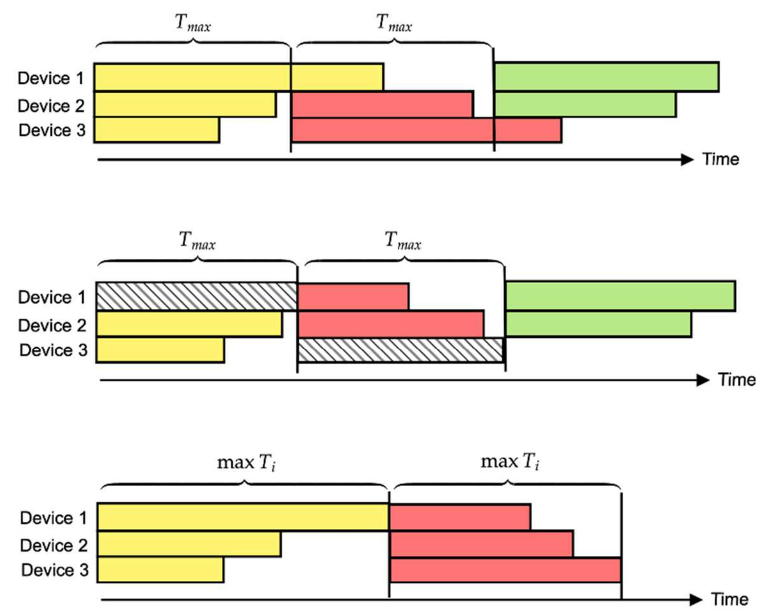
$$\underset{\theta_1,\ldots,\theta_m}{\text{minimize}}\ f(\theta_1,\ldots,\theta_m) := \frac{1}{m}\sum_{i=1}^{m} f_i(\theta_i)$$
$$\text{s.t.} \quad \theta_1 = \theta_2 = \cdots = \theta_m.$$

- Consensus constraint enforces *collaboration*.
  - Otherwise, the problem falls back to m parallel centralized (single-user) learning.
- Typically solved by iterative procedure that alternates between
  - local computation phase
  - one-hop neighbor communication phase

Asynchronous DSGD over Unreliable Wireless Networks

- Wireless network of 15 unreliable devices
- Collaborative training of a CNN to classify Fashion-MNIST images.
- Rayleigh fading and random computation time $T_{comp} = 0.25s + Exp(1)$ (shifted exponential)
- Strict delay barrier $T_{max} = \frac{4}{5}\mathbb{E}[T_{comp}]$



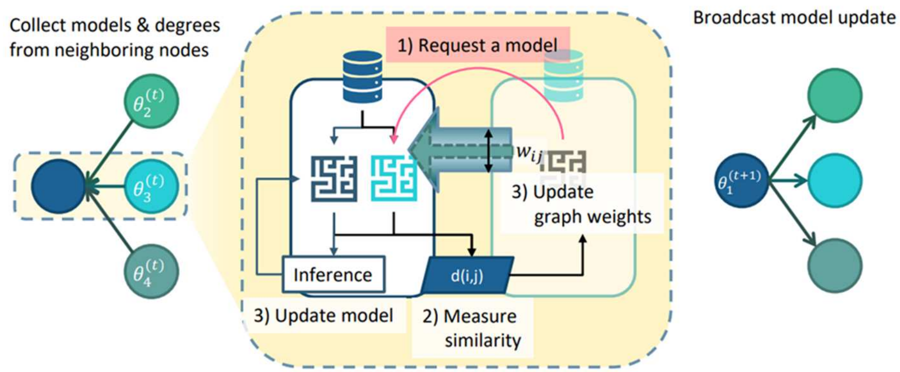


Reusing outdated gradient information from slow devices
is beneficial in asynchronous decentralized learning.

J. Eunjeong, M. Zecchin, MK, "*Asynchronous decentralized learning over unreliable wireless networks*", Proc. IEEE ICC'22

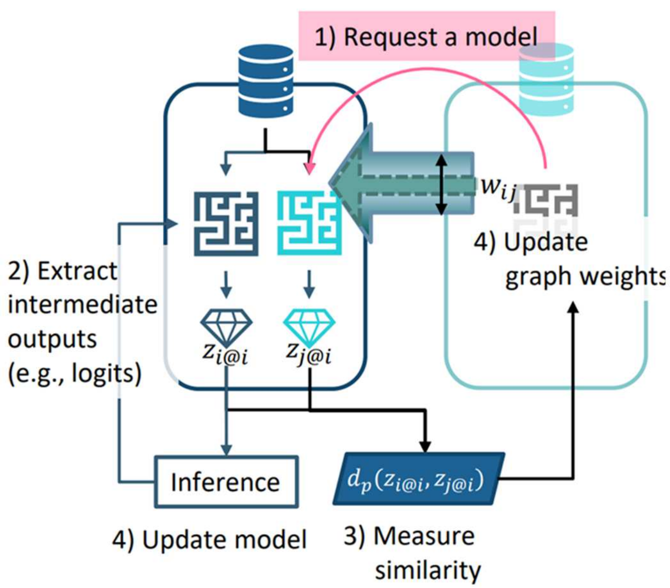
Personalized Decentralized Learning

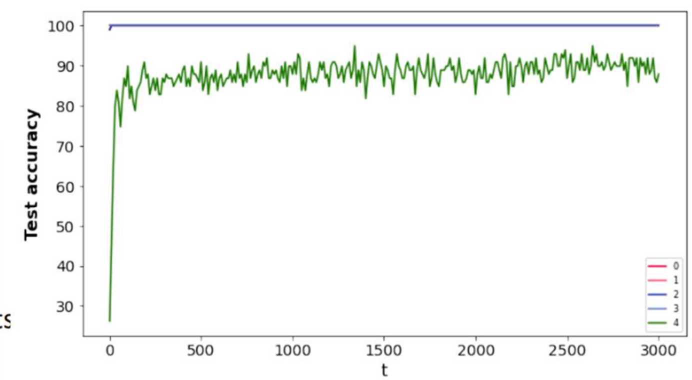
- *How can agents improve upon their locally trained model by communicating with other agents that have similar objectives?*



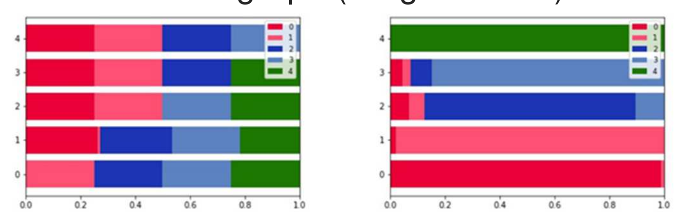
Statistical heterogeneity

- each user exhibits distinct preferences but has similarity with some of the other users' tasks
- information sharing is beneficial.

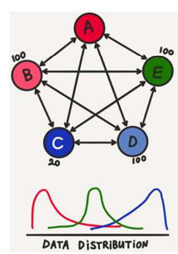


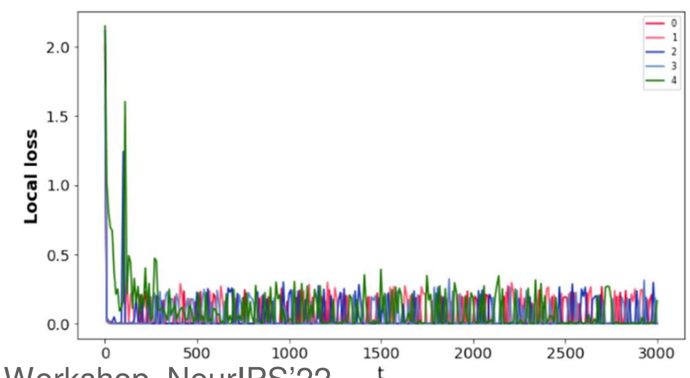


Collaboration graph (weight matrix)



(a) Weights at $t = 10$ (b) Weights at $t = 1000$

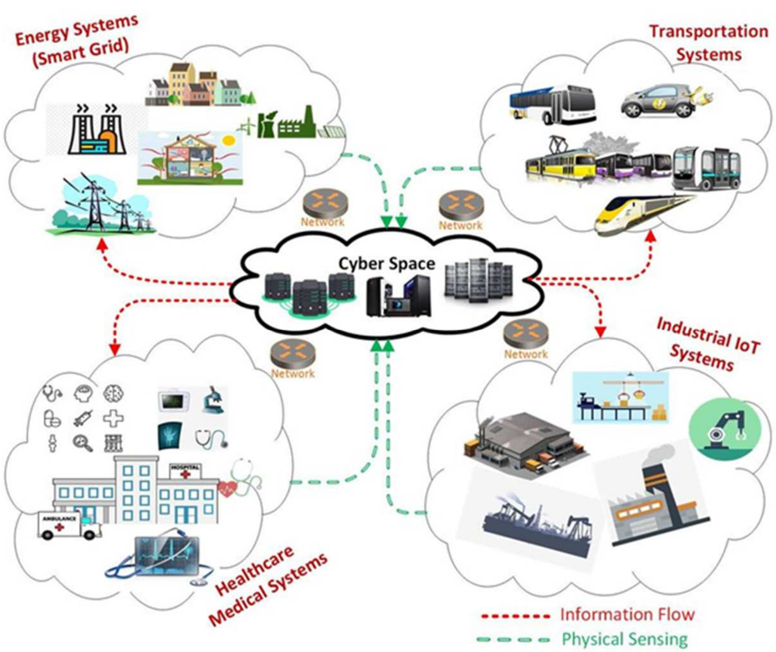


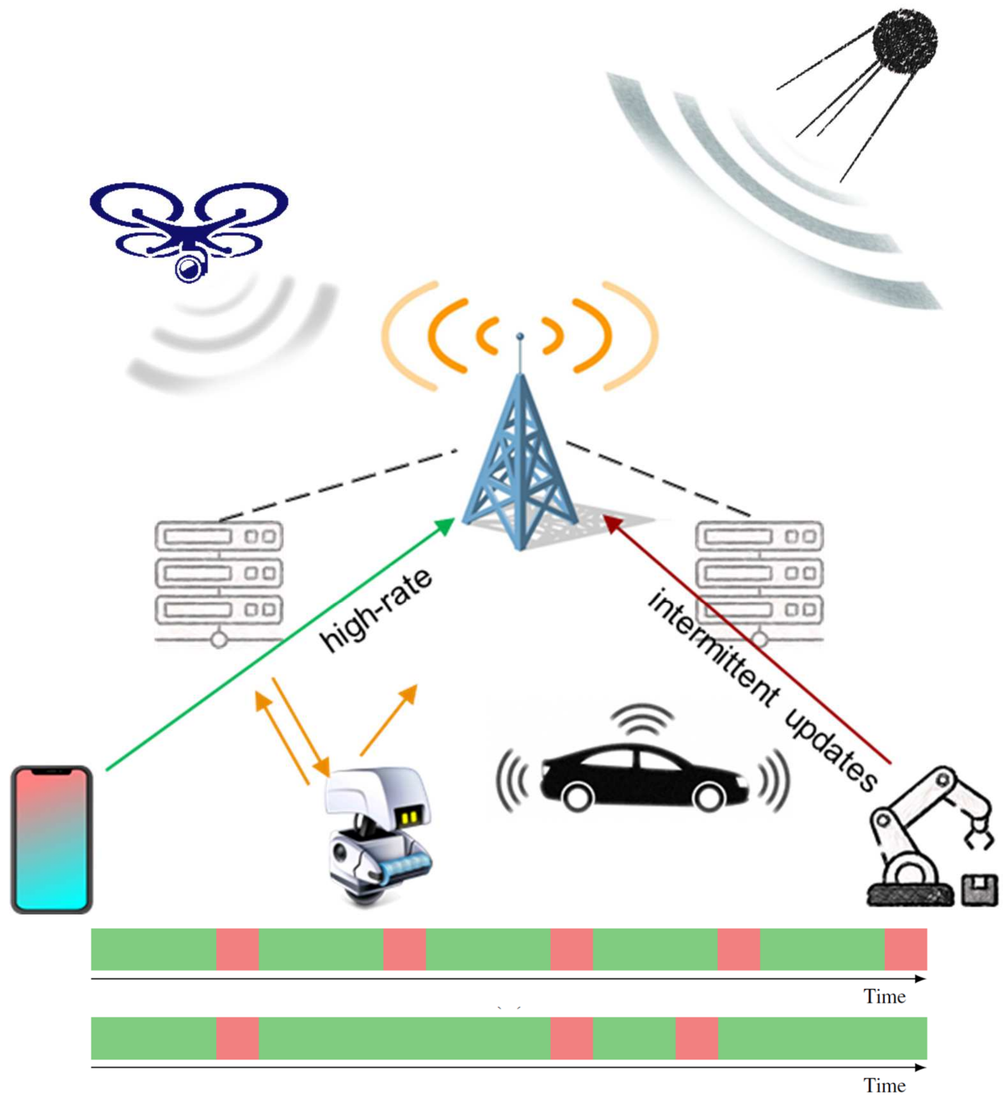


J. Eunjeong, MK, "*Personalized Decentralized Federated Learning with Distillation*", subm. FL Workshop, NeurIPS'22

Fast Forward to 2030

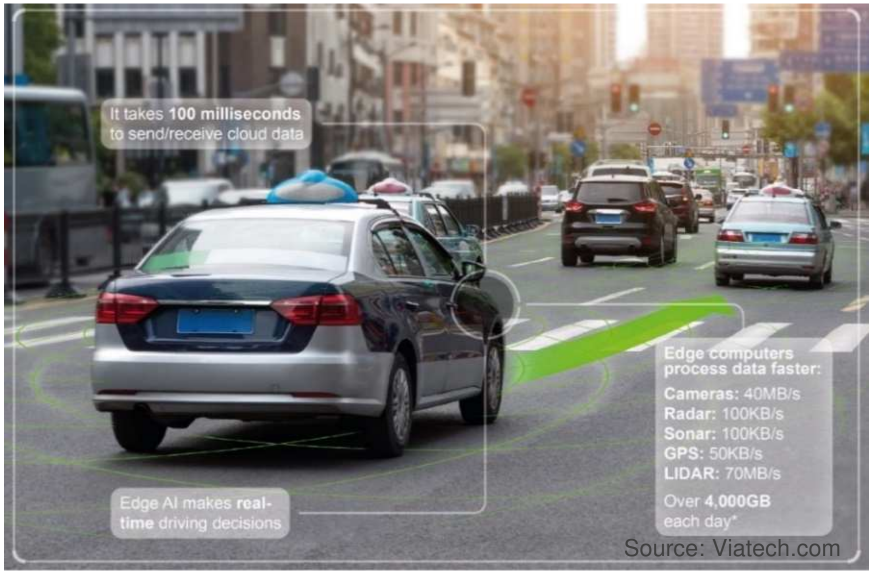
- **Cyber-Physical** and **Mission-Critical Interactive** Systems
  - swarm robotics, self-driving vehicles, smart IoT, …
- **Wireless Networked Intelligent Systems**
  - reliable ***real-time*** communication
  - *autonomous* interactions
  - automated *timely* decision making
  - on-device & in-network computation…





