EURECOM
Sophia Antipolis

# MALCOM
## Machine Learning for Communication Systems

# Lecture 5
# Information Theoretic Foundations
# *Interlude*

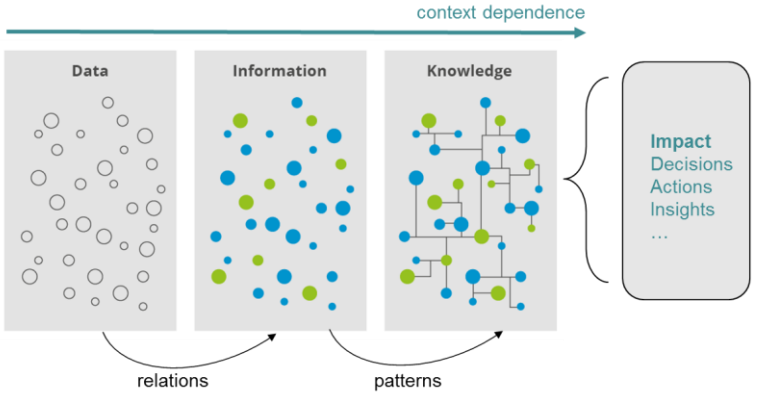
Slides: Marios Kountouris
Lecturer: Ayşe Ünsal

Spring 2024

# What is Information?

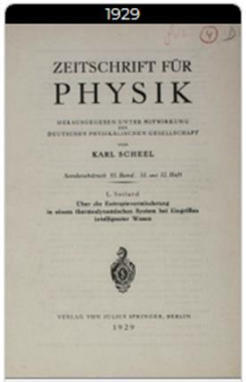
- How hard is it to estimate/discriminate distributions? ⟶ Fisher Information
  - Cramér-Rao lower bound
- How hard is it to compress data? ⟶ Shannon Information
  - Entropy- A lower bound on the number of binary digits for encoding our message
- How hard is it to classify *empirical* data?
  - (A tight lower bound on) Chernoff Information

- Relation to ***data*** and ***knowledge***

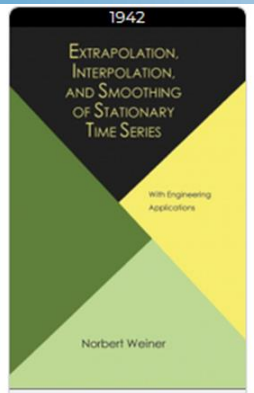


# Précis of Information


Hartley's Law



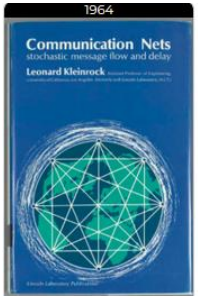

Szilard Models the Relationship between Information and Thermodynamics

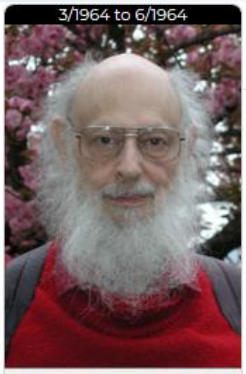



Norbert Weiner Formulates Communication Theory as a Statistical Problem:


Shannon's "A Mathematical Theory of Communication"

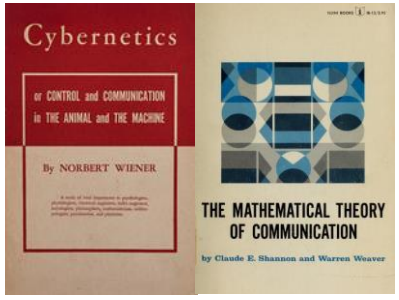

Leonard Kleinrock Writes the First Paper on Data Networking Theory




Solomonoff Begins Algorithmic Information Theory

Jeremy Norman's HistoryofInformation.com
Exploring the History of Information and




**Fisher Information** (1927)

Amount of information RV $X$ carries about $\theta$

$$\mathcal{I}(\theta) = \mathbb{E}\left[\left(\frac{\partial}{\partial\theta}\log f(X;\theta)\right)^2 |\theta\right]$$

**Algorithmic Information**

(Kolmogorov, Chaitin, Solomonoff)

Complexity is the size of the smallest program $p$ on a universal Turing machine $U$ that generates object $x$

$$K(x) = \min\{|p|: U(p) = x\}$$

| | |
|---|---|
| N. Wiener | F. Dretske |
| L. Brillouin | B. Skyrms |
| D. Gabor | C.F. von Weizsäcker |
| D. MacKay | J. Barwise |
| F. Dretske | D. Nauta |
| … | … |

Shannon entropy

**Shannon Information** (1948)

measured by the statistical uncertainty reduction upon receipt of a message

$$H(X) = -\sum_{x\in\mathcal{X}} P(x)\log P(x)$$

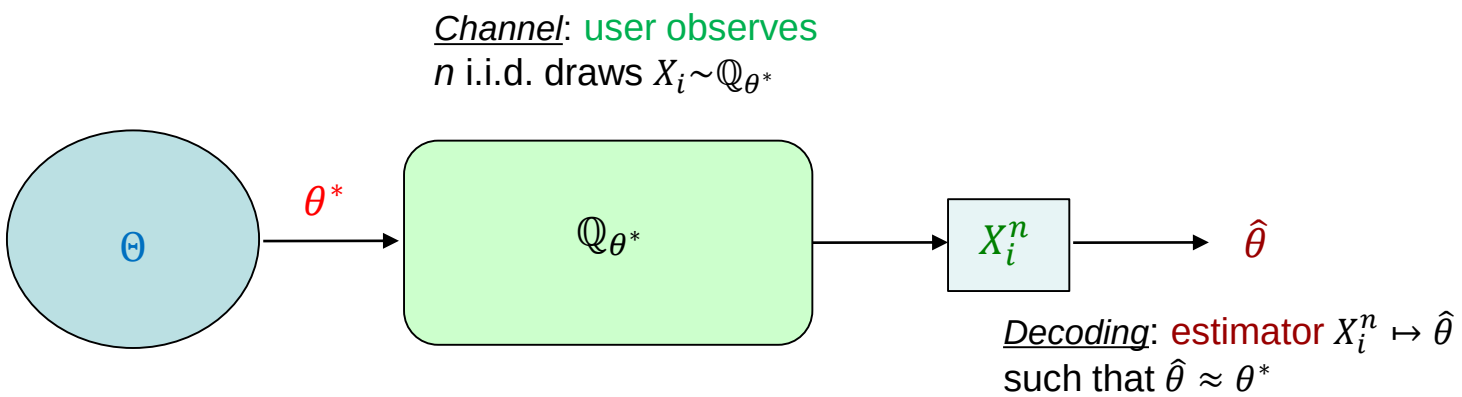


EURECOM Sophia Antipolis

# Information Theory, Statistics, and ML

**A rich intersection between information theory and statistics**

- Hypothesis testing, large deviations
- Fisher information, Kullback-Leibler divergence
- Metric entropy and Fano's inequality




Perspective dating back to Kolmogorov (1950s) with many variations:

- Codebooks/codewords: graphs, vectors, matrices, functions, densities....
- Channels: random graphs, regression models, elementwise probes of vectors/machines, random projections
- Closeness $\hat{\theta} \approx \theta^*$: exact/partial graph recovery in Hamming, $\ell_p$-distances, $L(\mathbb{Q})$ -distance

# Entropy and Mutual Information

- Entropy: a measure of the amount of uncertainty of an unknown/random quantity

$$H(X) = -\sum_x P(x) \log_2 P(x)$$ (to be explained later!)

Problem: You are invited to dinner at your friend's for the first time, but you do not know his apartment number. There are 8 apartments on each of the 4 floors, each equally likely. A neighbor you ran into (does not know exactly either) but tells you which floor certainly your friend lives.

- The *uncertainty* about his door number drops down to 8 choices (before it was 32 choices)
- Your friend's neighbor *conveyed* you 2 bits of information.

- Q: How do we quantify the amount of information? How do we measure it?

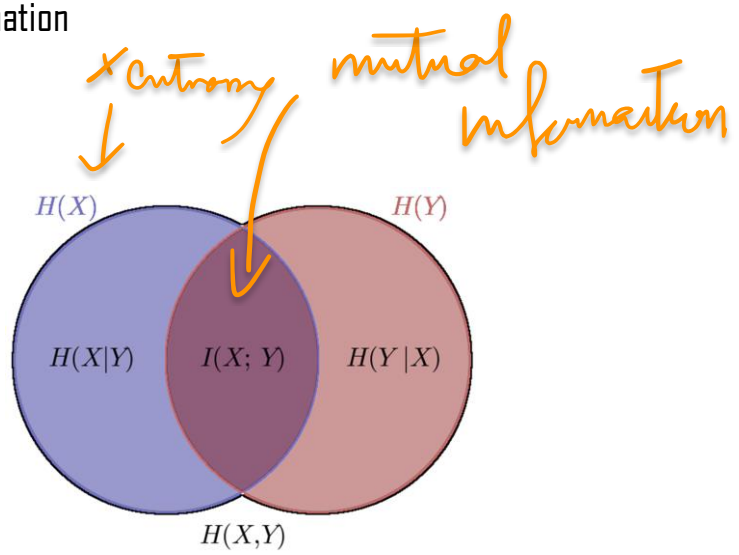
  A: Using information theory, we can quantify, manipulate and represent uncertainty.

