Goal-oriented Semantic Communication for Decentralized Intelligence

Marios Kountouris

Andalusian Research Institute in
Data Science and Computational Intelligence
Computer Science and AI Dept.
University of Granada, Spain

Communication Systems Dept.
EURECOM, France

I Brazilian Signal Processing Forum (BSP-Forum)

Rio de Janeiro, RJ, Brazil

June 21st, 2024

European Research Council
Established by the European Commission

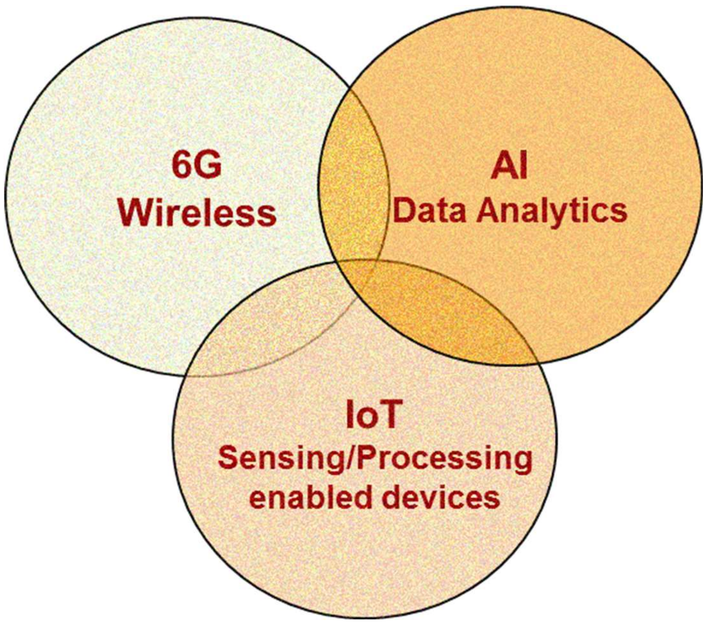
The Road to 6G

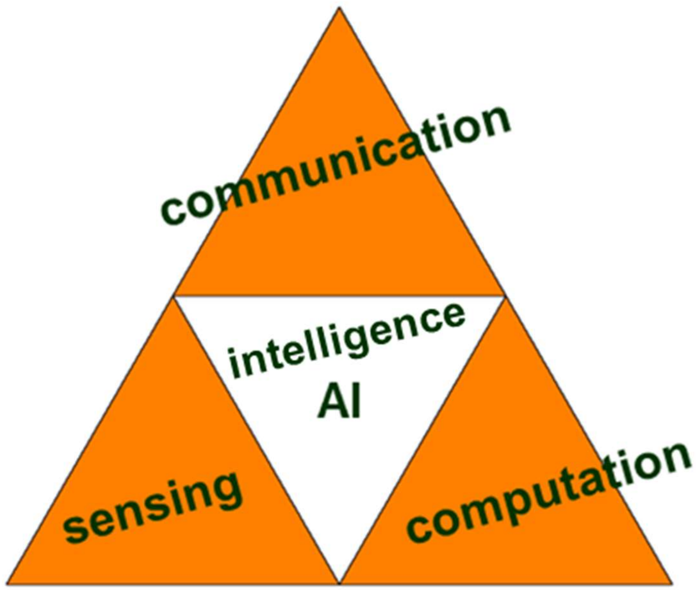
New Services & Use Cases

- Immersive, multisensory XR
- Hi-Res positioning, sensing, 3D mapping
- Holographic com., digital twins, metaverse
- eHealth, consumer robotics, tactile Internet

New Tech Enablers

- AI-native & Open Network Architectures
- Edge Intelligence
- Distributed Computing and Learning
- Joint Sensing and Communication



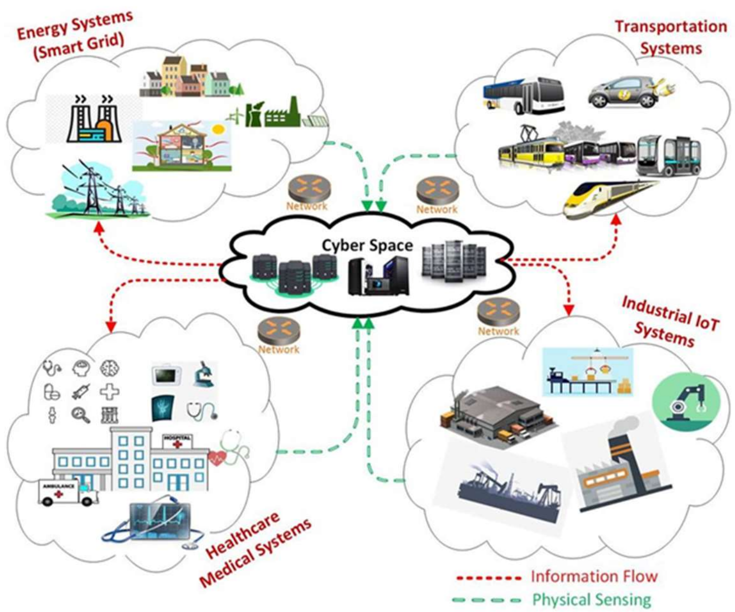


2G	3G	4G	5G	6G
Voice	Visio-phony	Mobile Internet	Wireless for Things	Wireless for ???

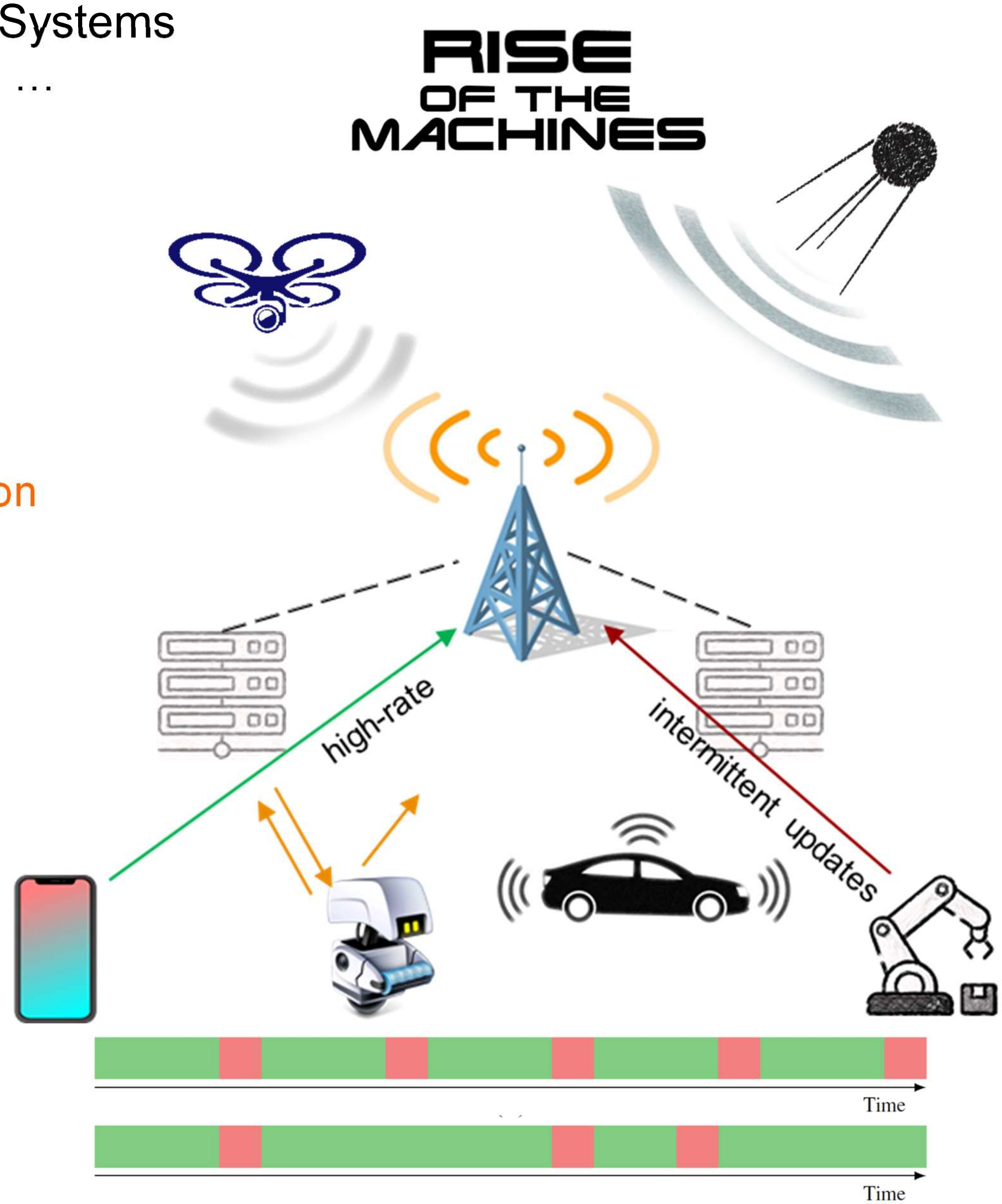
Fast Forward to 2030

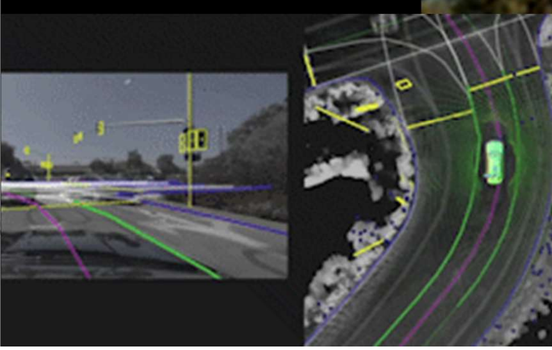
- **Cyber-Physical & Mission-Critical Interactive** Systems
  - swarm robotics, self-driving vehicles, smart IoT, …