From *Connected Things* to *Connected Intelligence*

- Future Wireless Networks **under pressure** … as always



Major Challenge

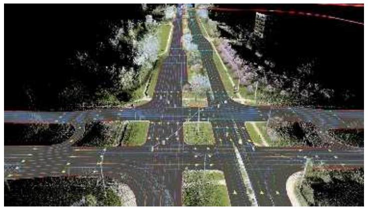
- *process and transport massive amounts of data*
  - generated by **countless** IoT connections
  - **constraints:** real-time, energy, security/privacy,…

Data

- multimodal, high-dimensional, and geo distributed

Let the numbers speak

- Edge Intelligence ~ 4 Tbps
- Autonomous transportation 4 TB/day
- Digital industry & robotics ≪1 ms

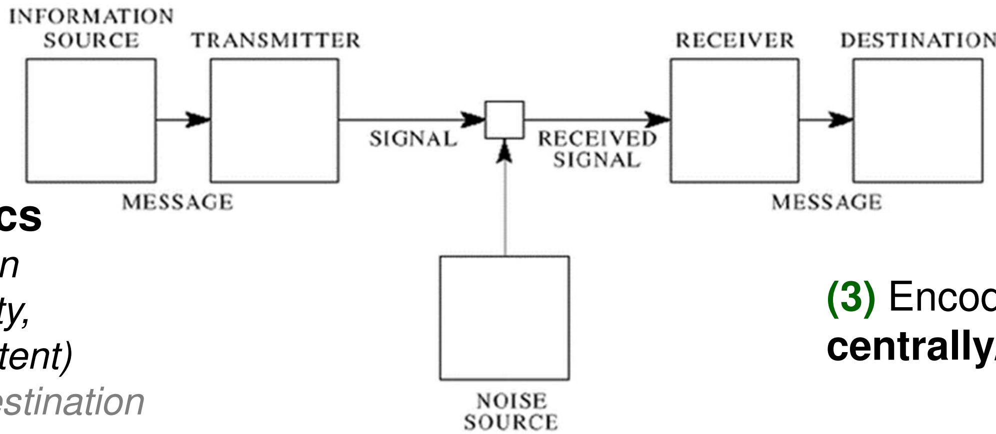


Do we know *how* to do it? Do we have the right *theory* & algorithms?

The Road So Far

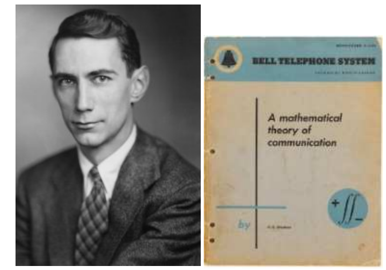
From Theory…

Shannon's model **(1) Reliable** transfer of information



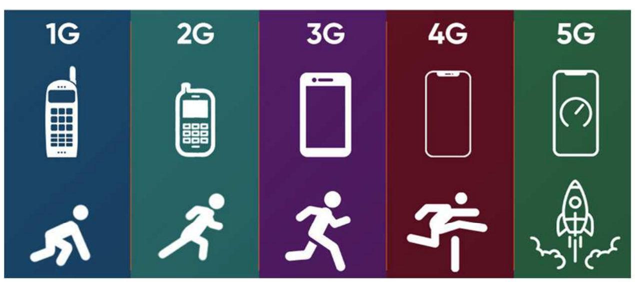
(2) No semantics
dichotomy between information quantity, and meaning (content) and its effect at destination

(3) Encoder and decoder **centrally/jointly** designed

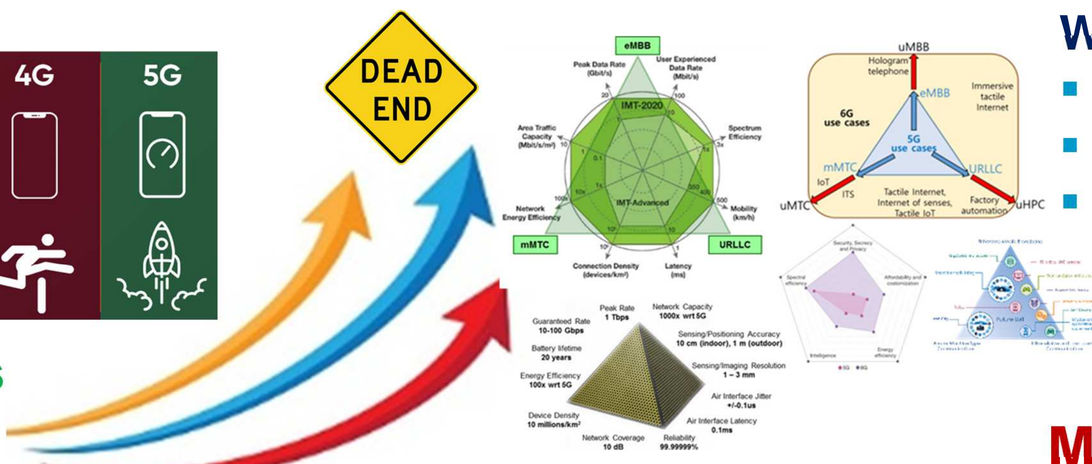


Focus on ***noise*** (& ***equivocation***) rather than ***signal***

… to Practice



CITIUS, ALTIUS, FORTIUS

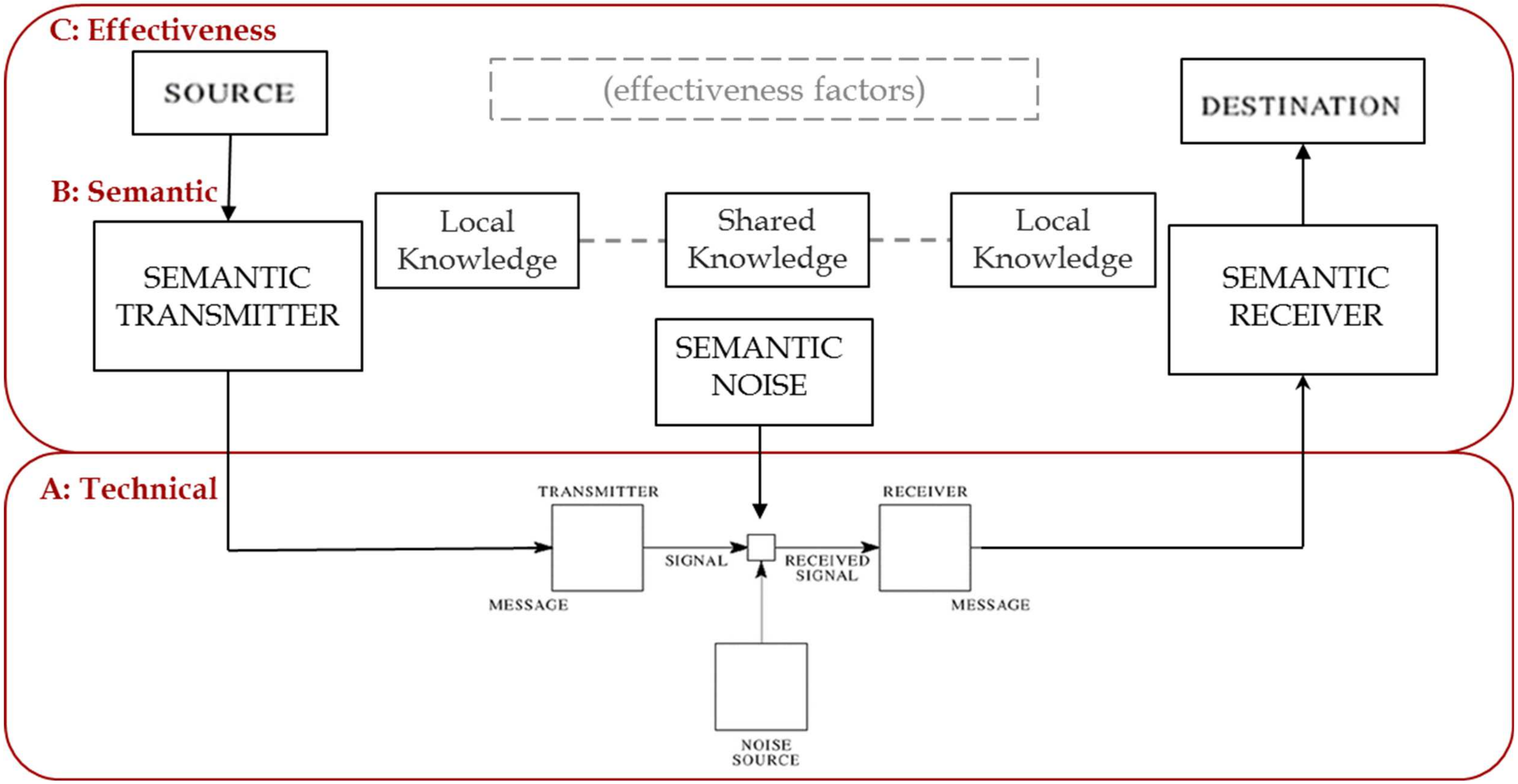


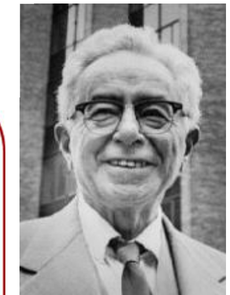
Wireless Systems Evolution
- Inflated requirements
- Overprovisioning
- Hard to meaningfully scale

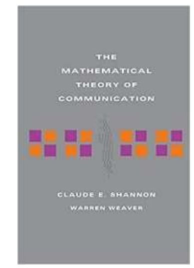
Maximalist approach

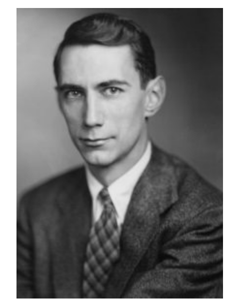
The Road Ahead

Augmenting Shannon's Communication Model & Theory









Shannon – Weaver Model & Vision

LEVEL A. How accurately can the symbols of communication be transmitted? (The technical problem.) — *Syntactic level*

LEVEL B. How precisely do the transmitted symbols convey the desired meaning? (The semantic problem.) — *Semantic level*

LEVEL C. How effectively does the received meaning affect conduct in the desired way? (The effectiveness problem.) — *Pragmatic level*

Effective Goal-Oriented Communications

Shannon's Communication Model (1948)

<u>Metrics</u>: *throughput*, *error rate, delay, (QoS/QoE)*

info source

……01000010101010

sequence of "random" bits

Opaque data pipe

……0101001X101010

Target

- **Reliable** transfer of information
- Reconstruct the entire information flow
- **Non-causal** signal reconstruction

- **Exogenous** traffic
- Data generation not included/controlled

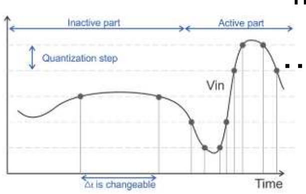
No semantics: *dichotomy between information quantity and its meaning, content, and effectiveness*

Effective Goal-Oriented Communication Model (202X)

Goal-oriented *unification* of data generation, transmission, and reconstruction

Joint design/Non-separable

importance-aware filtering & processing



…~~0100001~~0101010

101010~~10~~

……01010X10

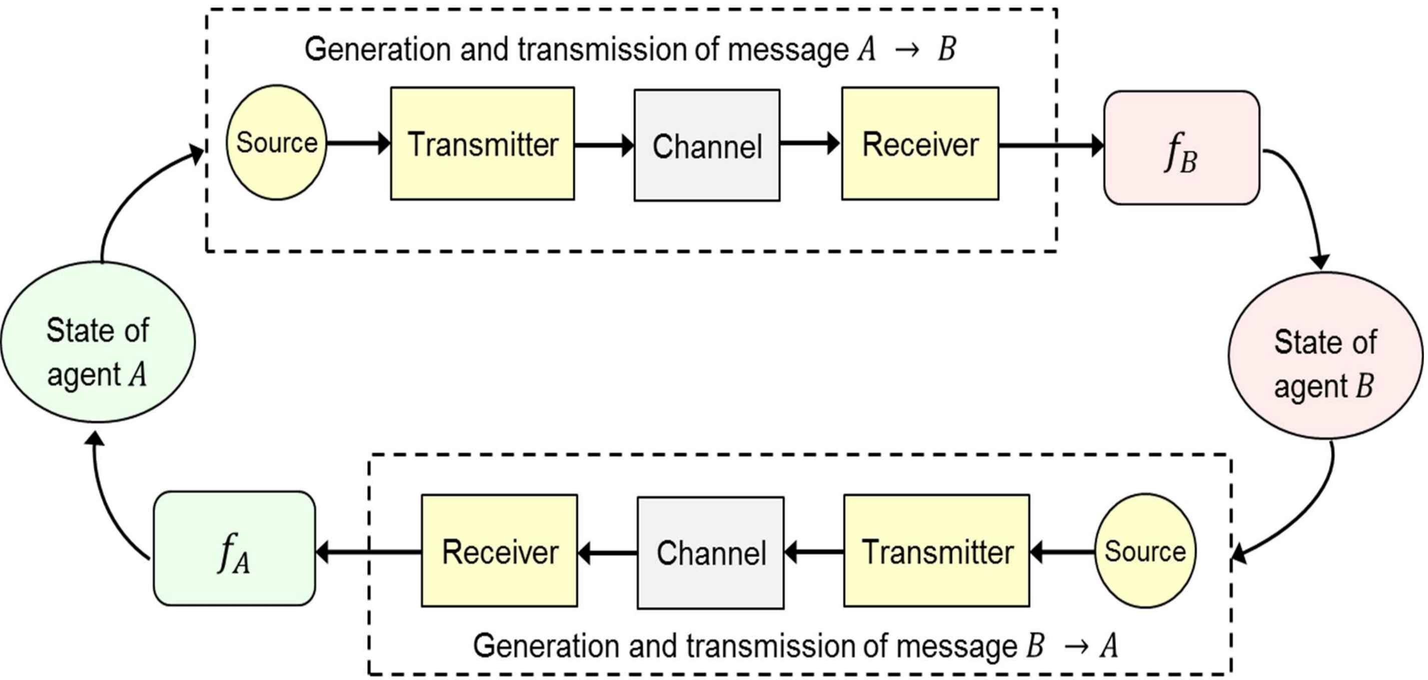
Data importance-aware PHY/MAC/RLC

- Active sampling
- Generation *at will*
- Source/process aware

Semantics of Information: *significance and effectiveness wrt. communication goal*