# Entropy and Mutual Information

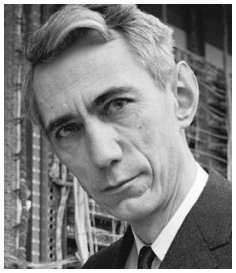
- Entropy is measured in bits.
- Example cont'd: Let $X$ be a random variable defined as the apartment where your friend lives, it can take up 32 values with equal probability, i.e. $p(x) = 1/32$ and $\log_2(1/32) = -5$, hence the entropy of $X$ is 5 bits.
- After the neighbor tells you which floor your friend lives, the probability of $X$ equals 0 for 24 out of 32 choices, and becomes 1/8 for other equally probable values. Entropy of $X$ becomes 3 bits.
- The neighbor conveyed you ***2 bits of information*** ⟶ Mutual Information

$$I(X;Y) = H(Y) - H(Y|X)$$
$$= H(X) - H(X|Y)$$

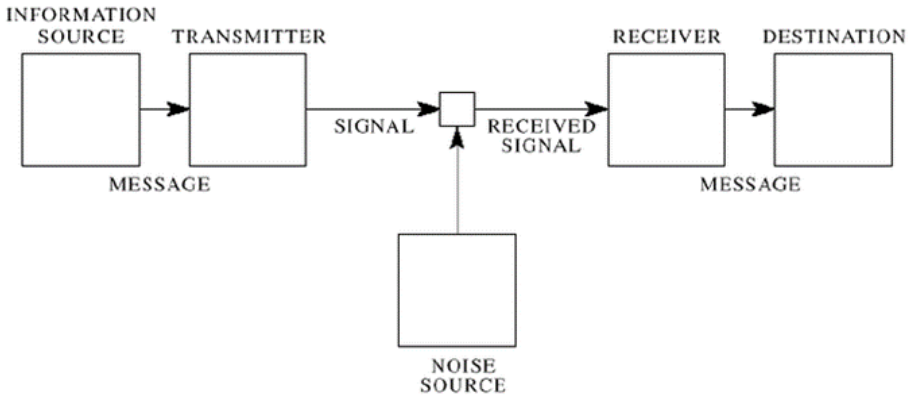
- $I(X;Y)$ is the intersection of information in $X$ with information in $Y$



# Information and Value



- Information as *uncertainty* and *surprise*
- *Information gain* – measuring the statistical uncertainty reduction upon receipt of a message
- Operational meaning to *entropy* $H(X) = -\sum_{x \in \mathcal{X}} P(x) \log P(x)$




*Decision-making information*

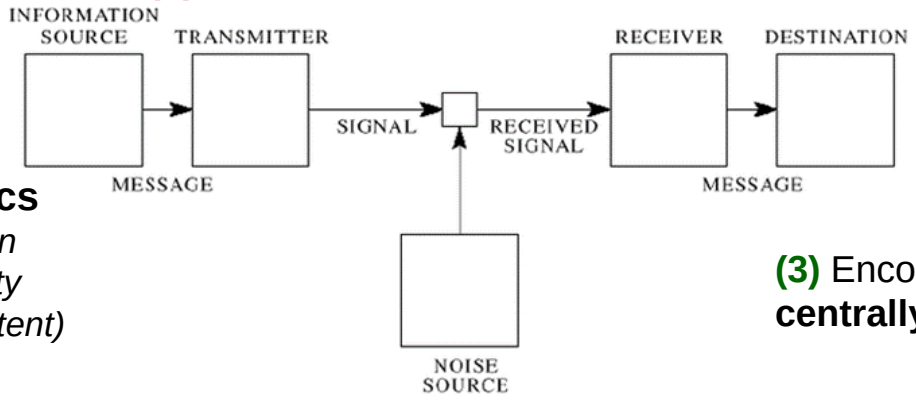
- Used by agents, observers, and info generators to make improved decisions (incl. better inference, predictions, etc.)
- Enable to select a course of action (or inaction) given all (known) possible choices of action
- Value, utility, importance, goals - *need for observer/receiver*

# The Communication Problem

1 The fundamental problem of communication is that of reproducing at one point either exactly or approximately a message selected at another point. 2 Frequently the messages have *meaning*; that is they refer to or are correlated according to some system with certain physical or conceptual entities. These semantic aspects of communication are irrelevant to the engineering problem. 3 The significant aspect is that the actual message is one *selected from a set* of possible messages.

*C. Shannon. A mathematical theory of communication. Bell System Technical Journal, 27:379-423, 625-56, 1948.*

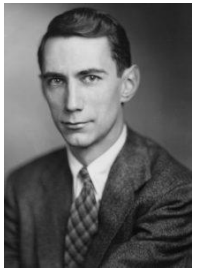
**(1) Reliable** transfer of information

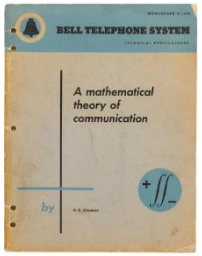



**Shannon's model**

**(2) No semantics**
*dichotomy between information quantity and meaning (content)*

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



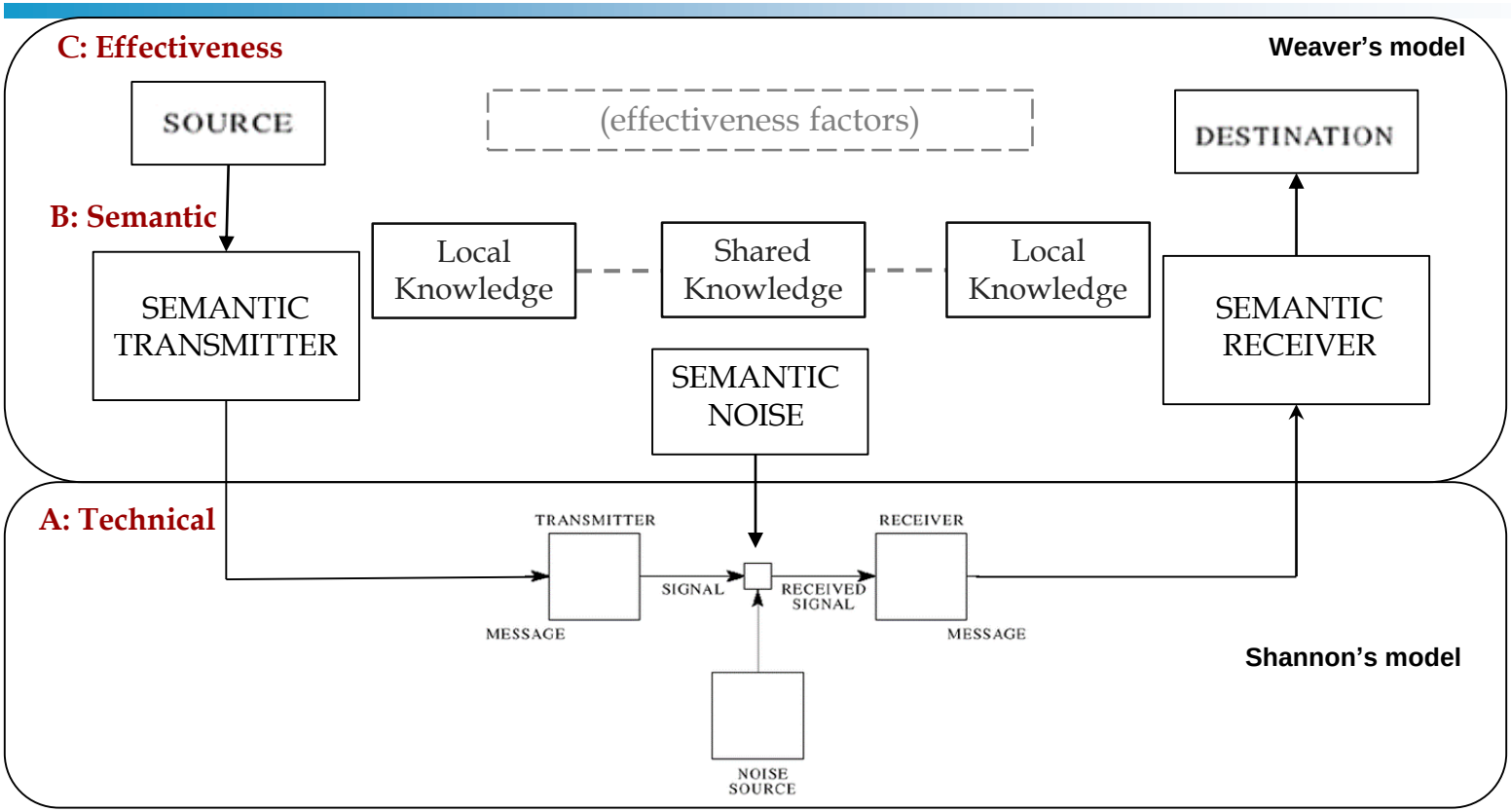


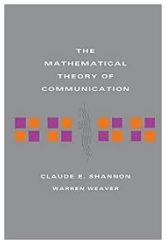

- **Lore**: theory of information quantity, but not of information content
- Focus on "*noise*" (and *"equivocation"*) rather than "*signal*"
- Suitable and instrumental for classical human-generated data communication

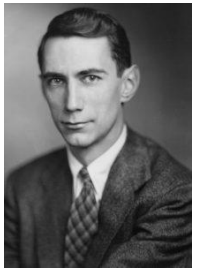


EURECOM Sophia Antipolis

# The General Problem



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

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

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

*Syntactic level*

*Semantic level*

*Pragmatic level*

# Measuring Information

- *"Information is the resolution of uncertainty"* (Shannon)
- Shannon's (and Hartley's) definition of the information obtained on being told the outcome $x_i$ of a probabilistic experiment $X$.

$$I(X = x_i) = \log\left(\frac{1}{P_X(x_i)}\right)$$

where $P_X(x_i)$ is the probability of the event $X = x_i$.