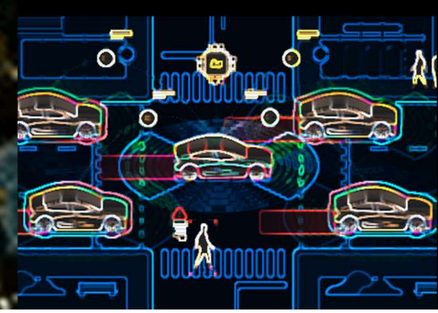
- **Networked Intelligent Systems**
  - reliable ***real-time*** communication
  - ***autonomous*** & ***automated*** interactions
  - ***timely*** & ***effective*** actuation
  - ***explainable*** & ***trustworthy*** decision making
  - on-device, in-network, decentralized computation



ArXiv: 1812.02282







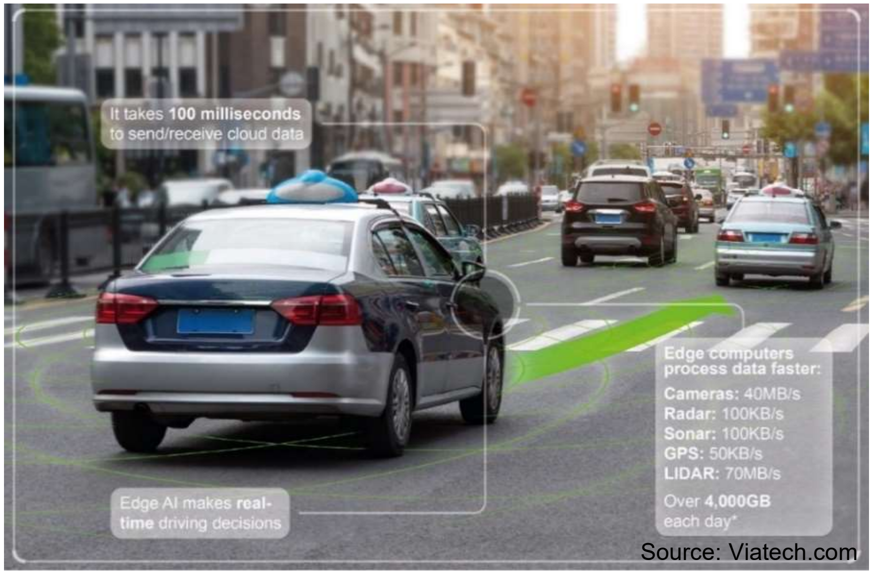
Hyper-connected Intelligence

Major Challenges

- *acquire, process, transport, fuse,…*
  *<u>massive</u> amounts of **data***
  - generated by countless IoT connections

Onerous Constraints & Requirements

- *Resources: energy, network, computational*
- *Security, privacy, sovereignty*
- *Explainability, trustworthiness, fairness*
- *Scalability*



Source: Viatech.com

high-dimensional

multimodal

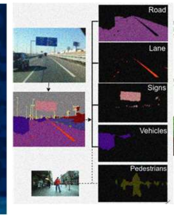
geo-distributed

Let the numbers speak

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







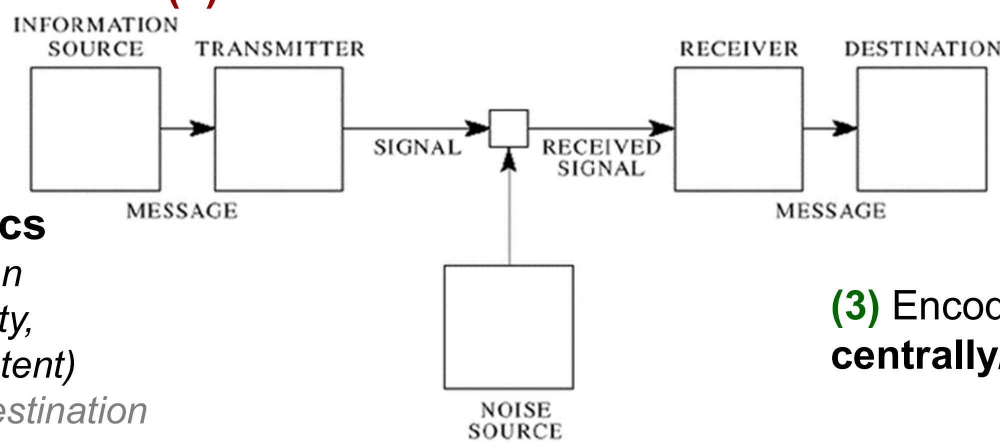
How **to do all that *efficiently*?** **Do we have the right *Theory* & *Algorithms*?**

The Road So Far

From Theory… (ad fontes)

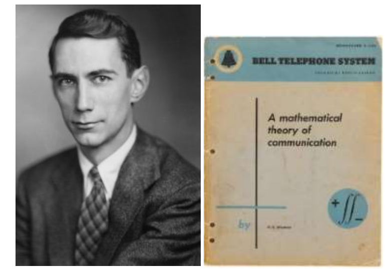
Shannon's model

(1) Reliable transfer of information



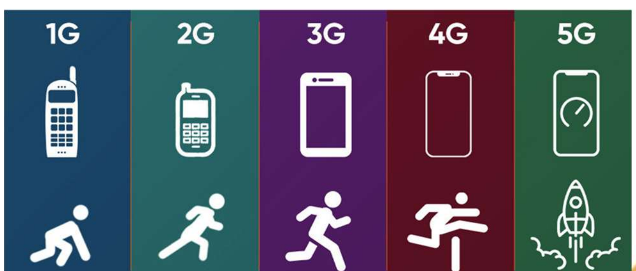
(2) No semantics
dichotomy between information quantity, and meaning (content) and its effect at destination

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

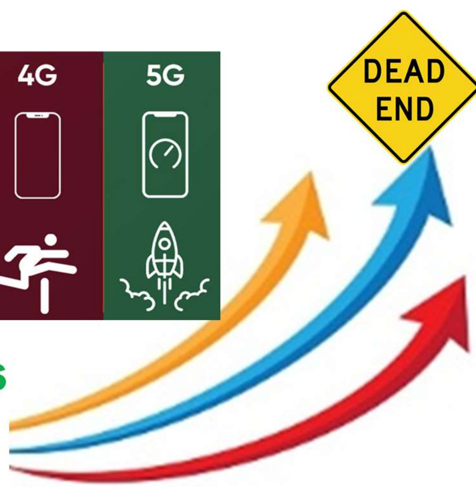


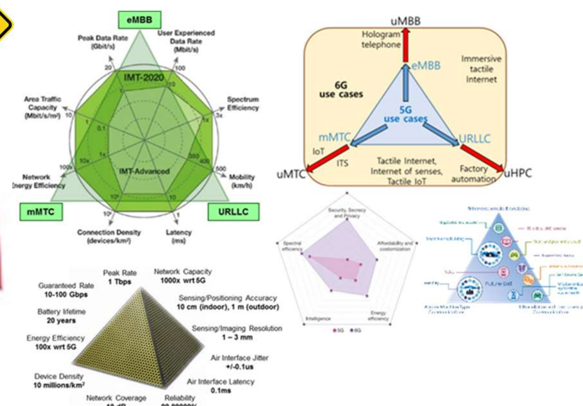
Focus on ***noise*** (& ***equivocation***) rather than ***signal***

… to Practice



CITIUS, ALTIUS, FORTIUS





Communication Systems *Evolution*

- Inflated requirements
- Overprovisioning
- Resource-hungry
- Scalability issues

Maximalistic approach

Λακωνίζειν: Less is More

Shannon's Communication Model (1948)

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

info source → ……01000010101010 → sequence of "random" bits → **Opaque** data pipe → ……0101001X101010