- *Quantitative* and *qualitative (context-aware)* approach
- Application/end-user requirements aware metrics
- **Real-time/causal** reconstruction

Effective Communication Diagram



- Agents (entities) A and B exchanging messages, which may affect one another's state
- Two-way comm. – feedback loops
- State change functions f
- Input/output Map: $\Omega \times \mathcal{M}^{\times n} \longrightarrow \Omega \times \mathcal{M}^{\times m}$
- $\mathcal{A}(\pi, x_1, \ldots, x_n) = (\sigma, y_1, \ldots, y_m)$
- *Decision-making and inference*

Various scenarios of interest

(i) **control-oriented** (e.g., remote control, actuation, real-time tracking,…)
(ii) **computation-oriented** (e.g., function computation, labelling, feature extraction)
(iii) **learning-oriented** (e.g. distributed/federated learning, generative model building,…)
(iv) **sensing/perception-oriented** (e.g., multi-view cameras, SLAM, situational awareness,…)
(v) **knowledge-oriented**

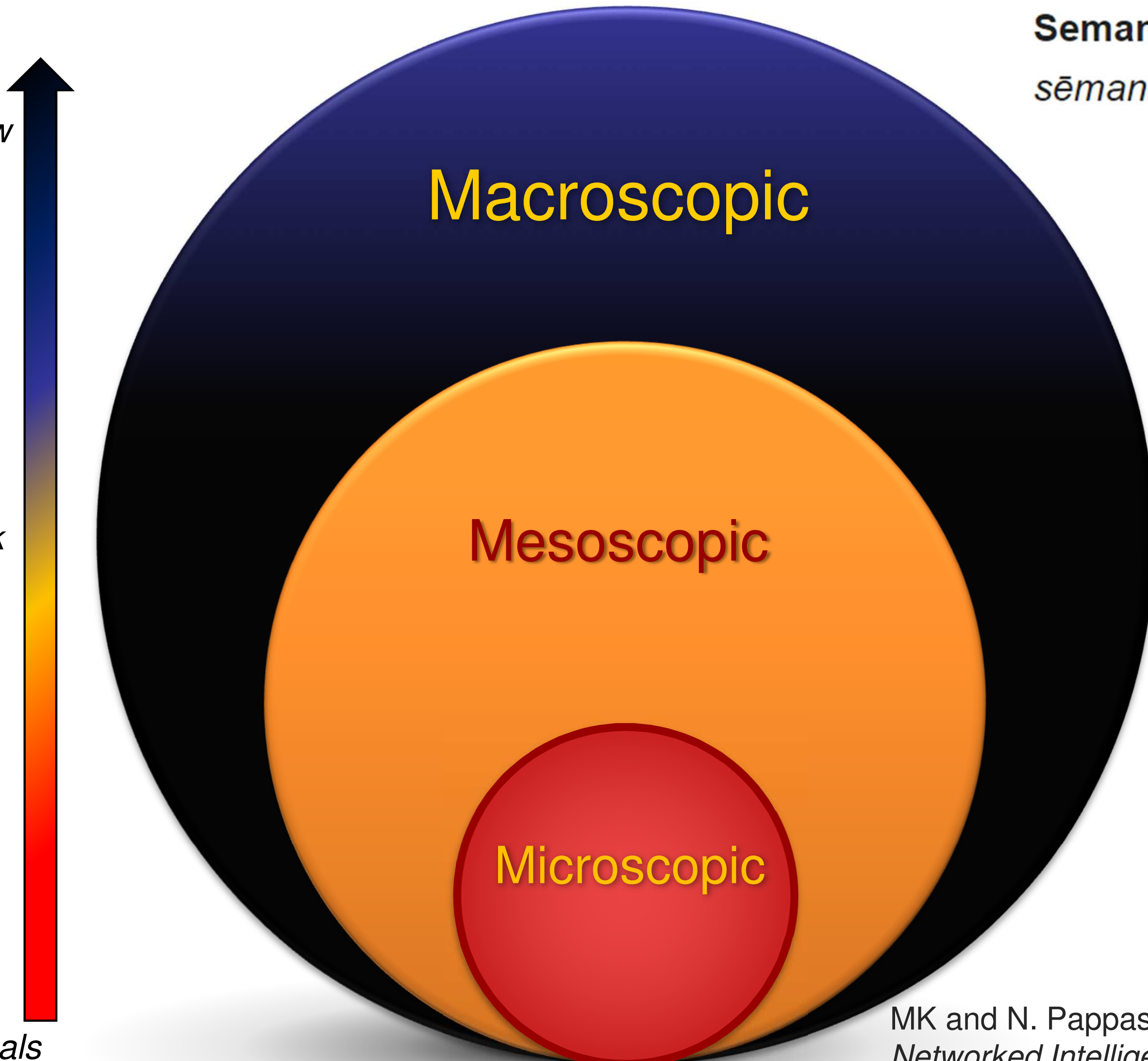
Defining Data Importance

How to *define* and *quantify* the **effectiveness** and **significance**?

System level - Information flow

Link – Network Multisource-Multiuser

Information sources and processes/signals



Semantics (from Ancient Greek: σημαντικός *sēmantikós*, "significant")

End-to-end, system state/timing "dilation"

Quantitative and *Qualitative* innate and contextual attributes of information

Relative importance of different outcomes/events/observations

MK and N. Pappas, "*Semantics-Empowered Communication for Networked Intelligent Systems*", IEEE Commun. Mag. 59(6), 2021

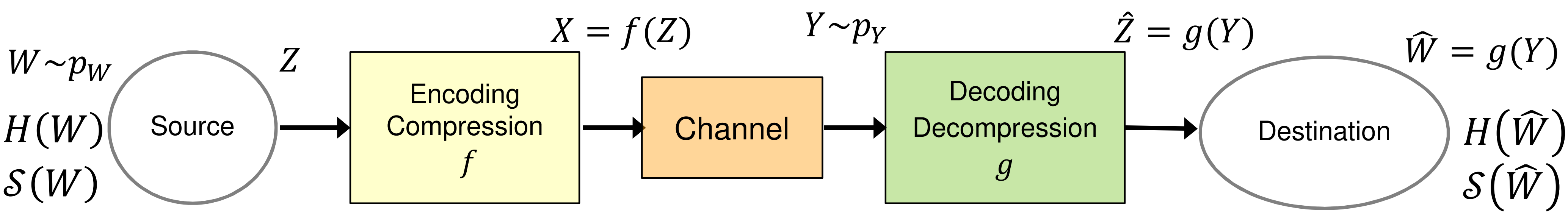
Information Importance Metrics

- **Rationale:** *incorporate disparity between outcomes of the same probability*
- Goal-oriented approach: occurrence of an event removes a double uncertainty
  - *Quantitative*: related to its probability of occurrence
  - *Qualitative:* related to its usefulness/utility for the fulfillment of the goal.
- Given a probability measure μ on data space $\mathcal{X}$ and a countable partition $\mathcal{P}$.
  Let $f\colon \mathbb{R} \to \mathbb{R}$ and $g\colon [0,1] \to \mathbb{R}$ be continuous functions

$$\mathcal{S}(\mu, \mathcal{P}) = \mathcal{S}(X) = f\left(\sum_{P \in \mathcal{P}} g(\mu(P))\right)$$

f is increasing, g is subadditive and concave
OR
f is decreasing, g is superadditive and convex

⚠ Satisfying axioms? (Khinchin's, Fadeev's, …)
⚠ Additivity? Operational meaning?



$$C = \sup_{p_X(x)} I(X;Y)$$
$$\text{s.t. } \mathcal{S}(W) - \mathcal{S}(W|\widehat{W}) \le \sigma$$

$$\min I(Z, \hat{Z}, \widehat{W})$$
$$\text{s.t. } d(Z, \hat{Z}) \le D_c$$
$$d(Z, \widehat{W}) \le D_s$$
$$\mathfrak{D}(p_W, p_{\widehat{W}}) \le \sigma$$

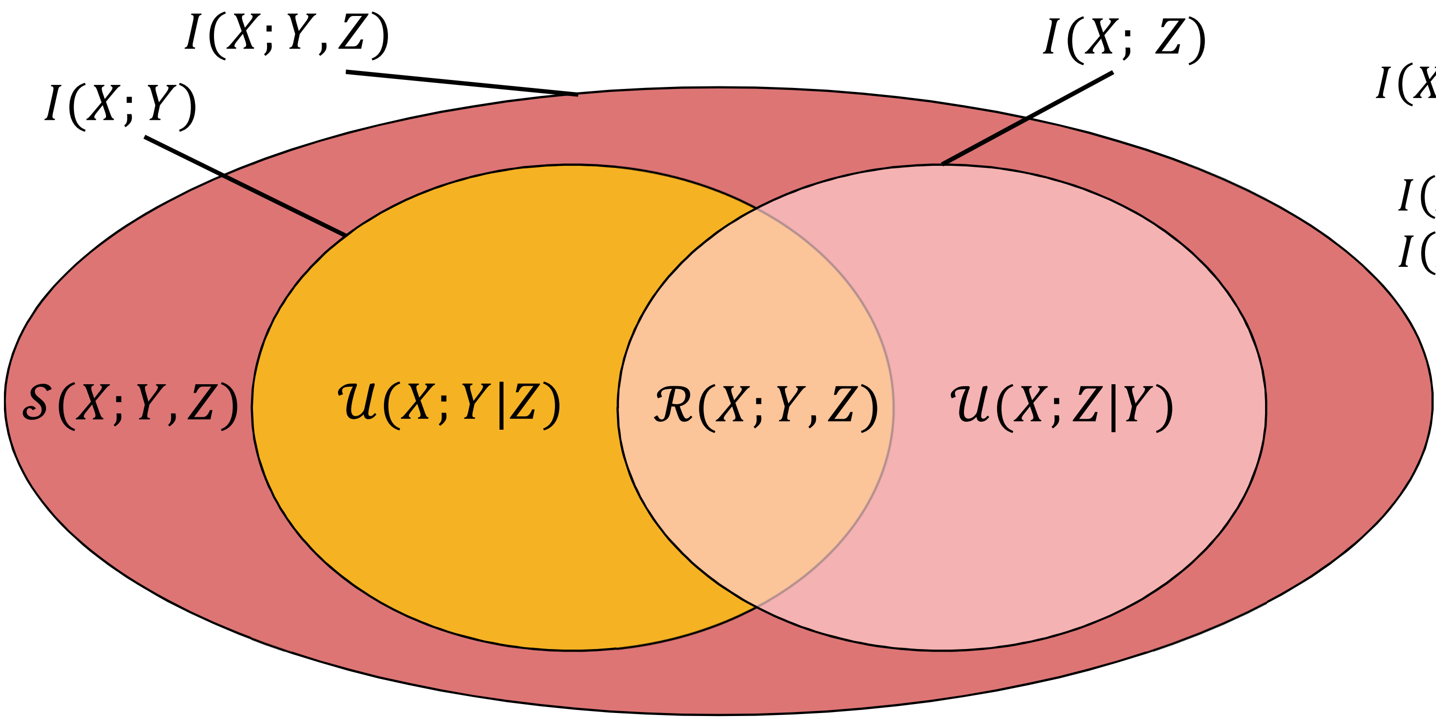
Various challenges

- Entropy measures average uncertainty – sensitivity to signal's nonstationary?
- Time-dependent info measures ? (capturing instantaneous signal changes).