- The unit of measurement (when the log is base-2) is the **bit** (**b**inary **i**nformation **u**nit – not the same as binary digit!).

- 1 bit of information corresponds to $P_X(x_i) = 0.5$ i.e HorT (coin)
- So, for example when the outcome of a fair coin toss is revealed to us, we have received 1 bit of information

Example

- We're drawing cards at random from a standard $N = 52$-card deck.
- Elementary outcome: card that's drawn, probability $1/52$, information $\log_2(52/1) = 5.7$ bits.
  Q: If I tell you the card is a 7, how much info?
  A: $N = 52, M = 4$, so info = $\log_2\left(\frac{52}{4}\right) = \log_2(13) = 3.7$ bits

# Properties of Information Definition

- A lower probability outcome yields higher information (bigger “surprise”)
- A highly informative outcome does *not* necessarily mean a more valuable outcome, only a more surprising outcome, i.e., no intrinsic value being assigned
  - can think of information as degree of surprise
- The information in independent events is additive
- Though independence is sufficient for additivity, it is not necessary, because we can have

$$P(ABC) = P(A)P(B)P(C)$$

  even when $A, B, C$, are not independent
- Independence requires the pairwise joint probabilities to also factor

# Entropy

- $X$: discrete random variable // Alphabet: $\mathcal{X} = \{x_1, x_2, \dots, x_N\}$ (of size $N$)
- For discrete random variables, we will write (interchangeably) $P(X = x)$, $P_X(x)$ or $p(x)$

- *Information content, self-information, surprisal*, or *Shannon information*:

$$I(X = x_i) \triangleq \log\left(\frac{1}{p(x_i)}\right)$$

## Entropy

- Entropy of a RV: average level of "information", "surprise", or "uncertainty" inherent to the variable's possible outcomes.

- The entropy of a discrete RV $X$ taking values in alphabet $\mathcal{X}$ is given by:

$$H(X) \triangleq \mathbb{E}[I(X)] = \mathbb{E}\left[\log\left(\frac{1}{p(X)}\right)\right] = \mathbb{E}\big[-\log\big(p(X)\big)\big] = -\sum_{x\in\mathcal{X}} p(x)\log p(x)$$

  where $X$ represents all $x$ values possible to the variable.

- The entropy is a property of the underlying distribution $P_X(x), x \in \mathcal{X}$ that measures the amount of randomness or surprise in the random variable.

# Entropy

- Imagine an information source that produces a value/result from a set of $n$ symbols $x_1, x_2, \dots, x_n$ with respective probabilities $p_1, p_2, \dots, p_n$.
- Ask yes/no questions to find the result produced by the information source as: "Is it equal to $x_1$?"
- The average number of yes/no questions necessary to find the right outcome is confined between $H$ and $H + 1$ given that the choice of questions is optimal, where $H$ is the entropy $-\sum p_i \log p_i$.
  - A question partitions the sample space into two subsets
  - The entropy of this partition is exactly equal to $H$
- $H$: the average amount of information required to determine the result produced by the source.

*Von Neumann told me, "You should call it entropy, for two reasons. In the first place your uncertainty function has been used in statistical mechanics under that name, so it already has a name. In the second place, and more important, no one really knows what entropy really is, so in a debate you will always have the advantage."*

C.E. Shannon (1916–2001)

# Properties of Entropy

- $H(X) \leq \log N$, with equality iff $X$ is uniformly distributed, i.e., $p(x) = \frac{1}{N}\ \forall x$

  Proof:

$$H(X) = \mathbb{E}\left[\log\left(\frac{1}{p(X)}\right)\right]$$
$$\leq \log \mathbb{E}\left[\frac{1}{p(X)}\right] \qquad \textit{by Jensen's inequality because log is concave}$$
$$= \log \sum_{x \in \mathcal{X}} p(x) \cdot \frac{1}{p(x)} = \log N$$

  Equality by Jensen's inequality iff $\frac{1}{p(X)}$ is deterministic, iff $p(x) = \frac{1}{N}\ \forall x$

- $H(X) \geq 0$, with equality iff $X$ is deterministic.

  Proof:

$$H(X) = \mathbb{E}\left[\log \frac{1}{p(X)}\right] \geq 0$$

  because $\log \frac{1}{p(X)} \geq 0$

  The equality occurs iff $\log \frac{1}{p(X)} = 0$ with probability 1 so $X$ must be deterministic.

# Significance of Entropy

- Entropy (in bits) tells us the average amount of information (in bits) that must be delivered in order to resolve the uncertainty about the outcome of a trial.
- This is a lower bound on the number of binary digits that must – on the average – be used to encode our messages.
- If we send fewer binary digits on average, the receiver will have some uncertainty about the outcome described by the message.
- If we send more binary digits on average, we're wasting the capacity of the communication channel by sending binary digits we don't have to.
- Achieving the entropy lower bound is the "gold standard" for an encoding (at least from the viewpoint of information compression).
- *Fixed vs. Variable-length encoding?*

# Relative Entropy

aka KL divergence

- A measure of distance between probability distributions is relative entropy:

$$D(p||q) \triangleq \sum_{x\in\mathcal{X}} p(x)\log\frac{p(x)}{q(x)} = \mathbb{E}\left[\log\frac{p(x)}{q(x)}\right]$$

- The relative entropy is always greater than or equal to 0, with equality iff $q = p$.
- Can be thought of as a measure of discrepancy between two probability distributions.

- For a PMF $q$ define $H_q(X) \triangleq \mathbb{E}\left[\log\frac{1}{q(X)}\right] = \sum_{x\in\mathcal{X}} p(x)\log\frac{1}{q(x)}$

  Then: $H(X) \leq H_q(X)$, with equality iff $q = p$.

  Proof:

$$H(X) - H_q(X) = \mathbb{E}\left[\log\frac{1}{p(X)}\right] - \mathbb{E}\left[\log\frac{1}{q(X)}\right] = \mathbb{E}\left[\log\frac{q(X)}{p(X)}\right]$$

$$\leq \log\mathbb{E}\left[\frac{q(X)}{p(X)}\right] = \log\sum_{x\in\mathcal{X}} p(x)\cdot\frac{q(x)}{p(x)}$$ ← KL divergence

$$= \log\sum_{x\in\mathcal{X}} q(x) = \log 1 = 0$$ ← probability distribution

  Thus, $H(X) - H_q(X) \leq 0$

Equality only holds when $\frac{q(x)}{p(x)}$ is deterministic, which occurs when $q = p$ (identical distributions)

# Information Measures

- Joint entropy: If $X_1, X_2, \ldots, X_n$ are independent random variables, then

$$H(X_1, X_2, \ldots, X_n) = \sum_{i=1}^{n} H(X_i)$$

- Conditional Entropy of $X$ given $Y$:

$$H(X|Y) \triangleq \mathbb{E}\left[\log \frac{1}{p(X|Y)}\right] = \sum_{y} p(y) H(X|Y=y)$$

$H(X|Y) \leq H(X)$ with equality iff $X$ and $Y$ are independent.

- Chain Rule:

$$H(X,Y) \triangleq \mathbb{E}\left[\log \frac{1}{p(X,Y)}\right] = H(X) + H(Y|X) = H(Y) + H(X|Y) \leq H(X) + H(Y)$$

with equality holding iff $X$ and $Y$ are independent

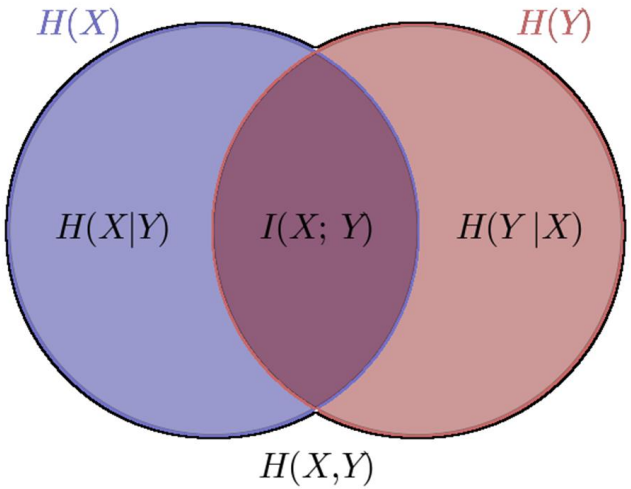
- More than two RVs:

$$H(X,Y,Z) = H(X) + H(Y,Z|X) = H(X) + H(Y|X) + H(Z|X,Y)$$

# Mutual Information

- We define the mutual information between random variables $X$ and $Y$ distributed according to the joint PMF $P(x,y)$:

$$
\begin{aligned}
I(X;Y) &\triangleq D\left(P_{x,y}||P_x \times P_y\right) \\
&= \mathbb{E}\left[\log \frac{p(X,Y)}{p(X)p(Y)}\right] \\
&= H(X) - H(X|Y) \\
&= H(Y) - H(Y|X) \\
&= H(X) + H(Y) - H(X,Y) \\
&= H(X,Y) - H(X|Y) - H(Y|X)
\end{aligned}
$$




Mutual Information

- quantifies the dependence of two random variables $X$ and $Y$
- captures non-linear statistical dependencies between variables (in contrast to correlation)
- tells how helpful one variable is at reducing uncertainty in the other.