- **Exogenous** traffic
- **Exogenous** data acquisition/generation

Content Agnostic

Targets

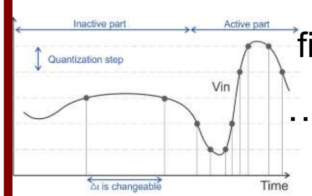
- ***Reliable*** *transfer of information*
- *Reconstruct the entire information flow*
- ***Non-causal*** *signal reconstruction*

Goal-Oriented Semantic Communication Model (202X)

Minimalistic approach

Goal-oriented *unification* of data generation, transport, and reconstruction/usage

Joint design/Non-separable



importance-aware filtering & processing

…~~0100000~~10101010 → 101010~~10~~ → ……01010X10

data valuation

Importance-aware transmission & medium access

- **Active** data acquisition
- **Event** triggered
- Generation ***at will***
- Source/process-aware

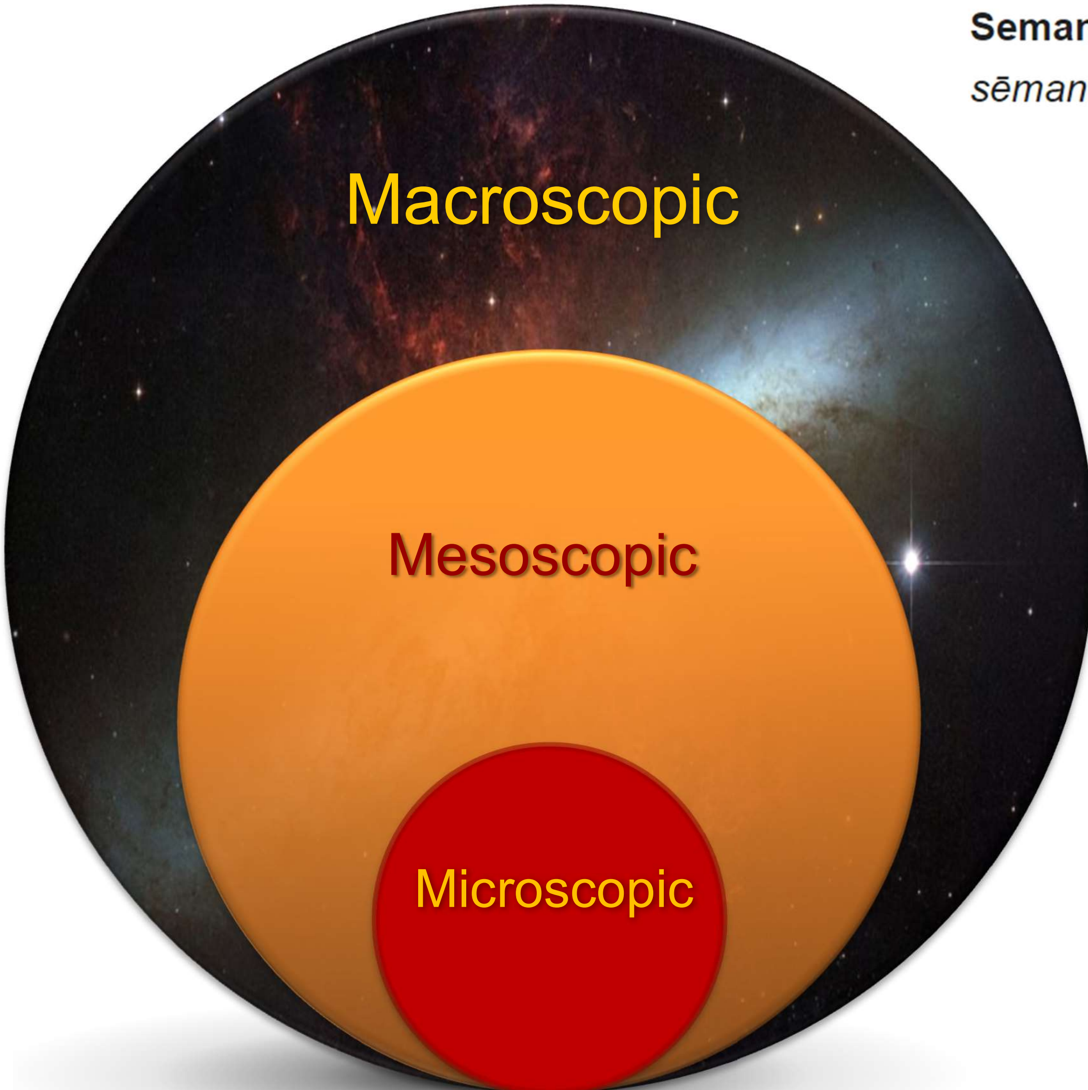
Semantics of Information

significance, usefulness, and effectiveness wrt. goal of data exchange

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

Search for *Semantics of Information*

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



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

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

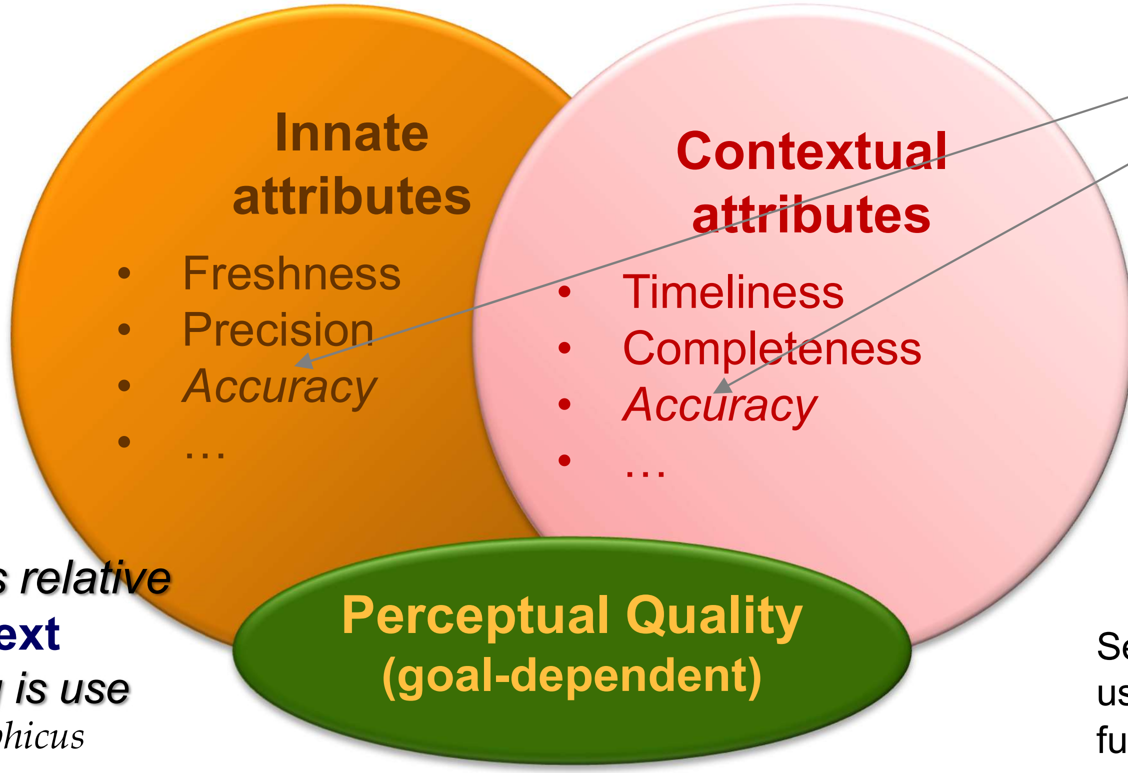
Quantitative and *Qualitative* innate and contextual attributes of information

Relative importance of different outcomes, events, observations

Defining Data Importance & Effectiveness

- 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 information is relative

Meaning is Context

Wittgenstein's *meaning is use*

Tractatus Logico Philosophicus

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

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

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

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

Semantics of Information

Semantics of Information (SoI)

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

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

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

A toy example

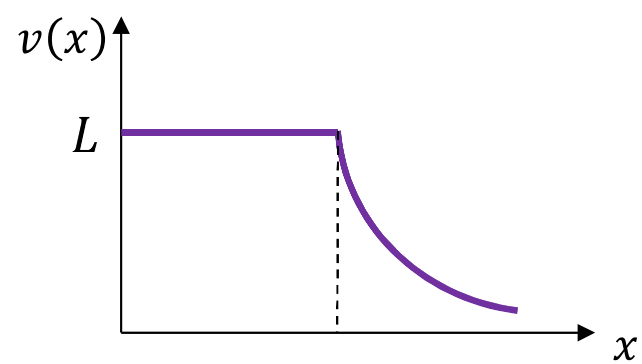
- Information Freshness/Age of Information (AoI): $\Delta_t = t - u_t$