Information Dependencies

Multivariate Information Decomposition

- **Q:** What is the information that two 'source' variables $\boldsymbol{Y}$, $\boldsymbol{Z}$ carry about a third 'target' variable $\boldsymbol{X}$
- Decompose total information into **redundant (shared)**, **unique** and **synergistic** components.



$$I(X;Y,Z) = \mathcal{U}(X;Y|Z) + \mathcal{U}(X;Z|Y) + \mathcal{S}(X;Y,Z) + \mathcal{R}(X;Y,Z)$$
$$I(X;Y) = \mathcal{U}(X;Y|Z) + \mathcal{R}(X;Y,Z)$$
$$I(X;Z) = \mathcal{U}(X;Z|Y) + \mathcal{R}(X;Y,Z)$$

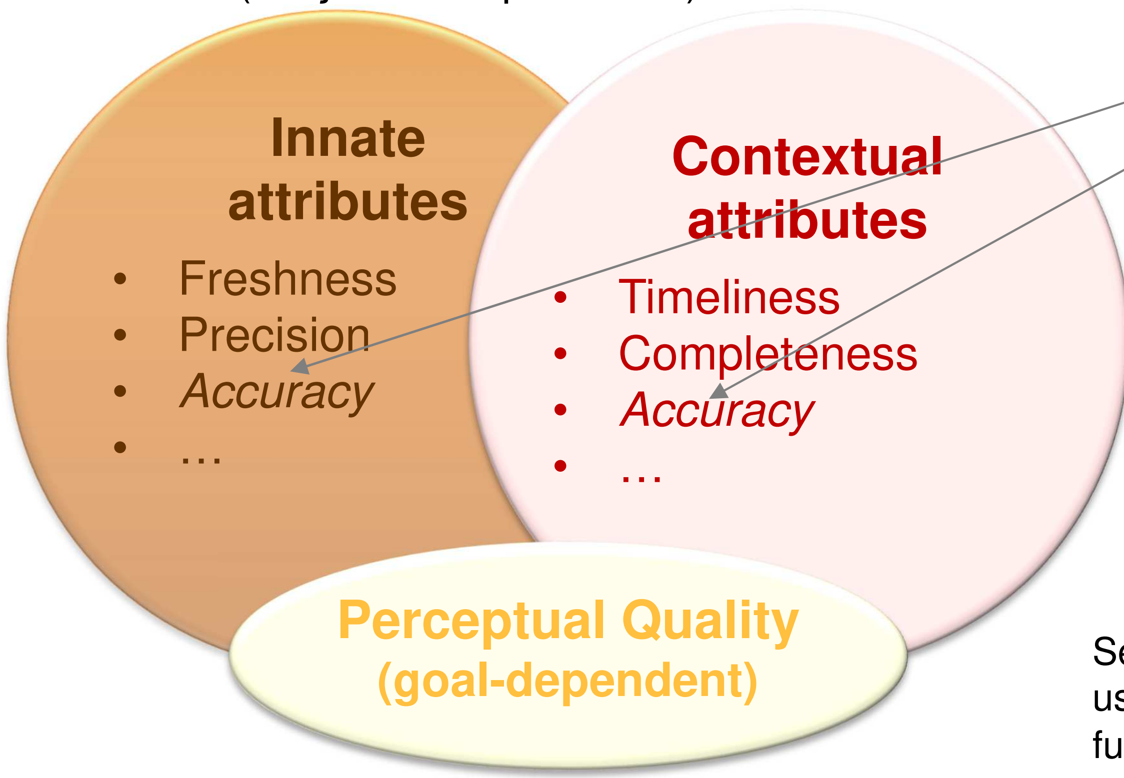
- Beyond Gács-Körner or Wyner common information?

- Need for **interaction info measures**…beyond $I_{\min}(Y_1,\dots,Y_n;X) = \sum_{x\in\mathcal{X}} P(x) \min_i D_{\mathrm{KL}}\left[P(Y_i|x)||P(Y_i)\right]$
- Information dynamics: stoch. processes $X_{\{0:k\}} = \{X_0,\dots,X_k\}$ and $Y_{\{0:k\}} = \{Y_0,\dots,Y_k\}$ and next state X_{n+1}at time $n<k$: $I\left(X_{n+1};X_{\{0:n\}},Y_{\{0:n\}}\right) = \underbrace{I\left(X_{n+1};X_{\{0:n\}}\right)}_{\text{active storage}} + \underbrace{I\left(X_{n+1};Y_{\{0:n\}}|X_{\{0:n\}}\right)}_{\text{transfer entropy}}$

Defining Effectiveness & Data Importance

- Let $\mathcal{V} \in \mathbb{R}^m$ denote the vector of m attributes of information, decomposed into:
  - $\mathcal{I} \in \mathbb{R}^n$ innate/intrinsic (*objective* - quantitative)
  - $\mathcal{C} \in \mathbb{R}^\ell$ contextual/extrinsic (*subjective* - qualitative)

$n, \ell \leq m$



information value **dualism**

Semantic quality measured using divergence or distance function $\mathfrak{D}(\cdot \,||\, \cdot)\colon \mathcal{D} \times \mathcal{D} \to \mathbb{R}$

Semantics of Information $\mathcal{S}_t = v(\psi(\mathcal{V}))$

$v\colon \mathbb{R}^z \to \mathbb{R}$: context-dependent, cost-aware function

$\psi(\mathcal{V})\colon \mathbb{R}^m \to \mathbb{R}^z, m \geq z$: nonlinear, multi-dim function of vector of information attributes $\mathcal{V}$

Semantics of Information

Semantics of Information (SoI)

$$\mathcal{S}_t = v\big(\psi(\mathcal{V})\big)$$

$v: \mathbb{R}^z \to \mathbb{R}$: context-dependent, cost-aware function

$\psi(\mathcal{V}): \mathbb{R}^m \to \mathbb{R}^z, m \geq z$: nonlinear, multi-dim function of vector of information attributes $\mathcal{V}$

A very simple example

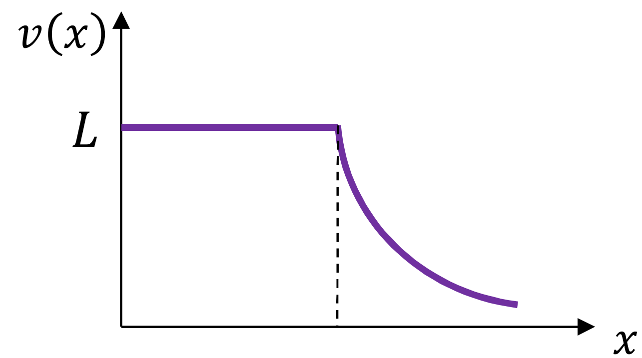
- Freshness (AoI): $\Delta_t = t - u_t$

  u_t: generation time of the newest sample that has been delivered at the destination by time instant t

- Accuracy (distortion): $\delta: \mathcal{X} \times \mathcal{X} \to \mathbb{R}_{\geq 0}$ e.g., $\delta\big(X_t, \hat{X}_t\big) = \big(X_t - \hat{X}_t\big)^2$

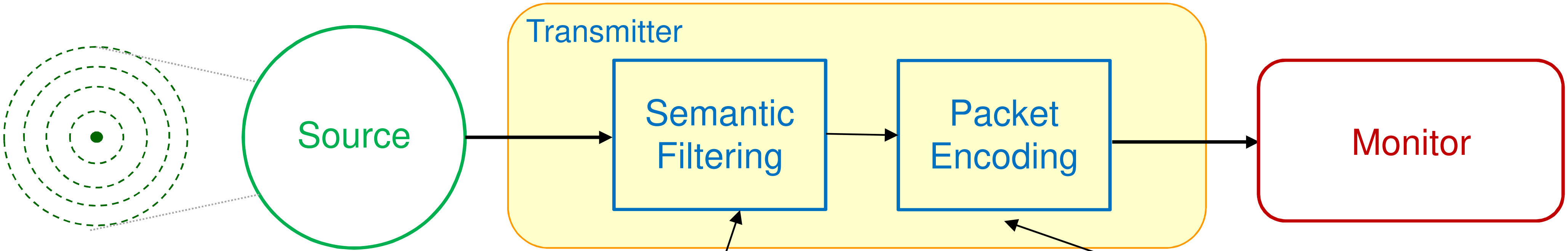
- $\psi(x, y) = Kxy$, so $\psi(\Delta_t, \delta) = K(t - u_t)\big(X_t - \hat{X}_t\big)^2$

- Timeliness: $v(\Delta_t) = \max\big(L, Le^{-\Delta_t}\big), x \geq 0$



- AoI (vanilla, nonlinear, AoII,..), VoI, QoI,… can be seen as special cases of SoI

Semantic Source Coding



$\mathcal{X} = \{x_1, x_2, \cdots, x_n\}$
discrete symbols

Probability of realization
$\widetilde{p}_i = P_X(x_i)$ – known PMF

wlog $\widetilde{p}_i \geq \widetilde{p}_j, \forall\, i \leq j$

Importance-based Selective Encoding

- Importance/semantic value = function of prob. occurrence.
- E.g., only k most (or least) probable realizations selected for TX
- remaining $n - k$ realizations censored

- TX encodes an admitted packet from the i-th realization using a prefix-free code based on the truncated distribution with conditional probabilities
$p_i = \frac{\widetilde{p}_i}{q_k}, \forall\, i \in \mathcal{I}_k \subset \mathcal{I}$, where $q_k = \sum_{i \in \mathcal{I}_k} \widetilde{p}_i$

- Information source generates status updates (packets) and forwards them to a TX
- TX encodes the packets and sends them to a remote monitor (RX)
- i.i.d. sequence of observations
- Packets generated ~ Poi(λ)
- TX is bufferless
- Error-free channel

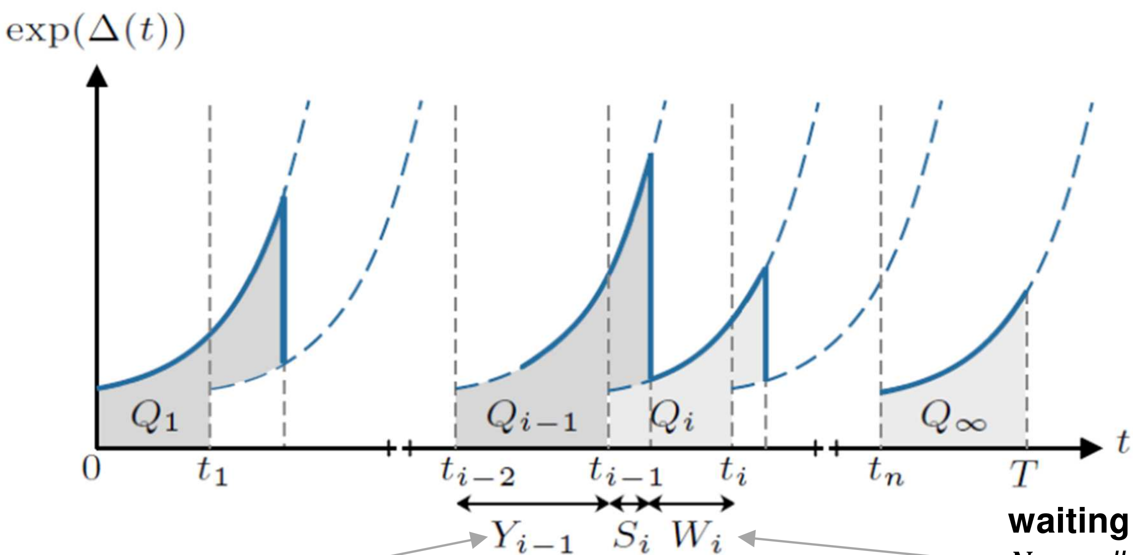
P. Agheli, N. Pappas, MK, "*Semantics-Aware Source Coding in Status Update Systems*", IEEE ICC'22

Semantic Source Coding

- Timeliness (SoI): $\mathcal{S}(t) = g\big(\Delta(t)\big)$ (time-varying stoch. process)
  - $g: \mathbb{R}_0^+ \to \mathbb{R}$ a non-increasing utility function of information freshness
  - AoI: $\Delta(t) = t - u(t)$

- Average SoI for an observation interval $(0, T)$: $\bar{\mathcal{S}} = \lim_{T\to\infty} \frac{1}{T}\int_0^T g(\Delta(t))\mathrm{d}t$

$$= \lim_{T\to\infty} \frac{1}{T}\left\{\sum_{i=1}^{\mathcal{N}(T)} Q_i + Q_\infty\right\} = \eta\mathbb{E}[Q]$$



Steady-state
average **arrival rate**
$\eta = \lim_{T\to\infty} \frac{\mathcal{N}(T) - 1}{T}$

waiting time $W_i = \sum_{k=1}^{N} Z_k$
N: r.v. # admitted arrivals generated before finding the channel idle
Z_k: time between two admitted arrivals, $\sim \exp(\gamma)$, $\gamma = 1/(\lambda q_k)$

$Y_i = t_i - t_{i-1}$
update interval between the i-th successive arrival and its next one

service time
$S_i = \ell_i$ time units
ℓ_i: length of the codeword assigned to x_i

Semantic Source Coding

"Optimal" Codeword Design

- **Aim**: Find the codeword lengths ℓ_i that optimize a weighted sum of the average SoI and the average length for a cost function $\varphi(\ell_i)$, i.e., $\sum_i p_i \varphi(\ell_i)$.

$$\min_{\{\ell_i\}} \quad L(\Delta) + w \sum_{i \in \mathcal{I}_k} p_i \phi(\ell_i)$$

$$\text{s.t.} \quad \sum_{i \in \mathcal{I}_k} 2^{-\ell_i} \leq 1$$

$$\ell_i \in \mathbb{Z}^+$$

integer constraint

Maximizing avg. SoI ≡ Minimizing avg. cost/penalty of lateness

$$L(\Delta) = \lim_{T \to \infty} \frac{1}{T} \int_0^T f(\Delta(t))\mathrm{d}t$$

$f: \mathbb{R}_0^+ \to \mathbb{R}$ a non-decreasing function

$$\phi(x) = \alpha x + \beta x^2, \ \alpha, \beta \geq 0$$

- Quadratic cost function for the codeword length under binary alphabetic
- φ convex: longer (shorter) codewords are penalized more (less) harshly than in the linear case (e.g., Huffman coding)