- While *relative entropy* is **not** symmetric, *mutual information* **is**.

# Information Measures for Continuous Variables

## Mutual information (MI)

- Mutual information for continuous random variables $X, Y \sim f$ is given by

$$I(X;Y) \triangleq \mathbb{E}\left[\log \frac{f(X,Y)}{f(X)f(Y)}\right]$$

$$= \int_{\mathcal{X}\times\mathcal{Y}} f(x,y) \log \frac{f(x,y)}{f(x)f(y)}\, dxdy = D_{KL}(f(x,y)||f(x)f(y))$$

## Differential Entropy

$$I(X;Y) = \mathbb{E}\left[\log \frac{f(X,Y)}{f(X)f(Y)}\right] = \mathbb{E}\left[\log \frac{f(Y|X)}{f(Y)}\right] = \mathbb{E}\left[\log \frac{1}{f(Y)}\right] - \mathbb{E}\left[\log \frac{1}{f(Y|X)}\right]$$

- For a continuous random variable $X \sim f$, the differential entropy is defined as

$$h(X) \triangleq \mathbb{E}\left[\log \frac{1}{f(X)}\right]$$

- The entropy of a continuous RV in its purest sense doesn't exist, since the number of random bits necessary to specify a real number to an arbitrary precision is infinite.
- $h(X)$ need not be non-negative for every continuous RV, bcz a PMF is always at most 1 but a density function can be arbitrarily large.
- $h(X)$ need not be label invariant. For $Y = aX$, $a \in \mathbb{R}$: $h(Y) = h(X) + \log|a|$

# Unified Treatment

- Measure theory formulation:
  - for discrete RVs, PMFs can be considered density functions with respect to the counting measure.
  - both the integral and the sum can be thought as integration on a measure space

  - A function $\mu$ on a field $F$ in a space $\Omega$ is called a measure if the following conditions hold:

1. $\mu(A) \in [0, \infty]$ for $A \in F$
2. $\mu(\emptyset) = 0$
3. If the sequence $F$ −sets $A_1, A_2, \ldots$ are a disjoint sequence of $F$ −sets where $\bigcup_{k=1}^{\infty} A_k \in F$ then;

$$\mu\left(\bigcup_{k=1}^{\infty} A_k\right) = \sum_{k=1}^{\infty} A_k$$

The pair $(\Omega, F)$ is called a measurable space if $F$ is a $\sigma$ −field in the sample space Ω.

Radon-Nikodym derivative: If a measure $P$ for $P(A)$ equals 0 whenever another measure $Q(A) = 0$, then $P$ is said to be dominated by another measure $Q$, denoted as $P \ll Q$. For $P \ll Q$, the Radon-Nikodym derivative of $P$ with respect to $Q$ is denoted by $dP/dQ$.

# Unified Treatment

## Kullback-Leibler divergence

- if $P$ and $Q$ are probability measures over a set $\mathcal{X}$, and $P$ is absolutely continuous with respect to $Q$, then the relative entropy from $Q$ to $P$ is defined as

$$D_{KL}(P||Q) = \int_{\mathcal{X}} \log\left(\frac{dP}{dQ}\right) dP = \int_{\mathcal{X}} p \log\left(\frac{p}{q}\right) d\mu$$

where $\frac{dP}{dQ}$ is the Radon–Nikodym derivative of $P$ with respect to $Q$ and

$\mu$ is any measure on $\mathcal{X}$ for which the densities $p = \frac{dP}{d\mu}$ and $q = \frac{dQ}{d\mu}$ exist ($P$ and $Q$ are absolutely continuous with respect to Lebesgue measure $\mu$)

## Mutual Information

$$I(X;Y) = \int_{\mathcal{X}\times\mathcal{Y}} \log \frac{d\mathbb{P}_{XY}}{d\mathbb{P}_X \otimes \mathbb{P}_Y} d\mathbb{P}_{XY} = D_{KL}(\mathbb{P}_{XY} \,||\mathbb{P}_X \otimes \mathbb{P}_Y)$$

where $\mathbb{P}_{XY}$ is the joint probability distribution, $\mathbb{P}_X = \int_{\mathcal{Y}} d\mathbb{P}_{XY}$, $\mathbb{P}_Y = \int_{\mathcal{X}} d\mathbb{P}_{XY}$ are the marginal

# $f$-Divergences

- Definition: let $f\colon \mathbb{R}_+ \to \mathbb{R}$ be a convex function such that $f(1) = 0$.

  For two PMF $P, Q$ we define the $f$-divergence between $P$ and $Q$ by

$$D_f(P||Q) \triangleq \sum_{x\in\mathcal{X}} Q(x) f\left(\frac{P(x)}{Q(x)}\right)$$

  $D_f(P||Q) = 0$ iff $P = Q$

  For $P$ and $Q$ two probability distributions on a measurable space

$$D_f(P||Q) \triangleq \mathbb{E}_Q\left[f\left(\frac{dP}{dQ}\right)\right]$$

- Examples:
  - Kullback-Leibler divergence: $f(u) = u\log u$
  - Symmetric Kullback-Leibler divergence: $f(u) = u\log u - \log u$
  - Total Variation: $f(u) = \frac{1}{2}|u-1|$
  - $\chi^2$ – divergence: $f(u) = (u-1)^2$
  - Hellinger-squared distance: $f(u) = (\sqrt{u}-1)^2$

# $f$–Divergences-Why $f$-divergences?

- Blackwell Theorem (1951): If a procedure A has smaller $f$-divergence than a procedure B (for some fixed $f$), then there exist some set of prior probabilities such that procedure A has a smaller probability of error than procedure B.
- $f$-divergences can be used as surrogates for probability of error or test error

- For two opposing hypothesis $H_0, H_1$ set to choose between two probability distributions $P$ and $Q$ respectively based on the observations, probability of false-alarm (Type I error) and mis-detection (Type II error) are defined by

$$P_{fa} = \Pr[Choose\ H_1 | H_0\ correct]$$
$$P_{miss} = \Pr[Choose\ H_0 | H_1\ correct]$$

- Optimal classifier obeys the following asypmtotics for an $M$ —dimensional random vector of observations:

$$\lim_{M\to\infty} \frac{P_{fa}}{M} = -D(Q||P)$$
$$\lim_{M\to\infty} \frac{P_{miss}}{M} = -D(P||Q)$$

# $f$–Divergences – Basic Properties

important

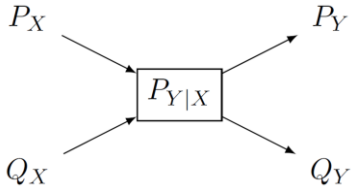
1. Non-negativity: $D_f(P||Q) \geq 0$ (If $f$ is strictly convex at 1, then $D_f(P||Q) = 0$ iff $P = Q$).
2. $D_{f_1+f_2}(P||Q) = D_{f_1}(P||Q) + D_{f_2}(P||Q)$
3. $D_f(P||P) = 0$
4. If $P_{X,Y} = P_X P_{Y|X}$ and $Q_{X,Y} = P_X Q_{Y|X}$, then
$$D_f(P_{X,Y}||Q_{X,Y}) = D_f(P_{Y|X}||Q_{Y|X}|P_X)$$
5. If $P_{X,Y} = P_X P_{Y|X}$ and $Q_{X,Y} = Q_X P_{Y|X}$, then
$$D_f(P_{X,Y}||Q_{X,Y}) = D_f(P_X||Q_X)$$
6. Monotonicity: If $P_{X,Y} = P_X P_{Y|X}$ and $Q_{X,Y} = Q_X P_{Y|X}$, then
$$D_f(P_{X,Y}||Q_{X,Y}) \geq D_f(P_X||Q_X)$$

(antecedency)

# $f$–Divergences – Basic Properties

- Data processing: Consider a channel that produces $Y$ given $X$ based on the conditional law $P_{Y|X}$. Let $P_Y$ (resp. $Q_Y$) denote the distribution of $Y$ when $X$ is distributed as $P_X$ (resp. $Q_X$):

$$D_f(P_Y||Q_Y) \leq D_f(P_X||Q_X)$$



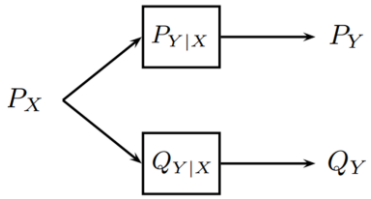

  - Due to monotonicity and Property 5.