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

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

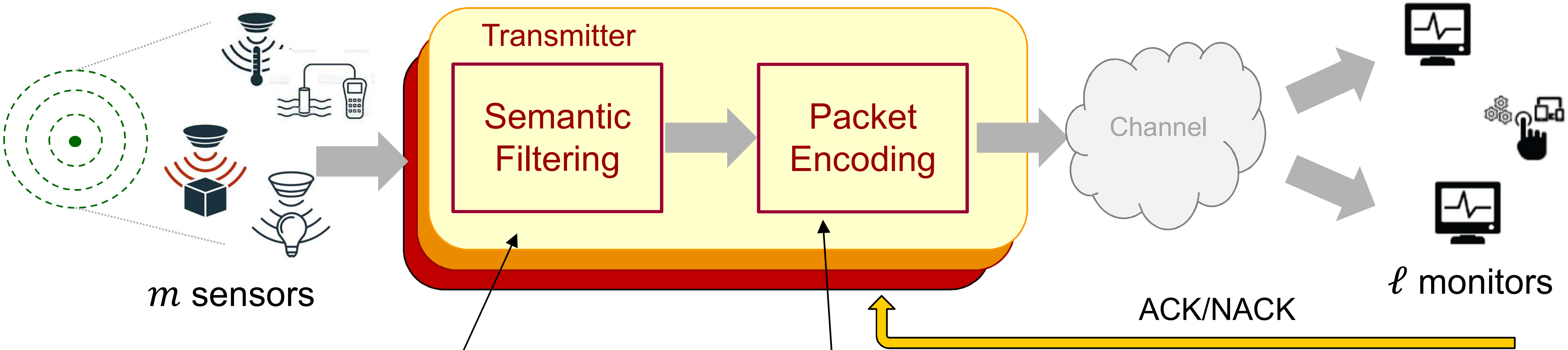
- $\psi(x, y) = Kxy$, so $\psi(\Delta_t, \delta) = K(t - u_t)\big(X_t - \hat{X}_t\big)^2$

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



- Special cases of SoI: AoI (vanilla, nonlinear, AoII,..), VoI, QoI,…

Semantic Source Coding in Multiuser Systems



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

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

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

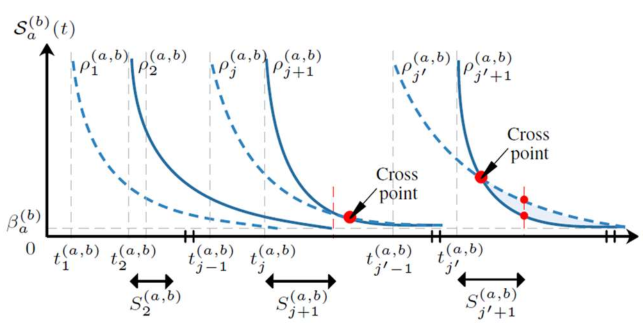
Admitted packets encoded using a prefix-free code based on the truncated distribution with conditional probabilities

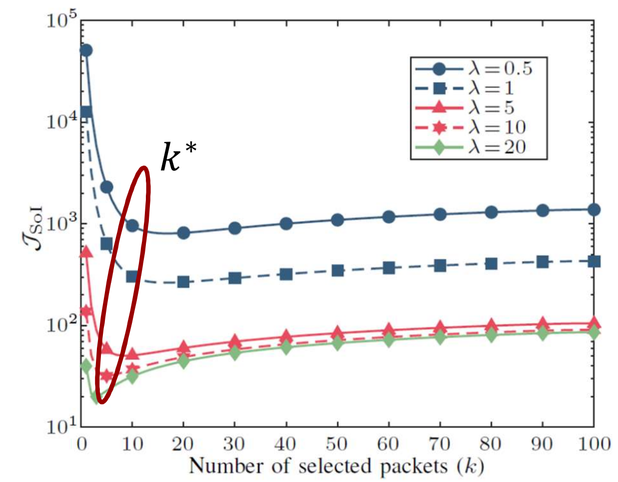
$$p_i = \frac{\tilde{p}_i}{\sum_{i \in \mathcal{I}_k} \tilde{p}_i}, \forall\, i \in \mathcal{I}_k \subset \mathcal{I}$$

Metric: SoI: $\mathcal{S}(t) = g(\Delta(t))$
g:non-increasing function
AoI: $\Delta(t) = t - u(t)$

Importance/Value assessment

- function of prob. occurrence & value of update packet
- Value: feature fusion (e.g., Choquet's discrete integral)
- "irrelevant" realizations censored





erc UNIVERSIDAD DE GRANADA EURECOM Sophia Antipolis

Interactive Communications

Feedback/Response X

Sensing Agent

$\mathcal{D}$ **data**

- Data $d_N \in \mathcal{D}$
- Prior knowledge: $\mathbb{P}(\mathcal{G} = g)$
- User behavior: $\mathbb{P}(Y = y | \mathcal{G} = g, X = x)$
- Bob queries/probes based on $\hat{\mathcal{D}}$
- Posterior knowledge: $\mathbb{P}(\mathcal{G} = g | X = x, Y = y)$

Partial Observability:

- Bob queries/probes based on $\hat{\mathcal{D}}$
- Alice responds based on $\hat{\mathcal{G}}$

End User
Observer

$\mathcal{G}$ **goal**

Input/Query/Command/Request Y

Alice's prior knowledge: $H(\mathcal{G})$

Alice's posterior/current knowledge: $H(\mathcal{G}|Y = y, X = x)$

Semantic Queries: high $H(\mathcal{G}) - H(\mathcal{G}|Y = y, X = x)$

Mutual Information

$I(\mathcal{G}; Y | X = x)$

what Bob wants and what he sends after seeing $X = x$

Bayesian Semantic Communication Model

Feedback/Response X

Bayesian Agent

Alice

$\mathcal{D}$ **data**

- Data $d_N \in \mathcal{D}$
- Prior knowledge: $\mathbb{P}(\mathcal{G} = g)$
- User behavior: $\mathbb{P}(Y = y|\mathcal{G} = g, X = x)$
- Bob queries/probes based on $\hat{\mathcal{D}}$
- Posterior knowledge: $\mathbb{P}(\mathcal{G} = g|X = x, Y = y)$

Partial Observability:
- Bob queries/probes based on $\hat{\mathcal{D}}$
- Alice responds based on $\hat{\mathcal{G}}$