Kraft-McMillan inequality

for the existence of a uniquely decodable code for a given set of codeword lengths

$$\min_{\{\ell_i\}} \quad \underbrace{\mathbb{E}[Q] + w \sum_{i \in \mathcal{I}_k} p_i (\alpha \ell_i + \beta \ell_i^2)}_{\mathcal{J}_{\text{SoI}}}$$

$$\text{s.t.} \quad \sum_{i \in \mathcal{I}_k} 2^{-\ell_i} \leq 1,$$

$$\ell_i \in \mathbb{R}^+$$

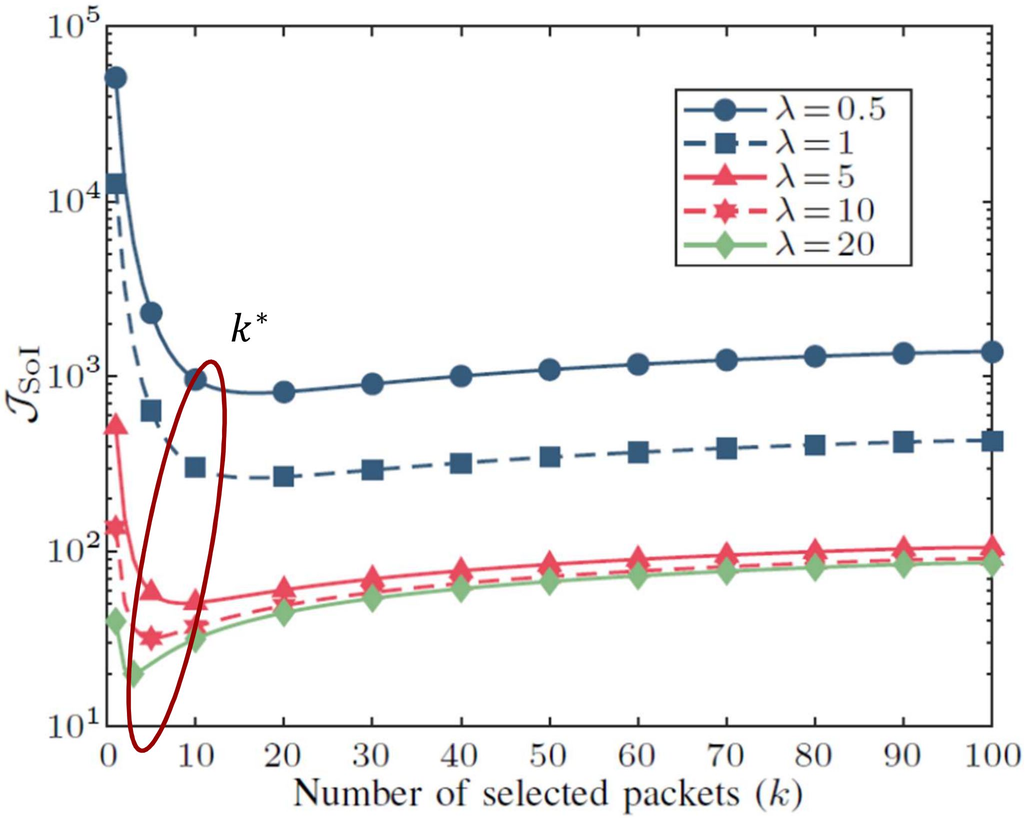
Relaxation: non-negative real valued codeword lengths

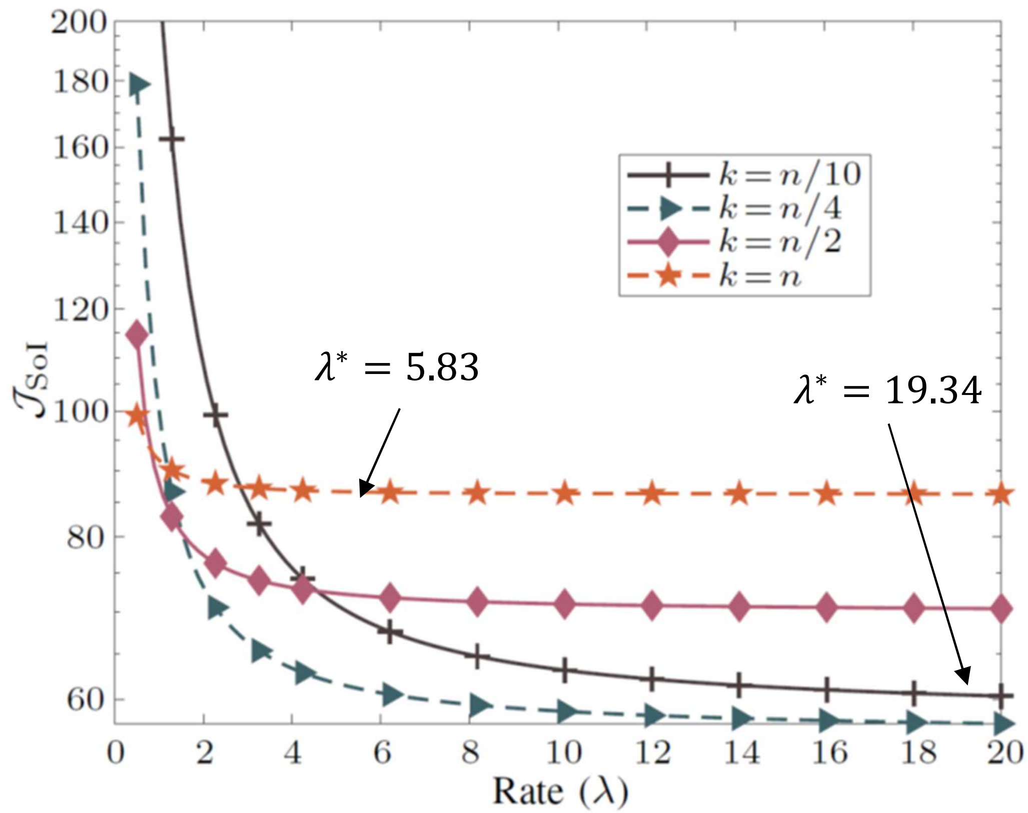
Semantic Source Coding

Zipf(n, s) distribution with pmf $P_X(x) = \frac{1/x^s}{\sum_{j=1}^{n} 1/j^s}$

$n = |\mathcal{X}| = 100$ and exponent $s = 0.4$

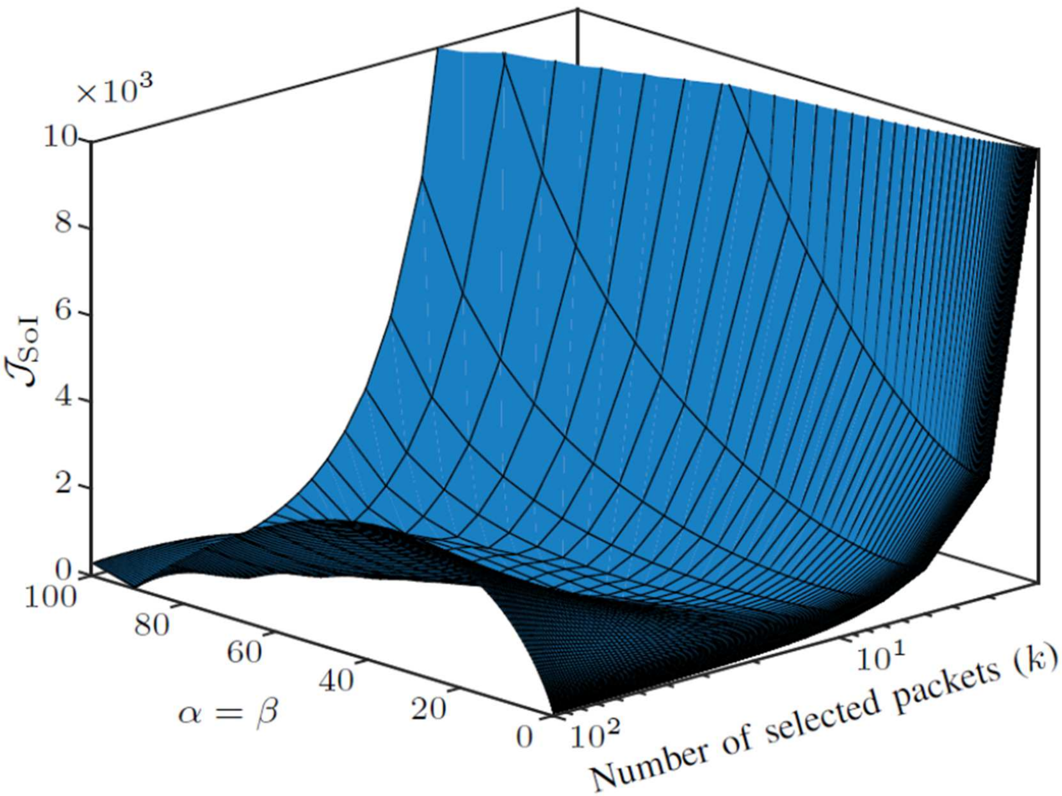
$s = 0$ uniform, $\nearrow s$ "peaky distribution"





Semantic Source Coding

Interplay among SoI, semantic filtering (k) and codeword length



λ	k	$\alpha = \beta$	λ	k	$\alpha = \beta$
0.5	20	1.26	10	5	2.5
1	18	1.58	20	2	12.59
5	10	1.99			

optimal parameters

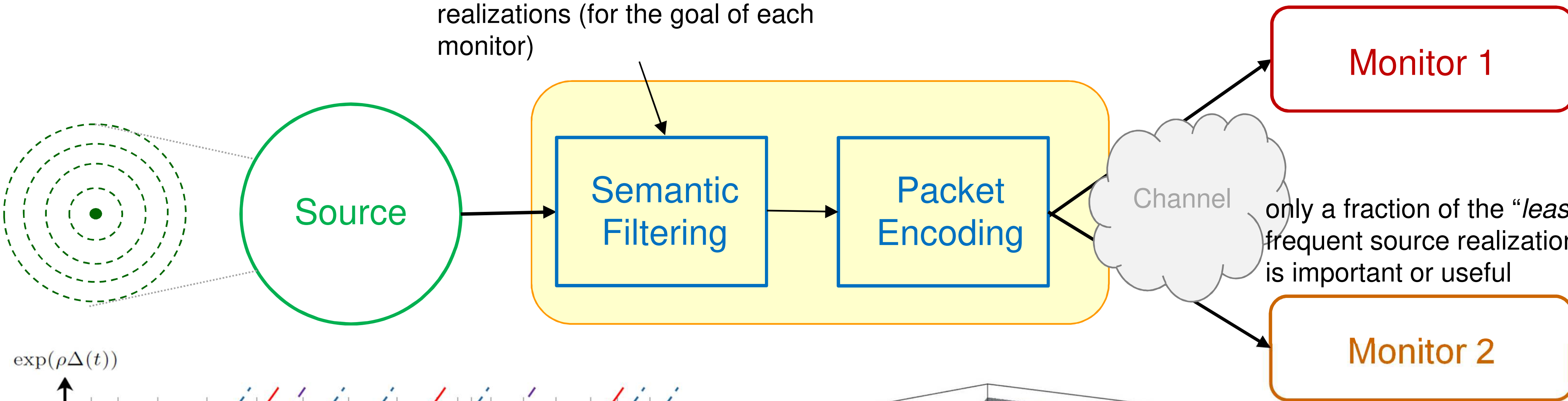
- Objective function continuously increases as cost parameters increase for small k
- For large k: increasing cost parameters causes the objective function to increase then decrease.
- Increasing the input rate (hence, decreasing k^*), optimal cost parameters increase.
- When input rate is high: larger penalties for the codeword length must be assigned

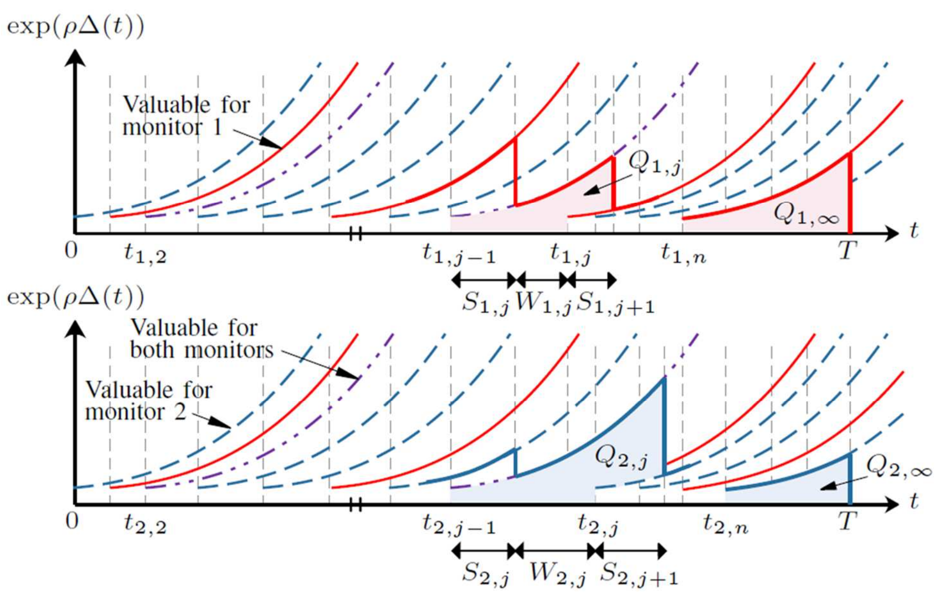
Semantic Source Coding in Multiuser Systems

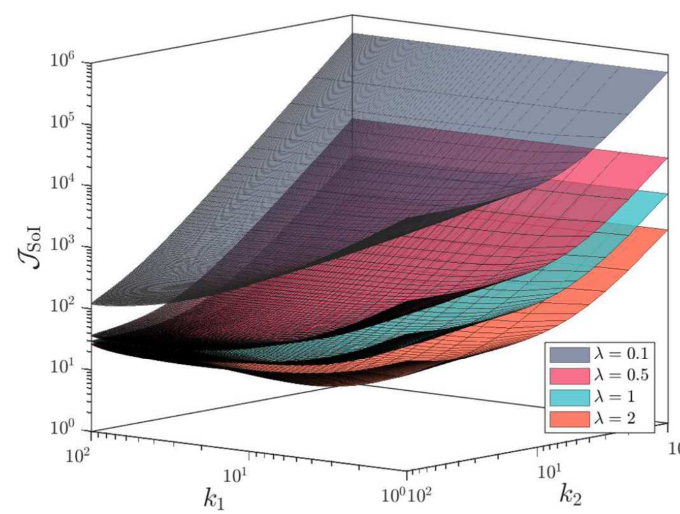
Two Receivers with Heterogeneous Goals

Selection of the most important realizations (for the goal of each monitor)

only a fraction of the "*most*" frequent source realizations is important or useful