- Conditioning increases $f$-divergence: Define $D_f(P_{Y|X}||Q_{Y|X}|P_X) \triangleq \mathbb{E}_{X\sim P_X}\left[D_f(P_{Y|X}||Q_{Y|X})\right]$

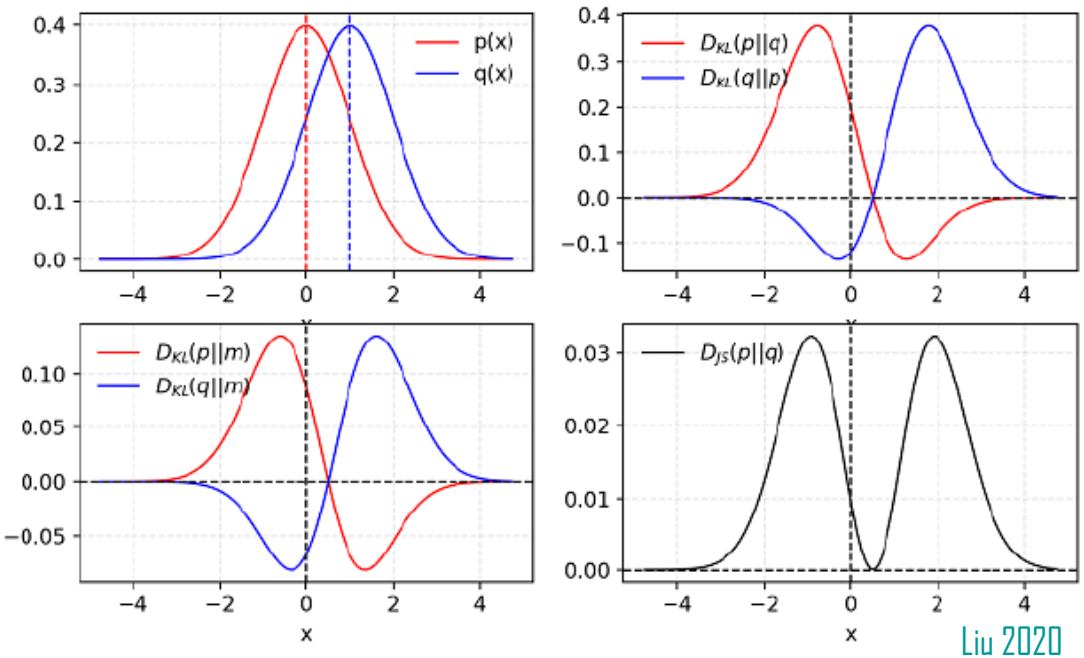
  Let $P_X \xrightarrow{P_{Y|X}} P_Y$ and $P_X \xrightarrow{Q_{Y|X}} Q_Y$, then

$$D_f(P_Y||Q_Y) \leq D_f(P_{Y|X}||Q_{Y|X}|P_X)$$

$H(X) \geq H(X|Y)$



# $f$–Divergences –Example




Liu 2020

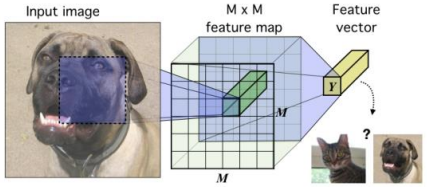
- Figure on the top left shows 2 Gaussian densities with the same variance and different mean, whereas the the on the top right is the KL divergence between the two. Note that KL is not symmetric!
- Figure on the bottom left shows the KL divergence between the Gaussian distributions and their average $m = \frac{p(x)+q(x)}{2}$
- Finally the one on the right bottom shows the Jensen-Shannon divergence which is the average of the two KL divergences on its left

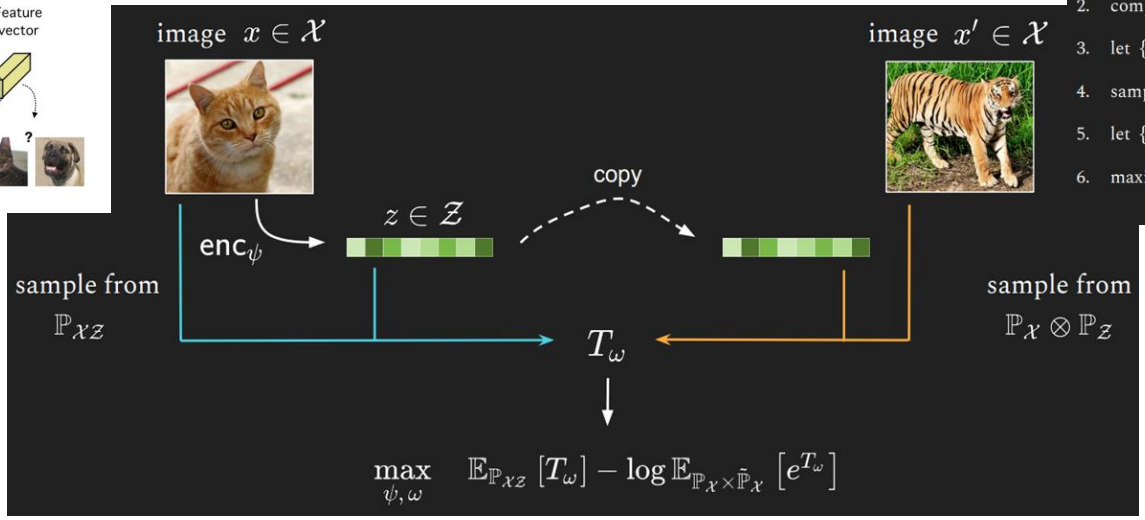
# Variational Representation

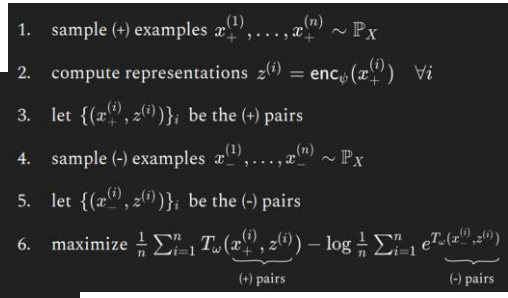
**Theorem** (Donsker-Varadhan). *Let $P, Q$ be probability measures on $\mathcal{X}$ and let $\mathcal{C}$ denote the set of functions $f : \mathcal{X} \to \mathbb{R}$ such that $\mathbb{E}_Q[\exp\{f(X)\}] < \infty$. If $D(P\|Q) < \infty$ then for every $f \in \mathcal{C}$ expectation $\mathbb{E}_P[f(X)]$ exists and furthermore*

$$D(P\|Q) = \sup_{f \in \mathcal{C}} \mathbb{E}_P[f(X)] - \log \mathbb{E}_Q[\exp\{f(X)\}] .$$

- Learning deep representations by MI estimation and maximization




- To maximize global MI, we use DV representation of KL divergence (lower bound of true MI), since we are unable to access the intractable, true joint and marginal distributions

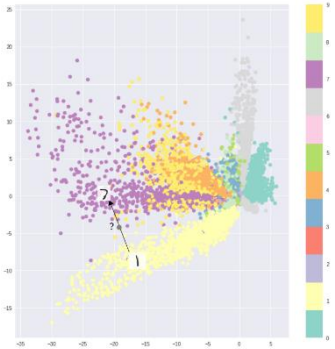
$$I(X;Z) = D_{KL}(\mathbb{P}_{XZ} \,||\mathbb{P}_X \otimes \mathbb{P}_Z) \geq \sup_{T:\Omega \to \mathbb{R}} \mathbb{E}_{\mathbb{P}_{XZ}}[T] - \log(\mathbb{E}_{\mathbb{P}_X \otimes \mathbb{P}_Z}[e^T])$$

where the supremum is taken over all functions $T$ such that both expectations are finite.