Bayesian Agent

End User
Observer

Bob

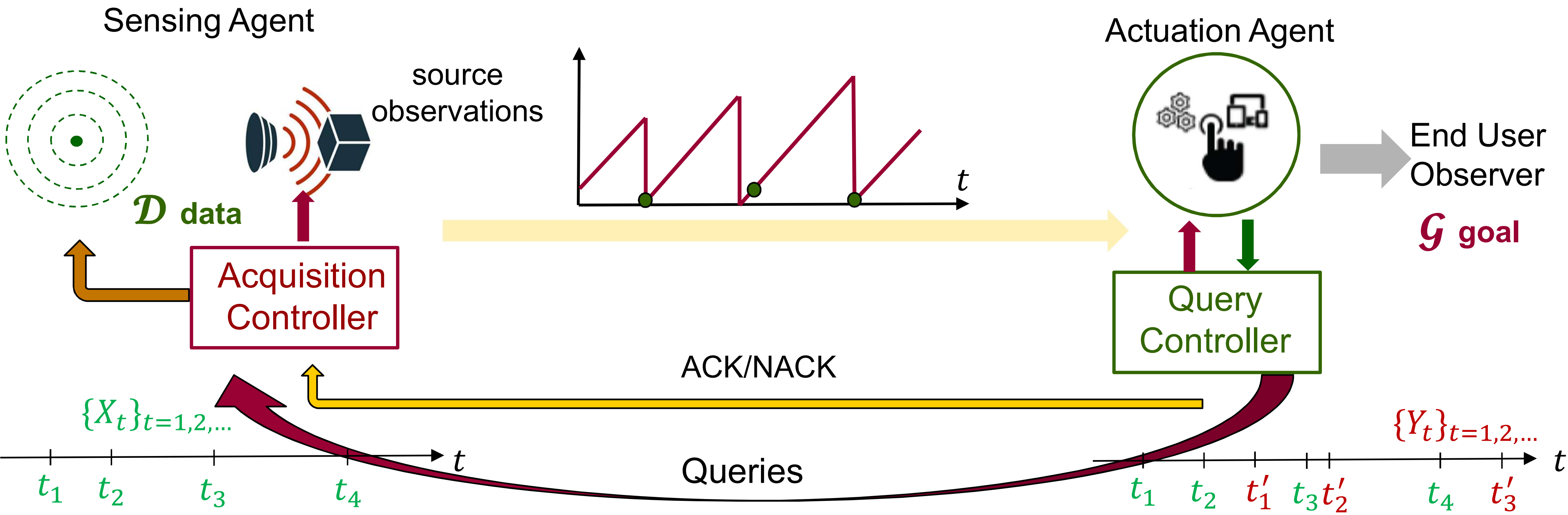
$\mathcal{G}$ **goal**

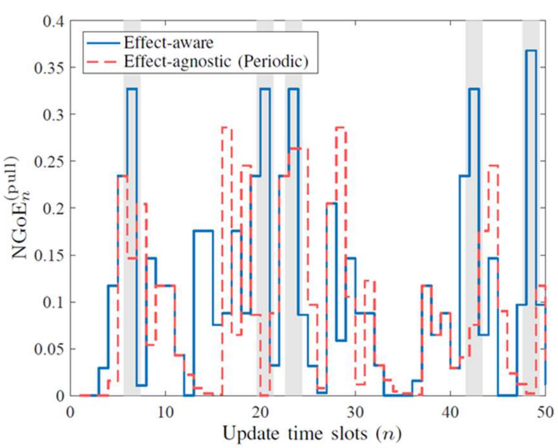
Input/Query/Command/Request Y

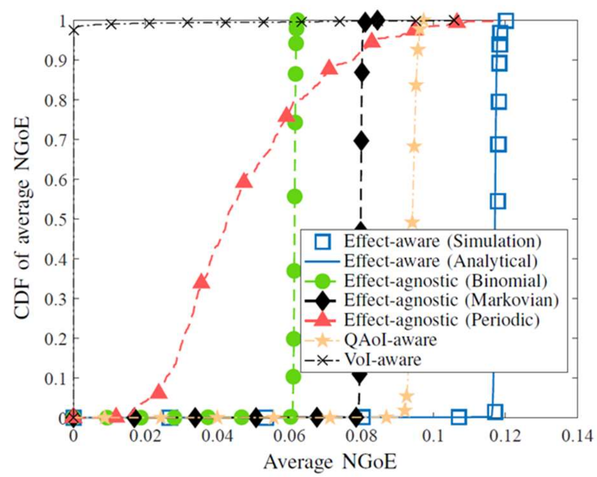
- Agent: belief $p_{\mathcal{G}}(x|\mathcal{G})$ (set of distributions) and prior $p_{\mathcal{G}}(\mathcal{G})$
- Finite data $d_N = \{x_1, x_2, \ldots, x_N\} \in \mathcal{D}$
- Bayesian entropy: $H_{\mathcal{G}}(X|d_N) = \sum_{x \in \mathcal{X}} p(x) \log p_{\mathcal{G}}(x|d_N)$
- Bayesian MI: $I_{\mathcal{G}} = H_{\mathcal{G}}(X|d_N) - H_{\mathcal{G}}(X|Y, d_N)$

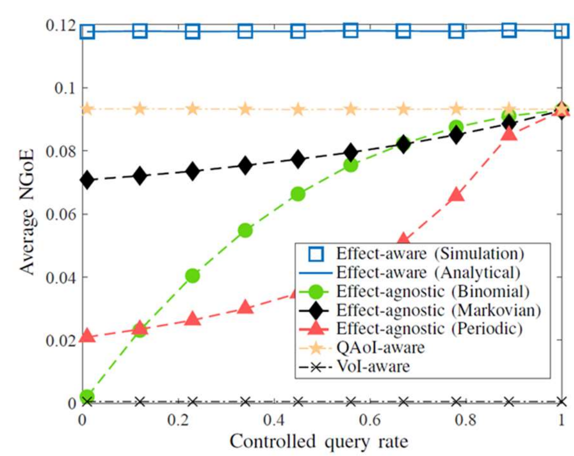
- No Data-Processing Inequality
  - *data can add information*
  - *processing can help*
- Information not always beneficial

Effective Pull-based Communication









Goal-*agnostic* Information Transmission

Rate R

Encoder f → Decoder g

$X \sim p_X$

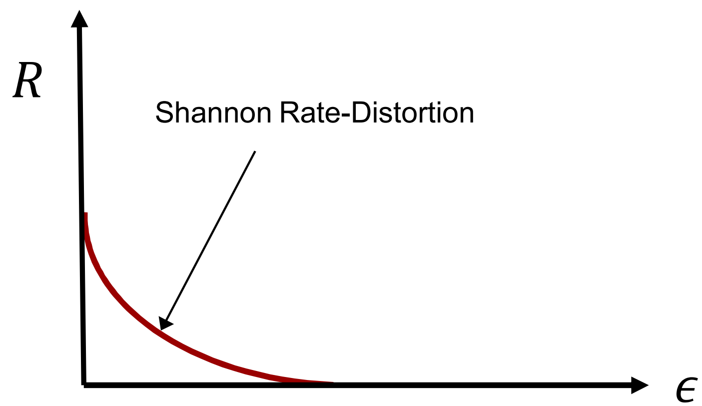
Distortion $\mathbb{E}_{p_{X,\hat{X}}}\left[\delta(X,\hat{X})\right]$

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

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

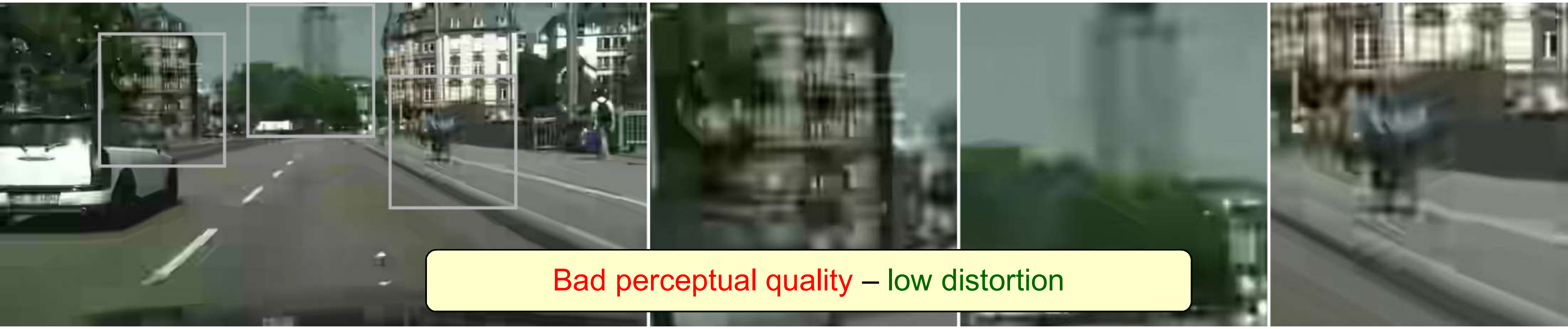
$$R(\epsilon) = \min_{p_{\hat{X}|X}} I(X,\hat{X})$$

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



Semantic Quality

Good perceptual quality – high distortion

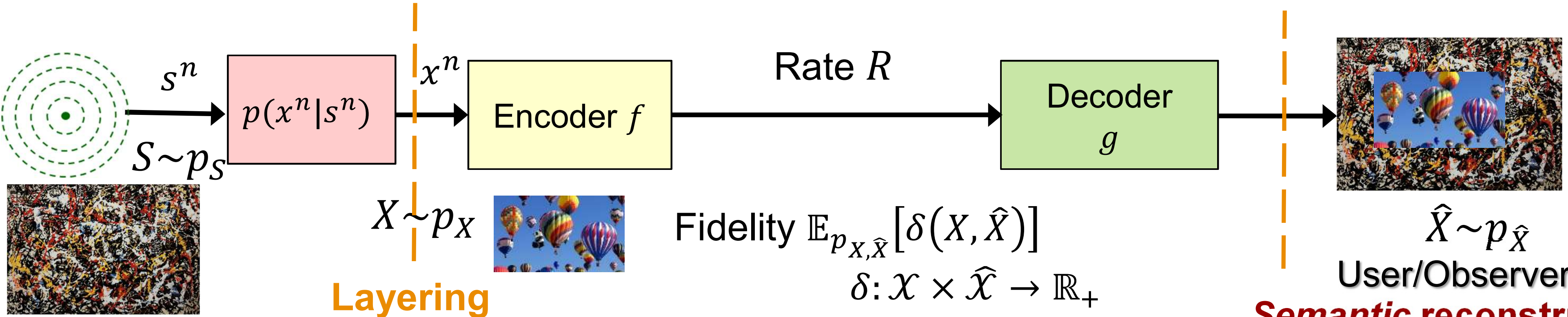


Bad perceptual quality – low distortion

Good perceptual quality ≠ low distortion

Agustsson et al. (2018)