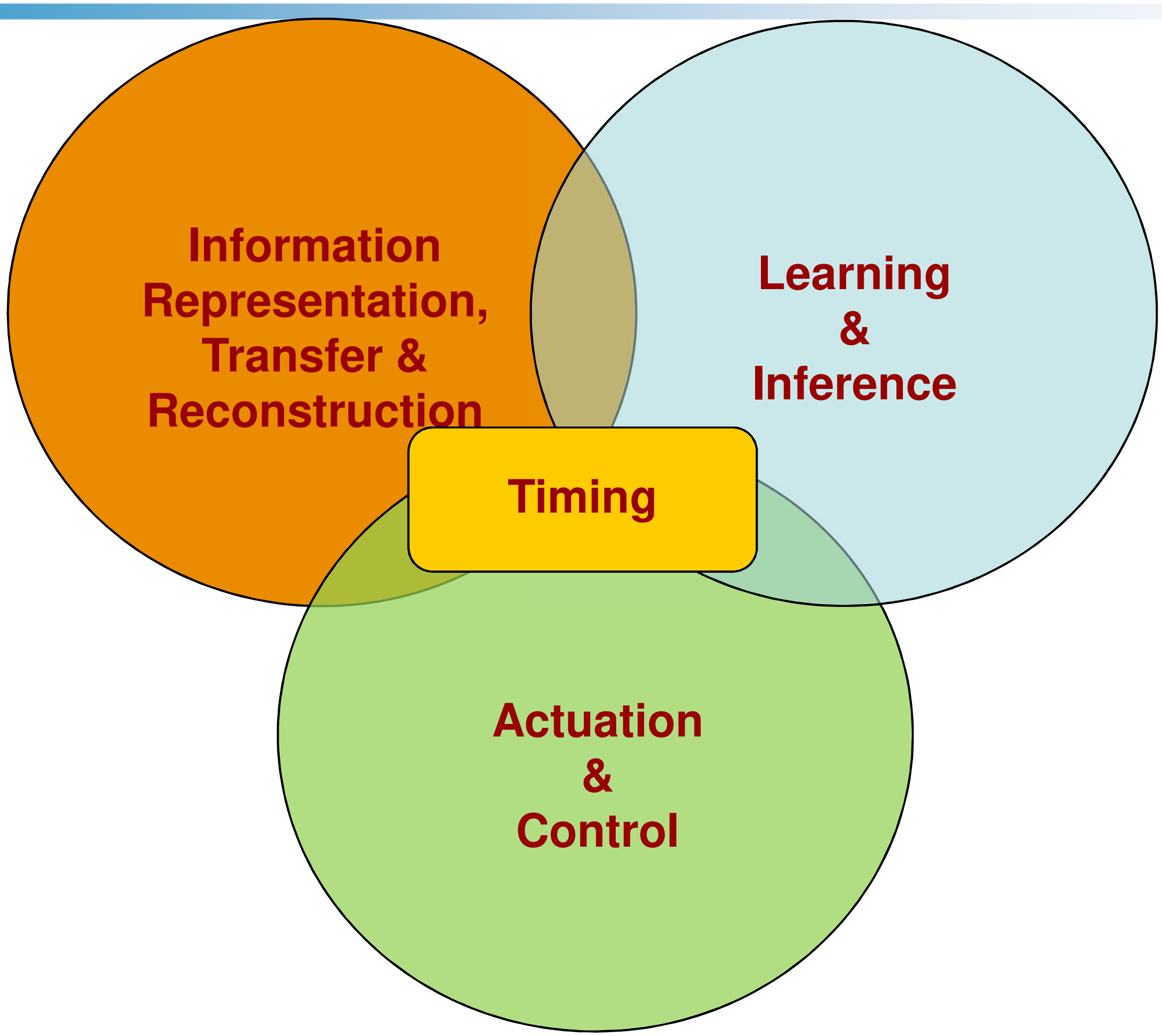




Arrival rate	k_1	k_2
$\lambda = 0.1$	100	94
$\lambda = 0.5$	100	45
$\lambda = 1$	45	14
$\lambda = 2$	45	10

P. Agheli, N. Pappas, MK, "*Semantic Source Coding for Two Users with Heterogeneous Goals*" IEEE Globecom'22

Redefining Effectiveness and Timing



- Effectiveness defined wrt. to the goal/use of the data exchange
- Knowledge/side info about the observer's state is <u>key</u>
- Effectiveness is related to timing in different communication scenarios

P. Popovski et al., "*A Perspective on Time Toward Wireless 6G,*" in Proc. of the IEEE, 110 (8), Aug. 2022

Information Representation & Reconstruction

Encoder f → Rate R → Decoder g

$X \sim p_X$

Source

Layering

Fidelity $\mathbb{E}_{p_{X,\hat{X}}}[\delta(X,\hat{X})]$

$\delta: \mathcal{X} \times \hat{\mathcal{X}} \rightarrow \mathbb{R}_+$

$\hat{X} \sim p_{\hat{X}}$

User/Observer

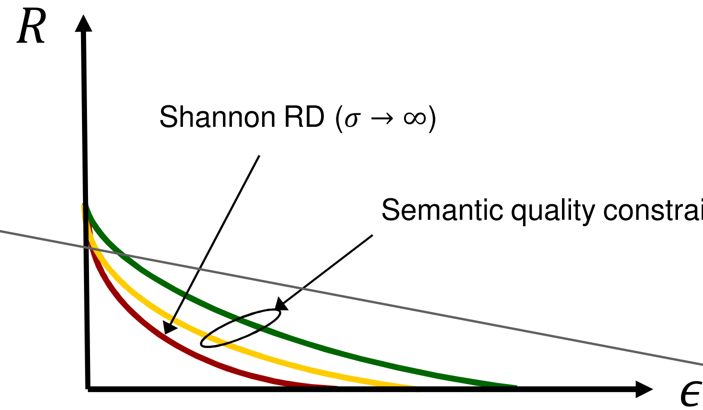
***Fast semantic* reconstruction**

(esp. images, patterns, model learning)

$$R(\epsilon,\sigma) = \min_{p_{\hat{X}|X}} I(X,\hat{X})$$

s.t. $\mathbb{E}_{p_{X,\hat{X}}}[\delta(X,\hat{X})] \leq \epsilon$

$$\mathfrak{D}(p_X, p_{\hat{X}}) \leq \sigma$$



monotonically non-increasing and convex in ϵ and σ

$R(\epsilon, 0) \neq R(\epsilon, \infty)$ (under assumptions)

Fidelity – Distortion metrics

$\delta(X,\hat{X}) = \sum_i \omega_i \left\| \mathcal{F}_i(X) - \mathcal{F}_i(\hat{X}) \right\|^2$

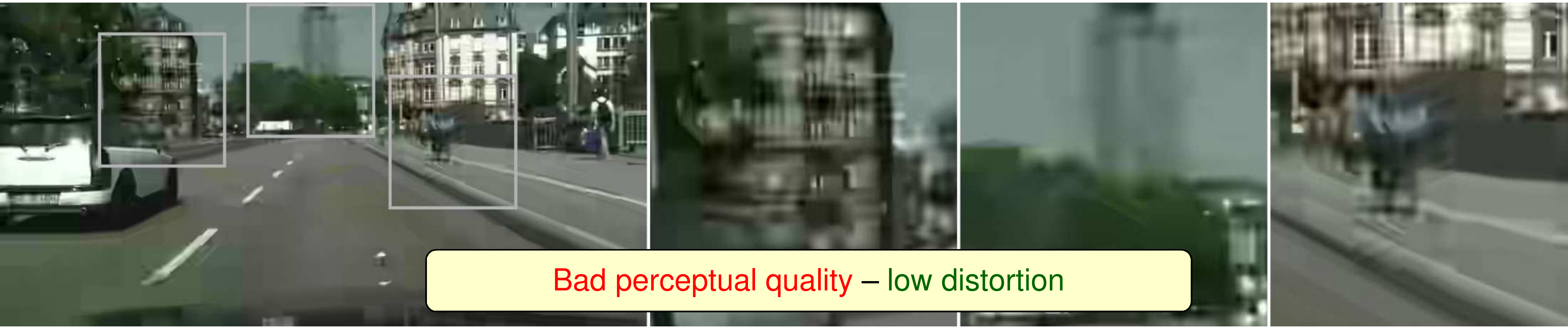
$\mathcal{F}_i$: feature-based mapping function

Semantic quality – Effectiveness

- divergence (Wasserstein, f-div, α-div, Rényi, …)
- generalized entropic measure $\mathcal{S}(X) = g(\int w(\mu(x)))$

$$\mathcal{S}(X) - \mathcal{S}(X|\hat{X}) \leq \sigma$$

Semantic Quality



Good perceptual quality ≠ low distortion

Agustsson et al. (2018)

Information Representation & Reconstruction

s^n → $p(x^n|s^n)$ → x^n → Encoder f → Rate R → Decoder g →

$S \sim p_S$

$X \sim p_X$

Fidelity $\mathbb{E}_{p_{X,\hat{X}}}[\delta(X,\hat{X})]$

$\delta: \mathcal{X} \times \hat{\mathcal{X}} \to \mathbb{R}_+$

$\hat{X} \sim p_{\hat{X}}$ $\quad \hat{S} \sim p_{\hat{S}}$

***Fast semantic* reconstruction**

(esp. images, patterns, model learning)

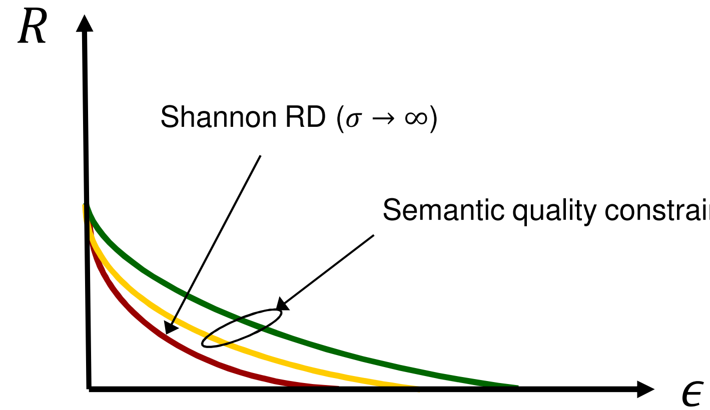
$$R(\epsilon,\sigma) = \min_{p_{\hat{X}|X}} I(X,\hat{X})$$

s.t. $\mathbb{E}_{p_{X,\hat{X}}}[\delta(X,\hat{X})] \le \epsilon$

$\mathbb{E}_{p_{S,\hat{S}}}[\delta(S,\hat{S})] \le \vartheta$

$\mathfrak{D}(p_S, p_{\hat{S}}) \le \sigma$

$\mathfrak{D}(p_X, p_{\hat{X}}) \le \varpi$



Semantic quality – Effectiveness

- divergence (Wasserstein, f-div, α-div, Rényi, …)
- generalized entropic measure $\mathcal{S}(X) = g(\int w(\mu(x)))$

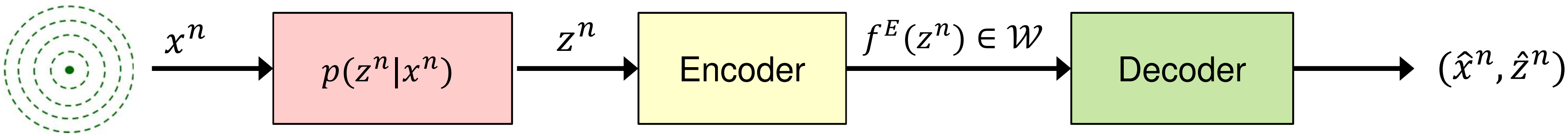
$$\mathcal{S}(X) - \mathcal{S}(X|\hat{X}) \le \sigma$$

Fidelity – Distortion metrics

$\delta(X,\hat{X}) = \sum_i \omega_i \left\| \mathcal{F}_i(X) - \mathcal{F}_i(\hat{X}) \right\|^2$

$\mathcal{F}_i$: feature-based mapping function

Goal-Oriented Communication: A Rate-Distortion Approach



$$f^E : \mathcal{Z}^n \to \mathcal{W} \qquad g_o^D : \mathcal{W} \to \widehat{\mathcal{Z}}^n,\ g_s^D : \mathcal{W} \to \widehat{\mathcal{X}}^n$$

$$\mathcal{W} \in \{1, 2, \ldots, M\}$$

- Memoryless *information source*: tuple $(\mathbf{x}, \mathbf{z})$ with prob. dist. $p(x, z)$ in product alphabet space $\mathcal{X} \times \mathcal{Z}$.
- $\mathbf{x}$: semantic or intrinsic information of the source
- $\mathbf{z}$: noisy observation of the source at the encoder side

- Semantic distortion function: $d_s: \mathcal{X} \times \widehat{\mathcal{X}} \mapsto [0, \infty)$
- Observation/communication distortion function: $d_o: \mathcal{Z} \times \hat{\mathcal{Z}} \mapsto [0, \infty)$
- Average per-symbol distortions: $d_s^n(x^n, \widehat{x}^n) = \dfrac{1}{n} \displaystyle\sum_{t=1}^{n} d_s(x_i, \widehat{x}_i)$

$$d_o^n(z^n, \widehat{z}^n) = \frac{1}{n} \sum_{t=1}^{n} d_o(z_i, \widehat{z}_i)$$

What is the role of the *fidelity* criterion
in a remote source coding problem with individual distortion measures?