# Information Bottleneck

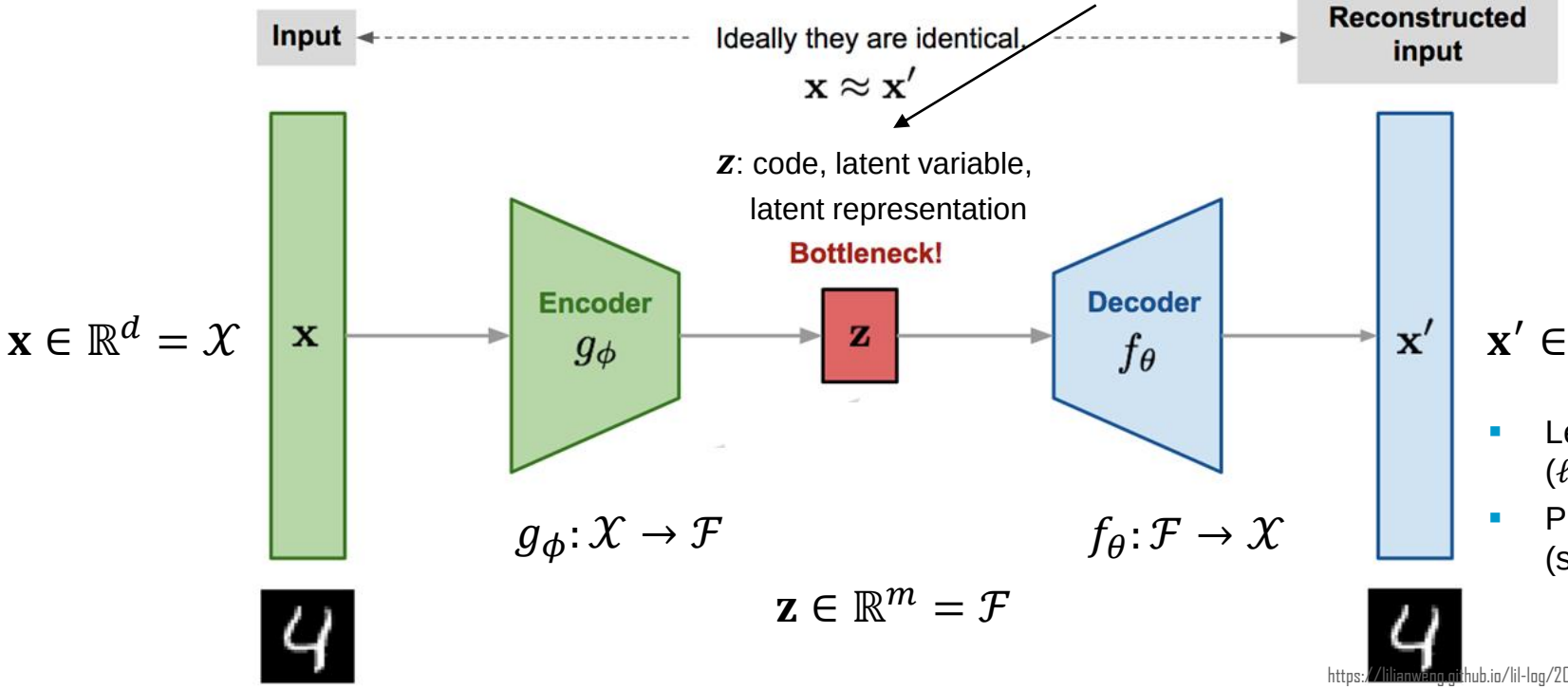
# Autoencoders

Encode the input into a representation (bottleneck) and reconstruct it with the decoder



Latent space learnt on MNIST

https://bit.ly/37ctFMS




- Learn features ($\ell_2$ loss)
- Predict labels (softmax, etc.)

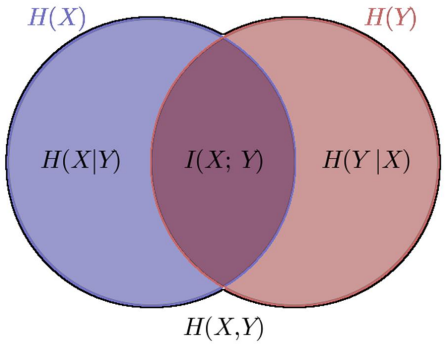
https://lilianweng.github.io/lil-log/2018/08/12/from-autoencoder-to-be vae.html



EURECOM Sophia Antipolis

# Information Measure in a Nutshell

- Entropy $H(X)$: the average number of bits to describe the outcome of a random variable $X$
- Conditional Entropy $H(Y|X)$: the average number of bits to describe the outcome of $Y$ given that the value of $X$ is known.
- Mutual Information $I(X;Y)$: the average number of bits that $X$ contains about $Y$




$$I(X;Y) = H(Y) - H(Y|X)$$

## Representation Desiderata

An optimal representation $Z$ of the data $X$ for the task $Y$ is a stochastic function $Z \sim p(Z|X)$ that is:

- Sufficient: $I(Z;Y) = I(X;Y)$
- Minimal: $I(X;Z)$ is minimal among sufficient $Z$
- Maximally disentangled: $TC(Z) = KL(p(Z)||\prod_i p(Z_i))$

# Rate Distortion Theory

Rate $R$

Encoder $f$

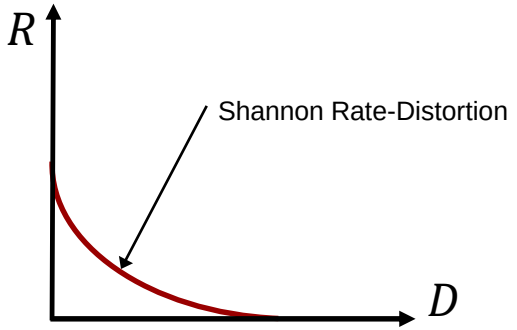
Decoder $g$

$X \sim p_X$

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

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

$$R(D) = \min_{p_{\hat{X}|X}} I(X,\hat{X})$$
$$= \min_{p_{\hat{X}|X}} H(X) - H(X|\hat{X})$$
$$\text{s.t. } \mathbb{E}_{p_{X,\hat{X}}}[d(X,\hat{X})] \leq D$$




Blahut-Arimoto algorithm

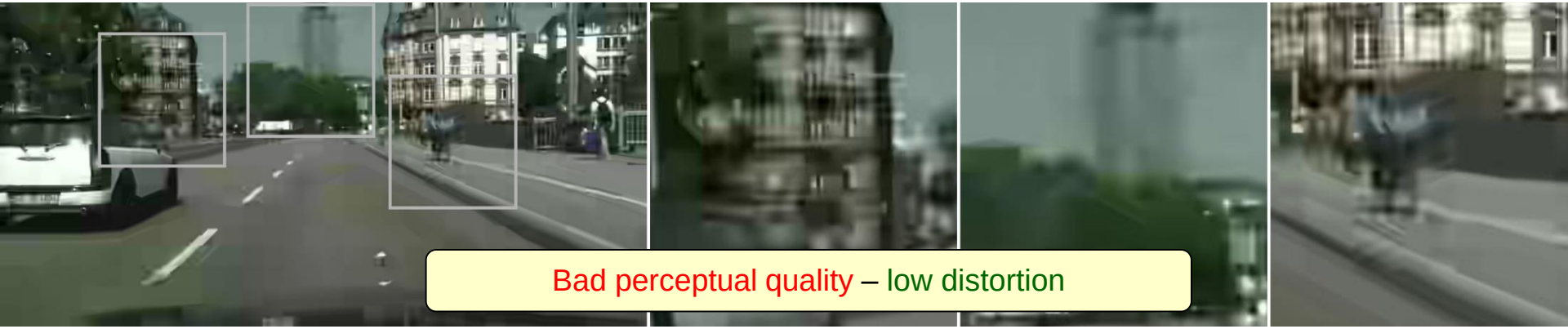
Lower the bit rate $R$ by allowing some distortion $d$ of the original “information”

$$\min_{p_{\hat{X}|X}} I(X,\hat{X}) + \beta \mathbb{E}[d(X,\hat{X})]$$

An alternating iterative algorithm to calculate $p(\hat{X})$ and $p(\hat{X}|X)$

# Perception Quality

Good perceptual quality – high distortion

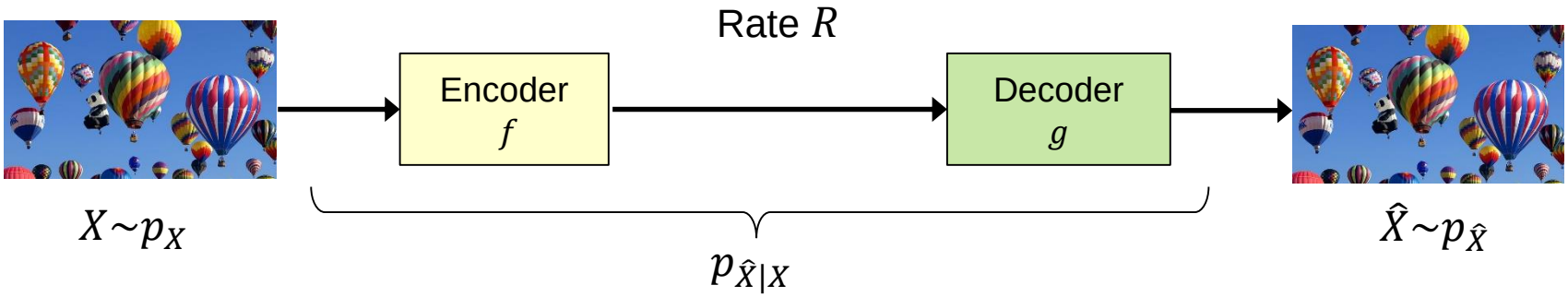


Bad perceptual quality – low distortion

Good perceptual quality ≠ low distortion

Agustsson et al. (2018)

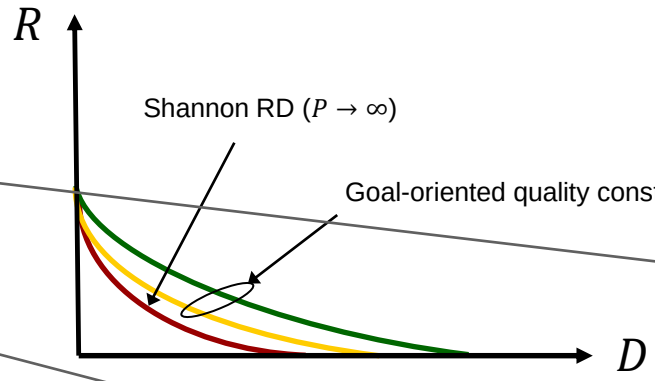
# Rate-Distortion-Perception (RDP)