Goal-oriented Information Handling



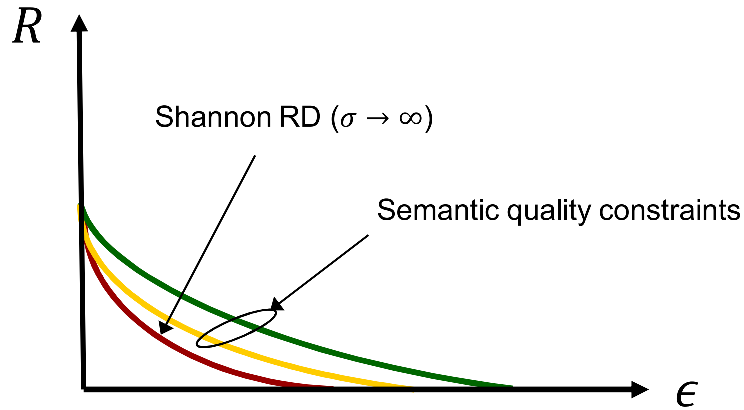
$$R(\epsilon,\sigma) = \min_{p_{\hat{X}|X}} I(X,\hat{X})$$

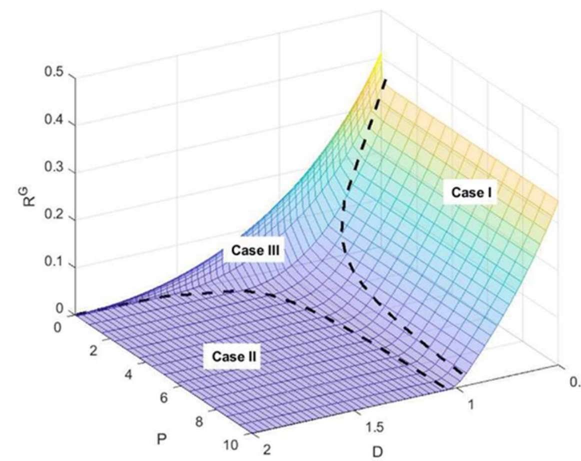
s.t. $\mathbb{E}_{p_{X,\hat{X}}}[\delta(X,\hat{X})] \leq \epsilon$

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

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

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





$R(D,P)$ for $X \sim \mathcal{N}(0,1)$ under direct KL div.

Semantic quality metrics

- divergence (Wasserstein, f-div, α-div, …)
- generalized entropic measures

Fidelity criteria (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

Alpha Divergence in Rate-Distortion-Perception Theory

Rate-Distortion-Perception (RDP)

$$R(D,P) \triangleq \inf_{p_{\hat{X}|X}} I(X,\hat{X})$$
$$\text{s.t.} \quad \mathbb{E}\left[d(X,\hat{X})\right] \le D$$
$$D(p_X \| p_{\hat{X}}) \le P$$

$X \sim \mathcal{N}(\mu, \sigma^2)$

jointly Gaussian

$\hat{X} \sim \mathcal{N}(\nu, \rho^2)$

$$R(D,P) \le R^G(D,P)$$

Gaussian RDP

$$R^G(D,P) = \min_{p_{\hat{X}|X},\nu,\rho^2} I(X,\hat{X})$$
$$\text{s.t.} \quad \mathbb{E}\left[(X-\hat{X})^2\right] \le D,$$
$$D_\alpha(p\|q) \le P.$$

Parametric jointly Gaussian RDP solution

$$R^G(D,P) = \begin{cases} \max\left\{\frac{1}{2}\log\frac{\sigma^2}{D}, 0\right\} & \textit{if } (D,P) \in \mathcal{S} \\ \frac{1}{2}\log\frac{2\rho^2\sigma^2}{\rho^2\sigma^2 - \left(\frac{\sigma^2+\rho^2-D}{2}\right)^2} & \textit{if } (D,P) \notin \mathcal{S} \end{cases}$$

where

$$\rho^2 = \begin{cases} \sigma^2 - D & \textit{if } (D,P) \in \mathcal{S} \\ \min\{r_0, r_1\} & \textit{if } (D,P) \notin \mathcal{S}. \end{cases}$$

$$\mathcal{S} = \Big\{(D,P) \in \mathbb{R}_+{}^2 : P > g(D,\sigma) \quad \wedge$$
$$(\alpha-1)\left(\left|1-\frac{D}{\sigma^2}\right| - \left(1-\frac{1}{\alpha}\right)\right) > 0\Big\}$$

$$g(D,\sigma) = \frac{1}{a(1-a)}\left(1 - \frac{\sigma^{1-\alpha}|\sigma^2-D|^{\alpha/2}}{\sqrt{\alpha|\sigma^2-D| + (1-\alpha)\sigma^2}}\right)$$

with r_0 and r_1 being the roots of $f(x) = x^\alpha - \alpha C x - (1-\alpha)C$

α- Divergence [Chernoff52, Amari82]

unique canonical divergences at the intersection of the f-divergences and Bregman divergences in a manifold of positive measures

$$D_\alpha(p\|q) = \frac{1}{\alpha(\alpha-1)}\left(\int_{-\infty}^{\infty} p(x)^\alpha q(x)^{1-\alpha}\,dx - 1\right)$$

$$\alpha \in \mathbb{R}\backslash\{0,1\}.$$

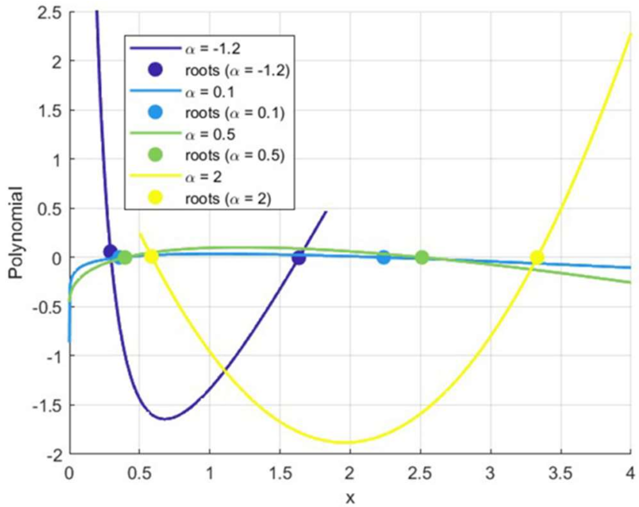
For two Gaussians:

$$D_\alpha(p\|q) = \frac{1}{\alpha(1-\alpha)}\left(1 - H_\alpha(p,q)\right)$$

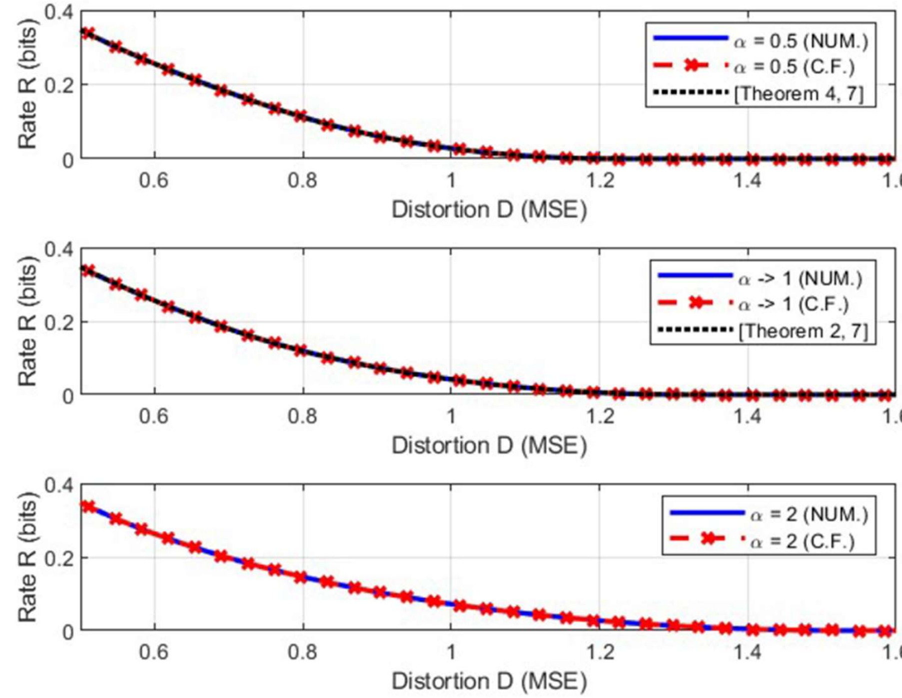
$$H_\alpha(p,q) = \frac{\rho^\alpha \sigma^{1-\alpha}}{\sqrt{\alpha\rho^2 + (1-\alpha)\sigma^2}} e^{-\frac{\alpha(1-\alpha)(\mu-\nu)^2}{2(\alpha\rho^2+(1-\alpha)\sigma^2)}}$$