Semantic Rate Distortion Function (SRDF)

Operational rates

- For a given $p(x)$ and $p(z|x)$, the SRDF is characterized as follows:

$$R(D_s, D_o) = \inf_{\substack{q(\widehat{z},\widehat{x}|z) \\ \mathbf{E}[\widehat{d}_s(\mathbf{z},\widehat{\mathbf{x}})]\leq D_s \\ \mathbf{E}[d_o(\mathbf{z},\widehat{\mathbf{z}})]\leq D_o}} I(\mathbf{z};\widehat{\mathbf{z}},\widehat{\mathbf{x}})$$

$$I(\mathbf{z};\widehat{\mathbf{z}},\widehat{\mathbf{x}}) \triangleq \mathbf{E}\left[\log\left(\frac{q(\widehat{\mathbf{z}},\widehat{\mathbf{x}}|\mathbf{z})}{\nu(\widehat{\mathbf{z}},\widehat{\mathbf{x}})}\right)\right]$$

$$\widehat{d}_s(z,\widehat{x}) = \sum_{x\in\mathcal{X}} p(x|z) d_s(x,\widehat{x})$$

Special case of multiple description source coding problem

Functional properties

- $R(D_s, D_o)$ is a non-increasing function of $D_s \in [0,\infty)$ and $D_o \in [0,\infty)$ and (jointly) convex with respect to (D_s, D_o)
- $I(\mathbf{z};\hat{\mathbf{z}},\hat{\mathbf{x}})$ is a convex functional of $q(\hat{z},\hat{x}|z)$ for a fixed $p(z)$
- If $R(D_s, D_o) < \infty$, then $R(\cdot)$ is continuous for $D_s \in [0,\infty)$ and $D_o \in [0,\infty)$

- Infimum is attained by a $q^*(\hat{z},\hat{x}|z)$ (compact constrained set and lower semi-continuous wrt $q(\hat{z},\hat{x}|z)$)

P. Stavrou and MK, "*A Rate-Distortion Approach to Goal-oriented Communication*," IEEE ISIT 2022

P. Stavrou and MK, "*The Role of Fidelity in Goal-Oriented Semantic Communication: A Rate Distortion Approach*," Techrxiv 20098970

Bounds and Conditions for Tightness

$$\underbrace{\max\{R(D_s), R(D_o)\} \le}_{R^L(D_s,D_o)} \quad R(D_s, D_o) \quad \underbrace{\le R(D_o) + R(D_s)}_{R^U(D_s,D_o)}$$

$$R(D_o) = \min_{\substack{q(\widehat{z}|z) \\ \mathbf{E}\{d_o(\mathbf{z},\widehat{\mathbf{z}})\} \le D_o}} I(\mathbf{z};\widehat{\mathbf{z}})$$

direct R-D problem
with an i.i.d. source $\mathbf{z}$

$$R(D_s) = \min_{\substack{q(\widehat{x}|z) \\ \mathbf{E}\{\widehat{d}_s(\mathbf{z},\widehat{\mathbf{x}})\} \le D_s}} I(\mathbf{z};\widehat{\mathbf{x}})$$

indirect R-D problem
with an i.i.d. remote source $\mathbf{x}$
and a noisy observation $\mathbf{z}$

Tightness

- $R^L(D_s,D_o)$ is tight *iff* $\mathbf{z} - \hat{\mathbf{z}} - \hat{\mathbf{x}}$ and $\mathbf{z} - \hat{\mathbf{x}} - \hat{\mathbf{z}}$ are concurrently satisfied
- $R^U(D_s,D_o)$ is tight *iff* $\hat{\mathbf{z}} - \mathbf{z} - \hat{\mathbf{x}}$ is satisfied

General Result (Theorem)

$$R(D_o, D_s) = \max_{\substack{s_1 \le 0 \\ s_2 \le 0}} \min_{\substack{q(\widehat{z},\widehat{x}|z) \ge 0 \\ \sum_{\widehat{z},\widehat{x}} q(\widehat{z},\widehat{x}|z)=1}} \left\{ I(\mathbf{z};\widehat{\mathbf{z}},\widehat{\mathbf{x}}) - s_1\left(\mathbf{E}\left[\widehat{d}_s(\mathbf{z},\widehat{\mathbf{x}})\right] - D_s\right) - s_2\left(\mathbf{E}\left[d_o(\mathbf{z},\widehat{\mathbf{z}})\right] - D_o\right)\right\}$$

- Parametric solutions for $R(D_s, D_o)$
- Used for analytical expressions & generalized Blahut-Arimoto algorithm

Binary Alphabets

- Binary alphabets with individual Hamming distortions

$$\mathcal{X} = \mathcal{Z} = \widehat{\mathcal{X}} = \widehat{\mathcal{Z}} = \{0, 1\}$$

$$p(x) = \begin{pmatrix} p(x=0) \\ p(x=1) \end{pmatrix} = \begin{pmatrix} \alpha \\ 1-\alpha \end{pmatrix},$$

$$p(z|x) = \begin{pmatrix} p(z=0|x=0) & p(z=0|x=1) \\ p(z=1|x=0) & p(z=1|x=1) \end{pmatrix}$$

$$= \begin{pmatrix} \beta & \gamma \\ 1-\beta & 1-\gamma \end{pmatrix}$$

where $(\alpha, \beta, \gamma) \in [0,1] \times [0,1] \times [0,1]$, $\beta \neq \gamma$ *and*

$$d_s(x, \widehat{x}) = \begin{cases} 0 \text{ if } & x = \widehat{x} \\ 1 \text{ if } & x \neq \widehat{x} \end{cases}, \quad d_o(z, \widehat{z}) = \begin{cases} 0 \text{ if } & z = \widehat{z} \\ 1 \text{ if } & z \neq \widehat{z} \end{cases}$$

- Rate-splitting bound is achievable:

$$R(D_s^*, D_o^*) = R^L(D_s, D_o)$$

- Equiprobable semantic remote source (i.i.d. Bernoulli$(1/2)$) - $p(x=0) = 1/2$
- Binary symmetric channel with crossover probability - $p(z=0|x=1) = 1-\beta, \beta \in [0, \frac{1}{2})$

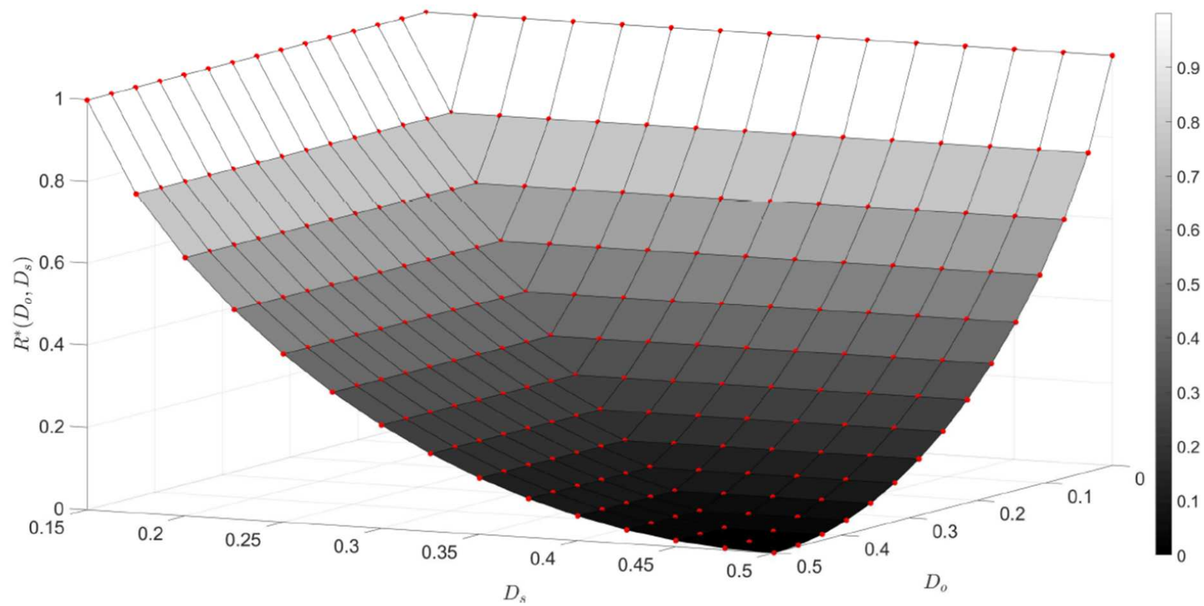
$$R(D_s^*, D_o^*) = \max\left\{[1 - H_b(D_o)]^+, \left[1 - H_b\left(\frac{D_s - \beta}{1 - 2\beta}\right)\right]^+\right\}$$

$H_b(p)$: binary entropy function

Binary Alphabets

Interplay (β, D_s, D_o)

- If $D_o > \frac{D_s - \beta}{1 - 2\beta}$: beneficial to encode only the semantic information (subject to a Hamming distortion) therefore the rate is $R(D_s^*)$

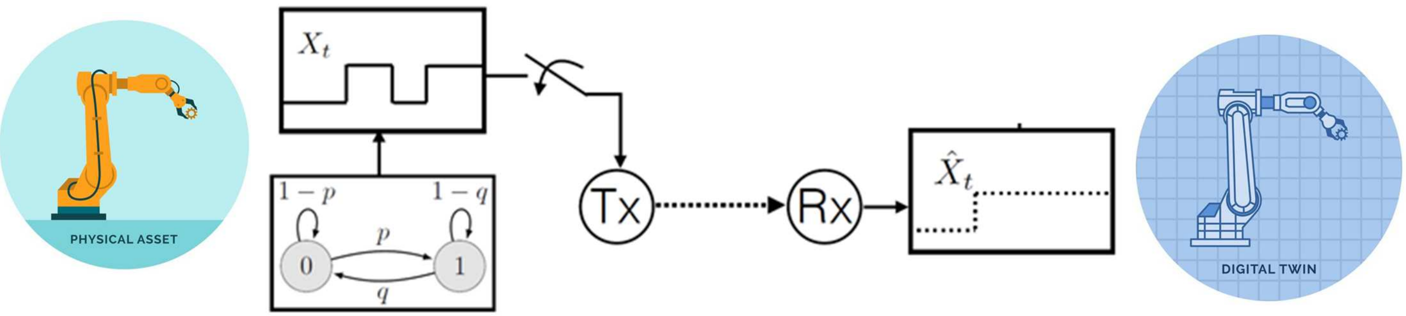


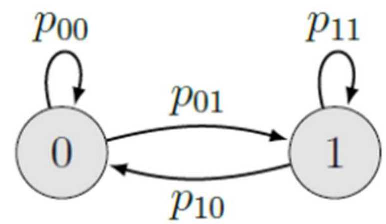
$R(D_s^*, D_o^*)$ for binary alphabets with an equiprobable semantic source and binary symmetric channel with β=0.15

- If $D_o < \frac{D_s - \beta}{1 - 2\beta}$: beneficial to encode the observable message (subject to its distortion) with rates $R(D_o^*)$.

- If $D_o = \frac{D_s - \beta}{1 - 2\beta}$: encoding either the semantic information or the observations does not offer any advantage for any value of the active distortion region.

Real-time Tracking





One-step transition probabilities

$$p_{ji} = \mathbb{P}\left(E_{t+1} = j | E_t = i\right), \forall i, j \in \{0, 1\}$$

- **E2E System**: robot monitors a two-state discrete-time Markov source (DTMC)
- Source initiates actions to a robotic object (Tx side)
- **Goal**: real-time actuation of digital twin (Rx side)

Metrics

Real-time reconstruction error

$$E_t = \mathbb{1}\left(X_t \neq \hat{X}_t\right) = \left|X_t - \hat{X}_t\right|$$

$$\bar{E} = \lim_{T\to\infty} \frac{\sum_{t=1}^{T} E_t}{T} = \lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} \mathbb{1}\left(X_t \neq \hat{X}_t\right)$$

Cost (penalty) of actuation error

$$\bar{C}_A = \pi_{(0,1)} C_{0,1} + \pi_{(1,0)} C_{1,0}$$

$C_{i,j}$: cost of being in i at Tx and $j \neq i$ at Rx

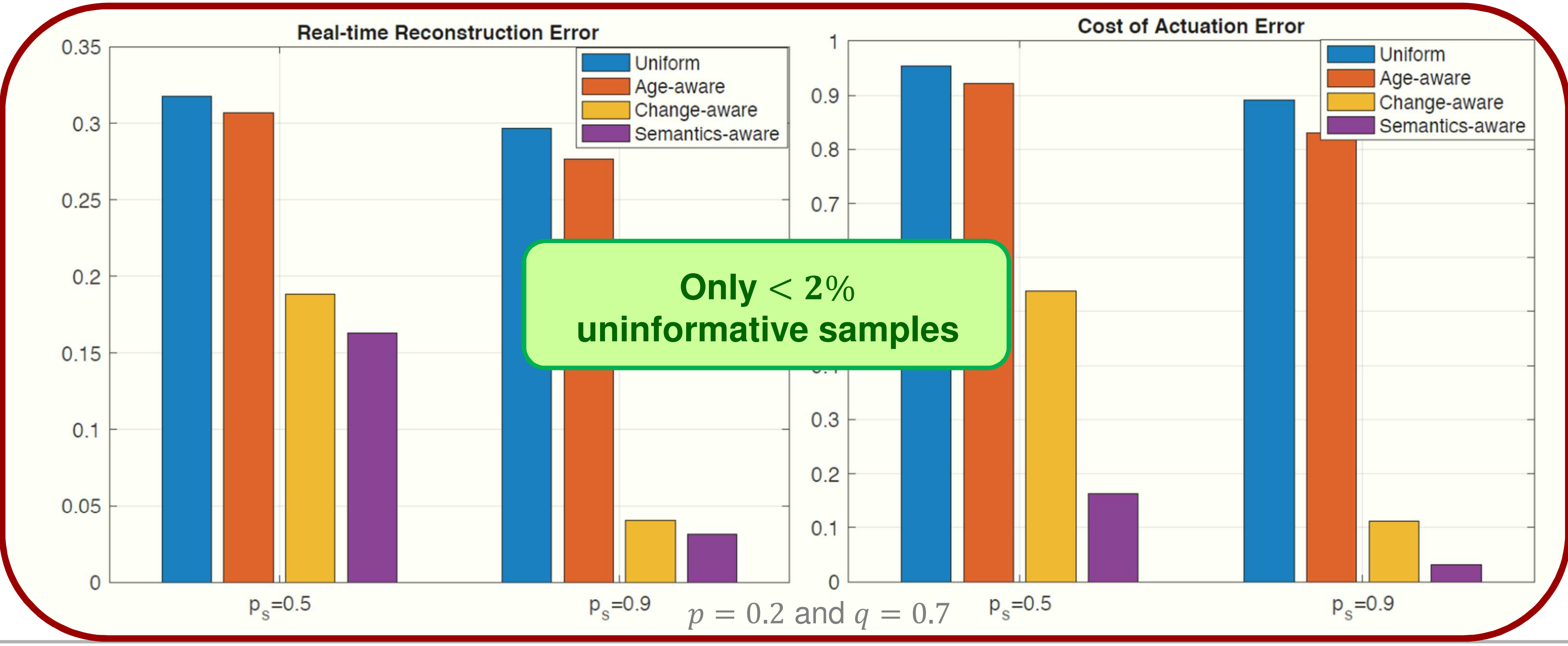
Non-commutative errors: $C_{0,1} \neq C_{1,0}$