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

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

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




monotonically non-increasing and convex in $\epsilon$ and $\sigma$

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

**Perception - Semantic quality**

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

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

**Distortion metrics**

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

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

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

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

# Perceptual lossy compression

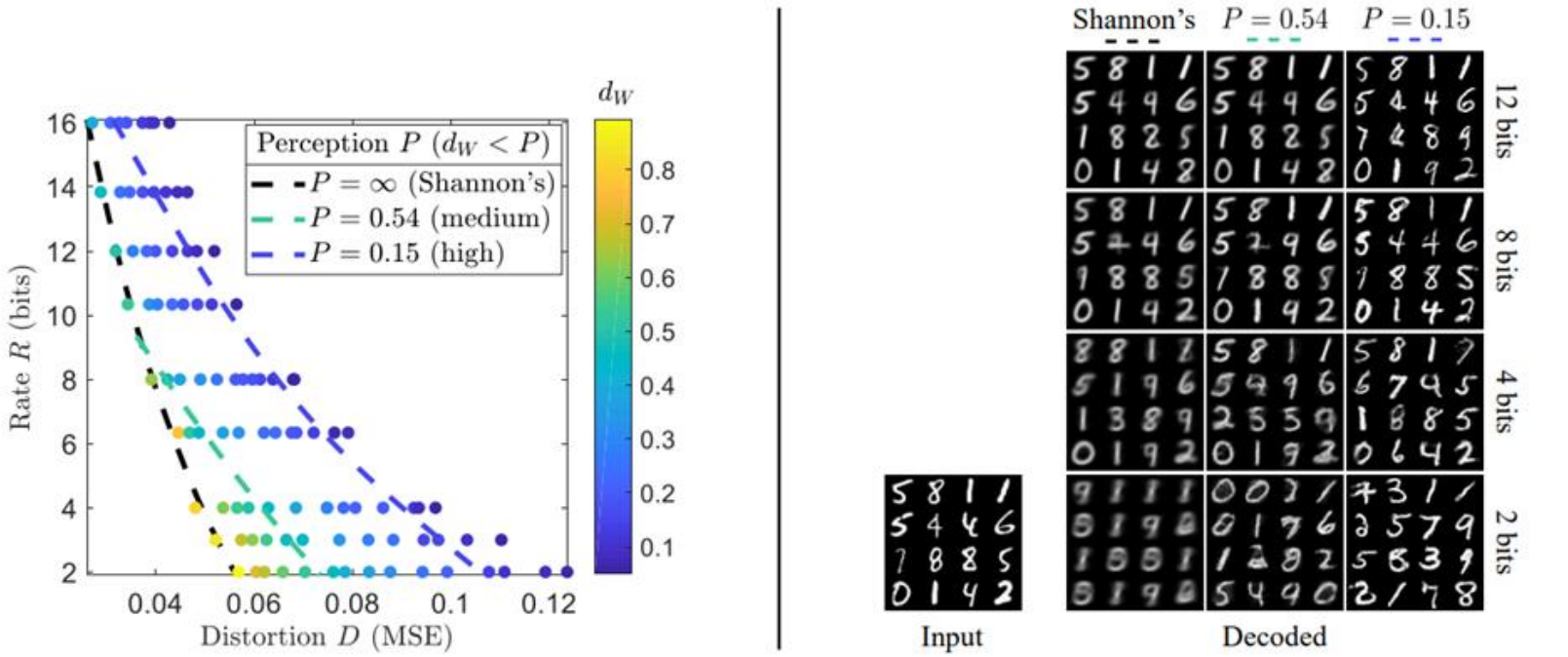
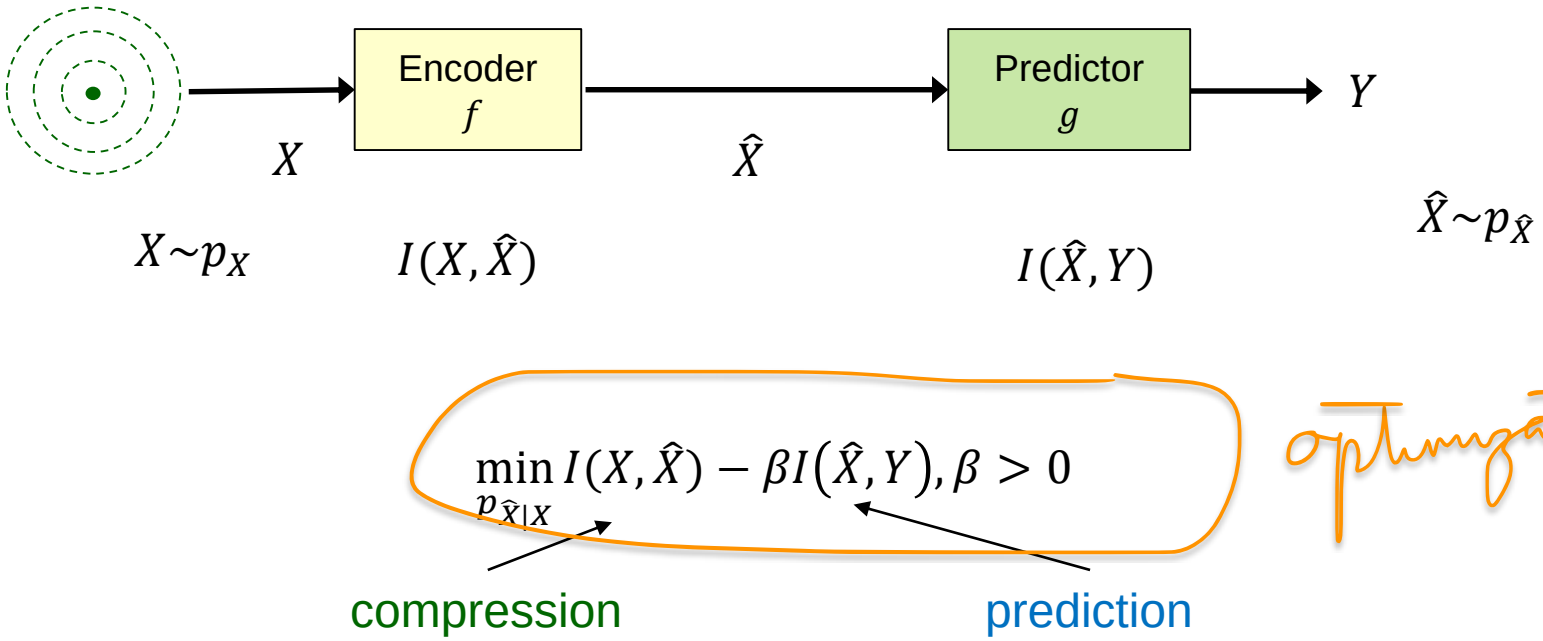



*Figure 6.* **Perceptual lossy compression of MNIST digits.** *Left*: Shannon's rate-distortion curve (black) describes the lowest possible rate (bits per digit) as a function of distortion, but leads to low perceptual quality (high $d_W$ values), especially at low rates. When constraining the perceptual quality to be good (low $P$ values), the rate-distortion curve elevates, indicating that this comes at the cost of a higher rate and/or distortion. *Right*: Encoder-decoder outputs along Shannon's rate-distortion curve and along two equi-perceptual-quality curves. As the rate decreases, the perceptual quality along Shannon's curve degrades significantly. This is avoided when constraining the perceptual quality, which results in visually pleasing reconstructions, even at extremely low bit-rates. Notice that increased perceptually quality does not imply increased accuracy, as most reconstructions fail to preserve the digits' identities at a 2-bit rate.

Arxiv: 1901.07821

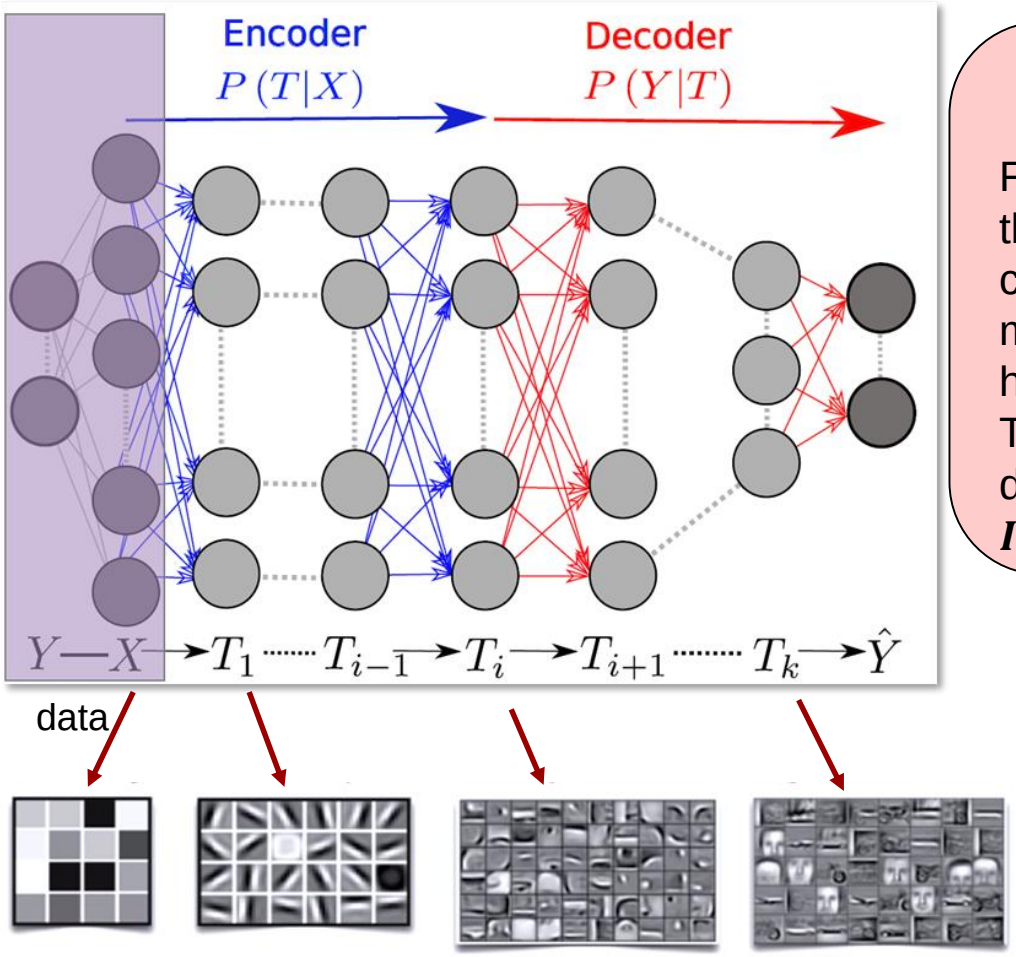
# Information Bottleneck

- IB extends RD theory to prediction, by measuring the quality of the encoding by its ability to predict another random variable.