$x = \frac{\rho^2}{\sigma^2}$ and $C = (1-\alpha(1-\alpha)P)^2$

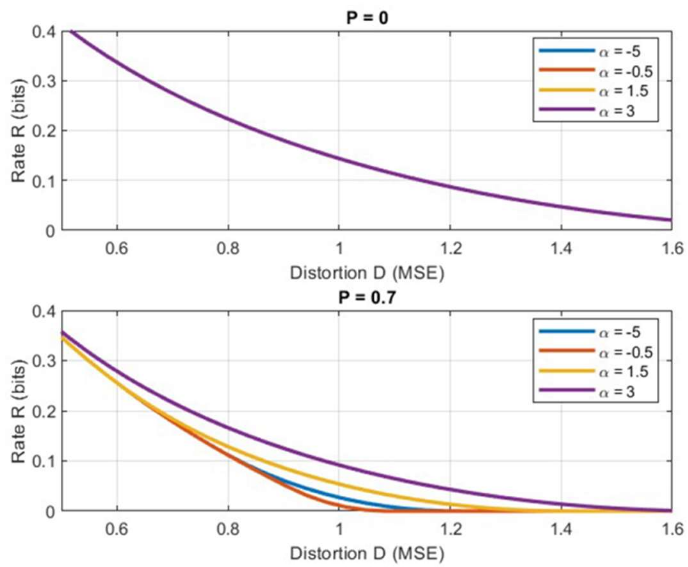
Alpha Divergence in Rate-Distortion-Perception Theory



Polynomial for $P = 0.2$



RD curves for fixed $P = 0.5$



RD curves for diff. α and P

Special cases

Hellinger div. ($\alpha = 0.5$)
KL div. ($\alpha \to 1$)
Reverse KL div. ($\alpha \to 0$)
Pearson div. ($\alpha = 2$)

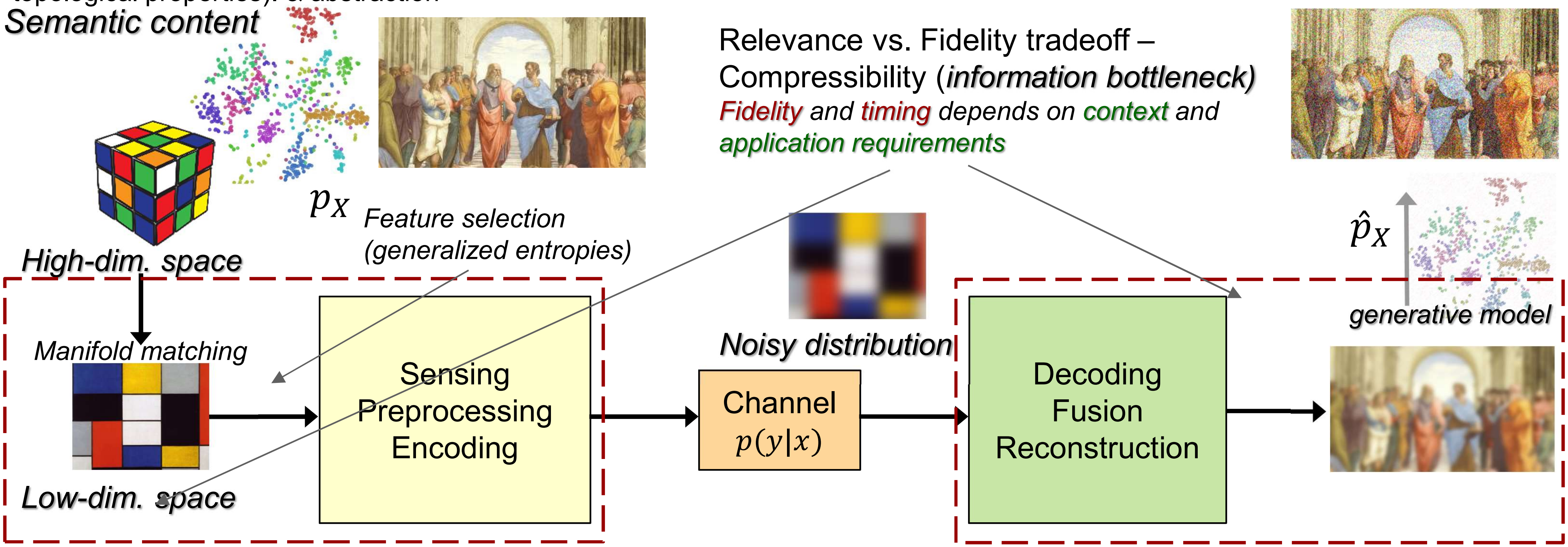
Pearson divergence ($\alpha = 2$)

$$\rho^4 - 2(1+2P)^2\rho^2\sigma^2 + (1+2P)^2\sigma^4 = 0$$

$$\rho^2 = \begin{cases} \sigma^2(1+2P)\left(1+2P+2\sqrt{P+P^2}\right) \\ \quad \text{if } \rho^2 - (1+2P)^2\sigma^2 > 0, \\ \sigma^2(1+2P)\left(1+2P-2\sqrt{P+P^2}\right) \\ \quad \text{if } \rho^2 - (1+2P)^2\sigma^2 \leq 0. \end{cases}$$

Exiting Plato's Cave

Complex data (feature richness, structural and topological properties). & abstraction



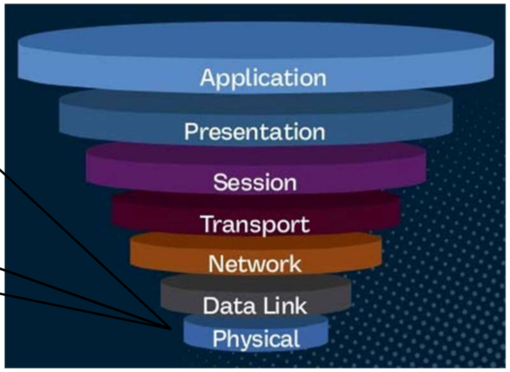
communication problem

optimal transport problem
& foundational learning problem

- Communicating *high-dimensional, multi-modal, multi-source rich* data
- Intriguing connections with optimal transport, generative models, decision-making, inference…
- Information Manifold: rate distortion perception manifolds for semantic information spaces

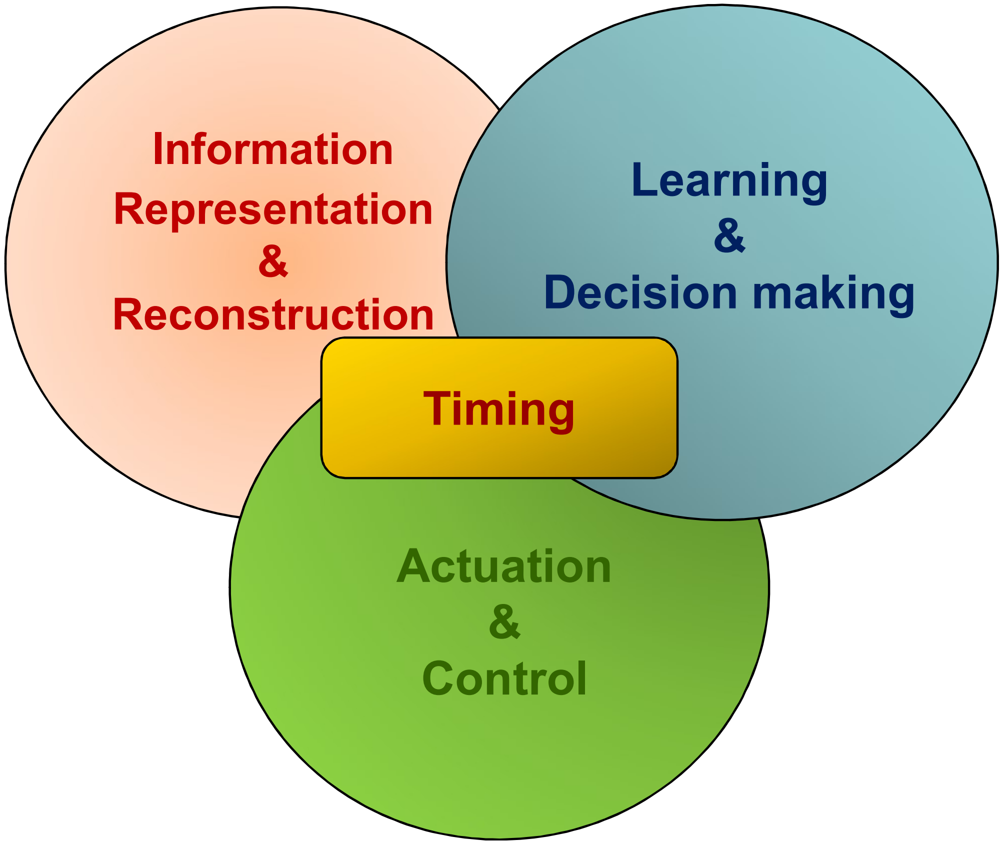
Plethora of Challenges

- Abandon *epistemology* and *doxastic logic* in the interest of Science
- Come up with a clear, crisp definition of information semantics/effectiveness
- Universal metrics vs. Subjectivity
- Formalize the notion of “subjectivity” (perception, context, …)
- Have a calculus for characterizing “goals” and “requirements”
- Tackle multi-source/multi-agent problems (multivariate information theory)
- Learn how to communicate/optimally transport rich data and distributions
- Risk-averse decision-making and *prospect*-based RRM
- Develop relevant AI/ML and Foundation Models for SemCom
- Reconciliate IT models with time (incl. (d)JSCC)
- See the Big Picture!