N. Pappas and MK, "*Goal-oriented Communication for Real-Time Tracking in Autonomous Systems*," *Proc. IEEE ICAS 2021*

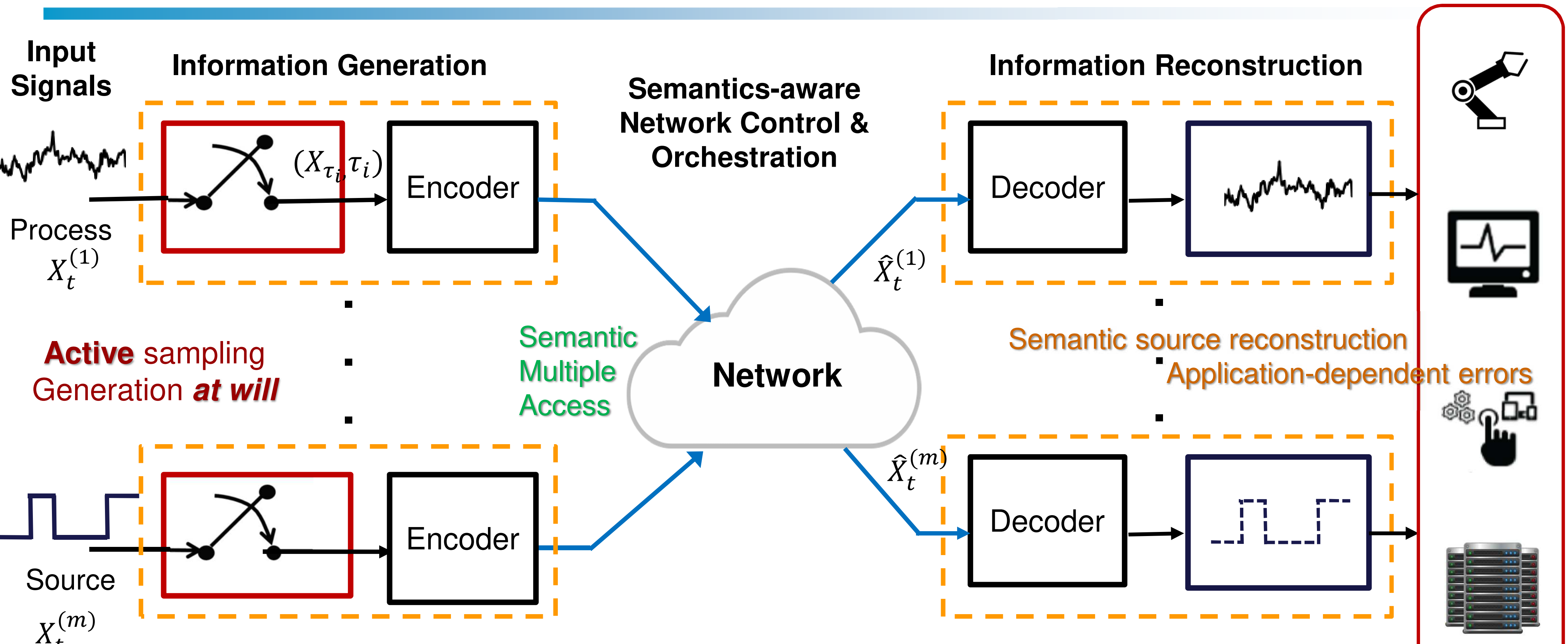
Synergetic Information Handling Gains

Goal-oriented Sampling and Effective Communication Policies

- **Uniform**: periodic, process-agnostic sampling
- **Age-aware**: sampling/transmission triggered when AoI exceed a threshold
- **Change-aware:** sample generation triggered at the Tx whenever a change at the source state is observed
- **Semantics-aware:** sample generation triggered whenever there is discrepancy between X_t and $\hat{X}_t$ (change tracked at both Tx and Rx)



Goal-oriented Data Networking



Key Semantics-Empowered Functionalities

Filtering/censoring
active data acquisition and sparse sampling

Preprocessing
e.g., feature extraction, embedding, segmentation

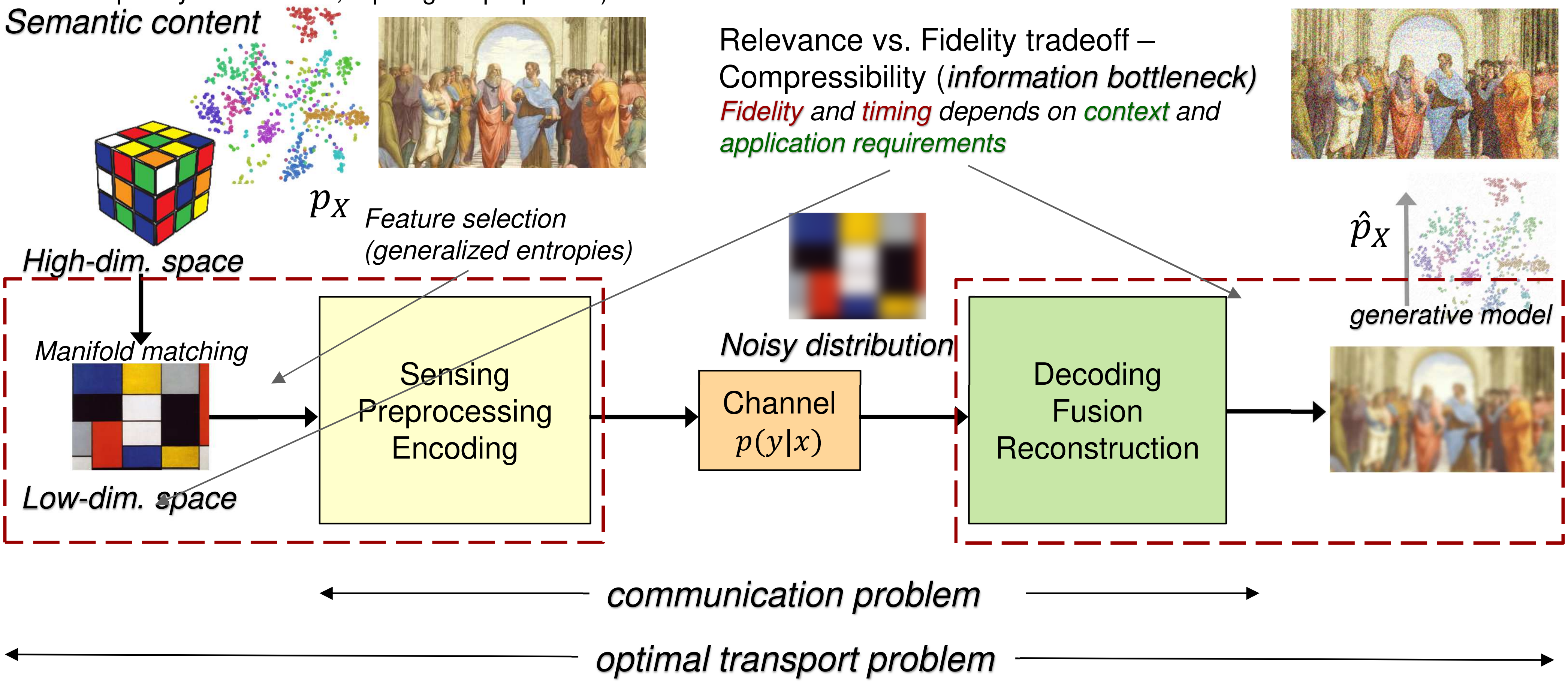
Reception
fast partial or approximate reconstruction, semantic data recovery, querying, …

Control
agile orchestration of multi-quality multi-modal information gathering

Non Separable Information Handling

The Bigger Picture

Complex data (feature richness, algorithmic complexity of inference, topological properties).

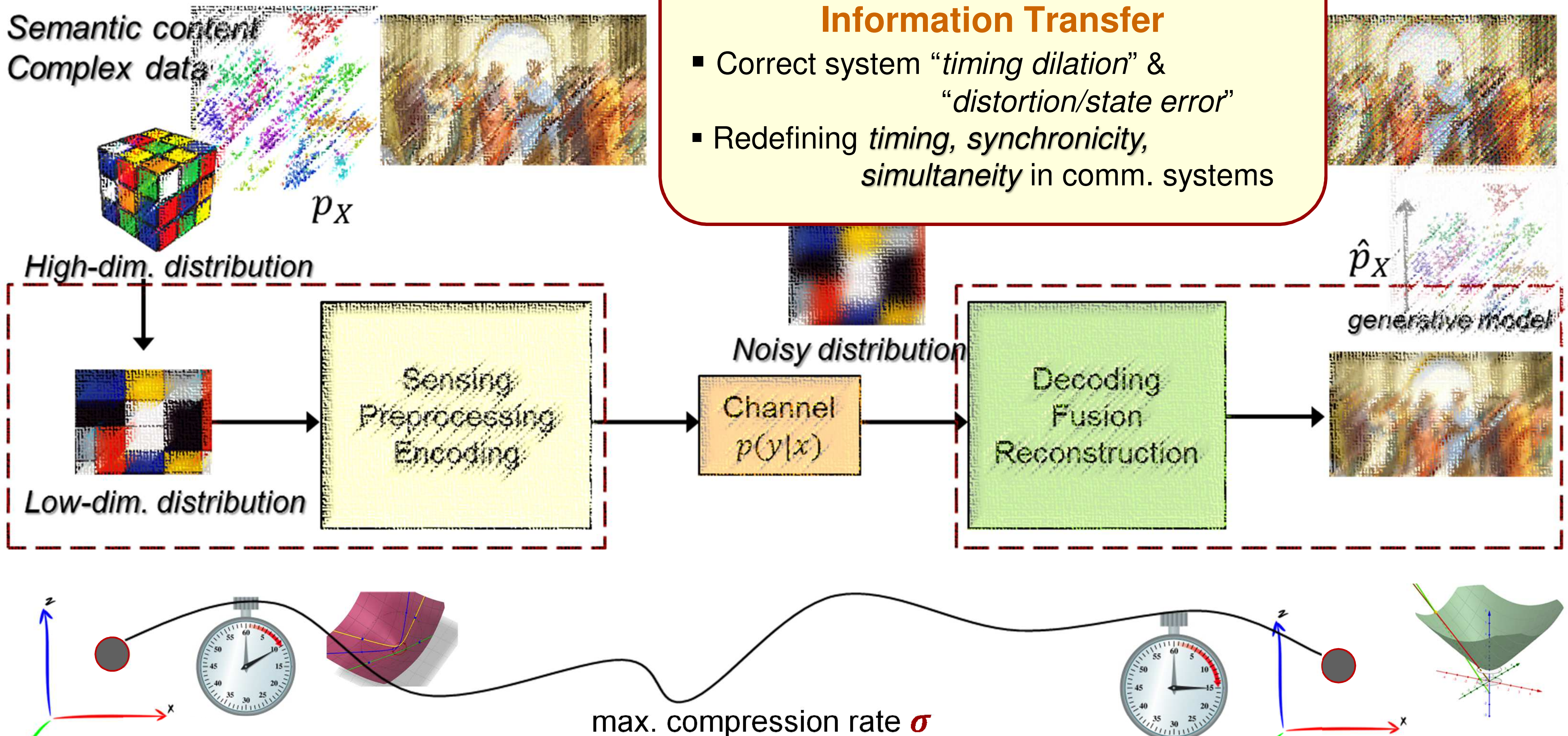


- Communicating *high-dimensional, multi-modal, multi-source rich* data
- Intriguing connections with optimal transport, generative models, decision making, inference…
- Fundamental tradeoffs: communication-computation/learning, rate-distortion-perception, …
- Information Manifold: rate distortion manifolds for extended/richer information spaces or sets/measures

How *Soon* is *Now* ?

"Relativistic" End-to-End Information Transfer

- Correct system "*timing dilation*" & "*distortion/state error*"
- Redefining *timing, synchronicity, simultaneity* in comm. systems



$\mathcal{A}_1 = (x_1, y_1, z_1)$ - t_1

Observer *A*
(physical world)

max. compression rate $\boldsymbol{\sigma}$
min. time duration $\boldsymbol{\tau}$

$$\|\varphi(\hat{X}_{t_2}) - \varphi(X_{t_1})\|^d \to 0$$
$$|t_2 - t_1| \to 0$$

"Twin paradox" ***analogy***
Frame dependent system state and evolution
"Poincaré-Lorentz"-type transformations wrt. observers

$\mathcal{A}_2 = (x_2, y_2, z_2)$ - t_2

Observer *B*
(virtual world)

After the Dust Settled

Common Questions and Misconceptions

- Goal-Oriented Semantic Communication (GSC) is beyond, post, new,… Shannon theory
- Everything in GSC is new, unheard, radical, …

  … or this has been done X years ago ☺
- GSC is a/the new 6G technology
- Where is the PHY in GSC?
- How is this different from End-to-End Learning for Communications
- ML/AI can solve all GSC problems and challenges?
- GSC is just JSCC (Joint Source & Channel Coding)
- GSC is just non-linear AoI (age of information)
- …

Epilogue

- Supporting connected intelligence and autonomous, real-time systems in future wireless networks necessitates
  - fundamental theoretical advances
  - transforming prevailing communication design paradigms

- Effective Goal-oriented Communications: a paradigm shift … not just hype!

- Grand Challenge: Goal-oriented *unification* of data generation/processing
  information transmission
  reconstruction

- Intriguing connections with learning, optimal transport, generative models, decision making … & many fundamental tradeoffs!

- Promising gains: significant improvement in
  - network resource usage
  - energy consumption
  - computational efficiency

  scalability

European Research Council
Established by the European Commission

EURECOM
Sophia Antipolis

Thank you