- Solution: Similar iterative algorithm as Blahut-Arimoto to calculate $p(\hat{X}|X)$, $p(\hat{X})$, and $p(Y|\hat{X})$.
- Trade-off: between sufficiency and minimality, regulated by the parameter $\beta$.

# DNN Layers and Encoder-Decoder Information

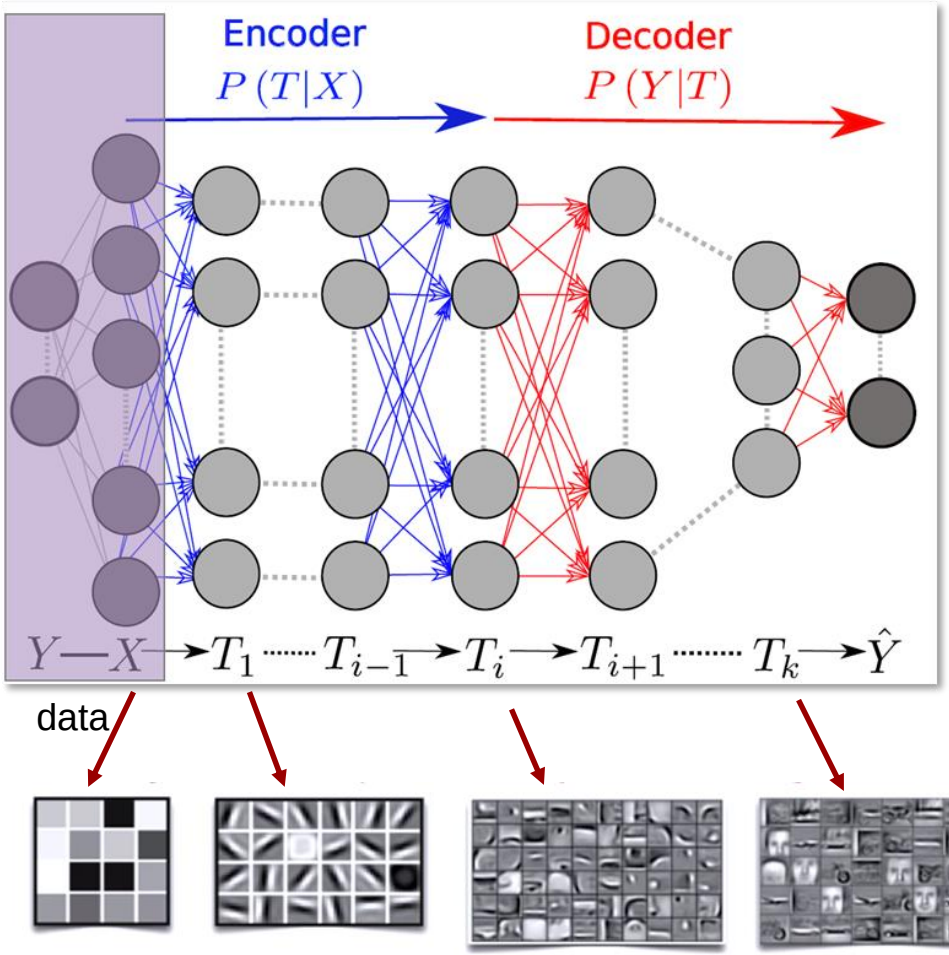



**Theorem (Information Plane)**

For large typical $\boldsymbol{X}$,
the sample complexity of a DNN is completely determined by the encoder mutual information $\boldsymbol{I(X;T)}$ of the last hidden layer.
The accuracy (generalization error) is determined by the decoder information, $\boldsymbol{I(T;Y)}$, of the last hidden layer.

- The complexity of the problem shifts from the decoder to the encoder, across the layers…

# What do DNN Layers represent?



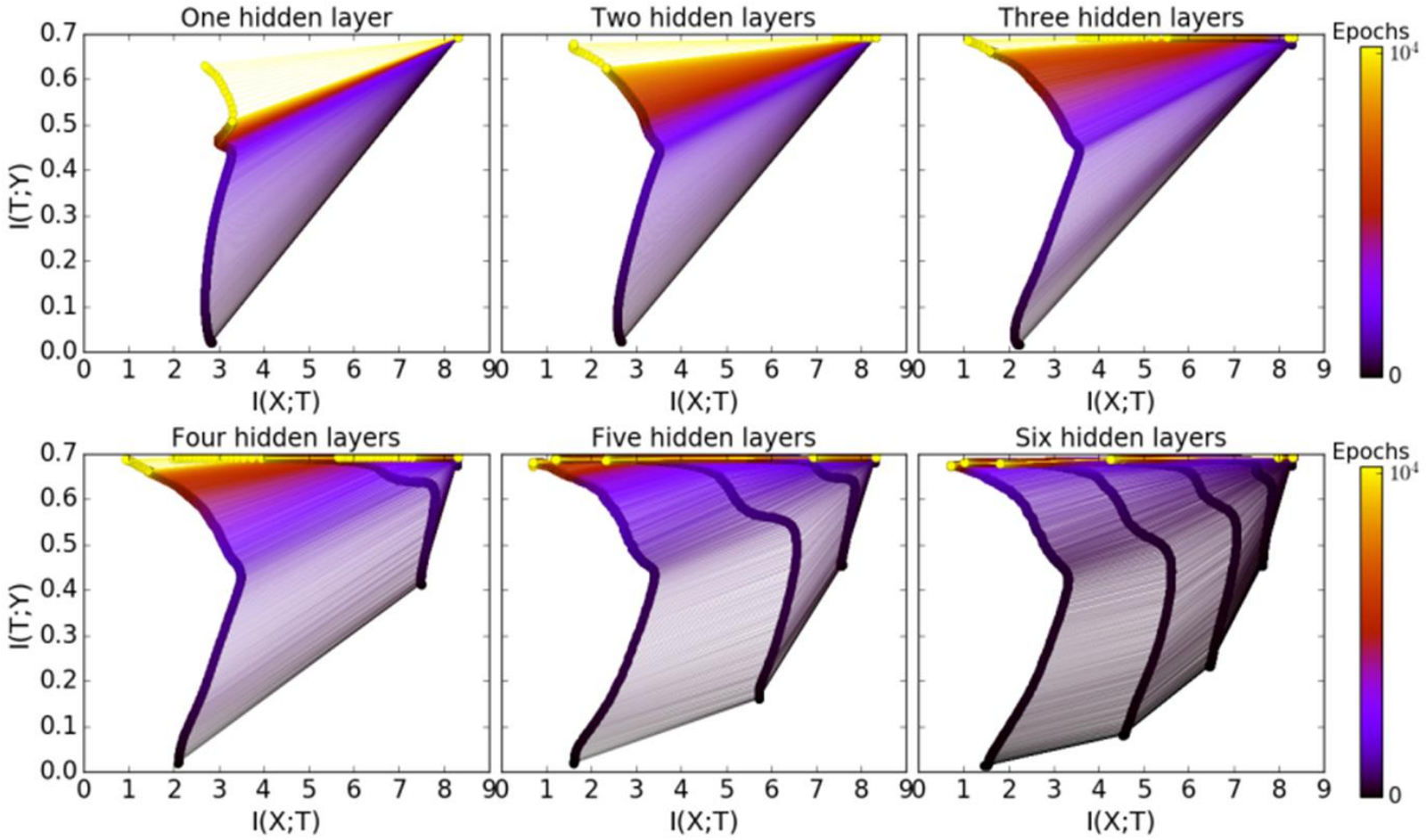

- A Markov chain of topologically distinct [soft] *partitions* of the input variable $\boldsymbol{X}$
- Successive Refinement of Relevant Information
- Individual neurons can be easily "scrambled" within each layer

Data Processing Inequalities:

$$H(X) \geq I(X;T_i) \geq I(X;T_{i+1}) \geq I(X;T_{i+2}) \geq \ \ldots$$

$$H(X;Y) \geq I(T_i;Y) \geq I(T_{i+1};Y) \geq I(T_{i+2};Y) \geq \ \ldots$$

# The Benefit of the Hidden Layers




- More layers take much fewer training epochs for good generalization
- The optimization time depend super-linearly on the compressed info for each layer