Redefining Effectiveness and Timing

- **Context** in ComSys: presupposed physical/comm dimensions (time, location, role)
- Background knowledge & side info is <u>key</u>



Wittgenstein's ***meaning is use***

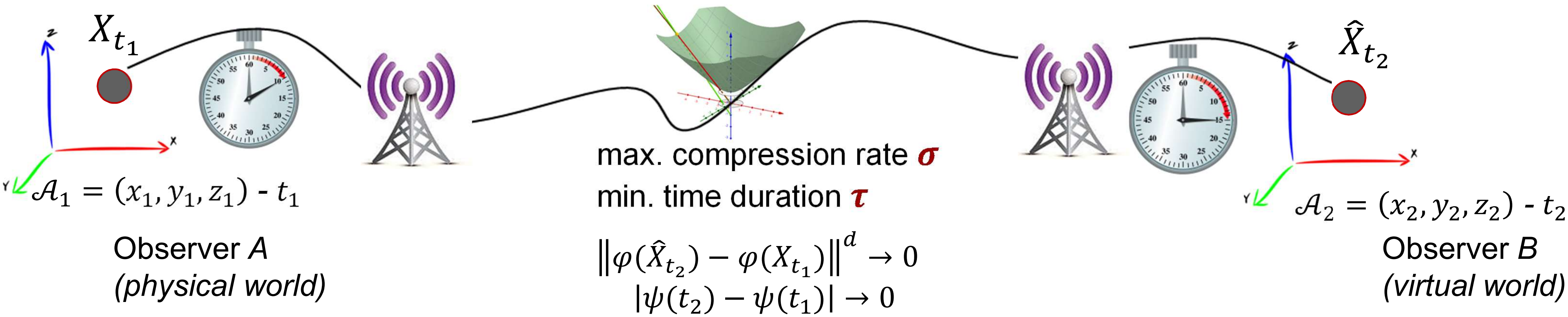
Tractatus Logico Philosophicus

Meaning is Context

- **Effectiveness**: measured wrt. to the goal/use of the data exchange (@observer side)
- *Semantic information is relative*
- **Timing** is related to effectiveness in different communication scenarios

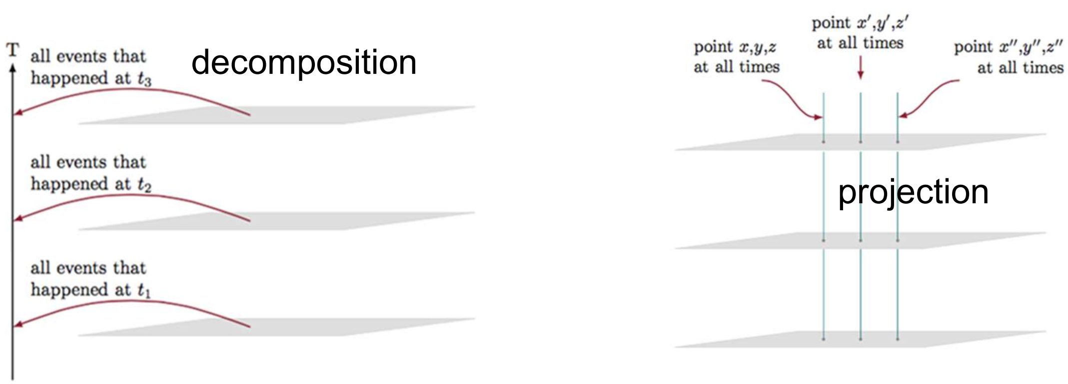
A Mathematical Theory of Timing

"Relativistic" & Relational Information Transfer



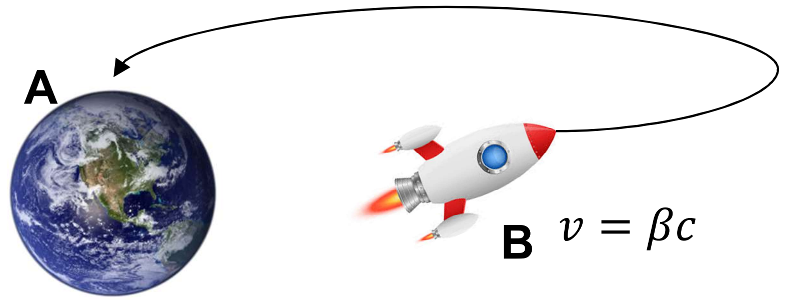
Frame of reference A

Frame of reference B



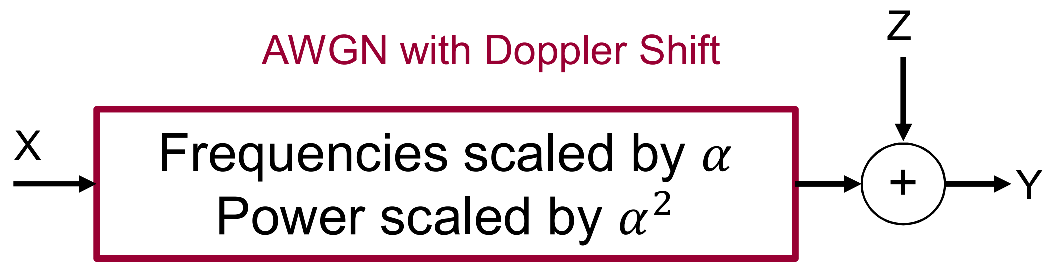
- "Twin paradox" analogy: ***frame*** dependent system state and evolution
- Correct system "timing dilation" & "distortion/state error" - Event invariance
- Redefining timing, synchronicity, simultaneity in comm. systems

Relativistic Information Transmission



A has aged by a factor of $\delta = (1-\beta^2)^{-1/2}$ more than **B**

AWGN with Doppler Shift



RX signal power $P' = \alpha^2 P$
RX signal bandwidth $W' = \alpha W$
RX signal rate $R' = \alpha R$

Doppler factor $\alpha = (1+\beta)\delta$

Cost of Asymmetry in Twin Problem
[Jarett&Cover'81]

For a given TX rate and bandwidth

- Traveler needs δ times the energy of the stationary spaceship to transmit $1/\delta$ times as much information
- Asymmetry in efficiency is thus δ^2
- Independent of acceleration and gravitational fields
- Shown for special cases, conjecture for arbitrary trajectory

If the RX sees the TX moving

Max. RX rate: $C' = W' \log\left(1 + \frac{P'}{NW'}\right)$

$$= \alpha W \log\left(1 + \frac{\alpha P}{NW}\right) \text{ bits/s}$$

Max. TX rate: $C = C'/\alpha$

Info Theoretic Analog of Twin Paradox

Asymmetry in Twin Paradox

- Max. # bits/s that **A** can transmit reliably to **B** > the one **B** can transmit to **A**
- [Jarett&Cover'81]: proved for the special cases of
  - purely circular (**B** moving on a circular orbit around **A**)
  - purely radial (**B** moving away from **A** along a straight line, and coming back the same way)

  constant-speed motion
- **We show that this is true for an arbitrary trajectory**

$$\frac{E_A/N_A}{E_B/N_B} = \frac{(\overline{P}_A T_A)/(\overline{C}_A T_A)}{(\overline{P}_B T_B)/(\overline{C}_B T_B)} = \frac{\overline{C}_B}{\overline{C}_A}$$

$$\overline{C}_A > \overline{C}_B \quad \Rightarrow \quad E_A/N_A < E_B/N_B$$

Key Inequality for Proof

Let $f : [0,1] \to \mathbb{R}$ *be a function satisfying* $|f(x)| \leqslant b < 1$ *and* $\int_0^1 f(x)dx = 0$

$$D_{\mathrm{KL}}(1 - f \,\|\, 1) - D_{\mathrm{KL}}(1 \,\|\, 1 + f) > \log(1 - b^3),$$

$$b := \max_{\tau} |\beta_r(\tau)| \;, \int_0^{T_B} \beta_r(\tau)\, d\tau = 0$$

Epilogue

- To support connected intelligence and autonomous (real-time) systems in future wireless networks
  - fundamental theoretical advances
  - augmenting prevailing communication design paradigms

- Goal-oriented Semantic Communications: a paradigm that redefines *importance* and *timing* in communication systems

- Taming "*subjectivity*" and achieving "*universality*" may pass through timing/time aspects

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

  } *scalability*

European Research Council
Established by the European Commission

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